

Semantic Communication for Video Coding and Transmission

Samarathunga Liyanage Don Buddhi Prabhath Samarathunga

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

AI and Multimedia Research Group
Department of Computer and Information Science (CIS)
University of Strathclyde, Glasgow

June 19, 2026

Declaration

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Signed: Prabhath Samarathunga

Date: June 19, 2026

Abstract

Video traffic dominates global internet bandwidth, and new applications such as extended reality, autonomous vehicle telemetry, and sixth-generation (6G) wireless networks are increasing the demands on video communication in bitrate, quality, robustness, content diversity, and dynamic range. Conventional standards such as H.266/Versatile Video Coding (VVC) are approaching the limits of manually designed coding tools, while end-to-end learned video codecs still suffer from unstable training, limited generalisation, and severe quality loss under channel impairments. Both approaches are semantically unaware because they aim for pixel-accurate reconstruction rather than preserving the meaning of the transmitted signal. This thesis argues that semantic communication, which sends task-relevant meaning instead of bit-accurate signal representations, offers a practical way forward, but its value for real video coding has not yet been developed in a single end-to-end framework.

This thesis develops that framework through original contributions in four areas: *hybrid coding*, *content adaptation*, *HDR extension*, and *robust wireless transmission*, supported by methods for Deep Neural Network (DNN) parameter efficiency. For hybrid coding, the thesis introduces a hybrid semantic–VVC video codec, extends it with hierarchical bidirectional frame prediction, and further develops it into a context-conditioned architecture based on a shared pretrained model with motion-aligned context generation. This system gives consistent rate–distortion gains over VVC and competitive performance against leading learned codecs. For content adaptation, the thesis restores per-Group-of-Pictures (GOP) specialisation by adjusting only a small part of the decoder parameters, which gives further gains across several datasets. For HDR extension, the thesis introduces an automatic and invertible tone-mapping step

that bridges the HDR-to-SDR gap, allowing an SDR-trained codec to handle HDR content without changing the model architecture. For robust wireless transmission, the thesis develops a Semantic Multiple Description Coding (SMDC) framework that removes cliff effects across fading channels for both image and video transmission. Two further contributions on DNN parameter efficiency, temporal prediction of decoder parameters and VVC-based weight compression, reduce parameter transmission overhead and support future work toward a fully semantic video codec.

Taken together, these contributions show that semantic communication can give consistent and reproducible gains over conventional and learned baselines in rate-distortion efficiency, channel robustness, content adaptation, and parameter-transmission efficiency. They provide a clear technical basis for semantic-driven video communication in next-generation wireless networks.

Contents

Declaration	i
Abstract	ii
List of Figures	x
List of Tables	xx
Acronyms and Abbreviations	xxv
Acknowledgements	xxxiii
Publications	xxxiv
1 Introduction	2
1.1 Background and Motivation	2
1.1.1 Learned Video Coding Using Deep Neural Networks	4
1.1.2 Semantic Communication: A Paradigm Shift	6
1.2 Problem Statement	12
1.3 Research Gap	14
1.4 Research Aim	16
1.5 Research Objectives	16
1.6 Research Questions	18
1.7 Contributions	19
1.7.1 Collective Significance and Scientific Impact	23
1.8 Thesis Structure	25

2	Background and Literature Review	28
2.1	Introduction	28
2.2	Evolution of Conventional Video Coding Standards	29
2.2.1	Origins and Early Standardization	29
2.2.2	H.264/Advanced Video Coding	30
2.2.3	H.265/High Efficiency Video Coding	31
2.2.4	H.266/Versatile Video Coding	32
2.2.5	AOMedia Video 1 and AOMedia Video 2 (AV2)	33
2.2.6	Key Coding Tools Across Generations	33
2.3	Challenges and Limitations of Conventional Video Coding	34
2.3.1	The Complexity Wall	34
2.3.2	Limitations for Emerging Content and Applications	35
2.3.3	Fundamental Rate–Distortion Limitations	36
2.4	Deep Learning-Based Image Compression	37
2.4.1	Early Neural Network Approaches	37
2.4.2	End-to-End Learned Image Compression with Autoencoders	37
2.4.3	Generative Models for Semantic Image Compression	38
2.4.4	Vision Transformers for Image Compression	38
2.5	End-to-End Learned Video Compression	39
2.5.1	From Image Compression to Video Compression	39
2.5.2	Optical Flow-Based Frameworks	40
2.5.3	Conditional Coding and Contextual Frameworks	41
2.5.4	Challenges in Learned Video Compression	43
2.6	Foundations of Semantic Communication	44
2.6.1	Shannon and Weaver’s Three-Level Framework	44
2.6.2	Renewed Interest Driven by Deep Learning	45
2.6.3	Deep Learning Implementations	45
2.6.4	Semantic-Aware Networking and Cross-Modal Communication	46
2.7	Semantic Communication for Text	47
2.8	Semantic Communication for Speech and Audio	48

Contents

2.9	Semantic Communication for Images	48
2.9.1	Deep Joint Source–Channel Coding	48
2.9.2	Region of Interest Semantic Coding	50
2.9.3	Task-Oriented Semantic Communication	51
2.9.4	Practical Wireless Semantic Image Communication Systems . . .	52
2.10	Semantic Communication for Video	53
2.10.1	Video Conferencing Applications	53
2.10.2	Video Compression for Machines	54
2.10.3	Hybrid Semantic and Conventional Video Coding	54
2.10.4	Deep Learning for Wireless Video Semantic Transmission	55
2.11	Survey of Recent Semantic Communication Systems	55
2.11.1	Multi-User and Network-Level Semantic Communication	56
2.11.2	Generative AI and Large Language Model-Enabled Semantic Com- munication	56
2.11.3	Federated Learning for Semantic Communication	56
2.11.4	Task-Oriented Communication and Semantic Information Theory	57
2.12	Hybrid Semantic and VVC-Based Video Coding	57
2.12.1	Motivation for Hybrid Architectures	57
2.12.2	Prior Work on Hybrid Learning and Conventional Coding	58
2.12.3	Bidirectional Hierarchical Prediction	59
2.13	Context-Aware and Content-Adaptive Semantic Video Coding	59
2.13.1	Pretrained Foundation Models for Video Coding	59
2.13.2	Content-Adaptive Parameter Overfitting	61
2.13.3	Differential Latent Encoding and DeepCABAC	62
2.13.4	Rate–Distortion Optimisation for Hybrid Semantic Video Coding	62
2.14	Semantic Video Coding for High Dynamic Range Content	63
2.14.1	High Dynamic Range Imaging	63
2.14.2	HDR Video Coding in Conventional Standards	63
2.14.3	Tone Mapping for HDR–SDR Bridging	64
2.14.4	Domain Gap Between HDR and SDR Semantic Codecs	64

Contents

2.15	Multiple Description Coding for Robust Transmission	65
2.15.1	Multiple Description Coding Theory	65
2.15.2	Practical MDC Methods	65
2.15.3	Semantic Multiple Description Coding	66
2.15.4	Fading Channel Models	67
2.16	DNN Parameter Efficiency for Semantic Video Coding	67
2.16.1	Neural Network Compression	67
2.16.2	Structured Pruning and Low-Rank Decomposition	68
2.16.3	Temporal Prediction of DNN Parameters	68
2.16.4	VVC-Based DNN Weight Compression	68
2.16.5	Neural Network Compression Standards	69
2.17	Use of the Term “Semantic Communication” in This Thesis	69
2.18	Evaluation Metrics for Video Coding Quality	71
2.18.1	PSNR and Bjøntegaard Delta Rate	71
2.18.2	Measuring Semantic Preservation and What the Hybrid Codec Preserves	72
2.18.3	PU-PSNR and PU-SSIM for HDR Content	75
2.19	Preliminary Work on Semantic Communication for Images	75
2.20	Summary of Research Gaps	77
3	Hybrid Semantic and VVC-Based Video Coding	80
3.1	Chapter Introduction	80
3.2	A Semantic Communication and VVC-Based Hybrid Video Coding System	81
3.2.1	Introduction	81
3.2.2	Methodology	82
3.2.3	Results and Discussion	91
3.2.4	Conclusion	103
3.3	Enhanced Hybrid Codec with Bidirectional Hierarchical Frame Prediction	104
3.3.1	Introduction	105
3.3.2	Methodology	106
3.3.3	Results and Discussion	122

Contents

3.3.4	Conclusion	135
3.4	Chapter Summary and Conclusion	137
4	Context-Aware Semantic Video Coding	139
4.1	Chapter Introduction	139
4.2	Context-Aware Coding with Pretrained Motion-Adaptive Models and Bidirectional Prediction	141
4.2.1	Introduction	141
4.2.2	Methodology	141
4.2.3	Results and Discussion	151
4.2.4	Conclusion	165
4.3	Content-Adaptive Parameter Overfitting for Per-GOP Specialisation . .	166
4.3.1	Introduction	166
4.3.2	Methodology	167
4.3.3	Results and Discussion	174
4.3.4	Conclusion	187
4.4	Chapter Summary and Conclusion	188
5	Semantic Video Coding for High Dynamic Range Content	191
5.1	Chapter Introduction	191
5.2	Content-Adaptive Parameter Overfitting for HDR Video Coding	194
5.2.1	Methodology	194
5.2.2	Results and Discussion	200
5.3	Chapter Summary and Conclusion	208
6	Robust Semantic Communication via Multiple Description Coding	210
6.1	Chapter Introduction	210
6.2	Semantic Multiple Description Coding for Robust Image Transmission .	214
6.2.1	Introduction	214
6.2.2	Methodology	215
6.2.3	Results and Discussion	224
6.2.4	Conclusion	234

Contents

6.3	Semantic Multiple Description Coding for Robust Video Transmission	235
6.3.1	Introduction	235
6.3.2	Methodology	235
6.3.3	Results and Discussion	242
6.3.4	Conclusion	249
6.4	Chapter Summary and Conclusion	251
7	DNN Parameter Efficiency for Semantic Video Coding	252
7.1	Chapter Introduction	252
7.2	Deep Neural Network Compression for Semantic Video Communications	255
7.2.1	Introduction	255
7.2.2	System Architecture	256
7.2.3	Results and Discussion	258
7.3	Temporal Prediction of Deep Neural Network Parameters for Efficient Video Coding	261
7.3.1	Introduction	261
7.3.2	System Architecture	262
7.3.3	Motivating Evidence: Temporal Correlation of DNN Parameters	266
7.3.4	Results and Discussion	266
7.4	Chapter Summary and Conclusion	270
8	Conclusions and Future Work	272
8.1	Summary of Contributions	272
8.2	Key Findings	276
8.3	Overall Conclusion	278
8.4	Implications for Research and Industry	279
8.5	Limitations	280
8.6	Future Work	282
8.7	Closing Remarks	286
	Appendix: Individual Rate Distortion Performance for Test Videos	287

Contents

A Individual Rate–Distortion Performance for Test Videos for GOP Sizes 8, 16, and 32	288
Bibliography	288

List of Figures

3.1	Block diagram of the proposed semantic communication and VVC-based hybrid video codec. The system comprises a semantic encoder/decoder pair realised via an autoencoder, a VVC-based context encoder/decoder for transmitting GOP key frames, and a VVC-based residual encoder/decoder for correcting semantic reconstruction errors. A scene transition detector triggers new GOPs at scene boundaries [1].	83
3.2	Autoencoder architecture used in the semantic encoder and decoder. The encoder layers compress each input frame to an 8×1 latent vector; the decoder layers reconstruct the frame from that vector [1].	83
3.3	Test video clips used for evaluation [1]: (a) Video 1 [2], (b) Video 2 [3], (c) Video 3 [4], (d) Video 4 [5], (e) Video 5 [6], (f) Video 6 [7], (g) Video 7 [8], (h) Video 8 [9], (i) Video 9 [10], (j) Video 10 [11], (k) Video 11 [12], (l) Video 12 [13], (m) Video 13 [14], (n) Video 14 [15], (o) Video 15 [16], (p) Video 16 [17], (q) Video 17 [18], (r) Video 18 [19], (s) Video 19 [20], (t) Video 20 [21].	88
3.4	Distribution of SI and TI for the 20 test video clips. Symbols indicate complexity category: $\circ$ low, $\triangle$ medium, $\square$ high [1].	89
3.5	Rate-distortion performance comparison of the proposed system against VVC, HEVC, and AVC for GOP size 8. Rows 1–3 show average PSNR, SSIM, and VMAF for low, medium, and high complexity categories. Row 4 shows BD-rates relative to each reference codec [1].	92

List of Figures

3.6	Rate-distortion performance comparison of the proposed system against VVC, HEVC, and AVC for GOP size 16 [1].	93
3.7	Rate-distortion performance comparison of the proposed system against VVC, HEVC, and AVC for GOP size 32 [1].	95
3.8	Effect of latent vector dimension on the average PSNR of the reconstructed frame at the semantic decoder output, evaluated on 320×180 input frames. A saturation point is reached at dimension 8×1 , beyond which further capacity yields no meaningful gain [1].	97
3.9	Frames reconstructed at (a) the transmitter AE decoder layers and (b) the remotely trained receiver AE decoder layers for an example high-complexity video. The outputs are visually indistinguishable, validating the remote training mechanism [1].	98
3.10	Example original frames and their semantic reconstruction residuals. Top row: original frames from (a) Video A (medium complexity), (b) Video B (high complexity), (c) Video C (low complexity). Bottom row: corresponding residuals prior to residual correction [1].	100
3.11	End-to-end view of the enhanced semantic communications and VVC-based hybrid video codec. Dashed lines denote training process flows and solid lines denote operational data flows [22].	107
3.12	Architecture of the semantic encoder and decoder, implemented as a fully connected autoencoder for block-based coding in the YUV420 colour space [22].	109
3.13	Bidirectional hierarchical prediction structure for a GOP of size 8. First-order frames (GOP midpoint) are predicted from the two key frames; subsequent orders are predicted from their adjacent references [22]. . . .	111
3.14	Training loss convergence curves for the first frame of a scene and three representative subsequent frames, averaged over a set of video scenes. Weight transfer from the first-frame training to subsequent frames reduces the number of epochs required to reach the early stopping criterion [22].	115

List of Figures

3.15	Cross-entropy between the semantic latent l_{sem} and the RIFE-predicted latent l_{RIFE} , evaluated as the mean over selected elements, confirming high mutual information between the two representations [22].	117
3.16	Visual and quantitative analysis of block boundary artefacts for a 4K video sequence. Top row: full-frame comparison of original, semantic prediction, and semantic+VVC prediction. Bottom row: zoomed block regions with localised PSNR/SSIM measurements at block boundaries and interior regions [22].	119
3.17	Classification of test videos from the HEVC Class B, HEVC Class E, MCL-JCV, and UVG datasets based on spatial information and temporal information, at 720p resolution [22].	120
3.18	Average rate–distortion performance of the proposed codec and VTM 17.0 for low-complexity test videos. Left to right: PSNR, SSIM, VMAF. Note: the y -axes are truncated for clarity [22].	122
3.19	Average rate–distortion performance of the proposed codec and VTM 17.0 for medium-complexity test videos. Left to right: PSNR, SSIM, VMAF. Note: the y -axes are truncated for clarity [22].	122
3.20	Average rate–distortion performance of the proposed codec and VTM 17.0 for high-complexity test videos. Left to right: PSNR, SSIM, VMAF. Note: the y -axes are truncated for clarity [22].	123
3.21	Average rate–distortion performance of the proposed codec and VTM 23.0 for 4K resolution test videos from the Netflix Open Content dataset. Left to right: PSNR, SSIM, VMAF. Note: the y -axes are truncated for clarity [22].	123
3.22	PSNR and bitrate trade-offs across residual orders. The monotonic increase in PSNR from 26.61 dB to 39.37 dB with increasing residual orders is accompanied by diminishing marginal PSNR gains: 1st order (+5.30 dB), 2nd order (+3.62 dB), 3rd order (+2.46 dB), 4th order (+1.38 dB) [22].	128

List of Figures

3.23	Prediction quality comparison illustrating the complementary roles of RIFE and the proposed semantic autoencoder. Combining RIFE with the per-GOP trained semantic autoencoder achieves a 1.95 dB PSNR improvement over RIFE alone [22].	132
3.24	Rate-distortion performance of low- ($\circ$), medium- ($\triangle$), and high- ($\square$) complexity videos under bit errors, for the proposed codec (solid lines) and VTM 17.0 (dotted lines) at GOP size 32. Left to right: (a) PSNR, (b) SSIM, (c) VMAF [22].	132
3.25	Distribution of semantic prediction residual values for different video content complexities, with overlaid fitted Gaussian curves. High-complexity content exhibits a wide, flat distribution indicating large semantic distortion; low-complexity content demonstrates a highly concentrated distribution around zero indicating minimal semantic distortion [22].	134
4.1	End-to-end architecture of the proposed context-conditioned hybrid video codec.	142
4.2	Network architecture of the proposed semantic autoencoder.	144
4.3	Kernel density estimate of the element-wise difference $\Delta l_t = l_t - l_t^{(C)}$ between the target-frame latent and the RIFE-predicted context-frame latent. The sharply peaked distribution centred near zero confirms the strong correlation between the two latent representations, motivating the differential encoding strategy.	149
4.4	Average rate-distortion performance for the UVG dataset. The horizontal axis is bitrate (kbps) and the vertical axis is quality (PSNR in dB, SSIM, or VMAF). Each curve connects four operating points at $QP \in \{27, 32, 37, 42\}$	152
4.5	Average rate-distortion performance for the MCL-JCV dataset.	153
4.6	Average rate-distortion performance for the HEVC Class B and Class E datasets.	153
4.7	Average rate-distortion performance for 4K content from the Netflix Open Content dataset, assessed against VTM 23.0.	153

List of Figures

4.8	Visual comparison of progressive reconstruction quality across four representative test frames (rows). Each column shows: (left) the RIFE-based bidirectionally interpolated prediction, (centre) the context-conditioned semantic communication (SC) prediction from the autoencoder, and (right) the final reconstruction after VVC residual correction. PSNR and SSIM values are reported per frame. Quality improves monotonically from left to right across all examples, validating the complementary roles of motion-aligned interpolation, semantic encoding, and residual correction.	162
4.9	End-to-end architecture of the proposed context-aware semantic video codec with content-adaptive parameter overfitting.	168
4.10	Network architecture of the proposed tapered semantic autoencoder. Red-highlighted blocks indicate the overfittable transposed-convolution bias parameters; blue-highlighted blocks indicate the per-channel Learnable Scale multipliers.	171
4.11	Average rate–distortion performance for the UVG dataset. The horizontal axis is bitrate (kbps) and the vertical axis is quality (PSNR in dB, SSIM, or VMAF). Each curve connects four operating points at $QP \in \{27, 32, 37, 42\}$	176
4.12	Average rate–distortion performance for the MCL-JCV dataset.	176
4.13	Average rate–distortion performance for the HEVC Class E dataset. . .	176
4.14	Average rate–distortion performance for the HEVC Class B dataset. . .	176
4.15	Average rate–distortion performance for the 3D animation dataset (average of 5 sequences), evaluated against VTM 23.0.	177
4.16	Average rate–distortion performance for the CCTV surveillance dataset (average of 3 sequences), evaluated against VTM 23.0.	177

List of Figures

4.17	Visual quality comparison across six representative frames spanning all three evaluated content classes. Each row shows a frame at four stages: (1) RIFE-only interpolation, (2) CASVC semantic prediction, (3) proposed parameter-updated semantic prediction, and (4) final hybrid reconstruction with the VVC residual. Captions report (PSNR in dB, SSIM). Rows 1–4 are natural-video content; Row 5 is a CCTV surveillance frame; Row 6 is a 3D animation frame. The per-GOP parameter update Δ_g consistently raises semantic prediction quality by 1.6–3.8 dB PSNR and up to 0.06 SSIM over CASVC before the VVC residual is added.	185
5.1	End-to-end architecture of the proposed context-aware semantic HDR video codec.	195
5.2	Average rate–distortion performance of the proposed method and VTM 23.0 on the HDM-HDRV dataset [23]. Subfigures (a)–(c) show SDR-domain metrics (PSNR, SSIM, VMAF) computed on the tone-mapped 10-bit SDR output; subfigures (d)–(e) show PU21-encoded HDR perceptual metrics (PU-PSNR, PU-SSIM) computed on the recovered linear HDR signal. Each curve connects four operating points at $QP \in \{27, 32, 37, 42\}$	202
5.3	Average rate–distortion performance of the proposed method and VTM 23.0 on the JVET-HDRV dataset [23]. Subfigures (a)–(c) show SDR-domain metrics computed on the tone-mapped 10-bit SDR output; subfigures (d)–(e) show PU21-encoded HDR perceptual metrics. Each curve connects four operating points at $QP \in \{27, 32, 37, 42\}$	203

List of Figures

5.4	Ablation of per-frame versus fixed tone-mapping parameters on four representative frames of an HDM-HDRV sequence. Top row: Fixed baseline ($e = 500$, $\gamma = 2.0$) applied uniformly to all frames. Frame 1 (high-luminance scene) is saturated to white because the fixed multiplier is approximately $78\times$ larger than the automatic estimate ($e_1 = 6.40$). Frames 21, 41, and 61 (low-luminance scenes) are underexposed because the fixed multiplier is roughly half the required value ($e_t \approx 950\text{--}1010$). Bottom row: Automatic per-frame estimator. Each frame is individually normalised to the correct mid-tone level, preserving highlight detail in the bright scene and shadow detail in the dark scenes.	205
6.1	End-to-end SMDC pipeline. The three colours represent the three independent channel paths for transmitting the three descriptions.	216
6.2	SMDC autoencoder architecture with interleaved channel-wise split into three descriptions. The encoder downsamples the input to a $9 \times 16 \times 24$ latent tensor; the decoder upsamples a concatenation (or zero-padded subset) of the three $9 \times 16 \times 8$ descriptions.	218
6.3	Performance comparison of the proposed SMDC framework against baseline methods across three channel models under three-path MDC transmission. (a)–(c) PSNR versus combined SNR for AWGN, Rayleigh, and Rician channels. (d)–(f) SSIM versus combined SNR.	225
6.4	Performance comparison across single-channel (non-MDC) configurations. (a)–(c) PSNR versus SNR for AWGN, Rayleigh, and Rician channels. (d)–(f) SSIM versus SNR.	226
6.5	Conventional pipeline used for the single-channel (non-MDC) baseline comparison. The entire image is source-coded, channel-coded with polar codes, and transmitted over a single path; the receiver applies the same codec to decode the received bitstream directly without any multi-description fusion.	226

List of Figures

6.6	Visual comparison of SMDC reconstruction quality under different reception scenarios. From left to right: (a) all three descriptions received; (b) any two descriptions received; (c) single description without bit errors; (d) single description at $\text{BER} = 0.001$; (e) single description at $\text{BER} = 0.01$	234
6.7	End-to-end SMDC-V pipeline for video transmission over three independent wireless channels. Boundary frames (F_m, F_{m+8}) are encoded by the MDC Intra-Frame Semantic Encoder; the seven intermediate frames are encoded by the MDC Inter-Frame Semantic Encoder conditioned on RIFE-interpolated context. Three descriptions per frame are independently entropy-coded, channel-coded, and transmitted.	236
6.8	MDC Inter-Frame Semantic Encoder and Decoder. The encoder computes the differential latent $\Delta l_{m+k} = l_{m+k} - l_{m+k}^C$ and splits it into three descriptions via interleaved channel-wise assignment. The context-conditioned decoder recovers $\hat{l}_{m+k}$ from the received descriptions and concatenates it with the locally recomputed context C_{m+k} to reconstruct $\hat{F}_{m+k}$	240
6.9	Detailed network architecture of the MDC Inter-Frame Semantic Encoder and context-conditioned decoder. The context-conditioning residual block at the final upsampling stage contributes approximately 62% of the total decoder FLOPs, for a decoder total of 52.22 GFLOPs. . . .	240
6.10	Multi-channel MDC performance of SMDC-V against VVenC MDC across three channel models. (a)–(c) PSNR versus combined SNR. (d)–(f) SSIM versus combined SNR. (g)–(i) VMAF versus combined SNR.	245
6.11	Single-channel (non-MDC) performance of SMDC-V against VVenC across three channel models. (a)–(c) PSNR versus SNR. (d)–(f) SSIM versus SNR. (g)–(i) VMAF versus SNR.	247

List of Figures

7.1	System architecture for VVC-based DNN parameter compression. The transmitter trains a two-model semantic autoencoder per GOP, reshapes the decoder weights and biases as image frames, and encodes them using VVenC. The receiver decodes the VVC bitstream to retrieve decoder parameters and reconstructs each video frame from the transmitted latent vectors [24].	256
7.2	Spatial arrangement of decoder (a) weight values and (b) bias values as grayscale image frames for one GOP. The structured spatial patterns confirm that the weight and bias matrices carry spatial correlation that VVC intra- and inter-prediction tools can exploit for efficient compression [24].	257
7.3	Average rate–distortion performance of the proposed VVC-based DNN parameter compression method against NNCodec over the ten test videos for (a) PSNR, (b) SSIM, and (c) VMAF [24].	259
7.4	Proposed system for temporal prediction of DNN parameters. The system comprises four components: context encoder/decoder (key frame extraction and HEIF compression), semantic encoder/decoder (per-frame overfitting autoencoder), temporal prediction and residual estimation, and NNCodec-based residual compression/decompression [25].	263
7.5	Network model for temporal DNN parameter prediction. <i>sequential_120</i> denotes the encoder layers and <i>sequential_121</i> the decoder layers [25]. .	264
7.6	Distribution of decoder weight and bias values when trained on two consecutive video frames. The near-identical distributions in (a, b) and (d, e), and the low-amplitude difference signals in (c, f) (plotted at higher resolution), confirm strong inter-frame temporal correlation in the DNN parameter space [25].	267
7.7	Per-video rate–distortion comparison of the proposed temporal DNN parameter prediction method against the NNCodec-only baseline for the ten test videos [25].	267

List of Figures

A.1	Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V01–V05, GOP size 8.	289
A.2	Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V06–V10, GOP size 8.	290
A.3	Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V11–V15, GOP size 8.	291
A.4	Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V16–V20, GOP size 8.	292
A.5	Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V01–V05, GOP size 16.	293
A.6	Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V06–V10, GOP size 16.	294
A.7	Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V11–V15, GOP size 16.	295
A.8	Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V16–V20, GOP size 16.	296
A.9	Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V01–V05, GOP size 32.	297
A.10	Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V06–V10, GOP size 32.	298
A.11	Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V11–V15, GOP size 32.	299
A.12	Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V16–V20, GOP size 32.	300

List of Tables

1.1	How Semantic Communication Mitigates the Limitations of Conventional and Learned Video Codecs.	27
2.1	Summary of identified research gaps and corresponding thesis contributions	79
3.1	Parameters used to configure the VVC, HEVC, and AVC baseline codecs [1].	90
3.2	Average BD-rates for the proposed system relative to AVC, HEVC, and VVC for GOP size 8 [1].	94
3.3	Average BD-rates for the proposed system relative to AVC, HEVC, and VVC for GOP size 16 [1].	94
3.4	Average BD-rates for the proposed system relative to AVC, HEVC, and VVC for GOP size 32 [1].	96
3.5	Comparison of average PSNR of reconstructions produced by the transmitter-side and receiver-side remotely trained AE decoder layers [1].	98
3.6	Comparison of average rate-distortion performance of the transmitter-side (TX) and remotely trained receiver-side (RX) AE decoders across different GOP sizes [1].	99
3.7	Effect of key frame QP on bitrate and quality for high SI/TI videos, with residual QP fixed at 22 [1].	99
3.8	Effect of key frame QP on bitrate and quality for low SI/TI videos, with residual QP fixed at 22 [1].	100
3.9	Average residual bitrate, PSNR, and SSIM for different GOP sizes used in the VVC residual encoding stage [1].	101

List of Tables

3.10	Effect of spatial resolution on the average PSNR of the proposed system [1].	101
3.11	Average rate–distortion performance comparison between the source and destination semantic decoders, confirming synchronisation under remote training [22].	114
3.12	BD-rate gains attributable to DeepCABAC-based latent entropy coding relative to VTM 17.0 [22].	116
3.13	PSNR BD-rate (%) comparison in YUV420 colour space with intra period 32 relative to VTM 17.0. Lower values indicate better rate–distortion performance [22].	125
3.14	Average composition of bits per frame for a 720p video from the proposed codec at GOP size 32, showing the relative contributions of the key frame, residual, and latent vector components [22].	126
3.15	Ablation study on the effect of residual frame order in the proposed codec for GOP size 32, with per-order marginal contributions [22]. . . .	127
3.16	Ablation study comparing unidirectional and bidirectional frame prediction modes relative to VTM 17.0 [22].	127
3.17	Ablation study on the hidden layer dimension of the semantic encoder [22].	129
3.18	Ablation study on the activation function of the semantic encoder [22]. .	129
3.19	Ablation study on the latent vector dimension of the semantic encoder [22].	129
3.20	Ablation study on rate–distortion and computational trade-offs at matched bitrate according to semantic block size [22].	130
3.21	Ablation study comparing joint YUV420 and separate channel processing strategies [22].	131
3.22	Comprehensive runtime and complexity comparison for 720p YUV420 video encoding and decoding [22].	134
4.1	Summary statistics of per-sequence PSNR BD-rates (%) within each 1080p dataset, relative to VTM 17.0. Every sequence achieves a negative BD-rate.	154

List of Tables

4.2	PSNR BD-rate (%) comparison in YUV420 colour space for 1080p datasets. Reference: VTM 17.0 (0%). Negative values indicate bitrate savings (lower is better). BD-rate figures for DCVC-DC, DCVC-FM, DCVC-FM fp16, and DCVC-RT fp16 are reproduced from [26]. Enc. fps and Dec. fps denote encoding and decoding throughput in frames per second on an NVIDIA A100 GPU.	155
4.3	Component-wise ablation: average PSNR BD-rate (%) relative to VTM 17.0. Negative values indicate bitrate savings.	158
4.4	Bitrate composition: percentage of total bitrate contributed by each coding component at different QP values.	158
4.5	Sensitivity analysis of the TI threshold for motion-adaptive model selection. Values are semantic PSNR (dB). Bold entries indicate the selected model for each content category.	159
4.6	Architecture ablation: semantic PSNR (dB) as a function of latent dimensionality and spatial block resolution. Bold entries indicate the proposed configuration.	160
4.7	Incremental contribution of hierarchical residual layers in a GOP of size 32, averaged across all 1080p test sequences at QP 32. Δ PSNR and Δ Bitrate denote marginal gains from each additional order; $\eta = \Delta$ PSNR/ Δ Bitrate is marginal coding efficiency (dB/Mbit).	161
4.8	PSNR BD-rate (%) comparison in YUV420 colour space for all 1080p datasets. Reference: VTM 17.0 (0%). Encoding and decoding throughput in frames per second (fps) on an NVIDIA A100 GPU. BD-rate figures for DCVC variants reproduced from [26].	177
4.9	SSIM BD-rate (%) comparison in YUV420 colour space for all 1080p datasets. Reference: VTM 17.0 (0%).	178
4.10	VMAF BD-rate (%) comparison in YUV420 colour space for all 1080p datasets. Reference: VTM 17.0 (0%).	178

List of Tables

4.11	BD-rate (%) comparison against VTM 23.0 for CCTV (average of 3 sequences) and 3D animation (average of 5 sequences) content. All sequences encoded at $QP \in \{27, 32, 37, 42\}$ with intra period = -1 (random access).	178
4.12	Bitstream composition ablation: per-component percentage change of the proposed codec relative to CASVC at each QP operating point (average of 2 sequences from each dataset). The $R_{\text{param}}/R_{\text{total}}$ column reports the parameter sub-stream as a fraction of the proposed codec's total bitrate (CASVC has no parameter sub-stream). Quality columns report percentage change relative to CASVC; positive values indicate improvement.	182
5.1	BD-rate (%) of the proposed method against VTM 23.0 on the HDM-HDRV dataset [23]. PSNR, SSIM, and VMAF are computed on the tone-mapped 10-bit SDR output; PU-PSNR and PU-SSIM are computed on the recovered HDR signal via PU21 encoding [27]. Negative values denote bitrate savings (lower is better).	202
5.2	BD-rate (%) of the proposed method against VTM 23.0 on the JVET-HDRV dataset [23]. PSNR, SSIM, and VMAF are computed on the tone-mapped 10-bit SDR output; PU-PSNR and PU-SSIM are computed on the recovered HDR signal via PU21 encoding [27]. Negative values denote bitrate savings (lower is better).	202
5.3	Average BD-rate (%) across HDM-HDRV and JVET-HDRV datasets against VTM 23.0. Negative values denote bitrate savings (lower is better).	203
5.4	Component-wise reconstruction quality on the evaluated HDR dataset in 10-bit YUV420 format. Higher PSNR is better.	206
6.1	Comparison of approaches for robust image transmission over wireless channels.	215
6.2	Minimum combined SNR (dB) to maintain PSNR > 30 dB and SSIM > 0.80. Lower values indicate greater robustness.	228

List of Tables

6.3	Performance comparison: interleaved versus contiguous splitting (AWGN). PSNR and SSIM reported as mean $\pm$ standard deviation over 100 test images.	231
6.4	Impact of number of descriptions on performance (AWGN). PSNR and SSIM reported as mean $\pm$ standard deviation.	232
6.5	Minimum SNR (dB) to achieve each quality threshold for SMDC-V and VVenC under multi-channel MDC and single-channel configurations. Lower values indicate greater channel robustness. Combined SNR is used for multi-channel; single-path SNR is used for single-channel. Values are approximate to the nearest dB.	246
7.1	BD-rates (%) of the proposed VVC-based DNN parameter compression method versus NNCodec. Negative values indicate bitrate savings at equivalent quality.	259

Acronyms and Abbreviations

3D Three-Dimensional

4G Fourth-Generation

6G Sixth-Generation

ACM Association for Computing Machinery

ACES Academy Color Encoding System

AE Autoencoder

ALF Adaptive Loop Filter

AMC Affine Motion Compensation

AMVP Advanced Motion Vector Prediction

AOM Alliance for Open Media

AV1 AOMedia Video 1

AV2 AOMedia Video 2

AVM AOMedia Video Model

AVC Advanced Video Coding

AWGN Additive White Gaussian Noise

BD Bjøntegaard Delta

Acronyms and Abbreviations

BDOF Bi-directional Optical Flow

BERT Bidirectional Encoder Representations from Transformers

BER Bit Error Rate

BPG Better Portable Graphics

BPSK Binary Phase Shift Keying

CABAC Context-Adaptive Binary Arithmetic Coding

CAPO Content-Adaptive Parameter Overfitting

CASVC Context-Aware Semantic Video Coding

CAVLC Context-Adaptive Variable-Length Coding

CCTV Closed-Circuit Television

CD-ROM Compact Disc Read-Only Memory

CCLM Cross-Component Linear Model

CDEF Constrained Directional Enhancement Filter

CIE International Commission on Illumination

CIF Common Intermediate Format

CNN Convolutional Neural Network

CPU Central Processing Unit

CTU Coding Tree Unit

CU Coding Unit

DBPSK Differential Binary Phase Shift Keying

DCI Digital Cinema Initiatives

Acronyms and Abbreviations

DCT Discrete Cosine Transform

DCVC Deep Contextual Video Compression

DCVC-DC Deep Contextual Video Compression with Diverse Contexts

DCVC-FM Deep Contextual Video Compression with Feature Modulation

DCVC-RT Real-Time Deep Contextual Video Compression

DeepCABAC Deep Context-Adaptive Binary Arithmetic Coding

DNN Deep Neural Network

DPCM Differential Pulse-Code Modulation

DVC Deep Video Compression

DVD Digital Versatile Disc

E2E End-to-End

eMBB Enhanced Mobile Broadband

ECM Enhanced Compression Model

EXR OpenEXR High-Dynamic-Range Image File Format

FEC Forward Error Correction

FLOP Floating Point Operation

FPGA Field-Programmable Gate Array

GAN Generative Adversarial Network

GOP Group of Pictures

GPM Geometric Partitioning Mode

GPT Generative Pre-trained Transformer

Acronyms and Abbreviations

GPU Graphics Processing Unit

HDTV High-Definition Television

HDR High Dynamic Range

HD High Definition

HEIF High Efficiency Image Format

HEVC High Efficiency Video Coding

HLG Hybrid Log-Gamma

HM HEVC Test Model

HMVP History-based Motion Vector Prediction

IDR Instantaneous Decoder Refresh

IoT Internet of Things

IoVT Internet of Video Things

ISP Intra Sub-Partition

IEEE Institute of Electrical and Electronics Engineers

IEC International Electrotechnical Commission

INR-MDSQC Implicit Neural Representation-Based Multiple Description Scalar Quantisation

ISO International Organization for Standardization

ITU International Telecommunication Union

ITU-R International Telecommunication Union Radiocommunication Sector

ITU-T International Telecommunication Union Telecommunication Standardization Sector

Acronyms and Abbreviations

JCCR Joint Coding of Chroma Residuals

JCT-VC Joint Collaborative Team on Video Coding

JPEG Joint Photographic Experts Group

JSCC Joint Source–Channel Coding

JVET Joint Video Experts Team

JVT Joint Video Team

LeakyReLU Leaky Rectified Linear Unit

LMDC Learned Multiple Description Coding

LPIPS Learned Perceptual Image Patch Similarity

LFNST Low-Frequency Non-Separable Transform

LLM Large Language Model

LS Learnable Scale

LTE Long-Term Evolution

M2H Machine-to-Human

M2M Machine-to-Machine

MAC Multiply-Accumulate Operation

MD5 Message-Digest Algorithm 5

MIMO Multiple-Input Multiple-Output

MDC Multiple Description Coding

MIP Matrix-based Intra Prediction

MLP Multilayer Perceptron

Acronyms and Abbreviations

mMTC massive Machine-Type Communication

MPEG Moving Picture Experts Group

MSE Mean Squared Error

MS-SSIM Multi-Scale Structural Similarity Index

MTS Multiple Transform Set

MV-HEVC Multiview HEVC

NR New Radio

NLP Natural Language Processing

NNC Neural Network Compression

NNR Neural Network Compression and Representation

OFDM Orthogonal Frequency-Division Multiplexing

PCA Principal Component Analysis

PQ Perceptual Quantizer

PROF Prediction Refinement with Optical Flow

PSNR Peak Signal-to-Noise Ratio

PSTN Public Switched Telephone Network

PU Perceptually Uniform

PU-PSNR Perceptually Uniform Peak Signal-to-Noise Ratio

PU-SSIM Perceptually Uniform Structural Similarity Index Measure

QCIF Quarter Common Intermediate Format

QP Quantisation Parameter

Acronyms and Abbreviations

QTMT Quad Tree with Nested Multi-Type Tree

ReLU Rectified Linear Unit

RAFT Recurrent All-Pairs Field Transforms

RD Rate–Distortion

RGB Red, Green, Blue

RIFE Real-Time Intermediate Flow Estimation

ROI Region of Interest

RX Receiver

SAO Sample Adaptive Offset

SC Semantic Communication

SCC Screen Content Coding

SDR Standard Dynamic Range

SEI Supplemental Enhancement Information

SHVC Scalable HEVC

SI Spatial Information

SMDC Semantic Multiple Description Coding

SMDC-V Semantic MDC Video

SNR Signal-to-Noise Ratio

SSF Scale Space Flow

SSIM Structural Similarity Index Measure

TI Temporal Information

Acronyms and Abbreviations

TU Transform Unit

TX Transmitter

UHD Ultra High Definition

URLLC Ultra-Reliable Low-Latency Communication

VAE Variational Autoencoder

VCEG Video Coding Experts Group

VCM Video Coding for Machines

VMAF Video Multi-Method Assessment Fusion

VQEG Video Quality Experts Group

VTM VVC Test Model

VVC Versatile Video Coding

WCG Wide Colour Gamut

XR Extended Reality

YUV Luma and Chrominance Colour Representation

YUV420 YUV 4:2:0 Chroma-Subsampled Format

Acknowledgements

I would like to express my sincere gratitude to my principal supervisor, Prof. Anil Fernando, for his invaluable guidance, expertise, and unwavering support throughout my PhD journey. His insightful feedback, thoughtful discussions, and constant encouragement have played a vital role in shaping my research and improving this thesis. I am deeply grateful for the opportunity to have worked under his supervision. I would also like to thank my co-supervisor, Prof. Feng Dong, for his valuable advice, guidance, and support throughout my studies.

My deepest gratitude goes to my parents for the countless sacrifices they have made to support my education and aspirations. Their unconditional love, encouragement, and belief in me gave me the strength to persevere through every challenge. I will always be grateful for everything they have done for me. I am also sincerely thankful to my sister for her constant encouragement, kindness, and support throughout this journey.

Finally, I would like to thank my colleagues and everyone who supported, encouraged, and contributed to my PhD journey. Their assistance, friendship, and encouragement made this experience both rewarding and memorable.

Publications

Journal Articles (Accepted as First Author)

1. **P. Samarathunga**, Y. Ganearachchi, T. Fernando, A. Jayasingam, I. Alahapperuma, and A. Fernando, “A Semantic Communication and VVC Based Hybrid Video Coding System,” *IEEE Access*, vol. 12, pp. 79202–79224, 2024, doi: 10.1109/ACCESS.2024.3399174.
2. **P. Samarathunga**, Y. Ganearachchi, T. Fernando, A. Jayasingam, T. Sivalingam, N. Rajatheva, and A. Fernando, “Enhanced Semantic Communications and VVC-Based Hybrid Video Coding System,” *IEEE Access*, vol. 14, pp. 50755–50778, 2026, doi: 10.1109/ACCESS.2026.3676702.
3. **P. Samarathunga**, Y. Ganearachchi, T. Fernando, A. Jayasingam, T. Sivalingam, N. Rajatheva, and A. Fernando, “Context-Aware Semantic Video Coding Using Pretrained Motion-Adaptive Models and Bidirectional Prediction,” *IEEE Access*, pp. 1–1, 2026, doi: 10.1109/ACCESS.2026.3686329.
4. **P. Samarathunga**, N. Pollwaththage, T. Fernando, U. Jayasinghe, N. Ihalagamage, Y. Ganearachchi, and A. Fernando, “Semantic and Technical Noise Modeling in Semantic Image Communication Over Wireless Channels,” *IEEE Access*, vol. 13, pp. 150872–150900, 2025, doi: 10.1109/ACCESS.2025.3600320.
5. **P. Samarathunga**, H. Rezaei, M. Lokumarambage, T. Sivalingam, N. Rajatheva, and A. Fernando, “An Autoencoder-Based Task-Oriented Semantic Communication System for M2M Communication,” *Algorithms*, vol. 17, no. 11, Art. no. 492, 2024, doi: 10.3390/a17110492.

Publications

6. **P. Samarathunga**, T. Fernando, V. Gowrisetty, T. Atulugama, and A. Fernando, “Autoencoder Based Image Quality Metric for Modelling Semantic Noise in Semantic Communications,” *Electronics Letters*, vol. 60, no. 4, Art. no. e13115, 2024, doi: 10.1049/ell2.13115.
7. **P. Samarathunga**, U. Jayasinghe, Y. Ganearachchi, T. Fernando, T. Sivalingam, N. Rajatheva, and A. Fernando, “Semantic Communication Based Multiple Description Coding for Robust Image Transmission over Wireless Channels,” *IEEE Access*, pp. 1–1, 2026, doi: 10.1109/ACCESS.2026.3716070.

Journal Articles (Under Review as First Author)

8. **P. Samarathunga**, Y. Ganearachchi, T. Fernando, A. Jayasingam, and A. Fernando, “Context-Aware Semantic Video Coding via Content-Adaptive Parameter Overfitting,” *IEEE Access*, 2026, *under review*.
9. **P. Samarathunga**, Y. Ganearachchi, T. Fernando, A. Jayasingam, and A. Fernando, “Context-Aware Semantic Video Coding via Content-Adaptive Parameter Overfitting for High Dynamic Range Videos,” *IEEE Access*, 2026, *under review*.
10. **P. Samarathunga**, U. Jayasinghe, Y. Ganearachchi, T. Fernando, T. Sivalingam, N. Rajatheva, and A. Fernando, “Semantic Communication Based Multiple Description Coding for Robust Video Transmission over Wireless Channels,” *IEEE Access*, 2026, *under review*.

Conference Papers (Accepted as First Author)

11. **P. Samarathunga**, T. Fernando, Y. Ganearachchi, N. Pollwatthage, and A. Fernando, “Semantic Communication and Residual Deep Learning-Driven Bit-plane Coding for Scalable Image Compression and Communications,” in *Proc. IEEE Int. Conf. Image Process. Workshops (ICIPW)*, 2025, pp. 568–573, doi: 10.1109/ICIPW68931.2025.11385999.

12. **P. Samarathunga**, I. Alahapperuma, T. Fernando, Y. Ganearachchi, and A. Fernando, “Deep Neural Network Compression for Semantic Video Communications,” in *Proc. IEEE Int. Conf. Consumer Electron. (ICCE)*, 2025, pp. 1–4, doi: 10.1109/ICCE63647.2025.10930011.
13. **P. Samarathunga**, U. Jayasinghe, Y. Ganearachchi, T. Fernando, N. Senanayake, and A. Fernando, “Semantic Communication Based Lossless Distributed Video Coding for Error Prone Channels,” in *Proc. IEEE Int. Conf. Consumer Electron. (ICCE)*, 2025, pp. 1–4, doi: 10.1109/ICCE63647.2025.10929820.
14. **P. Samarathunga**, Y. Ganearachchi, T. Fernando, I. Alahapperuma, and A. Fernando, “Semantic Communication Based Video Coding Using Temporal Prediction of Deep Neural Network Parameters,” in *Proc. IEEE Gaming, Entertainment, and Media Conf. (GEM)*, 2024, pp. 1–6, doi: 10.1109/GEM61861.2024.10585783.
15. **P. Samarathunga**, Y. Ganearachchi, T. Fernando, L. Thanippulige, and A. Fernando, “Semantic Communication Based Complexity Scalable Image Transmission System for Resource Constrained Devices,” in *Proc. IEEE Gaming, Entertainment, and Media Conf. (GEM)*, 2024, pp. 1–6, doi: 10.1109/GEM61861.2024.10585490.
16. **P. Samarathunga**, V. Gowrisetty, T. Fernando, Y. Ganearachchi, and A. Fernando, “Region of Interest Scalable Image Compression Using Semantic Communications,” in *Proc. IEEE Int. Conf. Consumer Electron. (ICCE)*, 2024, pp. 1–5, doi: 10.1109/ICCE59016.2024.10444357.
17. **P. Samarathunga**, Y. Ganearachchi, and A. Fernando, “Video Compression by Chroma Prediction Using Semantic Communications,” in *Proc. IEEE Int. Conf. Consumer Electron. (ICCE)*, 2024, pp. 1–4, doi: 10.1109/ICCE59016.2024.10444312.
18. **P. Samarathunga**, U. Jayasinghe, Y. Ganearachchi, A. Jayasingam, and A. Fernando, “Semantic Communication for 3-D Point Clouds via Channel-Aware

Publications

Sparse Convolutional Autoencoders,” in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, 2026, *in press*.

19. **P. Samarathunga**, Y. Ganearachchi, T. Fernando, and A. Fernando, “Error Resilient Rate-Scalable Image Compression Using Semantic Communications,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2026, pp. 1–6, doi: 10.1109/ICC59461.2026.11587078.
20. **P. Samarathunga**, I. Alahapperuma, Y. Ganearachchi, T. Fernando, and A. Fernando, “Neural Network Compression for Video Transmission Using Semantic Communications and VVC,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2026, pp. 1–6, doi: 10.1109/ICC59461.2026.11587780.

Other Publications (Accepted as Co-Author)

21. U. Jayasinghe, **P. Samarathunga**, N. Pollwaththage, T. Fernando, and A. Fernando, “Quantum Communication for Semantic Communication Based Image Transmission,” *IEEE Trans. Veh. Technol.*, pp. 1–10, 2026, doi: 10.1109/TVT.2026.3663843.
22. U. Jayasinghe, T. Fernando, Y. Ganearachchi, **P. Samarathunga**, and A. Fernando, “Quantum Communication-Based Image Transmission With Transmit and Receive Diversity in MIMO Communication Systems,” *IEEE Trans. Consumer Electron.*, vol. 71, no. 2, pp. 2500–2507, 2025, doi: 10.1109/TCE.2025.3559634.
23. U. Jayasinghe, N. Pollwaththage, Y. Ganearachchi, **P. Samarathunga**, T. Fernando, and A. Fernando, “Quantum Communication Based Image Transmission Over Error-Prone Channels with Three-Qubit Stabilizer Code,” in *Proc. IEEE Int. Conf. Consumer Electron. (ICCE)*, 2025, pp. 1–6, doi: 10.1109/ICCE63647.2025.10930032.
24. U. Jayasinghe, **P. Samarathunga**, N. Pollwaththage, Y. Ganearachchi, T. Fernando, and A. Fernando, “Quantum Communication for Video Transmission

Publications

- Over Error-Prone Channels,” *IEEE Trans. Consumer Electron.*, vol. 71, no. 1, pp. 1148–1155, 2025, doi: 10.1109/TCE.2025.3552930.
25. U. Jayasinghe, **P. Samarathunga**, Y. Ganearachchi, T. Fernando, and A. Fernando, “Adaptive Coding-Based Quantum Communication System for Image Transmission,” *Electronics Letters*, vol. 61, no. 1, Art. no. e70299, 2025, doi: 10.1049/ell2.70299.
26. U. Jayasinghe, **P. Samarathunga**, T. Fernando, Y. Ganearachchi, and A. Fernando, “Image Transmission Over Quantum Communication Systems With Three-Qubit Error Correction,” *Electronics Letters*, vol. 61, no. 1, Art. no. e70205, 2025, doi: 10.1049/ell2.70205.
27. U. Jayasinghe, **P. Samarathunga**, T. Fernando, and A. Fernando, “Transmit and Receive Diversity in MIMO Quantum Communication for High-Fidelity Video Transmission,” *Algorithms*, vol. 18, no. 7, Art. no. 436, 2025, doi: 10.3390/a18070436.
28. U. Jayasinghe, **P. Samarathunga**, Y. Ganearachchi, T. Fernando, and A. Fernando, “Quantum Communications for Image Transmission Over Error-Prone Channels,” *Electronics Letters*, vol. 60, no. 15, Art. no. e13300, 2024, doi: 10.1049/ell2.13300.
29. N. Pollwaththage, Y. Ganearachchi, **P. Samarathunga**, J. El Gemayel, and A. Fernando, “Multi-Frame Optimized Virtual Reference Frame for Low-Delay Predictive Video Compression,” in *Proc. IEEE Int. Conf. Consumer Electron. (ICCE)*, 2026, pp. 1–6, doi: 10.1109/ICCE67443.2026.11449762.
30. T. Fernando, N. Pollwaththage, U. Jayasinghe, **P. Samarathunga**, J. El Gemayel, and A. Fernando, “Beyond Disparity: A Semantic Communication Framework for Stereoscopic Video Coding,” in *Proc. IEEE Int. Conf. Ind. Inf. Syst. (ICIIS)*, 2026, pp. 85–90, doi: 10.1109/ICIIS69028.2026.11450610.

Publications

Chapter 1

Introduction

1.1 Background and Motivation

The global communications infrastructure is changing rapidly, driven by the fast growth of video data and the rising demand for bandwidth-intensive applications. Video traffic already accounts for more than 82% of all global internet bandwidth [28, 29], and this share is expected to grow further as emerging applications such as the metaverse, Extended Reality (XR), the Internet of Video Things (IoVT), holographic telepresence, autonomous vehicle telemetry, and sixth-generation (6G) wireless networks become mainstream [30, 31]. These applications require capabilities that go well beyond what current communication infrastructures can reliably deliver: ultra-high resolutions (8K and beyond), high frame rates (120–300 frames per second), sub-millisecond latency, massive device connectivity, and robust operation under severe channel impairments. As a result, the efficient, robust, and intelligent compression and transmission of video data, together with the closely related transmission of still images, has become one of the key technological challenges of this decade.

Video coding is central to digital visual communication. Its role is to compress raw video data into compact bitstreams for storage and transmission. Over more than three decades, two main organisations have led progress in this field: the International Telecommunication Union Telecommunication Standardization Sector (ITU-T), which developed the H.26x family starting with H.261 [32] (first ratified in 1988), and the

Moving Picture Experts Group (MPEG), which introduced the MPEG series beginning with MPEG-1 [33]. For many years, these two families developed independently, with ITU-T focusing on real-time communication (H.261, H.263 [34]) and MPEG targeting storage and broadcast (MPEG-1, MPEG-2 [35]). In the early 2000s, the two bodies joined forces to develop the H.264/Advanced Video Coding (AVC) standard [36], marking the start of a long-standing collaboration. Since then, ITU-T and MPEG have worked together on successive generations, leading to H.265/High Efficiency Video Coding (HEVC) [37] in 2013 and H.266/Versatile Video Coding (VVC) [38], standardised in 2020, which is the current leading conventional video coding standard. In parallel, the Alliance for Open Media has released the royalty-free AOMedia Video 1 (AV1) [39] codec as a competing standard. Each new generation has achieved significant coding gains, typically a 40–50% bitrate reduction at equivalent quality compared to its predecessor [37, 38], but with a notable increase in encoder complexity.

All conventional video coding standards from H.261 to H.266/VVC share a common hybrid block-based architecture [38, 40]. The encoder first partitions each frame into rectangular blocks. For intra-coded frames, it predicts each block from previously coded neighbouring pixels within the same frame through spatial prediction. For inter-coded frames, it performs motion estimation to find similar blocks in previously decoded reference frames and uses motion vectors to generate a temporal prediction. The encoder subtracts this prediction from the original block to form a residual, which is then transformed using the Discrete Cosine Transform (DCT) to concentrate energy into a small number of coefficients. These transform coefficients are quantized, a lossy step that discards perceptually less important information, and finally entropy coded using Context-Adaptive Binary Arithmetic Coding (CABAC) into a compressed bitstream [41]. This hybrid approach exploits spatial, temporal, and statistical redundancies in natural video and has been refined over decades to deliver remarkable compression efficiency.

Despite its success, conventional video coding has several key limitations [40–42]. First, linear transforms such as the DCT do not represent sharp edges, textures, and complex patterns in natural and synthetic images well [43, 44]. Second, encoding tools

are designed by experts and cannot adapt to all types of real-world content; each new standard adds dozens of such tools, causing rapid growth in encoder complexity. VVC, for example, incorporates over 100 coding tools [38] and its encoder is typically 5–10 times more complex than HEVC, making real-time software encoding increasingly difficult in practice [40]. Third, conventional entropy-coded bitstreams are very sensitive to channel errors: a single bit error can make an entire frame impossible to decode, producing the well-known cliff effect. Fourth, these codecs are static and cannot learn from data or adapt to the properties of a particular video sequence without explicit side information. Finally, all conventional codecs including VVC are entirely semantically agnostic: they allocate bits according to signal statistics alone, with no capacity to distinguish perceptually or task-relevant content from irrelevant background, or to optimise for downstream machine-inference objectives [41]. These limitations together have led researchers to rethink the video compression and transmission paradigm.

1.1.1 Learned Video Coding Using Deep Neural Networks

In response to the limitations of handcrafted, linear video codecs, a parallel research direction emerged in the mid-2010s: end-to-end learned compression [45,46]. Instead of designing transforms, motion estimators, and entropy models by hand, this approach replaces them with Deep Neural Networks (DNNs) trained directly from data using rate–distortion loss functions. The key idea is that a neural network can learn to represent natural video more efficiently than any fixed, handcrafted pipeline, and that this representation can be jointly optimised for both compression and reconstruction quality.

Early works focused on image compression, with architectures such as the variational hyperprior autoencoder [47] showing competitive performance against Joint Photographic Experts Group (JPEG) [48]. A notable early step for video was Deep Video Compression (DVC) [49], which replaced conventional motion estimation and residual coding modules with learned counterparts. Later work introduced scale-space flow for more accurate motion compensation and learned entropy models that adapt to the statistics of latent representations [50]. The Deep Contextual Video Compression

(DCVC) family [51] was a major step forward by introducing temporal context mining, where previously decoded frames provide useful information for predicting and coding current frames. DCVC and its successors, Deep Contextual Video Compression with Feature Modulation (DCVC-FM) [52] and Deep Contextual Video Compression with Diverse Contexts (DCVC-DC) [53], achieved rate–distortion performance that matches, and in some cases exceeds, H.266/VVC on standard benchmarks. Another notable approach is hierarchical recurrent learned video coding [54], which uses recurrent feedback to capture longer-range temporal dependencies that single-step prediction cannot access.

Despite this progress, end-to-end learned video codecs are still not a complete solution. Several key limitations remain:

1. **Training instability and poor generalization.** Learned video codecs are difficult to train. Training a single network to learn motion estimation, residual coding, and entropy modelling at the same time requires careful tuning of multiple loss terms and often leads to unstable training. Models trained on one dataset also tend to perform poorly on other content types, such as animations, screen recordings, or user-generated video with different properties.
2. **High computational complexity at inference.** Conventional codecs use more processing at the encoder and less at the decoder, but many learned codecs reduce or reverse this balance. The decoder, which runs on devices such as smartphones or smart televisions, can be nearly as slow as the encoder. Real-time decoding of models such as DCVC remains difficult on mobile and embedded devices, limiting practical use.
3. **Inconsistent performance across content types.** Learned codecs show uneven compression performance depending on the video content. For simple, predictable motion such as video calls, they can outperform VVC. However, for complex, fast, or heavily occluded motion such as sports footage, they often perform worse than conventional codecs refined over many years of research [53].
4. **Lack of standardisation and interoperability.** Conventional codecs use stan-

standardised bitstream formats, so any compatible decoder can decode any compatible bitstream. Learned codecs are tied to the specific model used for encoding, and a bitstream produced by one model cannot be decoded by a different model. This lack of interoperability is a major barrier to wide-scale deployment.

5. **Persistent channel fragility.** Most learned video codecs assume that the channel is error-free after entropy decoding. They do not address the cliff effect; a corrupted value or a lost packet can cause severe degradation in the decoded video. Many learned systems also lack the error-handling tools developed in conventional codecs, such as synchronisation frames, slice partitioning, and data partitioning.
6. **Limited semantic awareness.** Even advanced learned codecs aim to minimise pixel-level or perceptual error metrics such as Peak Signal-to-Noise Ratio (PSNR), Multi-Scale Structural Similarity Index (MS-SSIM), or Video Multi-Method Assessment Fusion (VMAF). They do not consider the meaning or task relevance of visual content. A learned codec may spend bits on a texture that is not needed for a downstream task while missing important details such as faces, text, or traffic signs. This is a key limitation in Machine-to-Machine (M2M) and IoVT applications, where the receiver is a machine rather than a human viewer [55].

These limitations show that neither conventional codecs nor learned codecs can fully meet the needs of low-bitrate, dynamic, and error-prone communication settings. The main problems are their lack of semantic awareness, their sensitivity to channel impairments, and the uneven performance of learned codecs across different content types. These shared weaknesses motivate a more basic change in direction: semantic communication, which focuses on preserving meaning rather than pixel-level accuracy and moves beyond the limits of both traditional and learning-based coding.

1.1.2 Semantic Communication: A Paradigm Shift

The limitations of conventional codecs and learned codecs come from the same basic assumption: both treat communication as the delivery of raw bit sequences. Their goal is to reproduce the source signal accurately at the bit or pixel level, rather than asking

which information really needs to be sent. Semantic communication challenges this assumption and has emerged as a practical alternative.

Shannon and Weaver distinguished accurate symbol transmission (Level A), preservation of intended meaning (Level B), and achievement of the intended effect (Level C) [56, 57]. Conventional and learned codecs primarily address Level A; the theoretical distinction and its later development are reviewed in Section 2.6. This thesis instead concentrates on Level B when bandwidth is limited, the channel is impaired, or the receiver requires task-relevant rather than bit-exact information.

Deep learning makes Level B operational by mapping an image, video frame, or three-dimensional (3D) point cloud to a compact semantic latent and reconstructing a semantically faithful rather than bit-exact signal [58–61]. The principal neural architectures and the role of shared decoder knowledge are reviewed in Sections 2.6–2.9; here, the important point is that statistically predictable or perceptually unimportant detail need not be carried in the latent.

This approach directly addresses the limitations of both conventional and learned codecs, as summarised in Table 1.1. Semantic communication replaces linear transforms with nonlinear learned representations that capture edges, textures, and complex patterns more effectively. It replaces hand-designed coding tools with end-to-end learned mappings that adapt to data statistics. It also avoids the rapid complexity growth seen in conventional codecs because semantic encoders and decoders use fixed processing structures whose complexity depends mainly on model size. Most importantly, semantic communication usually degrades smoothly under channel impairments: as channel noise increases, reconstruction quality falls gradually instead of collapsing suddenly, because semantic latents spread information across many dimensions rather than relying on fragile entropy-coded symbols [62, 63]. Semantic systems can also be designed for 3D point clouds, immersive 360-degree video, and High Dynamic Range (HDR) content. Compared with learned codecs, semantic communication can simplify training because the main goal is to preserve meaning and task relevance rather than optimise several coding tools at once. It also allows flexible system design, so either the decoder or the encoder can be made lightweight depending on the application. Because semantic

representations are task-aware, they often generalise better across content types than models trained only for pixel accuracy. Channel-aware training, which adds Signal-to-Noise Ratio (SNR)-adaptive noise during training, also makes semantic decoders more robust across a wide range of channel conditions without explicit channel coding. Most importantly, semantic communication models and preserves *meaning* rather than pixels, allowing systems to optimise for task accuracy, such as classification or detection, instead of only pixel fidelity, and this can lead to order-of-magnitude bitrate reductions for the same task performance in M2M applications [55, 64].

In recent years, semantic communication for visual data has attracted strong interest in the communications community. Deep Joint Source–Channel Coding (JSCC) methods [62, 63] have shown that one end-to-end trained neural network can map image pixels directly to channel symbols, avoid the traditional separation of source and channel coding, and degrade gracefully under channel noise. Task-oriented semantic communication systems have also been proposed for M2M and machine-to-human (M2H) applications, where the goal shifts from visual quality for human viewing to preservation of task-relevant features for later inference [55, 64]. Surveys [60, 61] have reviewed this rapid growth across different data types and identified semantic communication as an important technology for 6G networks.

This connection between semantic communication and 6G is part of a broader evolution in wireless network capabilities. Fourth-generation (4G) Long-Term Evolution (LTE) networks, standardised as IMT-Advanced by the International Telecommunication Union Radiocommunication Sector (ITU-R) [65], offer peak downlink rates up to 1 Gbps with typical end-to-end latency of approximately 30–50 ms [28]. This capacity supports rapid image transfer and compressed High Definition (HD) video at Standard Dynamic Range (SDR), and can carry compressed 4K or High Dynamic Range (HDR) video under favourable conditions. However, lower experienced rates and reduced reliability at cell edges under fading make consistent low-latency 4K/HDR delivery difficult. The network has no awareness of the semantic content it carries, although semantic coding can reduce the transmitted data when link capacity is limited or variable. Fifth-generation (5G) New Radio (NR) networks, standardised as IMT-2020 [66],

define three service categories: Enhanced Mobile Broadband (eMBB) targeting peak rates up to 20 Gbps, which supports faster high-resolution image transfer and makes live 4K and 8K HDR streaming feasible; Ultra-Reliable Low-Latency Communication (URLLC) targeting approximately 1 ms user-plane latency and 99.999% reliability for short packets under specified conditions; and massive Machine-Type Communication (mMTC) for dense device connectivity [30,66]. HDR delivery over 5G still depends on codecs such as VVC with the Perceptual Quantizer (PQ) or Hybrid Log-Gamma (HLG) transfer functions. The network remains semantically unaware, but its higher capacity and reliability provide a stronger platform for semantic image and video services. Sixth-generation (6G) networks, now being developed within the IMT-2030 framework and targeting commercial deployment around 2030, are projected to exceed 1 Tbps peak rates, provide sub-millisecond latency, and extend ultra-reliable operation to more demanding immersive and machine-oriented applications [30,31]. Research visions for 6G identify native intelligence and semantic communication as core design features. The network is intended to support goal-oriented transmission, where meaning is optimised rather than bit-level accuracy, with high-resolution visual sensing, holographic telepresence, immersive Extended Reality (XR), and HDR and Wide Colour Gamut (WCG) video as primary use cases. This directly motivates the research in this thesis: the bandwidth and reliability improvements of 5G and 6G create the deployment context for semantic image and video codecs, while the explicit semantic communication goals of 6G align with the thesis approach of transmitting meaning rather than pixels.

All of these use cases involve what this thesis calls high-quality video, a category of visually demanding content that is substantially more difficult to compress and transmit than lower-resolution SDR video. Throughout this thesis, the term refers primarily to content with a spatial resolution of HD or above, including Ultra High Definition (UHD) 4K (3840×2160) and 8K (7680×4320), together with one or more additional demands: frame rates from 20 to 120 frames per second; compressed delivery bitrates of approximately 5 to 80 Mbit/s, depending strongly on resolution, frame rate, content, and codec [28,38]; HDR signal formats using the Perceptual Quantizer (PQ) or Hybrid Log-Gamma (HLG) transfer function defined in ITU-R BT.2100 [67]; or WCG content

using BT.2020 colour primaries [68], which cover a substantially wider gamut than the BT.709 primaries used for HDTV and much SDR production. For the experimental results in Chapters 3–5, a final decoded output of approximately 38 dB or above in PSNR, 0.90 or above in SSIM, and 90 or above in VMAF is treated as a high-fidelity operating region. These are operational thresholds used to interpret the results of this thesis, not universal definitions of broadcast quality; perceived quality also depends on content, display conditions, viewing distance, and the metric implementation.

Such content is more challenging to compress and transmit for several reasons. Higher resolution and frame rate directly increase the raw data volume. For example, uncompressed 4K video at 60 frames per second with three 10-bit components and 4:4:4 sampling requires approximately 14.9 Gbit/s, excluding transport overhead; even 10-bit 4:2:0 sampling requires approximately 7.5 Gbit/s. Both values are four times those of 1080p video at the same frame rate and sampling format. HDR also increases bit depth and represents a much wider luminance range; PQ-encoded samples have different statistics from conventional gamma-encoded SDR samples, affecting prediction and quantisation behaviour [69]. Similarly, BT.2020 permits more saturated colours than BT.709, so chroma quantisation or subsampling errors can produce visible hue and saturation errors in wide-gamut regions. These properties increase the demands on motion estimation, transform coding, quantisation, entropy coding, colour processing, and perceptual quality assessment. The transmission challenge is greater still over bandwidth-limited or error-prone wireless channels because a larger source rate must be delivered while maintaining the required visual quality and reliability.

However, despite this growing body of semantic communication research, it has not yet been applied to high-quality video coding in the sense of producing broadcast-quality compressed video that can compete with conventional standards in rate–distortion performance for human viewers [60,61]. Most semantic communication systems are designed for image transmission or machine inference, where human viewing quality is not the main goal. Applying these methods to high-quality video coding is difficult for several reasons. First, semantic systems focus on preserving meaning rather than exact pixel values, which can introduce visible artefacts or lose fine detail at bitrates

needed for broadcast-quality video. Second, they are usually trained for specific channel conditions, so performance can drop when the real channel differs from the training setup. Third, unlike conventional codecs, semantic codecs can behave unevenly across different types of video content. For these reasons, a purely semantic approach cannot yet match traditional codecs for high-quality video viewed by humans.

This thesis therefore proposes hybrid coding as the solution. Instead of replacing conventional coding, the semantic communication framework first generates a base reconstruction that gives the decoder a strong starting point, and a conventional codec such as VVC then encodes the residual signal. In this way, the system can achieve broadcast-quality output at lower bitrates than either approach alone. Reaching this goal across different content types, dynamic range conditions, and fading wireless channels raises four main challenges, which define the four main parts of this thesis. First, *hybrid coding*: the semantic latents must provide a good base reconstruction that reduces residual energy and bitrate, while the VVC encoder must still be able to encode the remaining residual efficiently. Second, *content adaptation*: a decoder trained on general video data does not fully capture the properties of each video sequence or each Group of Pictures (GOP), but retraining the whole decoder for every GOP is too slow for practice. Third, *HDR extension*: current semantic video codecs are trained for Standard Dynamic Range (SDR) content and do not work well on HDR and Wide Colour Gamut (WCG) video. Fourth, *robust wireless transmission*: a single-path semantic system can fail completely when that path is lost in deep fading. This thesis addresses these four challenges with a hybrid architecture, lightweight per-GOP adaptation, an HDR-to-SDR tone-mapping bridge, and multiple description coding across separate paths.

This thesis addresses these four challenges through a connected programme of research built around the four pillars in the thesis title: *hybrid coding*, *content adaptation*, *HDR extension*, and *robust wireless transmission*. The main contributions focus on video coding for SDR content. Related contributions on semantic image transmission, especially the Semantic Multiple Description Coding (SMDC) image framework later extended to video, provide the earlier methods that share the same channel-aware

semantic latent design. The work progresses from a hybrid semantic-VVC video coding architecture to context-conditioned semantic video coding with pretrained motion-adaptive models, per-GOP content-adaptive parameter overfitting, HDR adaptation through invertible tone-mapping, and SMDC for images and video. It also includes two supporting methods for DNN parameter efficiency, temporal prediction and VVC-based weight compression, which reduce parameter transmission overhead and support future work toward a fully independent semantic video codec.

The broader motivation outlined above is now distilled, in the next section, into a single overarching problem statement and a set of interlocking sub-problems that the remainder of the thesis sets out to resolve.

1.2 Problem Statement

The central problem addressed by this thesis can be stated as follows:

Conventional video coding standards are reaching the limits of what manually designed tools can achieve in compression, yet the needs of new video applications in terms of bitrate, quality, robustness, content diversity, and dynamic range go well beyond what these standards, or early learned alternatives, can reliably deliver. Semantic communication can offer a new and promising approach, but its potential for semantic image transmission, and for the closely related problem of practical video coding, has not been fully achieved because several key problems remain unsolved.

More specifically, this thesis identifies and addresses the following connected problems:

1. **Bandwidth and Compression Efficiency.** Global video traffic is growing at a rate that outpaces improvements in network capacity [28,29]. While VVC delivers the best rate-distortion performance among conventional standards, it does so at high encoder complexity. End-to-end learned codecs demonstrate promising results but have not yet achieved consistent performance advantages across diverse

content classes and resolution scales. Hybrid semantic-conventional architectures that combine the compression efficiency of learned semantic encoding with the proven reliability of conventional residual coding offer a promising direction, but their design, optimisation, and evaluation are still at an early stage.

2. **Temporal Redundancy Exploitation in Semantic Video Coding.** Video sequences contain strong inter-frame temporal correlations that conventional codecs exploit through motion-compensated prediction to achieve substantially better compression than intra-frame coding alone. The semantic video coding systems studied in this thesis have shown that these correlations are often exploited only in a limited way. Achieving the full potential of semantic video coding requires better temporal modelling through hierarchical bidirectional prediction and motion-aligned semantic context generation.
3. **Content Adaptivity Without Computational Overhead.** DNNs trained on general data do not capture the specific statistical properties of individual video sequences, GOPs, or scenes. Per-GOP retraining from scratch, while effective, is too computationally expensive for real-time or near-real-time use. A low-cost method to adapt semantic decoders to each specific content, at a small fraction of the cost of full retraining, is needed to improve coding performance beyond what general-purpose models can achieve.
4. **High Dynamic Range Content.** HDR and WCG video, encoded with transfer functions such as the Perceptual Quantizer (PQ) and colour spaces such as BT.2020, has very different statistical properties from the SDR content on which most semantic communication systems are trained. Directly applying SDR-trained semantic codecs to HDR content leads to a significant drop in performance. An invertible adaptation method that closes the gap between HDR sources and SDR-trained semantic codecs is therefore needed.
5. **Robust Transmission Over Fading Wireless Channels.** Single-path semantic communication systems, while robust to random channel noise through their learned representations, remain vulnerable to deep fades, shadowing, and path

outages in Rayleigh and Rician fading environments. A complete path failure causes a total loss of all transmitted semantic information, reintroducing the cliff effect at the path level. Combining Multiple Description Coding (MDC) [70, 71] with semantic communication offers a way to achieve robustness through path diversity, but how to do this effectively for both image and video transmission over real wireless channels remains an open problem.

6. **DNN Parameter Efficiency and Transmission.** Semantic video coding systems that adapt their DNN parameters to each sequence or GOP need a method to send the updated parameters from the encoder to the decoder. Transmitting uncompressed DNN parameters adds significant overhead that can cancel out the compression gains of the semantic approach. Neural network compression and temporal prediction of parameter updates using inter-frame correlations are possible solutions, but how to integrate these methods into semantic video coding systems has not been fully studied.

Each of these six problems points to an area that has not been studied enough in earlier work. The next section maps them to a set of research gaps identified through a systematic review of the conventional, learned, and semantic communication literature.

1.3 Research Gap

A review of the literature reveals the following main gaps that this thesis addresses:

Gap 1: Hybrid Semantic–Conventional Video Coding with Bidirectional Hierarchical Prediction. While semantic communication has been proposed and studied in theory [56, 59, 60], the hybrid combination of semantic communication with a standardised codec such as VVC, in one system that uses semantic encoding for inter-frame prediction and VVC for intra-frame and residual coding, had not been studied in enough depth.

Gap 2: Context-Aware Semantic Video Coding with Pretrained Models.

The hybrid semantic video codecs in Gap 1 required per-GOP training of the full semantic autoencoder from scratch, which made them too computationally expensive for practical use. The use of one shared pretrained model with selective motion-adaptive fine-tuning, combined with hierarchical bidirectional temporal prediction and motion-aligned semantic context generation through real-time intermediate flow estimation, had not been studied enough.

Gap 3: Content-Adaptive Parameter Overfitting for Per-GOP Specialisation.

The pretrained semantic video codecs in Gap 2 produce general-purpose decoders that still leave residual energy for conventional codecs to encode. Per-GOP content specialisation through overfitting a small, differentially coded subset of decoder parameters (around 643 scalars, less than 0.1% of the full decoder) had not been studied as a practical way to reduce the rate–distortion gap between general and content-specific semantic coding.

Gap 4: Semantic HDR Video Coding. State-of-the-art hybrid semantic video codecs operate only in the SDR domain and degrade badly when applied to HDR content. An automatic per-frame invertible tone-mapping bridge that lets an SDR-trained semantic codec process HDR content without changing the model architecture, while sending the tone-mapping parameters as a small metadata side stream for closed-form HDR recovery at the receiver, had received limited attention.

Gap 5: Semantic Multiple Description Coding for Robust Wireless Transmission.

Single-path Deep JSCC systems are robust to additive noise but can degrade strongly under deep fading events. The integration of semantic communication with multiple description coding to provide path diversity and guaranteed minimum performance, achieved by interleaving and splitting semantic latents into multiple complementary descriptions, had not been studied for both image and video transmission.

Gap 6: Temporal Prediction of DNN Parameters for Semantic Video Coding. Using the temporal similarity between DNN parameters trained on consecutive video frames to predict decoder parameters from previously transmitted data, and thereby reduce parameter transmission overhead, had received limited attention in semantic communication-based video coding.

Gap 7: DNN Compression for Semantic Video Communications Using VVC. Using spatial correlations in DNN weight matrices, treated as two-dimensional signals, as the basis for VVC-based compression of neural network parameters in semantic video communication, and comparing this approach with the standardised NNCCodec method, had received limited attention.

Addressing these seven gaps together motivates the main research aim given below.

1.4 Research Aim

The overarching aim of this thesis is:

To design, implement, and evaluate a connected family of semantic communication systems for video coding that combines hybrid semantic-VVC coding, content-adaptive specialisation, HDR extension, and robust wireless transmission, covering SDR video over fading wireless channels and including related semantic image transmission methods with the same architectural philosophy, together with DNN parameter efficiency methods that support future fully independent semantic video coding.

This aim is pursued through a step-by-step progression from foundational semantic image communication to more advanced semantic video coding, robust wireless transmission, and related data modalities, with each contribution building on the previous ones.

1.5 Research Objectives

To achieve the research aim stated above, the following specific objectives are pursued:

1. **Objective 1: Hybrid Semantic–VVC Video Coding.** To design and evaluate a hybrid video coding system that combines semantic communication principles for inter-frame prediction with VVC for intra-frame and residual coding, demonstrating consistent rate–distortion improvements over H.264/AVC, H.265/HEVC, and H.266/VVC across video sequences with diverse complexity characteristics.
2. **Objective 2: Enhanced Hybrid Coding with Bidirectional Prediction.** To extend the hybrid semantic–VVC codec with bidirectional hierarchical frame prediction, achieving further improvements in rate–distortion performance through more effective exploitation of temporal redundancies, and demonstrating enhanced robustness to bitstream errors compared to conventional codecs.
3. **Objective 3: Context-Aware Semantic Video Coding with Pretrained Models.** To eliminate per-GOP retraining from hybrid semantic video coding by developing a shared pretrained foundation model with motion-adaptive specialisation, incorporating hierarchical bidirectional prediction and real-time intermediate flow estimation (RIFE)-based [72] motion-aligned context generation, and demonstrating consistent rate–distortion gains over VVC and competitive performance against state-of-the-art learned codecs.
4. **Objective 4: Content-Adaptive Parameter Overfitting.** To restore per-GOP content adaptivity to the pretrained semantic video codec at a fraction of the cost of full retraining, using a compact affine modulation of the decoder through per-channel learnable scale multipliers and biases, differentially encoded through a unified Principal Component Analysis (PCA)–Deep Context-Adaptive Binary Arithmetic Coding (DeepCABAC) pipeline.
5. **Objective 5: Semantic HDR Video Coding.** To extend the content-adaptive semantic video codec to HDR sources through an automatic, per-frame, fully invertible Reinhard tone-mapping bridge, enabling the SDR-trained codec to process linear OpenEXR (EXR) HDR content without architectural modification and to recover the linear HDR signal in closed form at the receiver.

6. Objective 6: Robust Semantic Multiple Description Coding for Images.

To propose a SMDC framework for robust image transmission that generates multiple semantically complementary descriptions via interleaved channel-wise splitting, eliminating cliff effects over Additive White Gaussian Noise (AWGN), Rayleigh, and Rician fading channels and establishing a guaranteed performance floor unreachable by conventional and learned separation-based baselines.

7. Objective 7: Robust Semantic Multiple Description Coding for Video.

To extend the SMDC framework to video transmission by designing dedicated intra-frame and inter-frame semantic MDC modules within a GOP-8 structure, with context-conditioned differential latent encoding for inter-frames exploiting RIFE-based temporal interpolation.

8. Objective 8: Temporal Prediction of DNN Parameters. To exploit the temporal correlation between DNN parameters trained on consecutive video frames to predict decoder parameters at the receiver from previously transmitted information, reducing parameter transmission overhead and demonstrating significant rate-distortion improvements over NNCodec.

9. Objective 9: DNN Compression for Semantic Video Communications.

To develop a neural network transmission framework that integrates VVC-based compression of DNN weights (treated as 2D signals with spatial correlations) with semantic communication principles, demonstrating consistent bitrate reductions over the standardised NNCodec approach.

The pursuit of these nine objectives is guided by four cross-cutting research questions, each spanning multiple objectives and corresponding directly to one of the four thesis pillars.

1.6 Research Questions

The following research questions guide the thesis and connect its contributions:

- RQ1:** *How can semantic communication be integrated with state-of-the-art VVC in a hybrid architecture that consistently achieves rate–distortion improvements over H.264/AVC, H.265/HEVC, and H.266/VVC, and how can the system be extended with bidirectional hierarchical prediction and pretrained models to improve temporal modelling efficiency and deployability?*
- RQ2:** *How can content-adaptive per-GOP specialisation of semantic video decoders be achieved at a fraction of the computational cost of full retraining, and how can this principle be extended to HDR content without modifying the codec architecture?*
- RQ3:** *How can semantic communication be made robust to single-path failures in wireless fading environments through the integration of multiple description coding, and what guaranteed performance floors can be established for image and video transmission over AWGN, Rayleigh, and Rician fading channels?*
- RQ4:** *How can DNN parameter efficiency be improved in semantic video coding systems through temporal prediction and VVC-based compression of parameter updates, and what rate–distortion gains over NNCodec-based approaches can be demonstrated?*

Each of these research questions is addressed by one or more of the nine original contributions listed below. For each contribution, the rationale, action, and impact map directly to the problems, gaps, objectives, and questions defined in the previous sections.

1.7 Contributions

This thesis makes the following original contributions to semantic communication systems for video coding, including the related semantic image transmission methods that support the video extensions. Each contribution is linked to peer-reviewed publications, as indicated. Each contribution is described in three parts: (i) the *rationale*, which is the research gap that motivated it, as discussed in Sections 1.2 and 1.3; (ii) the *action*, which is the technical method proposed, implemented, and evaluated in this thesis; and

(iii) the *impact*, which is the improvement over earlier work and what it means for the design of semantic communication systems for video coding. The overall importance of the nine contributions is discussed at the end of this section in Section 1.7.1, and their wider implications for research and industry are discussed in Section 8.4 of Chapter 8.

Contribution 1: Hybrid Semantic Communication and VVC-Based Video Coding System

A hybrid video coding system that combines autoencoder-based semantic communication for inter-frame prediction with VVC for intra-frame and residual coding is proposed [1]; as established by the systematic search described in Section 2.12, this is the first such hybrid system reported in the literature. The system introduces a novel mechanism for remote training of the semantic decoder that eliminates the need to transmit decoder parameters from encoder to decoder. Benchmarking across a diverse set of video sequences demonstrates consistent improvements over H.264/AVC and H.265/HEVC, and improvements over H.266/VVC for low- and high-complexity content. Bjøntegaard Delta (BD) rate analysis confirms the rate–distortion advantages of the proposed framework. Full details are presented in Chapter 3.

Contribution 2: Enhanced Hybrid Codec with Bidirectional Hierarchical Frame Prediction

Building on the hybrid semantic–VVC codec, an enhanced system is proposed that integrates bidirectional hierarchical frame prediction into the semantic encoding and decoding pipeline [22]. This approach more effectively exploits inter-frame temporal redundancies without requiring extensive per-GOP retraining or presharing of trained networks. The enhanced system achieves competitive rate–distortion performance relative to state-of-the-art conventional codecs and leading learned codecs across standard test sequences, and demonstrates improved robustness to bitstream transmission errors. Full details are presented in Chapter 3.

Contribution 3: Context-Aware Semantic Video Coding with Pre-trained Models

A context-conditioned semantic video coding framework is proposed that eliminates per-GOP retraining through a shared pretrained foundation model with selective motion-adaptive fine-tuning [73]. Hierarchical bidirectional prediction and RIFE-based motion-aligned context frame generation provide explicit spatiotemporal guidance for semantic decoding. Differential semantic latents are entropy-coded with DeepCABAC, and a VVC residual pathway guarantees reconstruction fidelity. The framework achieves an average BD-rate reduction of 22.03% over the VVC Test Model (VTM) and competitive performance against state-of-the-art learned codecs including DCVC-FM, across UVG, MCL-JCV, HEVC Class B, HEVC Class E, and 4K benchmarks. Full details are presented in Chapter 4.

Contribution 4: Content-Adaptive Parameter Overfitting for Per-GOP Specialisation

A content-adaptive parameter overfitting mechanism is proposed to specialise the semantic decoder to each GOP by updating only per-channel learnable scale multipliers and transposed-convolution biases (approximately 643 scalars per GOP, less than 0.1% of the full decoder). The parameter update is differentially encoded through a unified PCA–DeepCABAC pipeline alongside the differential semantic latents. A tapered, quantisation-aware autoencoder reduces per-block processing from 60.41 to 12.43 GFLOPs, a $4.9\times$ reduction relative to CASVC. Comprehensive evaluation on six datasets (including HEVC Class B/E, UVG, MCL-JCV, and out-of-distribution closed-circuit television (CCTV) surveillance and 3D animation) demonstrates average BD-rate reductions of 23.9%, 22.4%, and 21.4% in PSNR, Structural Similarity Index Measure (SSIM), and VMAF respectively over VTM. Full details are presented in Chapter 4.

Contribution 5: Context-Aware Semantic HDR Video Coding

The content-adaptive semantic video codec is extended to HDR and WCG content through an automatic per-frame invertible Reinhard tone-mapping pre-process. The per-frame exposure multiplier and gamma exponent are estimated in closed form from log-average luminance, mapping each linear HDR EXR frame into a 10-bit SDR frame whose statistics match the autoencoder’s training distribution. The two scalar tone-mapping parameters per frame are entropy-coded as a compact HDR metadata sidestream, enabling exact closed-form HDR recovery at the receiver. Evaluation on the HDM-HDRV and Joint Video Experts Team (JVET) HDR/WCG test sets against VTM 23.0 demonstrates average BD-rate reductions of 16.42% in PSNR, 14.14% in SSIM, 10.27% in VMAF, 15.31% in Perceptually Uniform (PU)-PSNR, and 11.12% in PU-SSIM. Full details are presented in Chapter 5.

Contribution 6: Semantic Multiple Description Coding for Robust Image Transmission

A SMDC framework is proposed for robust image transmission over wireless fading channels. An autoencoder with a novel interleaved channel-wise splitting strategy generates three semantically complementary descriptions, each independently compressed, channel coded with polar codes, and transmitted over separate paths. An intelligent fusion decoder reconstructs images from any received subset of descriptions, eliminating cliff effects. The system maintains PSNR above 30 dB and SSIM above 0.80 down to 1 dB combined SNR for AWGN, 5 dB for Rician, and 7 dB for Rayleigh fading channels, significantly outperforming JPEG [48], JPEG2000 [74], High Efficiency Image Format (HEIF) [75], and CompressAI [76] baselines. Full details are presented in Chapter 6.

Contribution 7: Semantic Multiple Description Coding for Robust Video Transmission

The SMDC framework is extended to video transmission through a dedicated context-conditioned inter-frame semantic MDC module. The framework decomposes video into intra-frames and inter-frames under a GOP-8 structure. A novel context-conditioned

inter-frame encoder encodes only the differential latent between each inter-frame and its RIFE-interpolated context, substantially reducing transmission entropy. Evaluation over 20 test sequences demonstrates 4–5 dB SNR advantages over a VVenC baseline across AWGN, Rayleigh, and Rician fading channels, with SMDC-V achieving VMAF above 70 at 2–3 dB lower combined SNR thresholds than VVenC MDC across all channel models. Full details are presented in Chapter 6.

Contribution 8: Temporal Prediction of DNN Parameters for Semantic Video Coding

A novel approach for reducing the overhead of transmitting DNN decoder parameters in semantic video coding is proposed, exploiting the strong temporal correlation between parameters trained on consecutive video frames [25]. The method predicts decoder parameters at the receiver using key frames and parameter displacement vectors, reducing the parameters that must be transmitted. Experimental results demonstrate PSNR gains of 3–25 dB and an average bitrate saving of 54% compared to NNCodec, depending on video complexity. Full details are presented in Chapter 7.

Contribution 9: DNN Compression for Semantic Video Communications Using VVC

A neural network transmission framework is proposed that treats DNN weight matrices as 2D signals and compresses them using VVC, leveraging spatial correlations in parameter space and semantic communication principles to optimise parameters for each video scene [24]. The framework consistently achieves lower bitrates than NNCodec while maintaining comparable video quality, providing a practical solution for reducing the bandwidth overhead of DNN parameter transmission in semantic video coding. Full details are presented in Chapter 7.

1.7.1 Collective Significance and Scientific Impact

Each of the nine contributions advances a specific part of the semantic communication pipeline for video coding. Taken together, they form one connected body of work with

a broader scientific impact. Three main field-level claims follow from the full evidence presented in this thesis.

First, the thesis shows *quantitatively* that semantic communication is not only an attractive idea, but also a practical alternative to both conventional hybrid codecs (H.264/AVC, H.265/HEVC, H.266/VVC) and end-to-end learned codecs (DCVC-FM, DCVC-DC, DCVC-RT). It demonstrates consistent BD-rate reductions relative to the VVC Test Model for both standard dynamic range and HDR video, with gains confirmed across several objective and perceptual metrics (PSNR, SSIM, VMAF, PU-PSNR, PU-SSIM) and across both in-distribution and out-of-distribution content. This evidence shows that semantic communication is a practical and reliable approach to video coding and provides a clear baseline for future work.

Second, the thesis shows that semantic communication *works across different content types, channel conditions, and deployment settings* that are often studied separately. The same main idea, preserving task-relevant meaning through channel-aware training, gives consistent benefits for standard dynamic range video (Contributions 1–4, 7, 8, 9), HDR and WCG video (Contribution 5), and robust image and video transmission over fading channels (Contributions 6–7), across AWGN, Rayleigh fading, and Rician fading channels, and under bandwidth-limited and parameter-limited deployment settings. This broad applicability supports the main claim of the thesis: semantic communication should be seen not as one isolated technique, but as a general design approach for next-generation video coding and transmission systems.

Third, the thesis introduces a set of *practical and reusable frameworks* that other researchers can build on directly. These include the hybrid semantic–VVC architecture (Contributions 1–4), the content-adaptive parameter overfitting method (Contribution 4), the invertible tone-mapping HDR adaptation bridge (Contribution 5), the SMDC framework with channel-wise interleaved splitting (Contributions 6–7), and the temporal and spatial DNN parameter compression methods (Contributions 8–9). Each framework has been implemented, evaluated against strong baselines, and described in enough detail to be reproduced from the corresponding chapters.

Together, these three claims, namely results that match or improve on strong base-

lines, applicability across content types and channels, and practical reusable frameworks, form the main scientific contribution of this thesis to communication engineering and to the wider 6G wireless research direction. The full implications of these contributions for academic research, 6G wireless network design, broadcasting, and streaming are discussed further in Section 8.4 of Chapter 8.

With the motivation, problems, gaps, aim, objectives, questions, and contributions of the thesis now established, the next section outlines how the remaining chapters are organised to present each contribution in technical detail.

1.8 Thesis Structure

The remainder of this thesis is organised as follows, with each chapter covering a different aspect of semantic communication for video coding:

Chapter 2 presents a comprehensive background and literature review covering the evolution of conventional video coding standards and their limitations, the principles and foundations of semantic communication, the state of the art in learned image and video compression, hybrid and context-aware semantic coding, DNN compression and parameter efficiency, HDR imaging, and multiple description coding for robust transmission. The chapter concludes by identifying the research gaps that motivate the contributions of this thesis.

Chapter 3 develops the hybrid semantic and VVC-based video coding systems. Section 3.2 presents the initial hybrid semantic–VVC video coding system (Contribution 1), and Section 3.3 presents the enhanced hybrid codec with bidirectional hierarchical frame prediction (Contribution 2). Rate–distortion analysis and comprehensive comparison against conventional and learned codecs are presented in Section 3.3.

Chapter 4 develops the context-aware semantic video coding framework. Section 4.2 presents the context-conditioned semantic video codec with pretrained motion-adaptive models and bidirectional prediction (Contribution 3), and Section 4.3 presents the content-adaptive parameter overfitting mechanism for per-GOP specialisation (Contribution 4).

Chapter 5 extends semantic video coding to HDR content. Section 5.2 presents the

Chapter 1. Introduction

HDR semantic video coding framework using invertible tone-mapping adaptation (Contribution 5), and experimental evaluation against VTM is presented in Section 5.2.2.

Chapter 6 develops robust semantic communication via multiple description coding. Section 6.2 presents the SMDC framework for robust image transmission (Contribution 6), and Section 6.3 presents the SMDC framework for robust video transmission (Contribution 7). Evaluation across AWGN, Rayleigh, and Rician fading channels is presented in Section 6.3.4.

Chapter 7 addresses DNN parameter efficiency for semantic video coding. Section 7.2 presents the VVC-based DNN compression framework (Contribution 9), and Section 7.3 presents the temporal prediction of DNN parameters (Contribution 8).

Chapter 8 presents conclusions and future research directions. Section 8.1 summarizes the key contributions of this thesis, discussing their individual and collective significance. Section 8.2 distils the principal empirical findings across all contributions. Section 8.4 discusses the broader implications of this work for the research community and for industry practitioners engaged in next-generation wireless systems. Section 8.5 acknowledges the limitations of the proposed systems and identifies boundary conditions on the reported results. Section 8.6 identifies open research directions stemming from this work, including potential extensions, applications, and theoretical investigations that could build on the foundations established in this thesis. Section 8.7 provides closing remarks.

Table 1.1: How Semantic Communication Mitigates the Limitations of Conventional and Learned Video Codecs.

Limitation	How Semantic Communication Addresses It
<i>Conventional Codecs</i>	
Linear transforms suboptimal for edges, textures, and complex patterns	Nonlinear learned representations capture complex statistics efficiently
Handcrafted tools cannot adapt to content diversity	End-to-end learned mappings adapt to data statistics automatically
Exponential encoder complexity growth with each new standard	Fixed, predictable computational graphs; complexity scales with model size
Cliff effect: a single bit error can destroy an entire frame	Graceful degradation: information distributed across latent dimensions; quality decreases smoothly with channel noise
Static codecs cannot learn from data or adapt to new content	Models can be continuously retrained and updated as new data becomes available
<i>Learned Codecs</i>	
Training instability from multiple competing objectives	Semantic objectives (preserve meaning, task accuracy) are simpler and more stable to optimize
High decoder complexity, asymmetric in the wrong direction for deployment	Enables flexible asymmetry: lightweight decoder possible with invertible latents, or lightweight encoder for uplink-constrained scenarios
Inconsistent performance across content types	Task-aware semantic features generalise consistently across diverse content
No standardisation; bitstreams tied to specific models	Growing community agreement on benchmarks, metrics, and reference architectures
Channel fragility persists; no inherent error resilience	Channel-aware training (adding SNR-adaptive noise) makes decoders robust without explicit channel coding
Limited semantic awareness: optimises pixels, not meaning	Explicitly models and preserves meaning; can optimise for task accuracy (classification, detection) rather than pixel fidelity

Chapter 2

Background and Literature Review

2.1 Introduction

This chapter provides the theoretical background and literature review needed for the research contributions presented in this thesis. Semantic communication for video coding sits at the intersection of several fast-growing fields: digital video coding, deep learning, JSCC, wireless communications, HDR imaging, multiple description coding, and neural network compression. Understanding the development, main results, and open challenges in these areas is important for placing the thesis contributions in context and for explaining the research gaps addressed.

The chapter is organised in a step-by-step way, moving from conventional baselines to deep learning and semantic communication, and then to the specific areas that support each thesis contribution. Section 2.2 reviews the development of conventional video coding standards from the 1980s to VVC, including the coding tools, performance, and limitations of each generation. Section 2.3 explains the main challenges faced by conventional coding for emerging visual applications and motivates the learned and semantic alternatives discussed later. Section 2.4 reviews deep learning-based image compression, from early autoencoder designs to current entropy-coded variational models. Section 2.5 extends this review to end-to-end learned video compression, covering

optical flow-based methods, conditional coding, and temporal modelling, and identifies the remaining limitations that motivate semantic communication. Section 2.6 presents the theoretical background of semantic communication, from Shannon and Weaver’s three-level framework to modern deep learning implementations. Sections 2.7–2.10 review semantic communication across text, speech, image, and video. Section 2.11 reviews recent system-level work in multi-user, generative, federated, and information-theoretic semantic communication that provides broader context for this thesis. Sections 2.12–2.16 then give focused reviews of the five areas that directly support the thesis pillars: hybrid semantic and conventional video coding (Section 2.12), context-aware and content-adaptive semantic coding (Section 2.13), HDR video coding (Section 2.14), multiple description coding for robust transmission (Section 2.15), and neural network compression and parameter efficiency (Section 2.16). Section 2.18 introduces the objective and perceptual quality metrics used throughout the thesis. Section 2.20 closes the chapter with a summary of the research gaps that motivate the thesis contributions.

2.2 Evolution of Conventional Video Coding Standards

2.2.1 Origins and Early Standardization

The history of digital video compression begins in the early 1980s, driven by the need to transmit digital video over the Public Switched Telephone Network (PSTN). The ITU-T ratified H.120 in 1984 as the first digital video compression standard, designed for videoconferencing over primary digital group transmission links [77]. Although limited in adoption, H.120 established a basic coding architecture using conditional replenishment, Differential Pulse-Code Modulation (DPCM), scalar quantisation, and variable-length coding; motion compensation was later added in the 1988 revision.

H.261 [32], ratified in 1988, was the first widely deployed digital video codec and introduced the core hybrid coding architecture that remains the basis of all mainstream codecs today. H.261 defined the Common Intermediate Format (CIF) and Quarter Common Intermediate Format (QCIF) frame formats, introduced macroblock-based motion compensation, DCT-based frequency-domain coding, and variable-length en-

trophy coding. Its successor H.263 [34] extended these concepts with half-pixel motion compensation, unrestricted motion vectors, and advanced prediction modes, achieving better performance at very low bitrates.

In parallel, the MPEG of the International Organization for Standardization and the International Electrotechnical Commission (ISO/IEC) developed MPEG-1 [33] in 1993, targeting storage on Compact Disc Read-Only Memory (CD-ROM) at bitrates around 1.5 Mbit/s, and introduced the concept of bidirectional (B-frame) prediction, which significantly improved coding efficiency by exploiting temporal correlation in both forward and backward directions. MPEG-2, released in 1995 and technically equivalent to ITU-T H.262, became the dominant standard for digital television broadcast, Digital Versatile Disc (DVD), and early digital satellite services [35]. It extended MPEG-1 with support for interlaced content, scalable profiles, and a hierarchical picture structure, and became one of the most widely deployed video coding standards in history.

2.2.2 H.264/Advanced Video Coding

The Joint Video Team (JVT), formed in 2001 by the ITU-T Video Coding Experts Group (VCEG) and ISO/IEC MPEG, led to the publication of AVC [36], also known as MPEG-4 Part 10, in 2003. H.264/AVC achieved a major improvement in coding efficiency, delivering approximately 50% bitrate reduction over MPEG-2 at equivalent quality. This was made possible by a set of new and improved coding tools.

The H.264/AVC standard introduced variable block-size motion compensation with block sizes ranging from 16×16 to 4×4 pixels, allowing more precise matching of complex motion patterns. Quarter-pixel motion compensation was adopted as the standard precision, compared to half-pixel in earlier standards. In-loop deblocking filtering was introduced to reduce blocking artifacts at the decoder, and the reference picture list mechanism allowed multiple reference frames to be used for inter prediction, significantly improving performance for scenes with complex or periodic motion. The CABAC and Context-Adaptive Variable-Length Coding (CAVLC) entropy coders were introduced, with CABAC in particular achieving notable bitrate savings by modelling the conditional probability of syntax elements using their surrounding context.

H.264/AVC quickly became the leading video coding standard across nearly every application area, from consumer video recording and streaming to professional broadcast and video conferencing. Its widespread adoption was further supported by the MPEG LA patent pool, which offered a single licensing arrangement for implementers. As of 2020, H.264/AVC remained the most widely used video codec in the world, encoding most video content on streaming platforms, social media, and consumer devices.

2.2.3 H.265/High Efficiency Video Coding

The rapid growth of high-definition video content and the rise of 4K and Ultra High Definition (UHD) television created the need for a new standard with better compression efficiency than H.264/AVC. The Joint Collaborative Team on Video Coding (JCT-VC), the successor to the JVT, developed HEVC [37], which was standardized in 2013 and represented the largest compression improvement achieved in a single generation of video coding standards.

H.265/HEVC introduced the Coding Tree Unit (CTU) as the fundamental coding unit, with a maximum size of 64×64 pixels, replacing the fixed 16×16 macroblock of previous standards. The CTU is divided into Coding Units (CUs), Prediction Units, and Transform Units (TUs) using a quadtree structure, allowing flexible spatial segmentation that better matches the structure of natural video content. The standard introduced 35 intra-prediction modes (33 directional modes plus DC and planar prediction, compared to 9 in H.264/AVC), along with improved inter-prediction tools including advanced motion vector prediction (AMVP) and merge mode. Sample Adaptive Offset (SAO) filtering was added as a second in-loop filter alongside the improved deblocking filter.

The compression efficiency of H.265/HEVC was evaluated against H.264/AVC using the BD-rate metric [78], showing average bitrate savings of approximately 50% at equivalent visual quality [79]. H.265/HEVC was also extended with several profiles and tiers: Range Extensions (RExt) for high bit depth and various chroma formats; Scalable HEVC (SHVC) for spatial, temporal, and quality scalability; Multiview HEVC (MV-HEVC) and 3D-HEVC for stereoscopic and volumetric video; and Screen Content

Coding (SCC) extensions for mixed content [37].

2.2.4 H.266/Versatile Video Coding

The JVET developed VVC [38] as the successor to H.265/HEVC, and it was standardized in July 2020. H.266/VVC was designed to support a wider range of content and applications than any previous standard, including 4K/8K UHD, HDR/WCG content, 360-degree omnidirectional video, and screen content. The standard achieved approximately 50% bitrate reduction over H.265/HEVC at equivalent quality, confirming that the hybrid coding approach remains effective [80].

H.266/VVC introduced many new coding tools compared to H.265/HEVC. The block partitioning scheme was improved with the Quad Tree with Nested Multi-Type Tree (QTMT) structure, which allows blocks to be split into binary and ternary partitions in addition to quadtrees, allowing better adaptation to different content types [81]. Inter prediction was improved with Affine Motion Compensation (AMC) for non-translational motion, Sub-block based Temporal Motion Vector Prediction (SbTMVP), and Bi-directional Optical Flow (BDOF) and Prediction Refinement with Optical Flow (PROF), which improve motion-compensated predictions using local gradient information [38]. Intra prediction was extended with 65 angular modes, Matrix-based Intra Prediction (MIP), the Intra Sub-Partition (ISP) tool, and the cross-component linear model (CCLM) for chroma prediction [82]. Transform coding was extended with the Multiple Transform Set (MTS), which supports multiple transform kernels including DCT-2, DST-7, and DCT-8, the Low-Frequency Non-Separable Transform (LFNST), and Joint Coding of Chroma Residuals (JCCR) [83]. The CABAC entropy coder was redesigned with improved context modelling and coding bypass mechanisms [38].

The VTM reference software is the official implementation of the standard and is used as the primary baseline for evaluating the compression performance of the systems developed in this thesis. An open-source encoder, VVenC [84], was developed by the Fraunhofer Heinrich Hertz Institute as a practical VVC encoder that achieves near-VTM compression quality with much lower computational cost.

2.2.5 AOMedia Video 1 and AOMedia Video 2 (AV2)

Alongside the ITU-T/ISO standardisation track, the Alliance for Open Media (AOM), a consortium of major technology and streaming companies, developed AV1 [39] as a royalty-free alternative to H.265/HEVC. AV1 was published in 2018 and achieves compression performance broadly comparable to H.265/HEVC, with a significantly more complex encoding process. AV1 introduced several new coding tools, including a superblock-based adaptive partitioning framework, the Constrained Directional Enhancement Filter (CDEF) for artifact reduction, a film grain synthesis mechanism, and an entropy coding framework based on multi-symbol arithmetic coding. AV1 has been widely deployed on streaming platforms (YouTube, Netflix) and video conferencing applications (Google Meet, WebRTC) due to its royalty-free licensing model. Its successor, AOMedia Video 2 (AV2) [85], is under active development with the goal of achieving performance comparable to H.266/VVC.

2.2.6 Key Coding Tools Across Generations

All mainstream video coding standards share a common hybrid coding architecture based on the following core tools. Block partitioning divides each frame into regions of different sizes to match the local structure of the content. Motion-compensated prediction reduces temporal redundancy by predicting each block in the current frame from one or more similar regions in previously decoded reference frames, and transmitting only the difference (residual) rather than the full block. Intra prediction reduces spatial redundancy within a single frame by predicting each block from its already-decoded neighbouring blocks using directional or model-based predictors. Transform coding converts the residual signal into frequency-domain coefficients using basis functions such as the DCT, concentrating most of the energy into a small number of values. Quantization reduces the precision of transform coefficients, introducing controlled distortion to achieve the target bitrate. Entropy coding losslessly compresses the quantized coefficients and other syntax elements by modelling their statistical structure using context-adaptive probabilities. In-loop filtering removes coding artefacts from decoded frames before they are stored as references, using tools such as deblocking filters, SAO,

and Adaptive Loop Filter (ALF).

The individual coding tools of H.266/VVC have been comprehensively documented in a dedicated special issue of Institute of Electrical and Electronics Engineers (IEEE) Transactions on Circuits and Systems for Video Technology [81–83, 86, 87]. Intra prediction in VVC is substantially richer than in H.265/HEVC, incorporating 65 directional angular prediction modes, MIP using compact learned linear transforms of reconstructed boundary samples, ISP coding that subdivides a block into narrow strips for prediction, and CCLM that derives chroma prediction from luma reconstructed samples [82]. The Geometric Partitioning Mode (GPM) [86] permits coding unit boundaries along arbitrary straight-line orientations rather than only axis-aligned splits, providing more flexible adaptation to diagonal edges and oblique motion boundaries. Transform coding is extended with MTS, which signals the selection from DCT-2, DST-7, and DCT-8 transform kernels, and with the LFNST that applies a secondary non-separable transform to the low-frequency primary transform coefficients, further decorrelating residual energy in the low-frequency region [83]. The ALF applies a spatially varying Wiener filter whose coefficients are designed by the encoder and signalled in the bit-stream [87], delivering perceptual quality gains beyond those achievable by the deblocking filter and SAO filter inherited from H.265/HEVC. History-based Motion Vector Prediction (HMVP) maintains a table of recently used motion vector candidates [38], reducing the signalling cost of motion information for sequences with repetitive motion patterns.

2.3 Challenges and Limitations of Conventional Video Coding

2.3.1 The Complexity Wall

Each successive generation of video coding standards has delivered approximately 50% bitrate reduction over its predecessor, but at the cost of greatly increasing encoder complexity. The VTM encoder for H.266/VVC is estimated to be approximately 10 to 100 times more computationally expensive than the HEVC Test Model (HM) encoder

for H.265/HEVC, which was itself several times more expensive than the JM encoder for H.264/AVC [40, 42]. This growth in encoding complexity comes from the large number of coding tool configurations that the encoder must search through to find the best coding decisions.

The VVC encoder must evaluate hundreds of candidate block partitions, dozens of intra-prediction modes, multiple motion search strategies, several transform kernel combinations, and various entropy coding configurations for each coding unit. This makes real-time software encoding of high-definition content with VVC impractical on general-purpose Central Processing Units (CPUs), requiring either dedicated hardware accelerators or a significant reduction in encoding quality through simplified encoder settings. This complexity wall is a major barrier to the wider adoption of VVC and reduces its competitiveness against deep learning-based approaches that can use parallel Graphics Processing Unit (GPU) computation.

2.3.2 Limitations for Emerging Content and Applications

Conventional video coding works well for the content types and scenarios it was designed for, but it faces important limitations when applied to emerging visual applications.

Ultra-high resolution content (8K and beyond) highlights the limits of the block-based approach, as the cost of searching for the best coding decisions grows faster than linearly with resolution. The spatial predictions and motion estimates that work well at lower resolutions also become less effective as scene complexity increases.

HDR and WCG content, encoded with transfer functions such as the PQ or Hybrid Log-Gamma (HLG) defined in the International Telecommunication Union Radiocommunication Sector (ITU-R) BT.2100 [67] using the wide colour primaries of BT.2020 [68], has very different statistical properties from the SDR content that conventional codec tools were designed for. The non-linear mapping between code values and light levels in HDR makes it harder to design effective prediction and quantisation tools, which were originally tuned for linear or gamma-compressed SDR signals [69].

Immersive 360-degree video and light field video represent the scene as a sphere or volumetric capture, and require projection formats and inter-prediction tools suited

to their geometric structure. While extensions such as OMAF (Omnidirectional Media Application Format) address some of these needs, the block-based architecture of conventional codecs is not well suited to spherical geometry.

3D point cloud data, which consists of unordered sets of 3D coordinates with associated attributes, cannot be directly processed by the frame-based architecture of conventional video codecs. MPEG has developed separate standards for point cloud compression, including G-PCC (Geometry-based Point Cloud Compression) for LiDAR-acquired outdoor scenes and V-PCC (Video-based Point Cloud Compression) for dense indoor scans [88], but these standards still struggle to achieve efficient compression for machine vision tasks.

2.3.3 Fundamental Rate–Distortion Limitations

Information theory shows that the rate–distortion performance of any coding system is bounded by the theoretical rate–distortion curve of the source [89]. Conventional video codecs minimise block-level distortion measures as a practical substitute for directly optimising this bound, and this substitution leaves a gap between the performance that is achieved and what is theoretically possible. In addition, the manually designed coding tools in conventional codecs are tuned to match the average statistical properties of standardised test sequences used in common test conditions, and cannot adapt to the specific statistics of individual sequences, scenes, or content types. This content-independent design leaves residual energy in the prediction signal that content-adaptive strategies could reduce.

Together, the complexity wall, the modality limitations, and this content-independent residual energy point to a fundamental limit that further refinement of the hybrid block-based approach is unlikely to overcome. This limit motivates the two directions surveyed in the remainder of the chapter: end-to-end learned compression, which replaces manually designed transforms and predictors with data-driven alternatives (Sections 2.4–2.5), and semantic communication, which shifts the coding goal from signal accuracy to preserving meaning (Sections 2.6–2.10).

2.4 Deep Learning-Based Image Compression

2.4.1 Early Neural Network Approaches

The use of neural networks for image compression goes back several decades before the modern deep learning era. Early work explored Multilayer Perceptron (MLP) autoencoders for image coding [90], showing that neural networks could learn compact image representations, but the limited computing power at the time meant these systems could not match conventional codecs in compression quality. Renewed interest in neural network-based image compression began around 2015–2016, driven by faster GPUs, larger training datasets, and better methods for training deep networks.

2.4.2 End-to-End Learned Image Compression with Autoencoders

The first practically competitive end-to-end learned image compression system [46] trained a convolutional encoder-decoder using a smooth approximation of the quantisation step and a fixed entropy model. This work showed that jointly training the encoder, decoder, and entropy model with a rate–distortion loss could match or exceed the compression quality of JPEG2000 for natural images, particularly as measured by MS-SSIM. A follow-up scale hyperprior model [47] added a side channel that transmits scale parameters of the latent prior, improving entropy coding by allowing the decoder to use spatially varying priors. This design established the core framework of modern learned image compression: an encoder, a decoder, a quantizer, and a learned entropy model that captures the statistical structure of the quantized latent features.

This framework was extended [91] with a joint autoregressive hierarchical prior, combining the hyperprior with a context model that predicts each latent variable from its already-decoded neighbours. This joint prior better captured the spatial structure of the quantized representation, achieving compression results that surpassed Better Portable Graphics (BPG), the HEVC-based image codec, for the first time on standard benchmarks, under both PSNR and MS-SSIM metrics. The entropy model was further improved [76] with attention modules and Gaussian mixture priors, achieving better rate–distortion performance with sharper reconstructions at low bitrates. These devel-

opments established end-to-end learned image compression as a strong competitor to conventional image coding standards.

2.4.3 Generative Models for Semantic Image Compression

An alternative approach to learned image compression uses generative models, particularly GANs, to reach very high compression ratios by generating visually plausible reconstructions from compact latent codes. A GAN-based extreme compression method [92] showed that plausible images can be reconstructed from very compact representations, producing better perceptual quality than conventional codecs at very low bitrates, but at the cost of reduced pixel-level accuracy. High-fidelity generative image compression [93] combined a learned encoder with a GAN-based decoder and showed that good perceptual quality can be achieved at bitrates well below those of conventional codecs.

These generative compression methods are closely related to semantic communication, as both transmit a compact semantic code rather than a pixel-accurate description, and rely on the decoder to fill in the visual details. However, generative decoders are not deterministic and can produce content that was not present in the original image, which limits their use in applications that require faithful reconstruction.

2.4.4 Vision Transformers for Image Compression

The transformer architecture, which uses self-attention to capture long-range dependencies, has been applied to learned image compression. Transformer-based image compression models [94,95] use attention to capture both fine local textures and global image structure, which CNNs handle less effectively, and have been shown to outperform CNN-based methods on standard benchmarks. These results reflect the wider shift in computer vision from convolutional networks to transformer-based models for tasks that require understanding the full image context.

2.5 End-to-End Learned Video Compression

2.5.1 From Image Compression to Video Compression

Image compression and video compression share the same core pipeline of transform coding, quantisation, and entropy coding, but they differ fundamentally in the source of redundancy they exploit. An image codec such as JPEG [48] or a learned image codec [47, 76] operates on a single frame in isolation and can only exploit spatial redundancy, which is the correlation between nearby pixels within the same image. Natural video sequences, however, contain an additional and often substantial source of redundancy: temporal redundancy across successive frames. Consecutive frames in a typical video sequence are highly similar because much of the scene, particularly the background, does not change between frames and moving objects often move by only a small distance per frame interval. Coding every frame independently therefore repeats information that could instead be predicted from reconstructed reference frames.

Video codecs exploit temporal redundancy through motion-compensated inter-frame prediction [37, 38]. The encoder divides the current frame into blocks and searches one or more reconstructed reference frames for regions that best match the current blocks; this search is motion estimation. The spatial displacement between a current block and its selected reference region is represented by a motion vector. Motion compensation uses the decoded motion information and reconstructed reference samples to construct a prediction of the current block. Instead of coding the full block, the encoder codes the prediction residual, which is the difference between the original block and its motion-compensated prediction. The residual usually contains less energy than the original block and can therefore be represented more compactly. Motion information must also be signalled, but for well-predicted content this overhead is normally outweighed by the reduction in residual rate.

Video codecs commonly organise pictures into Groups of Pictures (GOPs) [36, 38], which define a repeated prediction and reference structure over a sequence. An intra-coded picture uses only spatial prediction within that picture and does not depend on other pictures. Designated random-access intra pictures, such as Instantaneous De-

coder Refresh (IDR) or Clean Random Access (CRA) pictures, allow decoding to begin without reconstructing all earlier pictures and help limit error propagation. However, GOP size and random-access or intra period are distinct parameters: not every GOP must begin with an intra picture, and an open GOP may use references from an earlier GOP. A P-picture uses inter prediction from reference pictures in one reference list, usually pictures earlier in display order. A B-picture can use one or two reference lists and commonly combines pictures before and after it in display order. A later-displayed reference must be encoded and decoded first, so hierarchical B-picture coding requires picture reordering and can add delay. When both temporal directions provide accurate references, B-pictures often achieve a lower residual rate than P-pictures at similar quality. H.264/AVC, H.265/HEVC, and H.266/VVC support flexible GOP and reference structures [36–38]. Random-access configurations commonly arrange B-pictures in a temporal hierarchy between anchor pictures, with implementation-dependent GOP sizes such as 8, 16, or 32; selected B-pictures may themselves become references for higher temporal layers.

The compression advantage of inter-frame prediction over intra-only coding is large for temporally correlated content. Motion-compensated prediction can substantially reduce residual energy compared with intra prediction, which is why video codecs generally require fewer bits than independent image-by-image coding at the same quality. This principle also carries over to the inter-frame learned video codecs considered in this thesis. They replace one or more manually designed operations with trained neural networks while using reconstructed reference information to code a residual, latent correction, or conditionally represented current frame.

2.5.2 Optical Flow-Based Frameworks

Extending learned image compression to video requires effective modelling of the temporal similarity between frames. DVC [49], the first fully end-to-end trainable video compression framework, replaced the manually designed motion estimation and compensation of conventional codecs with learned optical flow estimation and warping. DVC uses a learned optical flow network to estimate dense flow fields between refer-

ence and current frames, a motion compression network to entropy-code the flow, and a residual compression network for the prediction error. All components are trained jointly using a single rate-distortion loss, allowing the motion and residual coding to be optimised together. DVC achieved performance on par with H.265/HEVC in terms of MS-SSIM, which was a major milestone for learned video compression.

Subsequent work improved upon DVC in several directions. Scale Space Flow (SSF) [50] proposed a scale-space optical flow representation that is better suited to entropy coding. FVC [96] proposed deformable convolution-based warping in the feature domain as an alternative to pixel-domain flow, allowing more flexible motion modelling. RLVC [54] introduced recurrent coding with feedback loops to capture longer-range temporal dependencies.

2.5.3 Conditional Coding and Contextual Frameworks

A major shift in approach came with the introduction of conditional video coding, where each frame is encoded using temporal context extracted from previously decoded frames. Rather than using motion only to form a pixel-domain prediction and then coding an explicit subtraction residual, conditional codecs encode the current-frame representation conditioned on temporally aligned context available to both the encoder and decoder.

Early conditional video coding methods were proposed in [50, 97]. The DCVC framework [51] established this approach on solid ground. It estimates and compresses motion, uses the decoded motion to align features extracted from the previously decoded frame, and supplies the resulting feature-domain contexts to the current-frame encoder, decoder, and entropy model. The current frame is therefore coded conditionally on this context rather than as an explicit pixel-domain residual. DCVC achieved performance that surpassed H.265/HEVC across standard benchmarks.

The DCVC family of codecs has been improved through several follow-up works. DCVC-TCM [98] improved temporal context by propagating multi-scale intermediate features between consecutive frames alongside reconstructed pixel-domain references, providing richer context for conditional encoding and enabling more compact represen-

tation of inter-frame differences. DCVC-HEM [99] proposed a hybrid spatial-temporal entropy model that combines autoregressive spatial context modelling with temporal prior propagation, exploiting both spatial and temporal correlations and achieving an 18.2% bitrate saving over H.266/VVC on the UVG dataset. DCVC-DC [53] increased context diversity through hierarchical quality patterns across frames, long-range temporal context reuse from intra-coded reference frames, group-based optical flow offsets for richer motion representation, and quadtree-based spatial partitioning that allows parallel entropy decoding. DCVC-FM [52] addressed two key practical limitations of earlier learned video codecs: the narrow quality range of a single model (extended from about 3.8 dB to 11.4 dB PSNR coverage using learnable quantisation scalers), and quality degradation over long inter-frame prediction chains (addressed through a periodic temporal feature modulation mechanism that prevents error accumulation). In a single-intra-frame setting, DCVC-FM achieves a 29.7% average bitrate saving over the previous state-of-the-art learned video codec and surpasses the Enhanced Compression Model (ECM), the research prototype that exceeds H.266/VVC, on standard benchmarks under low-delay conditions. DCVC-FM is used as the primary learned video codec comparative baseline throughout the evaluations presented in this thesis. Together, these developments establish learned video compression as a clearly superior alternative to H.265/HEVC across a wide range of content, with performance approaching H.266/VVC on standard benchmarks, while retaining the potential for content-adaptive specialisation that supports further gains in the hybrid semantic coding approach of this thesis.

The labels “optical-flow-based” and “DCVC-based” do not denote mutually exclusive families, because the original DCVC and later DCVC variants also use learned motion estimation and transmit entropy-coded motion latents [51–53]. The main distinction is how the motion-aligned reference information is used. In predictive residual codecs such as DVC [49] and SSF [50], the decoded flow warps a reference frame to produce a pixel-domain prediction, and a second coding path explicitly represents the subtraction residual between that prediction and the current frame. This closely mirrors conventional hybrid coding, in which motion information and transform residuals

are separately signalled. In DCVC, decoded motion instead aligns reference features to construct multi-scale temporal contexts. These contexts condition the current-frame transform and entropy model, so the transmitted current-frame latent represents what is needed given the context without requiring an explicit pixel-domain subtraction residual [51]. Later DCVC variants enrich this conditional context through multi-scale features, longer-term references, group-based offsets, or feature modulation. Both designs depend on sufficiently accurate motion information and can be affected by occlusion, scene changes, and quantised motion; conditional coding does not remove these problems, but its learned current-frame latent can compensate for context errors. Its richer feature-domain context can improve coding efficiency, at the cost of greater model complexity and tighter dependence on a shared neural decoder. The hybrid systems developed in this thesis are different from both families: a semantic decoder generates the base reconstruction, while the standard VVC path explicitly codes the remaining pixel-domain residual. They are nevertheless informed by DCVC’s broader principle of using learned temporal context to improve current-frame representation.

2.5.4 Challenges in Learned Video Compression

Despite their impressive rate–distortion performance, end-to-end learned video codecs still face several practical challenges. Training instability across a wide quality range is a recurring issue, as the rate–distortion loss function is non-convex and sensitive to the weighting between the rate and distortion terms. High computational latency at inference, particularly for encoding, limits the use of learned codecs in real-time applications. The fixed model architecture makes it difficult to adjust the codec for different computational requirements at the encoder or decoder. Perhaps most significantly, performance can drop substantially for content types not well represented in the training data, including animated video, screen captures, medical images, and document video. These generalisation challenges highlight the need for content-adaptive approaches that can tailor the codec to each input sequence.

Crucially, even the strongest learned video codecs still optimise pixel-level or perceptual quality metrics and remain unaware of the meaning of the transmitted signal,

just as conventional codecs do. Sensitivity to channel impairments, inconsistent generalisation across content types, and the lack of task awareness together motivate a fundamental reformulation of the communication objective itself. The following section traces the theoretical origins and practical developments of this reformulation, namely semantic communication, which provides the conceptual foundation for the contributions of this thesis.

2.6 Foundations of Semantic Communication

2.6.1 Shannon and Weaver’s Three-Level Framework

Building on the overview in Section 1.1, this section gives the theoretical detail needed by the later semantic rate–distortion formulations. In *The Mathematical Theory of Communication*, Shannon established the mathematical basis of information theory, defined channel capacity as the maximum rate for reliable transmission over a noisy channel, and proved the channel coding theorem that sets the basic limit for reliable communication [56, 57]:

$$C = W \log_2(1 + \text{SNR}) \quad (2.1)$$

where C is the channel capacity in bits per second, W is the channel bandwidth in hertz, and SNR is the signal-to-noise ratio.

Weaver’s introductory essay in the same volume described three levels of communication. The *technical problem* concerns how accurately symbols can be transmitted from sender to receiver, and this is the problem addressed by Shannon’s framework. The *semantic problem* concerns whether the transmitted symbols carry the intended meaning. The *effectiveness problem*, or pragmatic problem, concerns whether the received meaning produces the intended effect at the receiver.

For the next sixty years, communication engineering focused almost entirely on the technical problem. Information was treated as abstract bit sequences, and the main goal was reliable delivery. The semantic and effectiveness problems were mostly treated as outside engineering and left to fields such as language studies and philosophy. This focus was reasonable for building general communication infrastructure, because

a channel that delivers bits reliably can support many applications.

2.6.2 Renewed Interest Driven by Deep Learning

Renewed interest in semantic communication has been driven by two trends. First, the rapid growth of data traffic is pushing communication networks closer to the limits set by Shannon theory, which creates a strong need for systems that transmit meaning rather than raw bits [59]. Second, deep learning has matured enough to provide practical tools for extracting, representing, and reconstructing the semantics of complex signals such as images, video, speech, and text [55].

Together, these trends have made semantic communication a practical engineering approach rather than only a theoretical idea. A semantic communication system has a semantic encoder at the transmitter, a physical channel, and a semantic decoder at the receiver. The encoder extracts a compact representation of the meaning of the source signal, and the decoder reconstructs the signal from that representation. The efficiency gain comes from the fact that the semantic representation is usually much smaller than a bit-level representation, because the encoder can discard information that is redundant given the shared knowledge between the encoder and decoder.

Shared knowledge, or context, is therefore central to semantic communication. If the encoder and decoder share a common source model, vocabulary, or knowledge base, the encoder only needs to send the information that identifies the specific message within that shared model. This idea can be expressed using rate–distortion theory with decoder side information [100], which shows that the minimum rate required for a given distortion level becomes smaller when the decoder has correlated side information.

2.6.3 Deep Learning Implementations

Modern semantic communication systems usually implement the semantic encoder and decoder with DNNs trained jointly to minimise a combination of semantic distortion and transmission rate. The most common architecture is the autoencoder [101], which has an encoder that maps the source signal to a compact latent representation and a decoder that reconstructs the signal from that representation. When the model is

trained to minimise mean squared error or a task-specific loss, it learns to extract the semantically relevant features of the source.

CNNs are widely used for semantic encoding of visual signals because their hierarchical feature extraction matches the spatial structure of images and video well [102]. Transformer architectures [103], which use self-attention to capture long-range dependencies, have also been applied to many vision tasks and are increasingly used in semantic communication. Variational Autoencoders (VAEs) provide a probabilistic framework for learning compact representations with clear priors, which helps efficient entropy coding of the semantic representation.

2.6.4 Semantic-Aware Networking and Cross-Modal Communication

Deploying semantic communication at scale requires not only better semantic encoders and decoders, but also changes in network architecture and protocol design. Semantic-aware networking was introduced in [104], where network nodes consider the semantic content of transmitted data instead of treating all data as undifferentiated bit streams. In that framework, routing, resource allocation, and quality-of-service decisions are guided by semantic relevance rather than only technical measures such as throughput and Bit Error Rate (BER). A semantic fidelity metric was also proposed to replace BER as the main quality measure for semantic communication links, allowing end-to-end optimisation of network resources for semantic delivery.

An early empirical investigation [105] applied deep learning to semantic communication for text, using a transformer-based encoder–decoder architecture to transmit sentence meaning over wireless channels. It showed that semantic meaning can be preserved at low SNR values where conventional bit-level text transmission fails completely, and highlighted the benefit of channel-noise augmentation during training for robust generalisation to unseen channel conditions.

Cross-modal semantic communication [106] has also been studied for multi-modal services such as video, audio, and haptic signals, where cross-modal correlations among different signal types are exploited to reduce transmission overhead for immersive applications. By transmitting only the information needed to bridge the gap between

modalities at the receiver, these systems can achieve compression ratios that are not possible within a single modality. Context-based semantic communication via dynamic programming [107] was also proposed for text transmission, using a part-of-speech-based encoding strategy in which dynamic programming selects which words must be transmitted and which can be recovered from shared context at the receiver, balancing semantic fidelity and transmission overhead across different channel conditions.

2.7 Semantic Communication for Text

Text was the first domain in which semantic communication was explored and where early practical systems were developed. Early text-based semantic communication systems used Natural Language Processing (NLP) models to extract the meaning of messages and transmit compact semantic representations that could be decoded at the receiver.

Word embeddings such as Word2Vec [108] and GloVe [109] provided the first practical semantic representations of text by mapping words to dense vectors, where words with similar meanings are placed close together. Later context-aware language models such as ELMo [110], Bidirectional Encoder Representations from Transformers (BERT) [111], and Generative Pre-trained Transformer (GPT) [112] improved these representations by capturing how word meanings change with context.

DeepSC [55] is a transformer-based semantic communication system for text transmission that encodes whole sentences into compact channel symbols for JSCC. Instead of recovering every word exactly, DeepSC optimises semantic similarity between the transmitted and received sentences, which allows robust communication at low SNR values where conventional bit-level codecs fail. Task-oriented and multi-user variants [113] were later introduced to improve the semantic representation for specific downstream tasks such as machine translation.

Adaptive transformer architectures for semantic communication [114] have also been explored, showing that adapting the model to channel and task conditions can further improve communication efficiency.

2.8 Semantic Communication for Speech and Audio

Semantic communication for speech is motivated by the fact that many speech applications, such as voice assistants, automatic speech recognition, and machine translation, mainly care about the linguistic content rather than the exact waveform. A semantic speech communication system can therefore send a compact representation of the linguistic content and use a speech synthesis module at the receiver to generate intelligible speech.

Weng and Qin [115] proposed a semantic communication system for speech transmission that uses a DNN to compress speech into semantic features for wireless transmission, achieving large bitrate reductions over conventional speech codecs at similar intelligibility. The DeepSC-ST system [116] extends semantic communication to speech-to-text tasks by transmitting only the transcript and speaker identity features and then using neural speech synthesis at the receiver to reconstruct intelligible speech.

2.9 Semantic Communication for Images

2.9.1 Deep Joint Source–Channel Coding

Semantic communication for image transmission has produced the largest body of research, mainly because image transmission is important in wireless applications and large image datasets are available for training.

Deep JSCC [62], a pioneering convolutional autoencoder trained end-to-end as a joint source–channel codec for wireless image transmission, was the first practical semantic image communication system. Unlike the conventional approach, which separates source coding and channel coding, Deep JSCC maps image pixels directly to complex-valued channel symbols and adapts the representation to the channel conditions. This system removes the cliff effect seen in conventional coded systems at low SNR, where image quality suddenly collapses after decoding failure, and instead gives graceful quality degradation. Kurka and Gündüz [63] later extended Deep JSCC with feedback, using channel state information at the transmitter to allocate bandwidth

adaptively across image regions.

Later work extended Deep JSCC in several ways. One example is an SNR-adaptive Deep JSCC system [117] that uses channel-wise attention to adapt coding to changing channel conditions and works robustly across a wide SNR range with one model.

Despite these advantages, Deep JSCC and its variants have several limitations that distinguish them from the hybrid semantic video coding systems proposed in this thesis and make a direct codec comparison inappropriate. First, Deep JSCC operates in the domain of *channel uses per pixel* rather than *bits per pixel*: the encoder jointly learns source representation and channel protection by mapping each image directly to complex-valued channel symbols whose number is set by a fixed bandwidth compression ratio. It therefore has neither a separate source bitstream nor a separately identifiable channel-code bitstream. Mapping its channel uses onto a bits-per-pixel axis requires either an idealised channel-capacity assumption at a specified channel model and SNR, or a specified modulation, channel code, and target error rate. The resulting value is system-dependent and is not directly equivalent to the entropy-coded source rate reported for a video codec. Second, although SNR-adaptive variants can cover multiple channel qualities with one model, the bandwidth compression ratio remains an operating-point constraint; evaluating another ratio generally requires retraining or a specifically designed variable-bandwidth model. Third, the original Deep JSCC architecture was designed for single images and AWGN or slow-fading channels. Extending it to video requires temporal modelling, and video JSCC systems [118] still differ from the proposed hybrid codec in source type, channel assumptions, delay, and the absence of a standardised residual path for high-fidelity output. Fourth, Deep JSCC does not produce a standardised interoperable bitstream and requires the matching learned decoder. A controlled comparison would therefore require a video JSCC baseline evaluated with the same source sequences, bandwidth ratio, channel model, SNR, latency, and quality metrics; it would measure end-to-end transmission quality rather than conventional source-codec rate-distortion performance. For this reason, the hybrid video coding experiments in this thesis are not plotted directly against image-oriented Deep JSCC results. They instead use conventional rate-distortion curves in bits per pixel

against VVC under JVET common test conditions, which permits reproducible comparison with conventional and learned video codec baselines.

2.9.2 Region of Interest Semantic Coding

In the context of semantic multimedia communication, a Region of Interest (ROI) is defined as a spatially localised sub-region of an image or video frame that carries greater semantic significance than the surrounding content for the intended communication task. More formally, let Ω denote the spatial domain of an image or frame and let $s_{\mathcal{T}}(p)$ be the importance assigned to pixel location $p \in \Omega$ for task $\mathcal{T}$. For a selected threshold τ , the ROI is

$$\mathcal{R}_{\mathcal{T}}(\tau) = \{p \in \Omega : s_{\mathcal{T}}(p) \geq \tau\}.$$

The definition of ROI is therefore not absolute; it depends on the communication goal and on how the importance function $s_{\mathcal{T}}$ is obtained. Three broad approaches have been studied in the literature.

The first approach uses classical visual saliency detection, where importance is determined by low-level contrast, luminance discontinuities, and centre-surround differences that attract human visual attention. This is a bottom-up, stimulus-driven process that does not require semantic labels or task specification, and it is computationally efficient. However, it does not directly capture semantic relevance for machine inference tasks and can misidentify visually prominent but semantically unimportant regions as high-priority.

The second approach defines ROI based on object importance or semantic relevance, where specific categories of content such as faces, text, traffic signs, or moving objects are treated as semantically critical. This approach is task-driven and requires either a pretrained object detector or a semantic segmentation model to identify the relevant regions. It is well suited to goal-oriented communication [55], where the receiver needs specific semantic content rather than high pixel-level fidelity across the whole frame. The personalised saliency framework in [119] extends this idea to multi-user settings, where individual user interest profiles determine which regions receive higher quality.

The third approach uses learned attention mechanisms, where the semantic encoder

learns, through end-to-end training, to assign higher bit allocation or lower distortion to regions that carry more information for reconstruction or downstream task performance. The attention map may directly define a soft importance function, or a supplied ROI mask may condition the learned bit allocation. The ROI scalable image compression system in [120] illustrates ROI-conditioned learned coding: a semantic autoencoder first produces a base reconstruction of the full image, and then dedicated enhancement layers are applied selectively to the identified ROI to improve its quality, while non-ROI background areas retain only the lower-quality base reconstruction. This keeps the total bitrate the same as a uniform-quality system while improving perceptual quality in the regions that matter most to the viewer or downstream task. Thus, ROI is an umbrella concept rather than a single detection method; its implementation may be based on visual saliency, object or task relevance, learned attention, or a combination of these criteria.

2.9.3 Task-Oriented Semantic Communication

Task-oriented semantic communication systems optimise the transmitted representation for a specific inference task rather than for image reconstruction quality. A joint encoding and inference system for wireless image retrieval [121] used a CNN encoder to extract task-relevant features from images for remote retrieval over wireless channels and showed that sending only task-relevant features gives large savings compared with full image transmission. An autoencoder-based end-to-end semantic communication system for task-oriented M2M image transmission [122] trained the encoder and decoder jointly under different channel noise conditions and showed robust image reconstruction together with high classification accuracy. It outperformed JPEG in compression efficiency and achieved image-quality gains of up to 17 dB under channel impairments. These task-oriented frameworks support the hybrid and context-aware semantic video coding systems developed later in this thesis, which preserve task-relevant temporal context rather than exact pixel values.

2.9.4 Practical Wireless Semantic Image Communication Systems

Beyond the original Deep JSCC architectures, a growing body of work has studied practical semantic image transmission over realistic wireless channels. One end-to-end semantic image transmission system [58] transmits only a semantic segmentation map of each image over an AWGN channel protected by polar codes and uses a pretrained conditional GAN at the receiver to reconstruct realistic images from the received map. This system achieves about 20 times compression compared with JPEG by using only about 5% of the bandwidth required by JPEG, and it outperforms JPEG at low SNR where conventional JPEG decoding fails.

Domain knowledge-driven semantic communication for image transmission over wireless channels [123] used a dual-path framework that encodes both signal-level and concept-level semantics, with domain-specific conceptual knowledge guiding image reconstruction at the receiver. This approach showed that adding domain knowledge at the concept level improves image restoration quality over systems that rely only on signal-level redundancy reduction and remains robust across different channel conditions.

Generative JSCC for semantic image transmission [124] introduced two JSCC schemes that use StyleGAN to improve perceptual quality for wireless image transmission. GenerativeJSCC trains an encoder and a StyleGAN-based decoder end-to-end using Mean Squared Error (MSE) and Learned Perceptual Image Patch Similarity (LPIPS) loss functions, while InverseJSCC refines the output of a standard DeepJSCC decoder using the image priors of a pretrained StyleGAN. Both methods outperform standard DeepJSCC in perceptual quality at very low bandwidth compression ratios and low SNR, and GenerativeJSCC also reduces pixel-level distortion.

A joint framework for modelling semantic noise and technical noise in semantic image communication over wireless channels [125] established a noise model that separately describes distortion from the semantic encoding–decoding process and distortion from the physical channel. With turbo-coded channel protection, it enabled robust semantic image transmission across a wide SNR range. Related to this, an autoencoder-based image quality metric for semantic noise measurement [126] used the Euclidean

distance between the latent vector spaces of a reference autoencoder and its response to a semantically generated or received image as a measure of semantic fidelity. It achieved an 88% correlation with subjective quality assessment and showed that conventional image quality metrics do not fully capture semantic distortion. Although both works are the author’s own, the quality metric requires training a reference autoencoder for each test image, which is too computationally expensive for the video codec benchmarking used in this thesis. For that reason, PSNR, SSIM, and VMAF are used as the main evaluation metrics throughout the thesis.

An asymmetric autoencoder-based complexity-scalable semantic image transmission system [127] was proposed for resource-constrained Internet of Things (IoT) and wireless sensor network endpoints. It showed that placing more computation at the encoder and less at the decoder can reduce device-side processing while keeping competitive visual quality. A scalable image compression and communication framework that combines semantic communication with residual deep learning-driven bitplane coding [128] showed that iterative bitplane refinements can gradually improve reconstruction quality and enable flexible bitrate adaptation for bandwidth-constrained wireless links. Taken together, these studies show that semantic communication for images is a practical option for resource-constrained wireless applications, with robustness and perceptual quality available within one end-to-end framework.

2.10 Semantic Communication for Video

2.10.1 Video Conferencing Applications

Semantic communication has been applied to video transmission in several specific cases. One wireless semantic communication system for video conferencing [129] takes advantage of the predictable structure of talking-head video: the background is almost static, and the same speakers remain throughout the session, so the system transmits only compact facial keypoints to represent motions rather than full frame data. For this type of content, it dramatically reduces transmission resources compared with conventional codecs while maintaining high resolution.

This result shows the potential of semantic communication for constrained applications, but it depends on assumptions such as a static background and an unchanged speaker, which limits its use for general video content. Building on the hybrid semantic–VVC framework of Chapter 3 and the earlier noise modelling work [125], a lossless distributed video coding system for error-prone wireless channels was also developed [130]. It combines turbo channel coding with an autoencoder-based semantic layer to provide robust transmission and reliable lossless video recovery under the target channel conditions.

2.10.2 Video Compression for Machines

An important application area for semantic video communication is Video Coding for Machines (VCM), which compresses video for machine-vision tasks such as object detection, action recognition, and tracking rather than for human viewing quality [131]. VCM systems achieve high compression by keeping only the features needed for the target task and removing visual detail that is not useful for machine analysis. The ISO/IEC MPEG standards group is actively developing a VCM standard [132] to define a bitstream format for compressing both video and extracted features for multiple machine-vision tasks.

2.10.3 Hybrid Semantic and Conventional Video Coding

The hybrid video coding system in [1], which is, as established by the systematic search described in Section 2.12, the first to combine all three defining properties of semantic autoencoder inter-frame prediction, an unmodified VVC residual coding path, and parameter-free remote decoder training, uses an autoencoder for inter-frame semantic prediction and VVC for intra-frame coding and residual compression. This hybrid approach combines the strengths of both methods: compact semantic representations for temporal prediction and the reliable standardised decoding of VVC for accurate reconstruction. Benchmarking against H.264/AVC, H.265/HEVC, and H.266/VVC across diverse video sequences showed consistent BD-rate improvements for both low- and high-complexity content, with the VVC residual coding stage correcting semantic

reconstruction errors. This framework is discussed in detail in Chapter 3.

2.10.4 Deep Learning for Wireless Video Semantic Transmission

Extending semantic communication to general video transmission over wireless fading channels requires systems that can model temporal structure while remaining robust to channel noise. Wireless deep video semantic transmission [118] was proposed as a system that jointly trains a video semantic encoder and a channel-adaptive decoder for end-to-end transmission over wireless fading channels. It uses nonlinear transforms and a conditional coding architecture to extract semantic features from video frames and gives smooth quality degradation as SNR decreases across different video content, unlike the sudden failure of conventionally coded video at low SNR. Tests on practical fading channels showed that end-to-end semantic video codecs can maintain acceptable visual quality at SNR levels where conventional coded transmission fails, making this approach practical for bandwidth-constrained scenarios.

The wireless semantic communication framework for video conferencing [129] uses the predictable nature of talking-head video to achieve much better compression than general-purpose codecs, showing that domain-aware semantic encoding can provide large gains when the content type is well known to both encoder and decoder. Together, these systems cover a range of approaches, from domain-specific systems designed for one content type to general-purpose hybrid codecs that work across diverse content. The hybrid semantic-VVC framework developed in this thesis belongs to the second group, using semantic inter-frame prediction to reduce residual information while relying on VVC for standardised and reliable reconstruction.

2.11 Survey of Recent Semantic Communication Systems

Before turning to the five sub-fields that directly support the thesis contributions, this section reviews broader trends in semantic communication research that help place this thesis in context. These trends include multi-user and network-level operation, generative and Large Language Model (LLM)-enabled encoding, federated training, and

information-theoretic views of task-oriented communication. They are not addressed directly in this thesis, but they define the wider setting in which the proposed video coding systems are expected to operate.

2.11.1 Multi-User and Network-Level Semantic Communication

Beyond point-to-point systems, multi-user semantic communication for 6G networks has also been explored. A semantic communication framework for multi-user image transmission [119] used personalised saliency encoding based on individual user interests together with a game-theory-based resource allocation scheme for multi-user wireless channels. It showed improved communication quality and personalisation. This extension highlights the potential of semantic communication in dense network settings.

2.11.2 Generative AI and Large Language Model-Enabled Semantic Communication

The rise of LLMs and diffusion-based generative models has opened new possibilities for semantic communication. LLMs can use the broad knowledge stored in pretrained models to produce very compact semantic representations for transmission [133, 134]. Diffusion-based decoders can reconstruct high-quality images or video from very compact semantic codes [135]. These generative methods can achieve very high compression ratios, but they also introduce reconstruction uncertainty, which is a problem for applications that require accurate output. This motivates the hybrid design used in this thesis, where a residual coding stage corrects errors introduced by the semantic stage.

2.11.3 Federated Learning for Semantic Communication

Federated learning has also been applied to semantic communication so that semantic encoders and decoders can be trained across multiple wireless edge devices without sending raw source data to a central server [136]. In one such system, a wav2vec-based convolutional autoencoder extracts compact semantic features from audio signals, and federated learning coordinates model updates across devices. The system achieves close to a 100-fold reduction in mean square error compared with traditional coding.

2.11.4 Task-Oriented Communication and Semantic Information Theory

Several research groups have also worked on formalising task-oriented communication using information theory. A tutorial on the information-theoretic foundations, learning-based approaches, and practical system designs for context, semantics, and task-oriented communication was presented in [137]. It explained how semantic distortion, the information bottleneck, and goal-oriented communication can be built into system design. A rate-distortion framework for goal-oriented communication [138] was also developed, providing information-theoretic bounds for source coding under two distortion constraints. These studies provide an important theoretical basis for designing and analysing task-oriented semantic communication systems.

With this broader context established, the chapter now turns to focused reviews of the five sub-fields that directly support the thesis contributions, starting with hybrid semantic and VVC-based video coding.

2.12 Hybrid Semantic and VVC-Based Video Coding

2.12.1 Motivation for Hybrid Architectures

The strengths and weaknesses of learned video codecs and semantic communication systems complement each other, so a hybrid approach is a natural next step. End-to-end learned codecs achieve strong compression performance, but they lack interpretability, standardisation, and operational robustness. Semantic communication systems compress video by extracting meaning, but they cannot guarantee accurate reconstruction when used on their own. Conventional codecs, in contrast, provide proven and standardised coding of prediction residuals. A hybrid architecture that uses semantic communication for inter-frame prediction and a conventional codec for residual and key-frame coding can therefore combine the strengths of all three approaches. The author's earlier work on semantic communication for images directly motivated this design. Studies of task-oriented M2M image transmission [122], semantic and technical noise modelling [125], bitplane-based scalable image communication [128], and error-resilient

rate-scalable image compression [139] each showed that semantic encoders alone cannot guarantee pixel-accurate reconstruction, but work well when combined with a residual coding stage that corrects semantic reconstruction errors.

2.12.2 Prior Work on Hybrid Learning and Conventional Coding

Earlier work explored fully learned video codecs as alternatives to conventional coding pipelines. ELF-VC [140], for example, is a fully end-to-end learned flexible-rate video codec that uses a novel in-loop flow prediction scheme for motion modelling and surpasses H.264/AVC, H.265/HEVC, and AV1 across the rate-distortion curve under low-latency conditions (I- and P-frames only). However, these fully learned methods replace the conventional codec completely rather than combining learned semantic representations with the reliable residual coding tools of standard codecs. An extension of the hybrid semantic-VVC framework in Chapter 3 was also proposed in [141], where semantic communication is applied to chroma component prediction. In that system, chroma channels are predicted from the luma component using a deep neural network, improving rate-distortion performance over conventional chroma prediction within the same hybrid framework.

To determine whether any earlier work had combined all three defining properties of the proposed system, a systematic search of IEEE Xplore, the ACM Digital Library, and arXiv (categories cs.CV and eess.IV) was conducted prior to the submission of [1]. The defining properties were (i) an unmodified VVC intra and residual coding path, (ii) a semantic autoencoder replacing only the inter-frame prediction loop, and (iii) parameter-free remote decoder training via shared deterministic initialisation. The search used the keyword combinations “semantic communication AND video coding”, “hybrid semantic codec”, “autoencoder inter-frame prediction”, and “semantic communication VVC”. It identified three research directions as the closest candidates: collaborative-intelligence and split-computing hybrids [142, 143] that split a learned inference network between encoder and decoder but do not replace the inter-frame prediction loop of a standard video codec and do not produce a pixel-domain reconstruction; end-to-end learned video codecs such as the DCVC family [51–53] that replace the

entire conventional pipeline with neural networks, requiring a large shared pretrained decoder and giving no compatibility with the VVC bitstream; and deep JSCC image systems [62,63] that combine source and channel coding in a single learned mapping but transmit continuous channel symbols rather than an entropy-coded residual bitstream and include no standardised conventional residual path. None of the papers identified in this search combined all three defining properties simultaneously. The novelty claim for the proposed system therefore rests on this documented search.

2.12.3 Bidirectional Hierarchical Prediction

Bidirectional prediction, where each frame is predicted from both past and future reference frames, is a standard tool in conventional video coding that improves coding efficiency by using temporal correlation in both directions. In semantic video coding, it lets the semantic encoder use a richer set of reference frames for each encoded frame, which reduces the amount of information that must be transmitted. An enhanced hybrid semantic-VVC codec with bidirectional hierarchical frame prediction was proposed in [22], showing further rate-distortion gains over a simple interpolation baseline together with improved robustness to transmission errors. This system is described in Chapter 3.

2.13 Context-Aware and Content-Adaptive Semantic Video Coding

2.13.1 Pretrained Foundation Models for Video Coding

A key challenge in deploying semantic video coding systems is the high computational cost of per-GOP training. The hybrid semantic codec in [1] trained a new autoencoder for each GOP, which allowed content-specific optimisation but also caused very long encoding times. Using one pretrained model for all content removes this cost, but it also reduces the content-specific adaptation that improves compression.

A selectively adapted pretrained model offers a useful middle ground. In video coding, a model pretrained on a large and varied set of video data can learn a general

semantic representation of natural video. For sequences with specific properties, such as high motion or unusual textures, fine-tuning only a small subset of parameters can recover content-specific performance without the full cost of per-GOP training.

A context-conditioned semantic video coding framework [73] based on a shared pretrained model with motion-adaptive adaptation was proposed to remove per-GOP training while keeping competitive rate-distortion performance. The framework uses RIFE [72] to generate motion-aligned context frames that guide semantic decoding.

Three relevant approaches can be used to generate motion-aligned context: classical differential or variational optical flow, learned dense flow estimation such as RAFT [144], and task-specific video frame interpolation such as RIFE [72]. Classical methods estimate displacement from assumptions such as brightness constancy, local linear motion, and spatial smoothness. They require no training and can be attractive when model storage is constrained, but large displacement, motion boundaries, occlusion, illumination change, and non-rigid motion can violate their assumptions. Producing an intermediate frame also requires flow scaling, warping, blending, and an explicit strategy for occluded regions. RAFT builds all-pairs correlation volumes and iteratively refines dense endpoint-to-endpoint flow through a recurrent update operator, giving high accuracy on optical-flow benchmarks [144]. It is therefore a strong choice when the required output is a flow field. For the present task, however, RAFT’s output must be incorporated into a separate interpolation pipeline, normally including bidirectional flow estimation or approximation, temporal scaling, warping, and occlusion-aware blending. These additional operations are not a second neural-network inference by definition, but they must be designed and evaluated separately, and RAFT does not itself provide an arbitrary-time interpolated frame.

RIFE is trained for the required interpolation task and returns the synthesised intermediate frame in one network invocation [72]. Internally, its IFNet estimates bidirectional intermediate flows, warps the two input frames, and uses a learned fusion mask and refinement network to combine them; thus, RIFE integrates warping and occlusion-aware fusion rather than eliminating warping. Its temporal input encodes the target position as a fractional offset, allowing arbitrary-timestep interpolation. The systems

in this thesis use the midpoint case ($t = 0.5$) recursively within the hierarchical GOP, so the same model can be applied at every temporal level without separately scaling end-point flow. The RIFE paper reports more than 30 fps for $2\times$ interpolation of 720p video on an NVIDIA RTX 2080Ti [72]. This published throughput is not a controlled speed comparison with RAFT on identical hardware, but it demonstrates that RIFE was designed for efficient frame synthesis. Direct task output, integrated flow estimation and fusion, support for the required midpoint operation, and reported interpolation speed therefore motivated its selection. The trade-offs are that the fixed RIFE model and its computation are required at both encoder and decoder, and interpolation can still fail for severe occlusion, large temporal gaps, or non-linear motion; the VVC residual path corrects the remaining prediction error. The implementation and its ablation are presented in Chapter 4.

2.13.2 Content-Adaptive Parameter Overfitting

Adapting neural network models to specific content through fine-tuning is well known in transfer learning. In video coding, researchers have explored overfitting implicit neural representations to individual video frames [145], where each frame is represented by the quantized weights of a small coordinate-based network. This approach removes the need for a shared pretrained model between encoder and decoder, but it also causes high encoding latency because each frame must be optimised separately. Overfitting-based image compression [146] follows a similar idea by using a neural network that overfits to each image and then transmitting the network weights as the compressed representation.

A content-adaptive parameter overfitting mechanism for per-GOP adaptation achieves content-specific adaptation with low overhead by updating only a small set of per-channel learnable scale multipliers and transposed-convolution biases, about 643 scalars per GOP, which is less than 0.1% of the full decoder. The parameter updates are differentially encoded with a unified PCA [147]–DeepCABAC [148] pipeline alongside the differential semantic latents. This mechanism is presented in Chapter 4.

2.13.3 Differential Latent Encoding and DeepCABAC

The efficiency of semantic video coding depends strongly on how the semantic latent representations are entropy coded. Differential encoding sends the difference between the current frame’s latent representation and a predicted latent from earlier decoded frames. By using temporal correlation in latent space, it reduces the amount of information that must be transmitted. DeepCABAC [148] provides an arithmetic-coding framework for these differential latents by estimating their probability distribution from context and achieving near-optimal lossless compression.

2.13.4 Rate–Distortion Optimisation for Hybrid Semantic Video Coding

Designing a hybrid semantic video coding system requires a rate–distortion objective that includes the semantic latent rate, the DNN parameter update rate when per-GOP adaptation is used, and the conventional residual coding rate, while still optimising perceptual quality. A unified context-conditioned rate–distortion framework was proposed in [73]. It treats the semantic encoder, the motion-aligned context generator, the differential latent entropy coder, and the VVC residual path as jointly optimised components under one Lagrangian objective. In this formulation, the total rate is the sum of the semantic latent rate, entropy coded by DeepCABAC, and the residual rate, coded by VVC. The distortion is the pixel-domain difference between original and reconstructed frames weighted by a Lagrange multiplier. This framework allows systematic evaluation of the rate–distortion trade-off across different operating points and produces rate–distortion curves comparable to those of conventional codecs while reflecting the structure of the hybrid pipeline. When per-GOP parameter overfitting is used, the DNN parameter update rate is added as another rate term so that the optimisation can balance the cost of sending parameter updates against the compression gain from content adaptation. Full details are given in Chapter 4.

2.14 Semantic Video Coding for High Dynamic Range Content

2.14.1 High Dynamic Range Imaging

HDR imaging captures the full perceptual range of light intensities in natural scenes, from very dark regions to bright sunlight, without the clipping and banding that limit SDR imaging. HDR content is encoded with specialised transfer functions that map a wide luminance range to digital code values in a perceptually uniform way. The PQ transfer function [149], standardised in ITU-R BT.2100 [67], maps absolute luminance values up to 10,000 cd/m² to 10-bit or 12-bit code values using a curve that closely follows human contrast sensitivity. The HLG transfer function, also defined in ITU-R BT.2100, supports backward compatibility with SDR UHD displays by using a conventional gamma curve for the lower half of the signal range, so SDR-capable UHD displays can still reproduce a usable image without HDR metadata. However, HLG is not fully compatible with legacy BT.709 SDR displays because it uses BT.2020 colour primaries, which can cause visible colour desaturation on those devices. WCG content uses wider colour primaries, usually the BT.2020 colour space [68], which covers about 75.8% of the International Commission on Illumination (CIE) 1931 chromaticity diagram, compared with 35.9% for the BT.709 colour space used in HDTV.

2.14.2 HDR Video Coding in Conventional Standards

H.265/HEVC was extended to support HDR and WCG coding through range extensions and support for 10-bit and 12-bit coding with PQ and HLG transfer functions [69]. H.266/VVC natively supports HDR/WCG content through flexible bit-depth and colour-format support, and improvements for HDR content were a direct target during standardisation. Perceptually uniform metrics such as PU-PSNR [27] and PU-SSIM were also developed for HDR evaluation, taking into account the non-linear relationship between code values and perceived light intensity.

2.14.3 Tone Mapping for HDR–SDR Bridging

A common way to process HDR content with SDR-based tools is tone mapping, which compresses the HDR luminance range into the narrower SDR range while trying to preserve perceptual quality. The photographic tone-mapping operators in [150], including a global operator based on log-average luminance and a local dodging-and-burning operator, produce visually pleasing SDR versions of HDR images. The global Reinhard operator gives a simple and efficient mapping between HDR and SDR domains that can be inverted analytically if the per-frame exposure parameters are preserved.

The content-adaptive semantic video codec was extended to HDR sources through an automatic per-frame invertible Reinhard tone-mapping bridge. This allows an SDR-trained semantic codec to process linear HDR EXR content without changing the codec architecture. The method estimates the tone-mapping parameters for each frame in closed form from the log-average luminance and sends them as a small metadata side stream for exact HDR recovery at the receiver. This approach is described in Chapter 5.

2.14.4 Domain Gap Between HDR and SDR Semantic Codecs

A basic challenge in applying SDR-trained semantic codecs to HDR content is the mismatch between their source statistics. HDR content encoded with the PQ transfer function has a luminance histogram and spatial correlation structure that differs strongly from the gamma-compressed or linear SDR content on which autoencoder-based semantic codecs are usually trained. In particular, PQ encoding maps highlights and bright regions to the upper part of the 10-bit or 12-bit scale, where values are rare in SDR training data. As a result, the semantic encoder may produce latent representations that lie outside the training distribution and are decoded poorly by the semantic decoder.

The bridge introduced above treats this mismatch as a domain-adaptation problem outside the semantic codec, thereby preserving the SDR architecture. Chapter 5, particularly Sections 5.2.1–5.2.1, defines the forward and inverse mappings, transmitted metadata, quantisation limitation, and quantitative evaluation.

2.15 Multiple Description Coding for Robust Transmission

2.15.1 Multiple Description Coding Theory

MDC is a source-coding strategy designed for networks with multiple parallel paths, where each path may fail independently [70, 71]. In MDC, a source is encoded into two or more *descriptions*. Each description is sent over a different path and can be decoded on its own to produce a useful, though imperfect, reconstruction. When several descriptions are received together, the decoder combines them to obtain a higher-quality reconstruction.

The rate–distortion theory of multiple descriptions was established in two key works. El Gamal and Cover [71] derived the achievable rate–distortion region, or inner bound, for general two-description sources, while Ozarow [151] obtained the complete rate–distortion region for the Gaussian two-description case. Goyal [70] later surveyed the field and reviewed achievability and converse bounds as well as practical coding methods such as transform-based MDC, lattice vector quantisation-based MDC, and redundancy-based approaches.

2.15.2 Practical MDC Methods

Several practical MDC methods have been developed. Multiple description scalar quantisation [152] designs a scalar quantiser whose output is split into an index pair, one index for each description, with the assignment optimised to minimise reconstruction distortion under different loss patterns. Multiple description coding based on redundant frame expansions [153] uses overcomplete frame representations so that any received subset of frame coefficients still gives a useful reconstruction, and quality improves as more coefficients arrive. Multiple description lattice vector quantisation [154] uses lattice-based codebooks with designed index-pair assignments to generate descriptions with predictable side and central distortions.

For video, MDC is especially important because the temporal prediction structure of conventional codecs creates error-propagation chains: one lost packet can corrupt

all following frames until the next intra-coded refresh. MDC-based video coding limits this error propagation by ensuring that a separately decodable description is always available [155]. Choosing the number of descriptions and the amount of redundancy creates a fundamental trade-off. More redundancy lowers side distortion and improves reconstruction from a single description, but it also increases the total transmission rate because more bits are used for protection instead of new source information. The achievable rate–distortion region of two-description MDC [71] captures this trade-off and defines the inner bound within which practical MDC schemes must operate.

Recent learned methods have moved description generation from hand-designed quantisers and transforms into learned latent representations. LMDC jointly optimises side and central distortion, shares parameters between its autoencoder branches, and uses importance maps to allocate bits according to image content [156]. INR-MDSQC instead overfits a small implicit neural representation to each image and interleaves two hierarchical latent sets, allowing the redundancy level to be adjusted without training a shared model on a large dataset [157]. These methods improve image compression efficiency and recovery from missing descriptions, but do not jointly study semantic description generation with explicit channel coding and physical AWGN, Rayleigh, and Rician fading models. This distinction motivates the channel-aware SMDC evaluation in Chapter 6.

2.15.3 Semantic Multiple Description Coding

The combination of semantic communication and multiple description coding was proposed as a way to remove cliff effects in semantic communication over wireless fading channels. The main idea is that the feature maps in the latent space of a semantic autoencoder can be split into several complementary subsets, and each subset can be decoded independently from any received subset by a partial decoder. This channel-wise splitting creates descriptions that complement one another: each description carries a different subset of the source’s semantic features, and any received subset can still give a meaningful, though incomplete, reconstruction.

The SMDC framework proposed in this thesis keeps visual quality at useful levels

under severe channel conditions that cause complete failure in single-path semantic communication and conventional coded transmission. Full quantitative evaluation against JPEG, JPEG2000, HEIF, and learned compression baselines across AWGN, Rayleigh, and Rician fading channels is presented in Chapter 6.

2.15.4 Fading Channel Models

The performance of wireless communication systems over practical channels is strongly affected by multipath propagation. The Rayleigh fading model [158] assumes that the channel envelope follows a Rayleigh distribution, which corresponds to environments with many independent scatterers and no line-of-sight component, such as dense urban areas or non-line-of-sight indoor channels. The Rician fading model [158] includes a line-of-sight component in addition to scattered components and is characterised by the Rician K -factor, which is the ratio of the line-of-sight power to the scattered power. The AWGN channel model provides an idealised baseline for comparison.

2.16 DNN Parameter Efficiency for Semantic Video Coding

2.16.1 Neural Network Compression

Semantic video coding systems that adapt their DNN parameters to each video sequence or GOP need an efficient way to send updated parameters from encoder to decoder. Uncompressed DNN parameters require substantial storage and transmission bandwidth, which can cancel the compression gains of the semantic approach. Neural Network Compression (NNC) addresses this problem through quantisation, pruning, and encoding of neural network weights.

The NNCodec toolkit [159], developed by the Fraunhofer Heinrich Hertz Institute, is the open-source implementation of the ISO/IEC 15938-17 NNC standard. It compresses neural network parameters through quantisation, sparsification, and arithmetic coding to achieve high compression ratios with little loss in network performance. Deep-CABAC [148] is the arithmetic coding component used within NNCodec and applies

CABAC to the statistical properties of neural network weights.

2.16.2 Structured Pruning and Low-Rank Decomposition

Structured pruning reduces neural network size by removing entire filters, channels, or layers based on magnitude or gradient-based importance measures. This produces smaller networks that are cheaper to store and transmit. Low-rank decomposition factors weight matrices into products of smaller matrices, reducing the number of parameters while approximately preserving accuracy. A three-stage pipeline of pruning, quantisation, and Huffman coding [160] has been shown to reduce network size substantially with only small accuracy loss.

2.16.3 Temporal Prediction of DNN Parameters

A new strategy for reducing DNN parameter transmission overhead in semantic video coding uses the strong temporal correlation between parameters trained on consecutive video frames. Because consecutive GOPs in natural video often share similar statistical properties, the autoencoder parameters trained for them are also very similar, so only a small residual update, or parameter delta, needs to be transmitted. This principle was demonstrated in [25], which showed that temporal prediction of decoder parameters can reduce transmission overhead to a small fraction of the full parameter set while achieving PSNR gains of 3–25 dB and average bitrate savings of 54% compared with NNCodec. Full details are given in Chapter 7.

2.16.4 VVC-Based DNN Weight Compression

DNN weight matrices can be viewed as two-dimensional signals with spatial correlations, especially when the weights are arranged by filter index. This suggests that image compression tools can be reused to compress DNN weights. Treating DNN weight matrices as 2D signals and compressing them with VVC was proposed in [24]. This method uses the spatial correlations in the weight matrices and achieves consistent bitrate reductions compared with NNCodec while maintaining similar video quality. It therefore provides a practical way to reduce parameter transmission overhead in semantic video

coding. Building on the hybrid semantic–VVC framework of Chapter 3, a further extension of this idea was developed in the author’s own work [161]. That work combined VVC-based neural network compression with semantic communication for video transmission and showed that the spatial correlations in DNN weight matrices can be used together with the semantic coding pipeline for consistent bitrate reduction. Full details are presented in Chapter 7.

2.16.5 Neural Network Compression Standards

The standardisation of neural network compression has progressed in parallel with semantic video coding. The NNCodec toolkit introduced earlier in this section implements MPEG NNC, ISO/IEC 15938-17, and its DeepCABAC entropy-coding component [148, 159]; it therefore provides the standardised parameter-bitstream baseline used by the later experiments.

The distinction between this standard baseline and the spatial and temporal parameter-coding contributions is developed in Chapter 7; Section 7.2 provides the direct quantitative comparison with NNCodec.

Having reviewed the prior work across the five sub-fields that support the thesis contributions, the next section introduces the objective and perceptual evaluation metrics used throughout the thesis to measure the rate–distortion, perceptual, and HDR-specific performance of the proposed systems against the conventional and learned baselines reviewed above.

2.17 Use of the Term “Semantic Communication” in This Thesis

Throughout this thesis, the term “semantic communication” follows Shannon and Weaver’s Level B view of communication [57], where the goal is not exact symbol reproduction but preservation of meaning at the receiver. Applied to video coding, this means that the encoder should transmit only the information needed for the decoder to recover the meaningful content of the current frame, rather than sending a com-

plete pixel-accurate description. In practical terms, a semantic video encoder learns a compact latent representation that keeps the task-relevant and perceptually important structure of the frame while discarding details that can be inferred from shared knowledge already available at both ends.

In the systems developed in this thesis, that shared knowledge is the decoded video context, such as I-frames, previously reconstructed reference frames, or motion-aligned contextual information. The semantic encoder therefore does not need to describe the whole frame from scratch. Instead, it sends a low-rate latent that explains the current frame relative to this shared context. This is the main way in which semantic communication theory is used in the proposed video coding frameworks: communication efficiency comes from exploiting shared knowledge and transmitting only the information that is still semantically necessary after that shared knowledge is taken into account. This is fundamentally different from standard learned codecs, such as the variational hyperprior [47] or the DCVC family [51], where the transmitted latent must carry essentially all information needed for final reconstruction.

However, semantic communication also introduces an important limitation. Because the semantic encoder intentionally removes part of the source information, the reconstructed frame after semantic decoding is not identical to the original frame even when the physical channel is ideal. This distortion can be understood as *semantic noise*, that is, the loss introduced by semantic abstraction and reconstruction rather than by channel corruption. The distinction between semantic noise and technical channel noise was formalised in [125]. In video coding, this semantic noise appears as missing fine detail, imperfect textures, or reconstruction errors that remain after the semantic decoder has produced its low-bitrate approximation of the frame.

The central idea of the hybrid systems in Chapters 3-5 is therefore to use VVC to mitigate this semantic noise. The semantic module first generates a compact, meaning-preserving base reconstruction at very low bitrate, and VVC then encodes only the residual difference between this semantic reconstruction and the original frame. In this way, semantic communication provides the low-rate representation, while VVC removes the semantic noise that would otherwise limit reconstruction fidelity. This

combination directly applies semantic communication theory to practical video coding: the semantic stage preserves meaning efficiently, and the conventional residual codec restores the signal details needed for high-quality visual reconstruction. For this reason, the systems proposed in this thesis are semantic not because they ignore distortion, but because they first code meaning and then use VVC as a corrective stage to compensate for the semantic distortion that remains.

The same semantic principle also appears in Chapter 6, where the SMDC systems are trained to preserve useful content under channel uncertainty. There, the emphasis is on robustness to channel impairment, but the underlying idea is the same: the transmitted representation is designed to preserve meaning first, while graceful degradation or later reconstruction refinement handles the information that is not fully preserved in the initial semantic representation.

2.18 Evaluation Metrics for Video Coding Quality

2.18.1 PSNR and Bjøntegaard Delta Rate

PSNR is the most widely used objective video quality metric in the video coding research community, defined as:

$$\text{PSNR} = 10 \log_{10} \left(\frac{L^2}{\text{MSE}} \right) \text{ [dB]} \quad (2.2)$$

where L is the maximum signal value (255 for 8-bit content) and MSE is the mean squared error between the original and reconstructed frames. PSNR is widely used due to its simplicity and computational efficiency, though it has well-known limitations in predicting perceptual quality.

Beyond its perceptual limitations, PSNR has a deeper shortcoming in the context of semantic communication: it measures pixel-level distortion but gives no indication of whether semantic content has been preserved. This is particularly relevant for M2M applications, where the receiver is an automated inference system rather than a human viewer. No standard metric for evaluating semantic fidelity has yet been established in the literature. The author’s own work on task-oriented semantic communication for

M2M image transmission [122] noted explicitly that no image quality metric is available for semantic communication and therefore used a CNN-based image classifier alongside PSNR to assess performance. The results showed that reconstructed images with a PSNR of approximately 18–22 dB were sufficient to maintain classification accuracy above 90%, demonstrating that pixel-level PSNR alone is an incomplete indicator of semantic system quality. For this reason, PSNR, SSIM, and VMAF are used together as complementary metrics throughout this thesis, while acknowledging that task-accuracy measures such as classification accuracy would be required for a complete semantic quality evaluation in M2M settings.

The BD-rate metric [78, 162] provides a way to compare the rate–distortion performance of two codecs over a range of quality levels. BD-rate is computed as the average percentage bitrate difference needed to achieve the same quality (PSNR, SSIM, or VMAF) between a proposed system and a reference codec, averaged over a range of quality levels. A negative BD-rate indicates that the proposed system requires less bitrate than the reference for equal quality, representing a compression improvement.

2.18.2 Measuring Semantic Preservation and What the Hybrid Codec Preserves

A central challenge in evaluating semantic communication systems is that conventional distortion metrics such as PSNR and SSIM were designed to measure pixel-level or structural similarity, not the preservation of meaning. A system that perfectly preserves semantic content but introduces perceptually irrelevant texture softening will score lower on PSNR than one that reproduces unimportant background detail while losing task-relevant structure. Three approaches to measuring semantic preservation have been studied. Task-accuracy measurement runs a downstream inference task on the reconstructed output and reports accuracy relative to the original [55, 122]; this directly measures meaning preservation but cannot serve as a general codec benchmark because it requires a fixed downstream task. Latent-space similarity metrics use the distance between encoder representations of the original and the reconstruction as a proxy for semantic fidelity; the autoencoder-based metric in [126] achieves 88% correla-

tion with subjective quality assessment using this principle. Learned perceptual metrics such as LPIPS [163] measure feature-level similarity using a pretrained network and correlate better with human judgement than PSNR, but are not calibrated to semantic communication tasks and are not yet standard in codec benchmarks.

These alternatives measure different meanings of semantic fidelity and are not interchangeable. Task accuracy is appropriate when the receiver performs a defined machine-inference task, whereas latent-space and learned perceptual distances are proxies whose result depends on the selected feature extractor. The hybrid video codecs in Chapters 3–5 target reconstructed video for human viewing rather than a fixed downstream inference task; therefore, there is no task label against which a universal semantic accuracy can be reported. No dedicated semantic metric is claimed for their experimental results. Instead, PSNR measures pixel fidelity, SSIM measures structural similarity, VMAF provides a learned perceptual estimate, and BD-rate measures the bitrate difference at matched values of one of these quality metrics. These measures do not prove preservation of meaning, but they are appropriate for comparing the final human-viewing output because the semantic prediction is followed by a VVC residual correction stage. Task accuracy would be required for an M2M application, while latent-space similarity or LPIPS could provide complementary analysis of the semantic-only reconstruction in future work.

Semantic autoencoders trained with an MSE loss under a tight bottleneck tend to retain the low- and mid-frequency luminance structure that contributes most strongly to the squared-error objective. Consequently, coarse object shapes and strong boundaries generally remain recognisable. This should not be interpreted as perfect edge preservation: sharp edge transitions, fine texture, and subtle chrominance gradients contain higher-frequency or lower-energy information and can be smoothed by the bottleneck. In the hybrid codec used throughout this thesis, the VVC residual coding stage encodes the pixel-level difference between the original frame and the semantic reconstruction, reducing these losses by restoring texture, colour, and edge sharpness according to the selected quantisation level. The final frame is reconstructed by adding the VVC-decoded residual to the semantic prediction, as formalised in Eq. (3.16). This

is the fundamental division of work: the semantic stage provides a compact structure-preserving prediction, and VVC corrects the signal-level detail that the semantic stage cannot carry. At the final output (PSNR 30–42 dB), conventional metrics are informative because the residual stage has restored substantial pixel-level accuracy. At the semantic-only stage (PSNR 28–35 dB before residual correction), SSIM is more useful because its structural sensitivity better reflects the preservation of object boundaries and coarse shape even when fine texture is absent [164]. VMAF combines spatial quality features with a motion feature and therefore provides a broader perceptual indicator than framewise PSNR or SSIM [165]; however, it is not a dedicated measure of flicker or temporal consistency.

In the enhanced and context-aware frameworks, temporal information is preserved through three mechanisms. First, the hierarchical bidirectional GOP structure predicts each intermediate frame from decoded past and future references, preserving its temporal position and exploiting motion continuity from both directions; fully reconstructed intermediate frames then become references at the next hierarchical level, as described in Section 4.2.2. Second, RIFE estimates intermediate optical flows and a fusion mask from the two nearest decoded references to generate a motion-aligned context frame [72], as defined in Eq. (4.5). This context carries motion and scene continuity that do not need to be represented again in the semantic latent. Third, the VVC residual stage encodes the remaining framewise difference between the target and semantic prediction. Adding the decoded residual reduces spatial errors and temporal fluctuations left by semantic prediction, although the correction remains dependent on residual quantisation. Together, these mechanisms improve temporal coherence across a GOP while keeping the semantic bitrate low. The reported VMAF results provide indirect perceptual evidence because VMAF includes a motion feature, but no dedicated flicker or optical-flow-warping metric was evaluated; temporal consistency should therefore not be inferred from VMAF alone.

2.18.3 PU-PSNR and PU-SSIM for HDR Content

For HDR content evaluation, the PU encoding transforms HDR code values to a perceptually uniform scale before applying conventional metrics, so that quality differences are correctly captured across the full luminance range [27]. PU-PSNR and PU-SSIM apply standard PSNR and SSIM to PU-encoded values and give more reliable quality scores for HDR video than directly applying PSNR or SSIM to PQ-encoded code values.

2.19 Preliminary Work on Semantic Communication for Images

Before the video-focused contributions of this thesis were developed, a series of preliminary studies on semantic communication for still images established the main design principles and identified the key limitations that motivated the move to video. These image-domain studies are the author’s own work and are directly referenced throughout the thesis.

An autoencoder-based end-to-end semantic communication system for task-oriented M2M image transmission [122] showed that a jointly trained encoder–decoder pair can provide robust image reconstruction and high classification accuracy under different channel noise conditions. It outperformed JPEG in compression efficiency and produced image-quality gains of up to 17 dB under channel impairments. This study confirmed that semantic autoencoders are a practical basis for wireless image communication and directly motivated their use as the inter-frame prediction backbone in the hybrid video codecs developed in this thesis.

A region-of-interest scalable image compression system using semantic communication [120] showed that assigning higher quality to perceptually important image regions and stronger compression to background areas improves perceptual quality at the same overall bitrate. This selective compression strategy informed the design of the content-adaptive semantic video coding systems developed in later chapters.

A joint framework for modelling semantic noise and technical noise [125] established

a noise model that separately describes distortion introduced by the semantic encoder–decoder and distortion caused by the physical channel. With turbo-coded channel protection, it enabled robust image transmission across a wide SNR range. This work showed that semantic and channel distortions must be treated separately, and that principle underpins the hybrid coding and multiple description coding frameworks in this thesis.

An autoencoder-based image quality metric for semantic noise measurement [126] used the Euclidean distance between latent vector spaces to measure semantic fidelity and achieved an 88% correlation with subjective quality assessment. This work showed that conventional metrics such as PSNR and SSIM do not fully capture semantic distortion, which reinforces the need for task-aware evaluation in semantic video communication systems.

An asymmetric autoencoder-based complexity-scalable semantic image transmission system [127] for resource-constrained IoT and wireless sensor endpoints showed that placing more computation at the encoder while keeping the decoder lightweight can maintain competitive visual quality. This asymmetric design principle is carried forward in the encoder–decoder architecture of the semantic video codecs in this thesis.

A scalable image compression and communication framework using residual deep learning-driven bitplane coding [128] showed that iterative bitplane refinements can gradually improve reconstruction quality and allow flexible bitrate control for bandwidth-constrained wireless links. This idea of residual refinement is extended in this thesis to video, where VVC codes the residual left by the semantic inter-frame predictor.

An error-resilient rate-scalable image compression framework [139] further showed that combining a semantic layer with a residual coding stage can provide both quality scalability and robustness to packet loss. This directly motivated the SMDC framework for video introduced in this thesis.

Together, these preliminary studies established three consistent findings: (1) semantic autoencoders alone cannot guarantee pixel-accurate reconstruction but work well when paired with a residual coding stage; (2) semantic noise and channel noise must be handled separately for reliable wireless communication; and (3) asymmetric

complexity and bitrate scalability are important for practical deployment. These findings motivated the move from image-level semantic communication to the video-level systems developed in this thesis, where the same principles are applied to the more demanding problems of temporal prediction, HDR content, fading-channel robustness, and DNN parameter efficiency.

2.20 Summary of Research Gaps

The literature review in this chapter identifies the following main gaps in the existing literature, which together motivate the contributions of this thesis.

Gap 1: Hybrid Semantic–Conventional Video Coding. While semantic communication has been studied widely for specific applications such as video conferencing and surveillance, its integration with VVC in one hybrid architecture for general-purpose video coding remained underexplored. This thesis addresses this gap in Contributions 1 and 2 (Chapter 3).

Gap 2: Context-Aware Semantic Video Coding with Pretrained Models. Hybrid semantic video codecs required per-GOP training, which limited practical deployment. The use of a shared pretrained model with selective motion-adaptive fine-tuning and hierarchical bidirectional prediction remained underexplored. This gap is addressed in Contribution 3 (Chapter 4).

Gap 3: Content-Adaptive Per-GOP Parameter Overfitting. The idea of per-GOP content specialisation through overfitting a very small differentially coded subset of decoder parameters had not yet been fully developed as a practical approach. This gap is addressed in Contribution 4 (Chapter 4).

Gap 4: Semantic HDR Video Coding. State-of-the-art hybrid semantic video codecs operate only in the SDR domain. The design of an invertible tone-mapping bridge that allows an SDR-trained codec to process HDR content had received limited attention. This gap is addressed in Contribution 5 (Chapter 5).

Gap 5: Semantic Multiple Description Coding. The integration of semantic communication with multiple description coding for path-diversity-based robustness over wireless fading channels remained underexplored for both image and video transmission. This gap is addressed in Contributions 6 and 7 (Chapter 6).

Gap 6: Temporal Prediction of DNN Parameters. The use of temporal correlation in DNN parameters to reduce transmission overhead had received limited attention in semantic video coding. This gap is addressed in Contribution 8 (Chapter 7).

Gap 7: VVC-Based DNN Compression for Semantic Video. The use of VVC to compress DNN weight matrices treated as 2D signals had received limited attention in semantic video communication. This gap is addressed in Contribution 9 (Chapter 7).

Table 2.1 summarises these research gaps and the corresponding thesis contributions, and maps the existing literature to the original work presented in this thesis.

This thesis addresses each of these gaps through a progressive programme of research, starting with hybrid semantic–VVC video coding and moving through context-aware and HDR video coding, robust transmission through SMDC, and DNN parameter efficiency for semantic video communication. The following chapters present the technical details of each contribution together with comprehensive experimental evaluation.

Table 2.1: Summary of identified research gaps and corresponding thesis contributions

Research Gap	Addressed by	Chapter
General-purpose hybrid semantic-VVC video coding remains under-explored	Contributions 1 & 2	Chapter 3
Per-GOP training limits deployability of semantic video codecs	Contribution 3	Chapter 4
Lightweight per-GOP content-adaptive parameter overfitting remains understudied	Contribution 4	Chapter 4
HDR content support in semantic video codecs has been largely overlooked	Contribution 5	Chapter 5
Single-path semantic systems vulnerable to deep fades	Contributions 6 & 7	Chapter 6
Temporal prediction of DNN parameters for semantic video overhead remains underexplored	Contribution 8	Chapter 7
VVC-based DNN compression for semantic video has received limited attention	Contribution 9	Chapter 7

Chapter 3

Hybrid Semantic and VVC-Based Video Coding

3.1 Chapter Introduction

Conventional video coding standards, from AVC to HEVC and VVC, have improved compression efficiency through carefully designed pipelines of block-based prediction, transform coding, and entropy coding [38]. However, this approach is reaching its limits. Each new standard increases encoder complexity sharply while delivering smaller coding gains, so further progress becomes harder to achieve [41]. At the same time, end-to-end learned video codecs such as the DCVC family [51] have shown that neural networks trained for rate–distortion optimisation can match or exceed conventional codecs. Their limitations are the need for large pretrained models at both encoder and decoder, large training datasets, and limited generalisation across content types.

The principles and limitations of semantic communication are established in Chapter 1 and reviewed in Chapter 2. This chapter therefore focuses on applying those principles to inter-frame prediction and using a conventional residual path to restore fine spatial detail.

This chapter addresses Gap 1 identified in Section 2.12. That section documents the systematic novelty search and distinguishes the proposed method from collaborative-intelligence systems, fully learned codecs, and semantic-JSCC image systems. The

chapter therefore begins from the three properties established there: unchanged VVC intra and residual coding, semantic replacement of only the inter-frame prediction loop, and deterministic remote decoder training without parameter transmission.

Section 3.2 presents the foundational architecture and remote decoder training method reported in the first-author publication “A Semantic Communication and VVC Based Hybrid Video Coding System” [1]. Section 3.3 presents the bidirectional prediction, higher-resolution processing, and residual latent coding developed in “Enhanced Semantic Communications and VVC-Based Hybrid Video Coding System” [22]. Together, these two publications establish the hybrid coding foundation extended in the later chapters. Section 3.4 summarises their combined contribution and connects it to the thesis objectives.

3.2 A Semantic Communication and VVC-Based Hybrid Video Coding System

Publication Notice

The research presented in this section has been published in the following peer-reviewed publication:

P. Samarathunga, Y. Ganearachchi, T. Fernando, A. Jayasingam, I. Alahaperuma, and A. Fernando, “A Semantic Communication and VVC Based Hybrid Video Coding System,” *IEEE Access*, vol. 12, pp. 79202–79224, 2024, doi: 10.1109/ACCESS.2024.3399174.

This material is reproduced and adapted under the terms of the IEEE Copyright policy for use in a doctoral thesis. The content has been revised and extended for inclusion and integration in this doctoral thesis, including additional explanations, expanded discussion, and alignment with the overarching research narrative.

3.2.1 Introduction

The motivation and positioning of hybrid semantic coding are established in the Chapter Introduction and Section 2.12. This section therefore focuses on the first implemen-

tation: an autoencoder provides semantic inter-frame prediction, VVC codes key frames and residuals, and synchronised Glorot initialisation [166] lets the receiver reproduce the trained decoder without parameter transmission. The evaluation compares this design with VVC, HEVC, and AVC on twenty sequences spanning different structural information and temporal information levels.

3.2.2 Methodology

System Architecture Overview

Figure 3.1 shows the proposed hybrid video codec, which combines semantic communication with VVC for rate-efficient video transmission. The system has four main parts: a semantic encoder, a semantic decoder that acts as the frame predictor, a context encoder/decoder, and a residual encoder/decoder. Semantic encoding and decoding use the encoder and decoder layers of an autoencoder (AE), while context encoding and decoding use VVC intra coding. A residual path compensates for quality loss by computing and transmitting the pixel-level difference between the original frame and the semantic decoder output. A scene transition detector controls GOP boundaries when abrupt or gradual scene changes occur in the input video. The system operates on a GOP basis in the luma and chrominance YUV 4:2:0 chroma-subsampled format (YUV420), using the same GOP settings as the VVC, HEVC, and AVC reference baselines.

Semantic Encoder

The semantic encoder maps each video frame to a compact latent vector using the encoder layers of an undercomplete AE. Figure 3.2 shows the AE architecture used for both the semantic encoder and the semantic decoder.

Each input frame has width W and height H . In YUV420 format, this corresponds to a total of $N = 3WH/2$ samples per frame, so the frame is flattened to a vector of size $N \times 1$ and passed to an input layer that is fully connected to a hidden layer of size 1024×1 with a LeakyReLU activation and nonlinearity factor $\alpha = 0.01$. This hidden layer is fully connected to a bottleneck layer of size 8×1 , again using LeakyReLU,

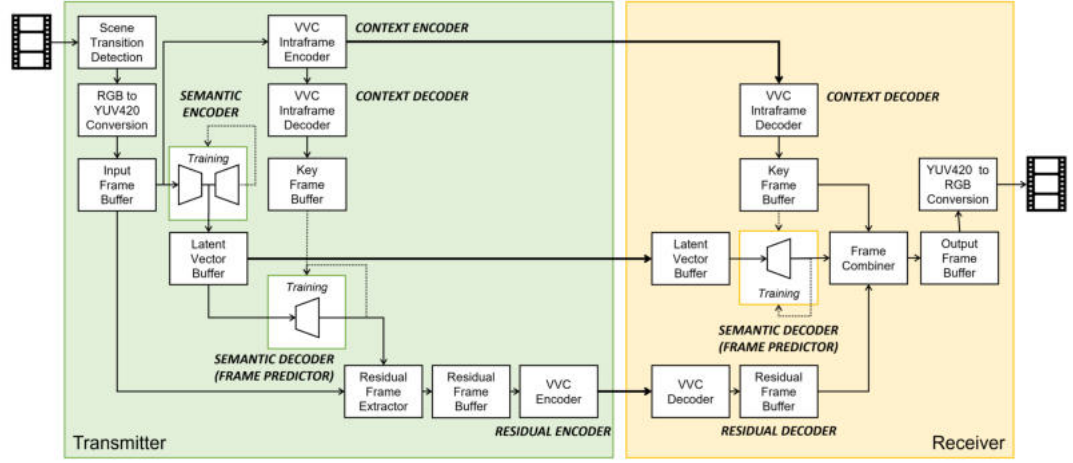


Figure 3.1: Block diagram of the proposed semantic communication and VVC-based hybrid video codec. The system comprises a semantic encoder/decoder pair realised via an autoencoder, a VVC-based context encoder/decoder for transmitting GOP key frames, and a VVC-based residual encoder/decoder for correcting semantic reconstruction errors. A scene transition detector triggers new GOPs at scene boundaries [1].

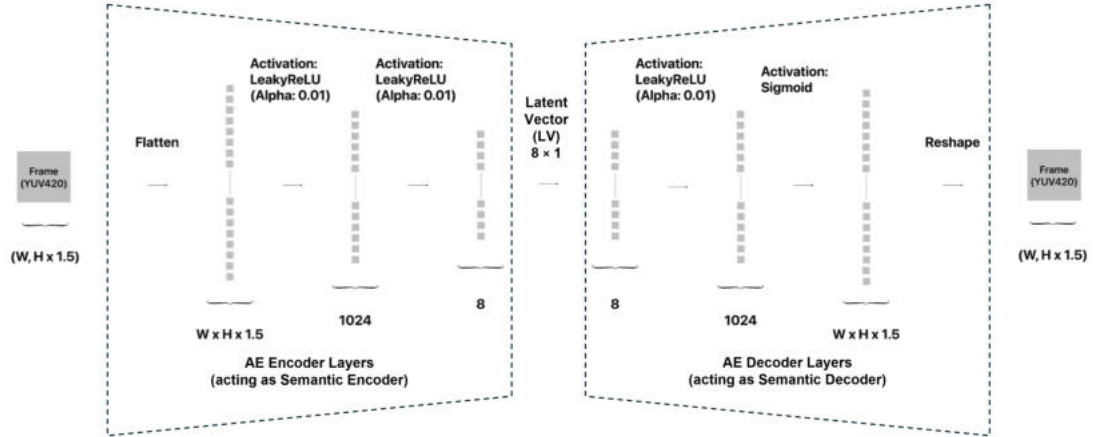


Figure 3.2: Autoencoder architecture used in the semantic encoder and decoder. The encoder layers compress each input frame to an 8×1 latent vector; the decoder layers reconstruct the frame from that vector [1].

and this bottleneck output is used as the latent vector. The decoder mirrors the same structure in reverse: a 1024×1 hidden layer followed by an output layer of size $N \times 1$ with a sigmoid activation, which keeps reconstructed pixel values in the range $[0, 1]$ and allows reshaping to the original frame size. A single hidden layer was found to be the best choice through architecture search, and the AE is trained by minimising the MSE between the input and the reconstruction:

$$L(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{x}_i)^2 \quad (3.1)$$

where $\mathbf{x} \in \mathbb{R}^{N \times 1}$ is the flattened input frame vector, $\hat{\mathbf{x}} \in \mathbb{R}^{N \times 1}$ is the AE reconstruction, and x_i and $\hat{x}_i$ denote their i th elements.

The latent vector size of 8×1 was chosen after evaluating average PSNR for sizes from 1×1 ($2^0 \times 1$) to 256×1 ($2^8 \times 1$), as discussed in Section 3.2.3. The size 8×1 is the empirical saturation point, beyond which a larger latent space gives no meaningful improvement. It is therefore the smallest representation that avoids semantic loss while keeping per-frame overhead low.

The AE is trained separately for each GOP, using all frames in that GOP as the training set so that the model intentionally overfits to the temporal content of the current scene. This per-GOP overfitting takes advantage of the strong temporal correlation between consecutive frames in a scene. For each GOP, the AE is initialised with a Glorot uniform initialiser [166] using a fixed seed. This makes the initial weights deterministic and identical at both the transmitter and the remote receiver decoder, which is necessary for successful remote decoder training, as explained in Section 3.2.2. Training runs for 250 epochs, which was found to be enough to overfit the model to the limited variation in one GOP without adding unnecessary computation.

Semantic Decoder (Frame Predictor)

The semantic decoder plays the same role as a frame predictor in a conventional video coding system. It has the same structure as the AE decoder shown in Figure 3.2 and is placed at the receiver. Instead of sending trained decoder parameters from transmitter

to receiver at every GOP boundary, which would add a large metadata overhead and reduce the coding gain, the proposed system uses a remote decoder training approach. The receiver trains its own decoder independently using the received latent vectors and the VVC-decoded key frame, without requiring a feedback channel [167].

The mathematical basis of this approach is as follows. Let $\mathbf{x} \in \mathbb{R}^{N \times 1}$ be a flattened input frame, where $N = 3WH/2$ is the number of YUV420 samples in the frame. The encoder compresses $\mathbf{x}$ to a latent vector $\mathbf{h} \in \mathbb{R}^{p \times 1}$ (where $p \ll WHC$) as:

$$\mathbf{h} = f(\mathbf{W}_{\text{enc}} \mathbf{x} + \mathbf{b}_{\text{enc}}) \quad (3.2)$$

where $\mathbf{W}_{\text{enc}} \in \mathbb{R}^{p \times N}$ and $\mathbf{b}_{\text{enc}} \in \mathbb{R}^{p \times 1}$ are the encoder weight matrix and bias vector, and $f(\cdot)$ denotes the encoder nonlinearity. The transmitter AE decoder reconstructs $\hat{\mathbf{x}} \in \mathbb{R}^{N \times 1}$ as:

$$\hat{\mathbf{x}} = f(\mathbf{W}_{\text{dec}} \mathbf{h} + \mathbf{b}_{\text{dec}}) \quad (3.3)$$

where $\mathbf{W}_{\text{dec}} \in \mathbb{R}^{N \times p}$ and $\mathbf{b}_{\text{dec}} \in \mathbb{R}^{N \times 1}$ are the decoder weight matrix and bias vector. At the receiver, the remotely trained decoder has weight matrix $\overline{\mathbf{W}}_{\text{dec}}$ and bias $\overline{\mathbf{b}}_{\text{dec}}$ that differ from the transmitter values by additive zero-mean Gaussian perturbations $\mathbf{N}_W$ and $\mathbf{n}_b$ with variances σ_w^2 and σ_b^2 applied elementwise, respectively:

$$\overline{\mathbf{W}}_{\text{dec}} = \mathbf{W}_{\text{dec}} + \mathbf{N}_W, \quad \mathbf{N}_W \sim \mathcal{N}(\mathbf{0}, \sigma_w^2 \mathbf{I}) \quad (3.4)$$

$$\overline{\mathbf{b}}_{\text{dec}} = \mathbf{b}_{\text{dec}} + \mathbf{n}_b, \quad \mathbf{n}_b \sim \mathcal{N}(\mathbf{0}, \sigma_b^2 \mathbf{I}) \quad (3.5)$$

The receiver reconstruction is $\mathbf{y} = f(\overline{\mathbf{W}}_{\text{dec}} \mathbf{h} + \overline{\mathbf{b}}_{\text{dec}})$, and the accumulated deviation $\boldsymbol{\epsilon} = \mathbf{y} - \hat{\mathbf{x}}$ is:

$$\boldsymbol{\epsilon} = f(\overline{\mathbf{W}}_{\text{dec}} \mathbf{h} + \overline{\mathbf{b}}_{\text{dec}}) - f(\mathbf{W}_{\text{dec}} \mathbf{h} + \mathbf{b}_{\text{dec}}) \quad (3.6)$$

When the receiver minimises the MSE between received latent vectors and the VVC-decoded key frame, the reconstruction deviation tends to zero:

$$\|\boldsymbol{\epsilon}\|_2 \rightarrow 0 \quad (3.7)$$

$$\Rightarrow f(\overline{\mathbf{W}}_{\text{dec}} \mathbf{h} + \overline{\mathbf{b}}_{\text{dec}}) \approx f(\mathbf{W}_{\text{dec}} \mathbf{h} + \mathbf{b}_{\text{dec}}) \quad (3.8)$$

This convergence is strengthened by initialising the decoder layers at both transmitter and receiver with the same Glorot seed [166]. This ensures identical initial states and reduces the starting deviation in (3.6). The receiver decoder is trained for 100 epochs, fewer than the 250 epochs used for the encoder, because it is optimised only against the key frame rather than all GOP frames. This lower epoch count is enough to achieve close alignment while avoiding memorisation of a single frame, which would increase the MSE for later reconstructions. Because training is done GOP by GOP, any remaining deviation at the receiver is limited to the current GOP and does not propagate further.

Context Encoder/Decoder

Context sharing between transmitter and receiver is achieved by sending the key frame of each GOP using VVC intra coding, which is a state-of-the-art method for single-frame compression [38]. Sending raw key frames would create too much bandwidth overhead and remove the savings from semantic compression, so the key frames are compressed with VVC before transmission. In the proposed system, frame reconstruction depends mainly on the latent vector rather than the key frame, so the key frame does not need to be encoded at very high quality. A higher quantisation parameter (QP) can therefore be used to minimise its bitrate contribution. The key-frame QP is chosen empirically by studying the effect of QP on bitrate and quality for both high-complexity and low-complexity video, as reported in Section 3.2.3. Based on that analysis, QP 37 is used for all key-frame encoding because it gives the lowest key-frame bitrate without causing visible degradation in the final reconstruction.

Residual Encoder/Decoder

Although the semantic and context encoding subsystem gives reconstruction quality that is sufficient for machine-vision applications, the residual between the original frame and the semantically reconstructed frame is still large, especially for high-complexity content. This causes visible quality loss for human viewing. To solve this, the proposed system uses a residual encoder/decoder that computes and transmits the frame-level difference signal with VVC intra coding. At the transmitter, a local semantic decoder

and context decoder with the same structure and settings as the receiver-side versions reconstruct a predicted frame. This transmitter-side decoder is initialised with the same Glorot seed as the encoder and trained using the same remote training procedure described in Section 3.2.2, so its output is almost identical to the receiver output. The reconstructed frame is subtracted from the original to produce a signed residual, and each residual pixel is represented as a 9-bit signed binary value to include the sign bit. The residual is then encoded with VVC using an input and internal bit depth of 10 bits. At the receiver, the decoded residual is added back to the semantic decoder output to produce the final high-quality reconstruction.

A key difference from conventional VVC residual coding is that no motion prediction or motion compensation is applied to the residual. The temporal correlation between frames has already been captured by per-GOP AE training, so VVC operates in intra-prediction mode for the residual signal. This reduces motion-estimation complexity in the residual coding stage. The GOP size and QP values for residual coding are chosen by analysing the trade-off between bitrate and reconstruction quality across three GOP sizes, as reported in Section 3.2.3. Based on this analysis, a GOP size of 32 is selected for residual encoding. The residual coding QP is not fixed. Instead, it is chosen in the range [8, 37] according to video complexity and matched to the QP values used for the reference baselines, so that fair rate-distortion comparisons can be made at equivalent quality points.

Scene Transition Detection

Scene changes introduce frames that are semantically dissimilar from the current GOP key frame, causing the semantic decoder to produce highly inaccurate reconstructions and substantially increasing the residual entropy. A scene transition detection algorithm [168] capable of identifying both abrupt and gradual changes in uncompressed video is therefore applied at the system input. Each detected transition triggers a new GOP, ensuring that the key frame always provides a valid semantic context for all frames within the GOP and limiting the propagation of any semantic prediction error.

Dataset and Test Videos

Twenty video clips of spatial resolution 320×180 pixels and frame rate 20 frames per second (fps) are used to evaluate the proposed system. Figure 3.3 shows representative frames from all twenty clips.

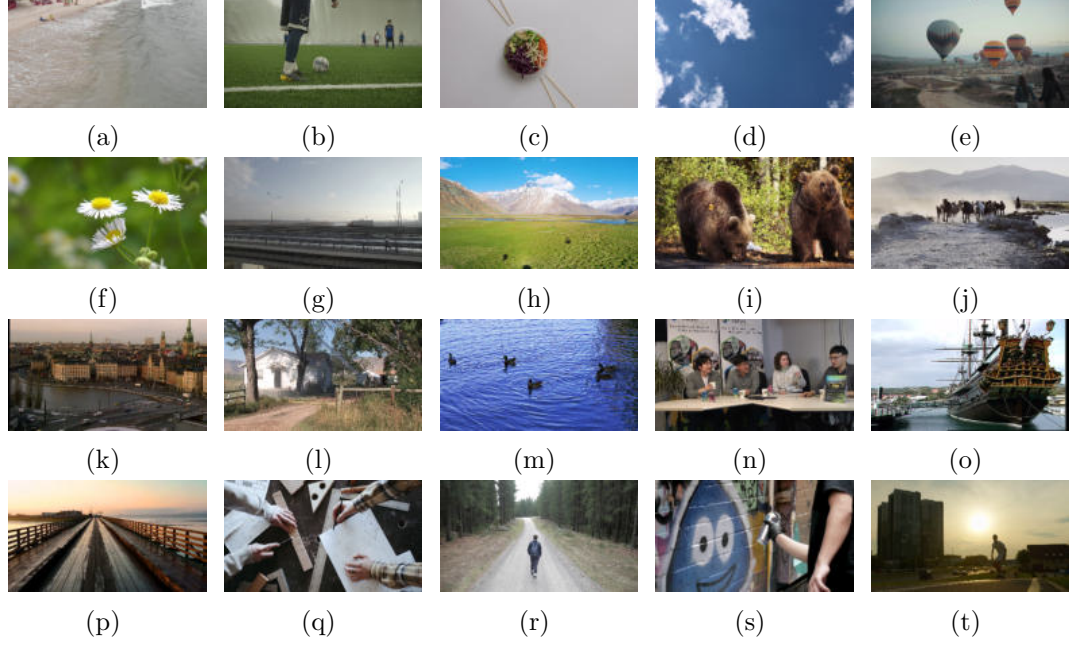


Figure 3.3: Test video clips used for evaluation [1]: (a) Video 1 [2], (b) Video 2 [3], (c) Video 3 [4], (d) Video 4 [5], (e) Video 5 [6], (f) Video 6 [7], (g) Video 7 [8], (h) Video 8 [9], (i) Video 9 [10], (j) Video 10 [11], (k) Video 11 [12], (l) Video 12 [13], (m) Video 13 [14], (n) Video 14 [15], (o) Video 15 [16], (p) Video 16 [17], (q) Video 17 [18], (r) Video 18 [19], (s) Video 19 [20], (t) Video 20 [21].

The videos were selected to represent a range of spatial and temporal complexity, quantified using Spatial Information (SI) and Temporal Information (TI) computed with the method of [169] and visualised in Figure 3.4. For analysis purposes the 20 clips are grouped into three complexity categories: low complexity (Videos 1, 3, 4, 5, 7, and 8 with low SI and low TI), medium complexity (Videos 6, 9, 10, 12, 13, 18, and 20 with medium SI and medium TI), and high complexity (Videos 2, 11, and 19 with high TI; and Videos 14, 15, 16, and 17 with high SI).

The spatial resolution of 320×180 pixels was adopted because training and evaluating the semantic encoder and decoder across multiple GOP iterations places high

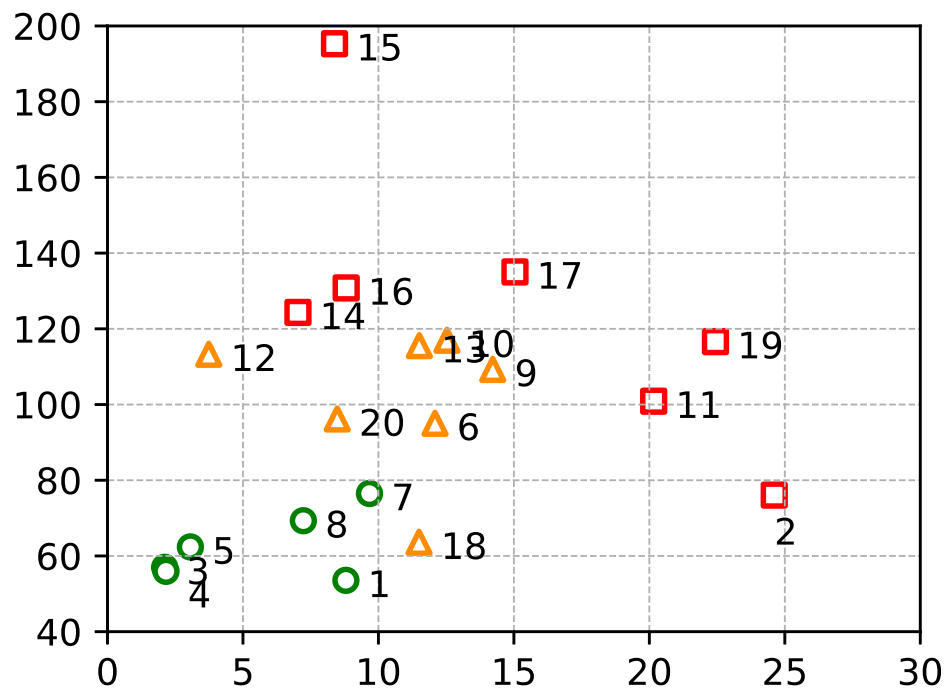


Figure 3.4: Distribution of SI and TI for the 20 test video clips. Symbols indicate complexity category: $\circ$ low, $\triangle$ medium, $\square$ high [1].

demands on computational resources; this resolution provides a practical working point. The resolution does not, however, impose a fundamental limitation on the findings, as verified in Section 3.2.3.

Baseline Codecs and Benchmark Configuration

Three conventional video coding systems serve as performance baselines: VVC (implemented using VVenC 1.8 [84]), HEVC (implemented using `ffmpeg` [170]), and AVC (also using `ffmpeg` [170]). The reference codecs are configured as shown in Table 3.1.

Table 3.1: Parameters used to configure the VVC, HEVC, and AVC baseline codecs [1].

Parameter	Value
Colour space	YUV420
Internal bit depth	8
Frame rate	20 fps
Resolution	320×180
GOP	8, 16, 32
QP	22 / 27 / 32 / 37
Preset	Slow
Pixel format	YUV420 (4:2:0)

The QP values used for the proposed system’s residual coding (22, 27, 32, and 37) are matched to those of the reference codecs, enabling a rate–distortion comparison at equivalent quality levels.

Evaluation Metrics

Rate–distortion performance is assessed using three complementary objective video quality metrics. PSNR [171] measures pixel-level reconstruction fidelity as the logarithmic ratio between the squared peak pixel value and the mean squared error, and is the most widely used metric for lossy video compression. SSIM [164] measures structural consistency in the spatial domain by comparing the luminance, contrast, and local structure of corresponding frame regions. VMAF [165] is a perceptual quality metric that combines several basic quality features into a score that correlates more closely with human visual judgement, and it is particularly sensitive to temporal quality

changes. BD-rate [162] measures the average bitrate saving (negative value) or overhead (positive value) of the proposed system relative to each reference codec across all four QP operating points on the rate–distortion curve.

3.2.3 Results and Discussion

Rate-Distortion Performance Comparison

Figures 3.5, 3.6, and 3.7 show the rate–distortion comparisons of the proposed system against VVC, HEVC, and AVC for GOP sizes of 8, 16, and 32, respectively. Each figure shows average PSNR, SSIM, and VMAF curves as functions of bitrate for each complexity category, together with the corresponding BD-rate bar charts. A key characteristic of the proposed system is that the latent vector adds a constant 2.5 kbit/s per frame to the total bitrate, regardless of QP or content complexity, because the 8×1 latent vector has a fixed size; almost all bitrate variation across QP values therefore comes from the VVC-encoded residual.

For GOP size 8 (Figure 3.5 and Table 3.2), the proposed system achieves PSNR BD-rates of -32.28% , -36.27% , and -31.62% versus AVC and -21.37% , -19.16% , and -19.69% versus HEVC for low, medium, and high complexity content, respectively. Against VVC, the BD-rates are -17.49% , -13.93% , and -12.74% , confirming that the proposed system matches or exceeds VVC coding efficiency at a GOP size of 8. The SSIM and VMAF BD-rates follow the same trend: the proposed system achieves negative BD-rates against all three baselines across all complexity categories at GOP 8, demonstrating consistently superior rate–distortion performance. For high-complexity content, VVC marginally outperforms the proposed system in PSNR and SSIM, while the proposed system retains a VMAF advantage, indicating that the semantic approach preserves perceptual temporal quality even when pixel-level reconstruction accuracy is slightly below VVC.

For GOP size 16 (Figure 3.6 and Table 3.3), the improvements over HEVC and AVC are more pronounced than at GOP 8 across all complexity categories. PSNR BD-rates versus AVC range from -46.53% (medium) to -55.71% (high), and versus HEVC from -36.75% to -50.88% . Against VVC, the proposed system achieves BD-rates of

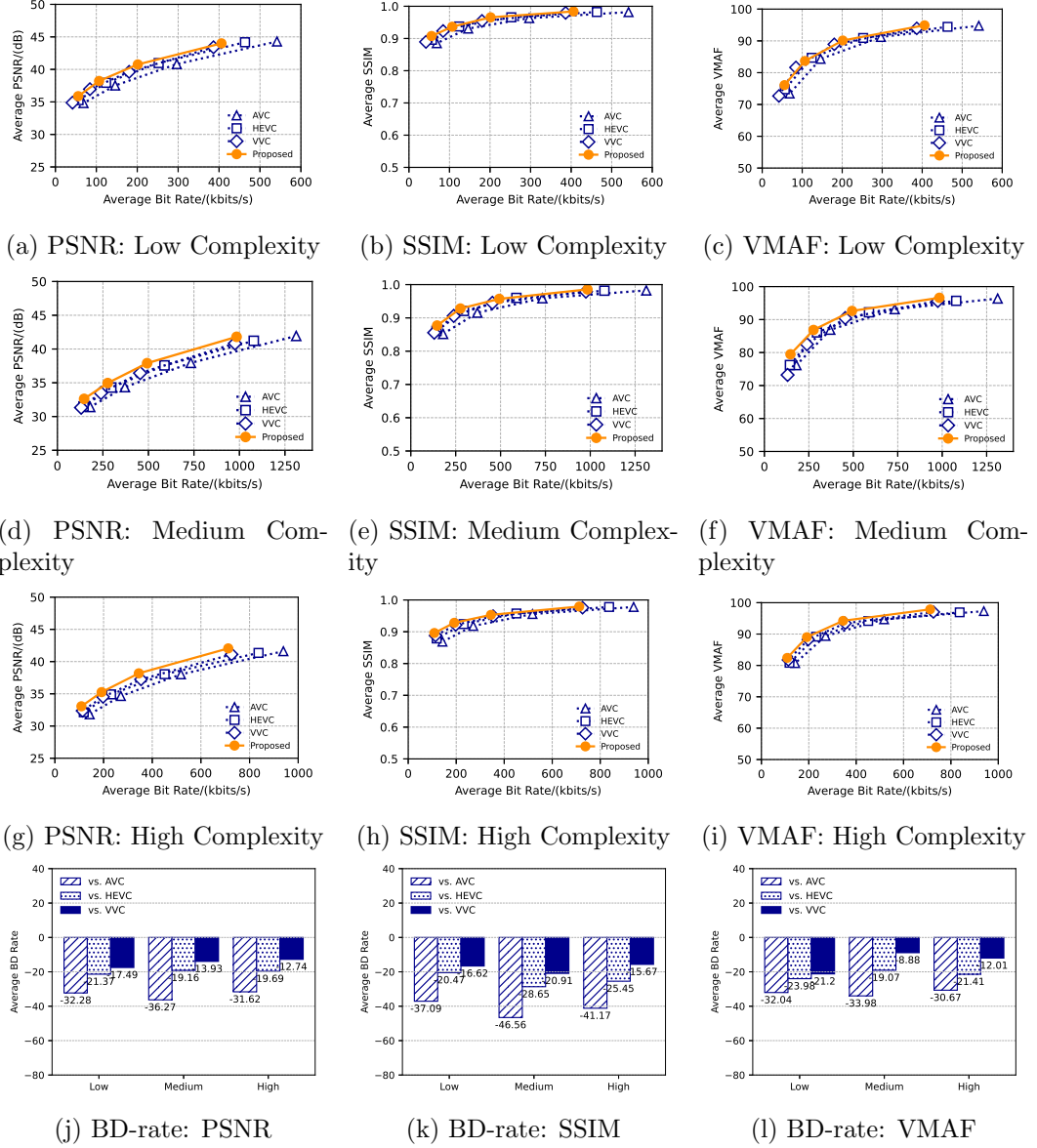


Figure 3.5: Rate-distortion performance comparison of the proposed system against VVC, HEVC, and AVC for GOP size 8. Rows 1–3 show average PSNR, SSIM, and VMAF for low, medium, and high complexity categories. Row 4 shows BD-rates relative to each reference codec [1].

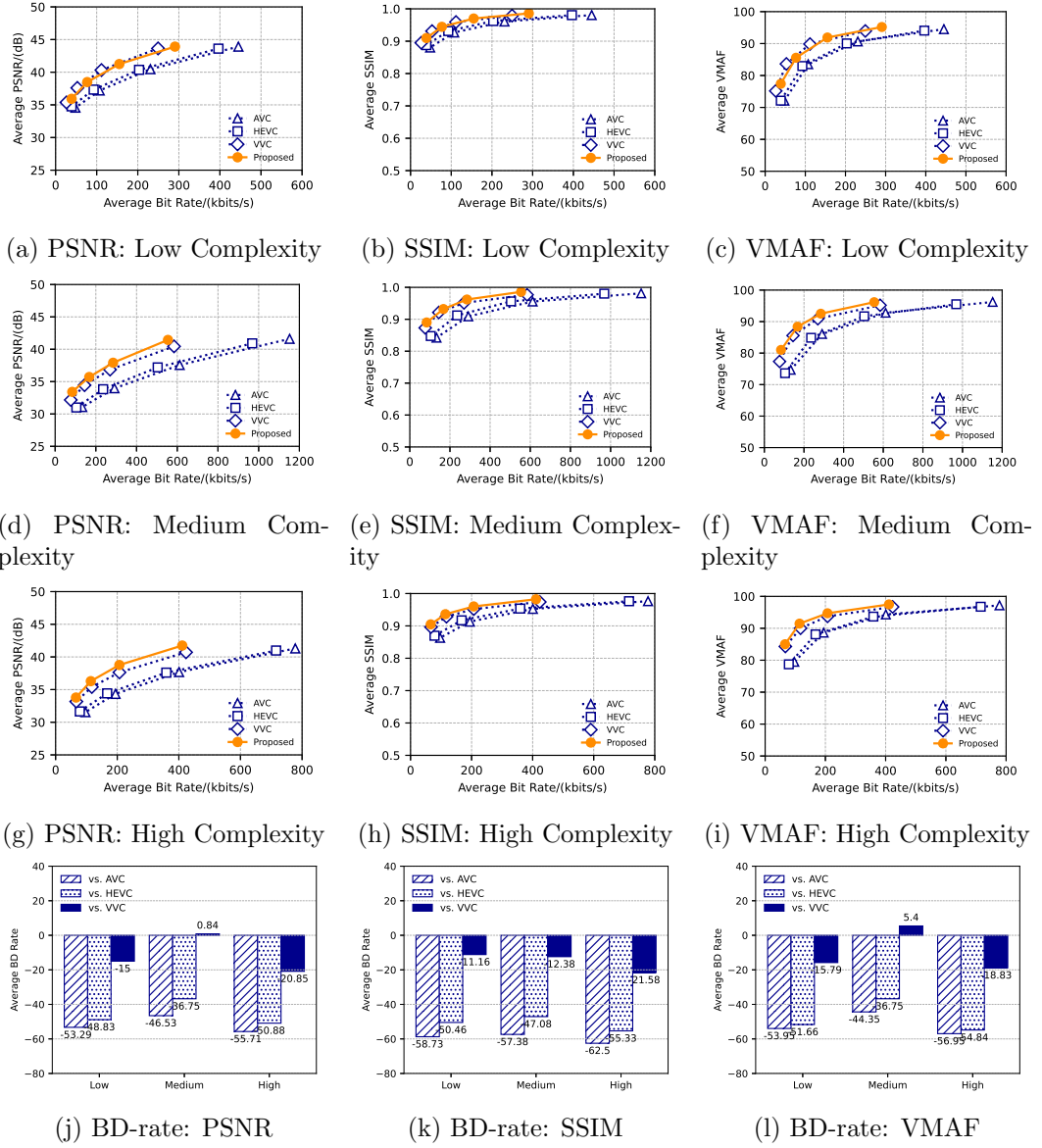


Figure 3.6: Rate-distortion performance comparison of the proposed system against VVC, HEVC, and AVC for GOP size 16 [1].

Table 3.2: Average BD-rates for the proposed system relative to AVC, HEVC, and VVC for GOP size 8 [1].

Metric	Complexity	vs. AVC	vs. HEVC	vs. VVC
PSNR	Low	−32.28%	−21.37%	−17.49%
PSNR	Medium	−36.27%	−19.16%	−13.93%
PSNR	High	−31.62%	−19.69%	−12.74%
SSIM	Low	−37.09%	−20.47%	−16.62%
SSIM	Medium	−46.56%	−28.65%	−20.91%
SSIM	High	−41.17%	−25.45%	−15.67%
VMAF	Low	−32.04%	−23.98%	−21.20%
VMAF	Medium	−33.98%	−19.07%	−8.88%
VMAF	High	−30.67%	−21.41%	−12.01%

−15.00% and −20.85% in PSNR for low and high complexity content, confirming continued competitive performance. For medium complexity content at GOP 16, the PSNR BD-rate versus VVC rises slightly to +0.84%, indicating a marginal VVC advantage; the VMAF BD-rate is also slightly positive (+5.40%) while SSIM remains negative (−12.38%). This mixed outcome is discussed further in the context of the GOP 32 results below.

Table 3.3: Average BD-rates for the proposed system relative to AVC, HEVC, and VVC for GOP size 16 [1].

Metric	Complexity	vs. AVC	vs. HEVC	vs. VVC
PSNR	Low	−53.29%	−48.83%	−15.00%
PSNR	Medium	−46.53%	−36.75%	+0.84%
PSNR	High	−55.71%	−50.88%	−20.85%
SSIM	Low	−58.73%	−50.46%	−11.16%
SSIM	Medium	−57.38%	−47.08%	−12.38%
SSIM	High	−62.50%	−55.33%	−21.58%
VMAF	Low	−53.95%	−51.66%	−15.79%
VMAF	Medium	−44.35%	−36.75%	+5.40%
VMAF	High	−56.95%	−54.84%	−18.83%

For GOP size 32 (Figure 3.7 and Table 3.4), the proposed system achieves its largest gains over HEVC and AVC: PSNR BD-rates versus AVC of −57.14%, −48.20%, and −58.57% for low, medium, and high complexity, and versus HEVC of −54.19%,

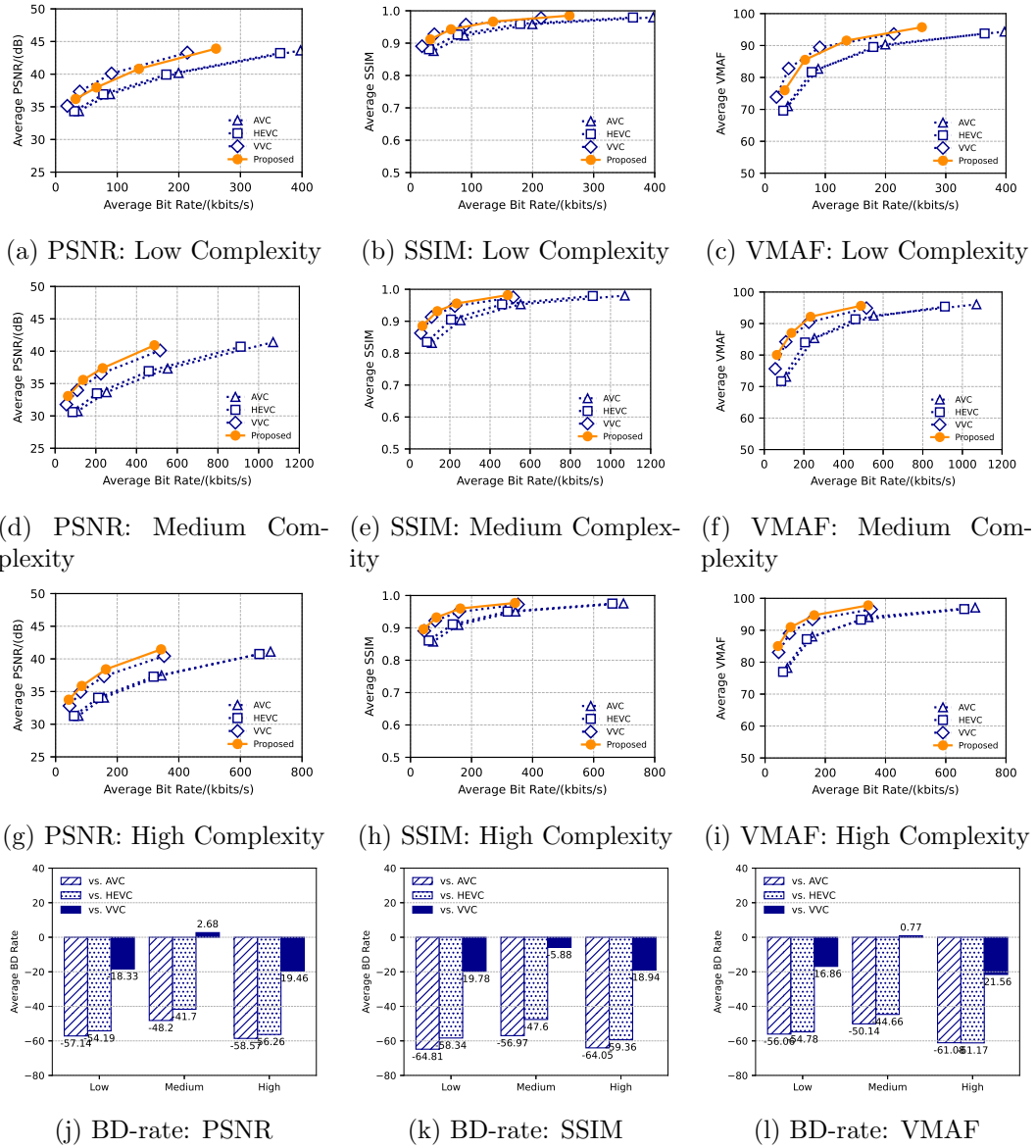


Figure 3.7: Rate-distortion performance comparison of the proposed system against VVC, HEVC, and AVC for GOP size 32 [1].

−41.70%, and −56.26%. Against VVC, the proposed system achieves PSNR BD-rates of −18.33% and −19.46% for low and high complexity, and SSIM BD-rates of −19.78% and −18.94%. For medium complexity content at GOP 32, the PSNR BD-rate versus VVC is +2.68% and the VMAF BD-rate is +0.77%, while SSIM remains negative (−5.88%), continuing the pattern observed at GOP 16.

Table 3.4: Average BD-rates for the proposed system relative to AVC, HEVC, and VVC for GOP size 32 [1].

Metric	Complexity	vs. AVC	vs. HEVC	vs. VVC
PSNR	Low	−57.14%	−54.19%	−18.33%
PSNR	Medium	−48.20%	−41.70%	+2.68%
PSNR	High	−58.57%	−56.26%	−19.46%
SSIM	Low	−64.81%	−58.34%	−19.78%
SSIM	Medium	−56.97%	−47.60%	−5.88%
SSIM	High	−64.05%	−59.36%	−18.94%
VMAF	Low	−56.06%	−54.78%	−16.86%
VMAF	Medium	−50.14%	−44.66%	+0.77%
VMAF	High	−61.08%	−61.17%	−21.56%

The marginal VVC advantage for medium complexity content at larger GOP sizes arises from VVC’s mature motion estimation and motion compensation toolchain, which has been optimised over three decades of standardisation to model the predictable inter-frame motion patterns characteristic of medium-complexity scenes [38]. VVC’s block-based motion vector pipeline is specifically tuned for this class of content, producing inter-predicted residuals with lower entropy than those generated by the AE-based prediction. In contrast, when content complexity is high, the nonlinear AE transformation provides a decisive advantage: the semantic bottleneck compresses high-motion and high-texture frames to the same constant latent bitrate regardless of scene complexity, while VVC’s motion estimation cost increases substantially with content complexity [40]. At low complexity, the proposed system benefits from both small semantic residuals and the inherently high compressibility of the VVC-encoded residual for simple scenes.

Latent Vector Dimensionality Analysis

Figure 3.8 shows the effect of latent vector dimension on the average PSNR of semantically reconstructed frames for input videos of resolution 320×180 .

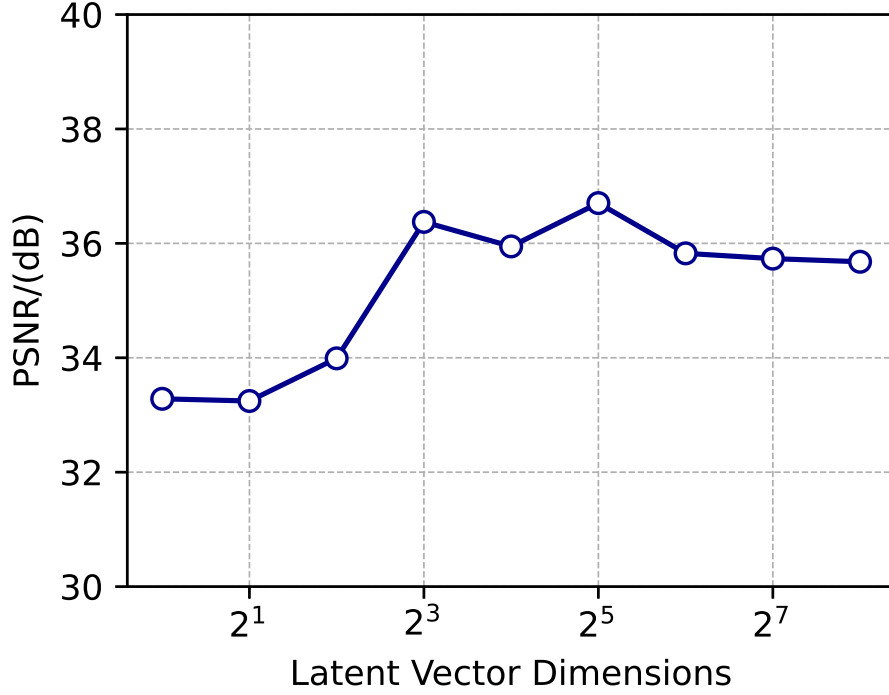


Figure 3.8: Effect of latent vector dimension on the average PSNR of the reconstructed frame at the semantic decoder output, evaluated on 320×180 input frames. A saturation point is reached at dimension 8×1 , beyond which further capacity yields no meaningful gain [1].

PSNR increases steeply as the latent dimension grows from 1 to 8, after which the improvement saturates. Increasing the dimension beyond 8×1 yields no significant additional gain, confirming that this is the minimum dimensionality at which the semantic bottleneck captures the perceptually significant content of each frame without imposing avoidable bitrate overhead. This saturation is consistent with the high temporal correlation within a single GOP: the semantic content of all frames in the GOP can be compactly described by a low-dimensional manifold that the 8×1 latent vector is sufficient to represent.

Remote Decoder Training Validation

Table 3.5 reports the average PSNR of video reconstructions produced by the transmitter-side AE decoder layers and the remotely trained receiver-side AE decoder layers for seven video sequences spanning medium and high SI/TI content.

Table 3.5: Comparison of average PSNR of reconstructions produced by the transmitter-side and receiver-side remotely trained AE decoder layers [1].

Video	Avg. PSNR (Transmitter)	Avg. PSNR (Receiver)
Video A	39.69 dB	39.69 dB
Video B	39.79 dB	39.79 dB
Video C	39.78 dB	39.78 dB
Video D	39.76 dB	39.76 dB
Video E	39.85 dB	39.85 dB
Video F	39.85 dB	39.85 dB
Video G	39.77 dB	39.77 dB

The PSNR values at the transmitter and receiver agree to sub-decimal-place accuracy for all seven sequences, confirming that the Glorot-seeded remote training procedure converges to a decoder that is functionally equivalent to the transmitter-side decoder. Figure 3.9 provides visual confirmation for a high-complexity example: frames reconstructed by the transmitter decoder layers and the remotely trained receiver decoder layers are perceptually indistinguishable.



Figure 3.9: Frames reconstructed at (a) the transmitter AE decoder layers and (b) the remotely trained receiver AE decoder layers for an example high-complexity video. The outputs are visually indistinguishable, validating the remote training mechanism [1].

Table 3.6 extends this validation across three GOP sizes, demonstrating that the

remote decoder consistently matches the transmitter decoder regardless of the temporal span of the GOP.

Table 3.6: Comparison of average rate-distortion performance of the transmitter-side (TX) and remotely trained receiver-side (RX) AE decoders across different GOP sizes [1].

GOP	PSNR (TX)	SSIM (TX)	PSNR (RX)	SSIM (RX)
8	20.27 dB	0.759	20.27 dB	0.759
16	26.64 dB	0.844	26.64 dB	0.844
32	35.88 dB	0.988	35.88 dB	0.988

TX and RX values are identical for all three GOP sizes, confirming convergence under remote training. A notable additional observation is the strong dependence of semantic reconstruction quality on GOP size: increasing the GOP from 8 to 32 raises the average PSNR from 20.27 dB to 35.88 dB and SSIM from 0.759 to 0.988. This improvement occurs because a larger GOP provides more training frames from the same scene, enabling the per-GOP AE overfitting to more accurately model the content manifold and reduce the per-frame semantic prediction error.

Key Frame QP Analysis

Tables 3.7 and 3.8 report the effect of the VVC intra coding QP used for key frames on the total system bitrate and quality metrics, for high SI/TI and low SI/TI video content respectively, with the residual coding QP held fixed at 22.

Table 3.7: Effect of key frame QP on bitrate and quality for high SI/TI videos, with residual QP fixed at 22 [1].

Key Frame QP	Key Frame Rate	PSNR	SSIM	VMAF
22	1,748.9 kbit/s	32.72 dB	0.7866	80.94
27	1,046.5 kbit/s	32.72 dB	0.7853	80.47
32	520.1 kbit/s	32.94 dB	0.7872	79.82
37	210.1 kbit/s	33.04 dB	0.7824	78.08

In both complexity regimes, PSNR and SSIM remain broadly stable as the key frame QP increases from 22 to 37, while the key frame bitrate decreases by a factor of approximately $8\times$ for high-complexity content (from 1,748.9 to 210.1 kbit/s) and

Table 3.8: Effect of key frame QP on bitrate and quality for low SI/TI videos, with residual QP fixed at 22 [1].

Key Frame QP	Key Frame Rate	PSNR	SSIM	VMAF
22	947.9 kbit/s	42.71 dB	0.9862	89.80
27	607.2 kbit/s	41.84 dB	0.9819	89.01
32	345.9 kbit/s	40.19 dB	0.9732	86.39
37	187.5 kbit/s	38.17 dB	0.9591	81.02

approximately $5\times$ for low-complexity content (from 947.9 to 187.5 kbit/s). The VMAF reduction at QP 37 is modest: from 80.94 to 78.08 for high complexity and from 89.80 to 81.02 for low complexity. These results demonstrate that the semantic decoder’s primary reliance on the latent vector for reconstruction renders the system relatively insensitive to key frame quality, justifying the selection of QP 37 for all key frame encoding to minimise bandwidth consumption.

Residual Encoding Optimisation and Visual Residual Analysis

Figure 3.10 shows example frames from three test videos of differing complexity alongside their semantic reconstruction residuals before residual correction is applied.

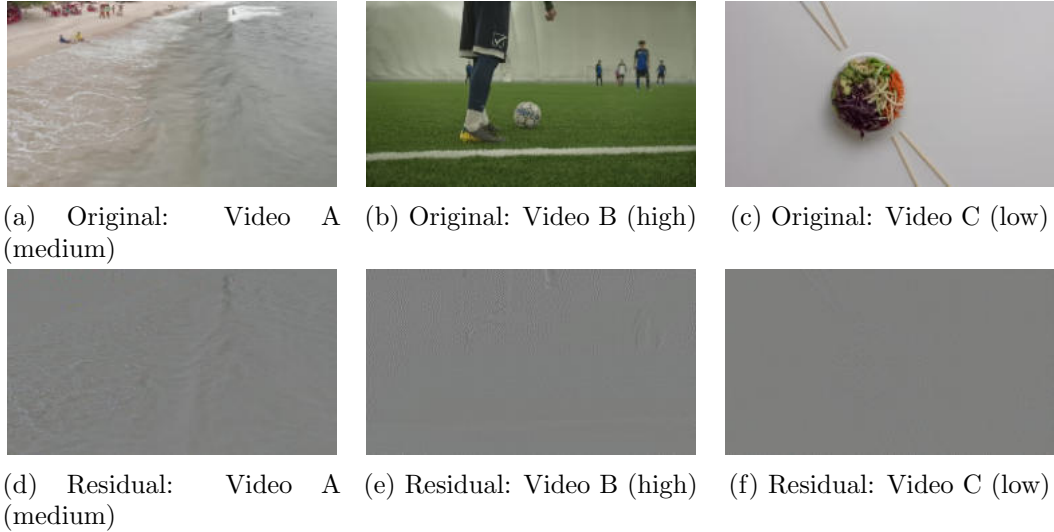


Figure 3.10: Example original frames and their semantic reconstruction residuals. Top row: original frames from (a) Video A (medium complexity), (b) Video B (high complexity), (c) Video C (low complexity). Bottom row: corresponding residuals prior to residual correction [1].

The residuals are largest for high-complexity content (Figure 3.10e), where the 8×1 latent vector is unable to capture all spatial detail present in a high-SI/TI frame. For low-complexity content (Figure 3.10f) the residual is much smaller, reflecting the lower information density that the AE can represent within its latent capacity.

Table 3.9 reports the average residual bitrate, PSNR, and SSIM for three GOP sizes at the VVC residual coding stage.

Table 3.9: Average residual bitrate, PSNR, and SSIM for different GOP sizes used in the VVC residual encoding stage [1].

GOP	Avg. Bitrate	Avg. PSNR	Avg. SSIM
8	43.7 kbit/s	30.22 dB	0.916
16	39.9 kbit/s	30.37 dB	0.916
32	32.6 kbit/s	30.40 dB	0.911

Increasing the GOP from 8 to 32 reduces the average residual bitrate by approximately 25% (from 43.7 to 32.6 kbit/s) with negligible change in PSNR (30.22 to 30.40 dB) and only a marginal SSIM reduction (0.916 to 0.911). The bitrate reduction at larger GOP sizes occurs because more training frames per GOP enable a more accurate per-GOP AE overfit, reducing the semantic prediction error and therefore the entropy of the residual. GOP size 32 is therefore selected for all residual encoding experiments.

Resolution Scalability Verification

Table 3.10 reports the average PSNR of the proposed system when an example video is encoded at five spatial resolutions spanning nearly $20\times$ in pixel count.

Table 3.10: Effect of spatial resolution on the average PSNR of the proposed system [1].

Resolution	Total Pixels	Average PSNR
256×144	110,592	34.32 dB
320×180	172,800	34.33 dB
420×240	306,720	34.33 dB
1280×720	921,600	34.10 dB
1920×1080	2,073,600	34.01 dB

PSNR varies by less than 0.35 dB across the full resolution range. This near-invariance occurs because PSNR and the semantic reconstruction quality are both scale-invariant with respect to the MSE-normalised training objective: the AE captures relative structural content rather than absolute pixel values, and the normalisation by $N = WHC$ in (3.1) ensures that the loss does not depend on resolution. The result confirms that the use of 320×180 for training and evaluation does not limit the validity of the findings, and the proposed system scales to higher resolutions without architectural modification.

Computational Complexity Analysis

For video at spatial resolution 320×180 with YUV420 format (1.5 colour channels), the computational cost of the semantic encoder and decoder is estimated in terms of FLOPs and multiply-accumulate operations (MACs). The input flatten layer contributes zero FLOPs. The first dense layer connecting the 86,400-dimensional input to the 1024-dimensional hidden layer requires approximately 176,947,200 FLOPs, and the hidden layer to bottleneck connection ($1024 \rightarrow 8$) requires 16,384 FLOPs. The decoder layers follow the mirror arrangement, contributing a similar count. The total FLOPs for the complete semantic encoder and decoder are approximately 353,927,168, corresponding to approximately 176,963,584 multiply-accumulate (MAC) operations.

A practical complexity comparison is provided by comparing the per-frame encoding time of VVenC with and without the semantic pre-processing stage. On identical hardware, the average time to encode the VVC-coded residual of the semantic output is 0.765 s/frame, compared with 2.026 s/frame to encode the same video directly with VVenC at matched quality. This represents a $2.65\times$ reduction in per-frame VVenC encoding time, mainly due to the reduced motion estimation workload: because the semantic encoder captures inter-frame temporal correlation during per-GOP training, the residuals contain very little inter-frame redundancy, allowing VVenC to operate mainly in intra-prediction mode with much less motion estimation computation [40]. The additional cost of per-GOP AE training must be balanced against this saving; the net encoding time therefore depends on the content, and future work on improved

training strategies and dedicated neural inference hardware is expected to reduce the AE training overhead.

3.2.4 Conclusion

This section has presented the first end-to-end hybrid video codec that combines autoencoder-based semantic communication for inter-frame prediction with VVC intra coding and VVC residual coding for key frame context and quality correction [1]. The key findings are summarised below.

The remote decoder training mechanism based on Glorot uniform initialisation with a shared fixed seed produces the same results as local training: the difference in average PSNR between transmitter-side and receiver-side semantic decoder outputs across seven test sequences is negligible, confirming that the shared seed synchronisation avoids significant semantic noise. This finding removes the need to transmit network parameters, which has been a major obstacle to deploying DNN-based video codecs.

The latent vector dimension of 8×1 is confirmed as the point at which increasing the bottleneck size gives no meaningful improvement in average PSNR, establishing the minimum representation size for the YUV420 320×180 frame format. A key frame quantisation parameter of 37 is identified as the best operating point that keeps the context bitrate low without reducing semantic prediction quality, because the residual coding stage corrects the quality loss caused by the coarse key frame quantisation.

Rate-distortion evaluation across three GOP sizes (8, 16, and 32) on twenty test sequences confirms consistent and large BD-rate improvements over AVC and HEVC across all video complexity classes. Relative to AVC, the proposed system achieves average PSNR BD-rates ranging from -31.6% to -58.6% depending on complexity class and GOP size, and relative to HEVC from -19.2% to -56.3% . Relative to VVC, the system achieves negative average PSNR BD-rates (i.e., bitrate savings compared to VVC) for low and high complexity content across all three GOP sizes, except for medium complexity content where VVC outperforms the proposed system by up to $+2.68\%$ PSNR BD-rate. This limitation for medium complexity video is because VVC's linear motion compensation is well suited to content with structured, moderate motion,

where the nonlinear semantic representation does not provide enough gain to offset the residual bitrate cost. For high complexity content, the semantic encoder keeps the latent bitrate constant at 2.5 kbit/s regardless of scene complexity, while conventional codecs need higher bitrates as motion increases because their linear transforms cannot efficiently represent fast-moving content.

Resolution invariance is confirmed across spatial resolutions from 256×144 to 1920×1080 , with average PSNR varying by less than 0.32 dB across the range, confirming that the rate–distortion results from 320×180 test content carry over to full-resolution use. The $2.65\times$ reduction in per-frame VVenC encoding time when residual coding is used instead of direct VVC encoding shows that semantic pre-processing significantly reduces the VVC motion estimation workload.

This section directly addresses Gap 1 identified in Chapter 2 by presenting a validated hybrid semantic-conventional video coding architecture that is the first of its kind reported in the literature, as confirmed by the systematic search described in Section 2.12, showing that compact semantic inter-frame prediction and standard intra and residual coding can be combined without sending decoder parameters. The limitation for medium complexity content and the resolution restriction to 320×180 pixels are the open problems addressed in the enhanced system presented in Section 3.3.

3.3 Enhanced Hybrid Codec with Bidirectional Hierarchical Frame Prediction

Publication Notice

The research presented in this section has been published in the following peer-reviewed publication:

P. Samarathunga, Y. Ganearachchi, T. Fernando, A. Jayasingam, T. Sivalingam, N. Rajatheva, and A. Fernando, “Enhanced Semantic Communications and VVC-Based Hybrid Video Coding System,” *IEEE Access*, vol. 14, pp. 50755–50778, 2026, doi: 10.1109/ACCESS.2026.3676702.

This material is reproduced and adapted under the terms of the IEEE Copyright

policy for use in a doctoral thesis. The content has been revised and extended for inclusion and integration in this doctoral thesis, including additional explanations, expanded discussion, and alignment with the overarching research narrative.

3.3.1 Introduction

The basic hybrid codec presented in Section 3.2 showed that combining autoencoder-based semantic prediction with VVC context and residual coding is feasible, and confirmed the remote decoder training approach [1]. However, it had two main limitations that reduced its compression efficiency. First, the system was tested only at 320×180 spatial resolution, limiting its use for standard and high-definition content that is common in real video applications. Second, the temporal prediction used simple interpolation, where each frame was reconstructed from the GOP key frame using the semantic autoencoder trained on the full GOP, without using the bidirectional temporal context available from frames on both sides of the target frame within the GOP. This limitation was identified in the conclusions of that work as a key open problem, with hierarchical bidirectional prediction suggested as a natural next step.

Further analysis shows that the semantic autoencoder alone achieves reconstruction quality in the range of 24–34 dB PSNR, which is sufficient for machine vision tasks such as scene classification and object detection but leaves a large residual for the VVC residual coder. Without improved temporal prediction, the residual coding dominates the total bitrate, which limits the compression benefit of the semantic approach. Ablation experiments confirm that a system with unidirectional prediction gives a PSNR BD-rate of +4.68% relative to VTM 17.0, showing that unidirectional prediction does not achieve a rate–distortion advantage over the conventional codec.

This section addresses these limitations through three main improvements [22]. First, a bidirectional hierarchical frame prediction structure is introduced, where RIFE [72] is used to generate intermediate frame interpolations guided by optical flow, which serve as better semantic predictions for non-keyframe positions within a GOP. The hierarchical structure uses both past and future reference frames at multiple time scales,

similar to B-frame prediction in conventional codecs, and limits the prediction window to prevent error accumulation beyond a GOP boundary. Second, a block-based processing strategy is introduced that splits input frames of any spatial resolution into non-overlapping 320×180 blocks, allowing the fully connected autoencoder to be used at any resolution without changing its architecture. Third, residual latent entropy coding using DeepCABAC [148] is introduced to further compress the latent vector bitstream, adding to the RIFE-based reduction in residual energy. The enhanced system is compared against VTM 17.0 and a wide range of conventional and learned codec references across the HEVC class B through E test sequences [172], the UVG dataset [173], and the MCL-JCV dataset [174], and against VTM 23.0 for high-resolution test sequences including 4K content.

3.3.2 Methodology

System Overview

The three changes motivated in the preceding subsection are combined in the architecture of Fig. 3.11: RIFE-assisted hierarchical prediction, block-based resolution scaling, and DeepCABAC latent coding. All components inherited from the basic codec retain the definitions in Section 3.2; the following subsections specify only the enhanced processing.

Like its predecessor, the enhanced framework transmits three bitstream components: intra-coded key frames providing shared context, latent vectors forming the semantic channel payload, and VVC-coded residuals forming the correction layer. This three-part structure follows Shannon’s three-level communication model [57] within a single codec: the VVC-based key frame and residual encoders address the technical problem (Level A), the autoencoder-based latent representation addresses the semantic problem (Level B), and the overall system addresses the effectiveness problem (Level C) by enabling reconstruction quality that is sufficient for both machine vision tasks and human viewing.

An important feature of the framework, preserved from the original design, is that it does not require extensive offline pre-training on large video datasets, nor does it

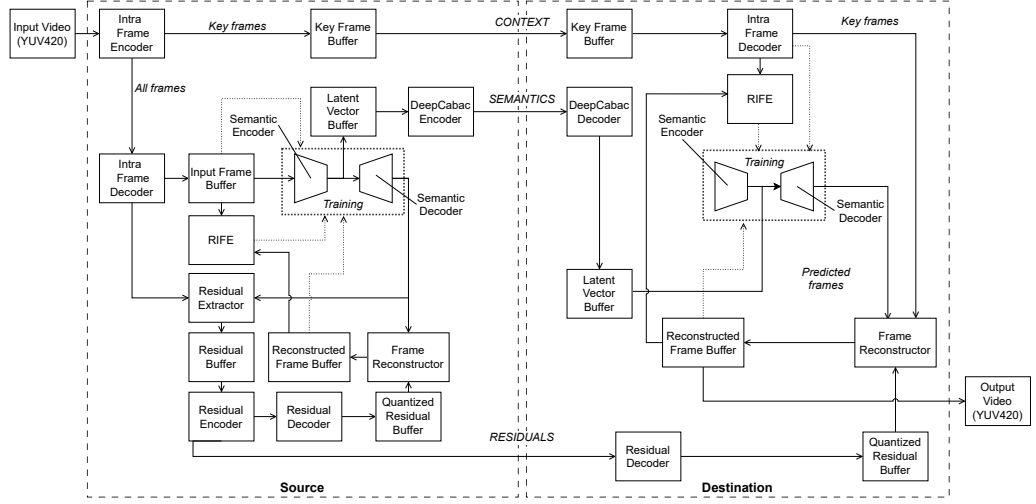


Figure 3.11: End-to-end view of the enhanced semantic communications and VVC-based hybrid video codec. Dashed lines denote training process flows and solid lines denote operational data flows [22].

require a pre-shared decoder model at the receiver. Instead, semantic encoder and decoder parameters are trained independently at both endpoints per GOP, with synchronisation achieved through deterministic weight initialisation and an asynchronous remote training procedure. This design contrasts with learned video codecs such as the DCVC family [51–53, 98, 99], which require weeks of training on the Vimeo-90k dataset [175] and the deployment of the full trained model at the decoder.

Semantic Encoder Architecture

Let a video sequence be denoted $\mathcal{F} = \{F_t\}_{t=0}^T$, where $T+1$ is the total number of frames in the sequence. Each frame $F_t \in \mathbb{R}^D$ denotes the flattened YUV420 representation with frame height H , frame width W , and total sample count $D = 3HW/2$. The semantic encoder, parameterised by θ_e , is defined as the nonlinear mapping

$$\phi_{\theta_e} : \mathbb{R}^D \rightarrow \mathbb{R}^L, \quad (3.9)$$

which transforms each quantised input frame to its latent representation

$$h_t = \phi_{\theta_e}(Q_i(F_t)), \quad (3.10)$$

where $L \ll D$ is the latent dimensionality and $Q_i(\cdot)$ denotes input quantisation with quantisation parameter i . The semantic decoder, parameterised by θ_d , performs the inverse mapping

$$\hat{F}_t = \psi_{\theta_d}(h_t). \quad (3.11)$$

The semantic encoder and decoder are implemented as the encoding and decoding layers of a shallow fully connected autoencoder with a single hidden layer of width $d_h = 1,024$ and a bottleneck of dimension $L = 8$. The architecture is illustrated in Figure 3.12. Input frames in YUV420 format are first flattened to a vector of dimension $HW \times 1.5$; for a block of size 320×180 , this yields an input dimensionality of 86,400. LeakyReLU activations with $\alpha = 0.1$ are applied in the hidden and bottleneck layers, while the decoder output layer employs a sigmoid activation to constrain reconstructed pixel values to the unit interval.

The choice of a fully connected architecture is deliberate. By flattening all spatial dimensions before encoding, the autoencoder learns global correlations across the entire frame, capturing overall semantic content rather than only local spatial patterns. This also permits deterministic synchronisation between the source and destination decoders: unlike convolutional layers, which can introduce non-determinism through implementation-specific algorithm selection (such as Winograd versus FFT-based convolution) and vendor-specific kernel optimisations, fully connected layers involve only matrix multiplications and element-wise additions that are well-defined under IEEE 754 floating-point arithmetic [176] across hardware platforms. This determinism is required for the asynchronous remote training procedure described in Section 3.3.2.

The encoder and decoder are jointly trained by minimising the mean-squared reconstruction loss over a GOP of N frames:

$$L(\hat{F}, Q_i(F)) = \frac{1}{N} \sum_{k=m}^{m+N-1} \|Q_i(F_k) - \hat{F}_k\|_2^2, \quad (3.12)$$

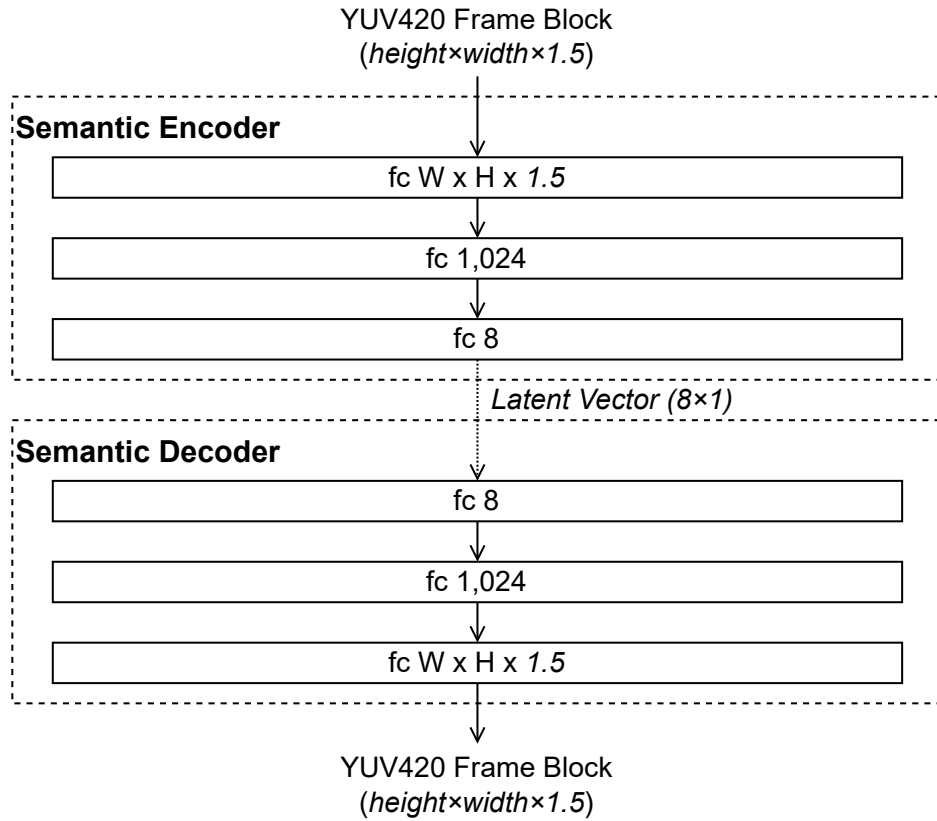


Figure 3.12: Architecture of the semantic encoder and decoder, implemented as a fully connected autoencoder for block-based coding in the YUV420 colour space [22].

where m denotes the GOP start index. This objective can be expressed through a semantic rate–distortion function

$$R_s(D_s) = \min_{\substack{p(\hat{F}|F): \\ \mathbb{E}[d_s(F, \hat{F})] \leq D_s}} I(F; \hat{F}), \quad (3.13)$$

where $d_s(F, \hat{F}) = \mathbb{E}[\|\phi_{\theta_e}(F) - \phi_{\theta_e}(\hat{F})\|_2^2]$ quantifies distortion in the semantic space, with the expectation taken over the empirical frame distribution of the training or evaluation set. This formulation extends Shannon’s classical rate–distortion theory [56, 89] to semantic compression, where the latent bottleneck of dimension $L = 8$ acts as a fixed-rate constraint and minimising training loss under this bottleneck minimises semantic distortion at the rate determined by the bottleneck dimensionality.

The ablation study in Section 3.3.3 confirms that $d_h = 1,024$ and $L = 8$ represent the optimal trade-off between reconstruction quality, encoding time, and memory efficiency. Increasing d_h beyond 1,024 yields only marginal PSNR improvements at substantially increased computational cost and memory requirements, and increasing L beyond 8 raises the semantic bitrate with negligible reconstruction benefit. The use of implicit channel weighting in the YUV420 flattening procedure, wherein Y-channel pixels outnumber U/V pixels four to one, causes the MSE loss to be dominated by luminance reconstruction error, effectively prioritising the perceptually dominant channel without requiring explicit hyperparameter tuning.

Bidirectional Hierarchical Frame Prediction and RIFE Integration

The central innovation of the enhanced framework is the replacement of simple interpolation semantic prediction with a bidirectional hierarchical frame prediction structure integrated with the RIFE intermediate flow estimation network [72]. Bidirectional prediction has long been a cornerstone of conventional video coding [177], where B-frames reference both past and future frames to exploit temporal redundancies more effectively than P-frames alone. The proposed framework applies this principle within a semantic coding context, where the semantic autoencoder trained per GOP serves as the bidirectional predictor.

Let m be the starting index of a GOP interval of length $N = 2^n$, $n \in \mathbb{Z}_0^+$, bounded by key frames F_m and F_{m+N} . The intermediate frames within this interval are denoted F_{m+k} for $k \in \{1, 2, \dots, N-1\}$. The semantic prediction of an intermediate frame is modelled as

$$\hat{F}_{m+k} = f_{\text{sem}}(F_m, F_{m+N}; \tau_k, \theta_f) + \epsilon_{m+k}, \quad \tau_k = \frac{k}{N}, \quad (3.14)$$

where f_{sem} is the bidirectional semantic predictor realised using the RIFE network and the semantic autoencoder, τ_k is the temporal interpolation position of frame F_{m+k} within the GOP interval, and ϵ_{m+k} is the residual error. The residual for an intermediate frame at GOP midpoint is computed as

$$R_{m+N/2} = Q_i(F_{m+N/2}) - \hat{F}_{m+N/2}, \quad (3.15)$$

and the reconstructed frame after residual decoding is

$$F'_{m+N/2} = \hat{F}_{m+N/2} + Q_r(R_{m+N/2}), \quad (3.16)$$

where $Q_r(\cdot)$ denotes the residual quantiser. Figure 3.13 illustrates the bidirectional dependencies for a GOP of size $N = 8$.

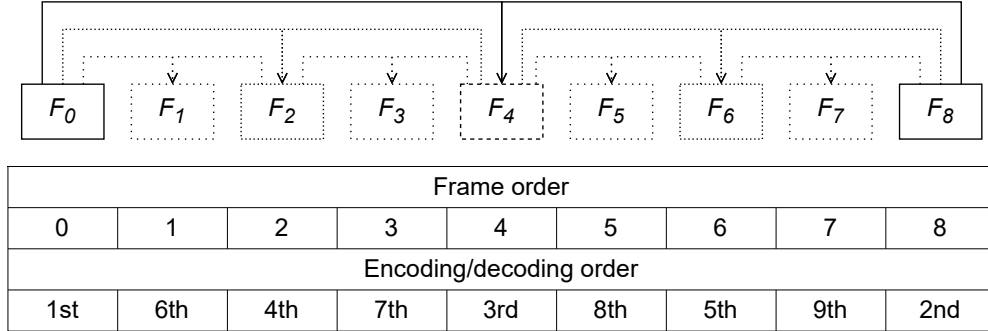


Figure 3.13: Bidirectional hierarchical prediction structure for a GOP of size 8. First-order frames (GOP midpoint) are predicted from the two key frames; subsequent orders are predicted from their adjacent references [22].

The hierarchical structure proceeds as follows. The first-order intermediate frame at position $m + N/2$ is predicted from the two key frames F_m and F_{m+N} . The second-

order frames at positions $m + N/4$ and $m + 3N/4$ are then predicted from their adjacent references, and this process recurses until all frames in the GOP have been predicted. At each level of the hierarchy, the semantic autoencoder trained on the entire GOP provides global scene understanding, while the RIFE network provides pixel-level motion alignment between reference frames. The combination of these two complementary components achieves 2 to 4 dB higher prediction PSNR than either the semantic autoencoder alone or RIFE alone, as demonstrated in the ablation study.

The RIFE module is a lightweight RIFE network that employs a teacher-student distillation strategy [72]. During training, a teacher network with privileged access to the middle frame supervises a lightweight student network to predict the middle frame from two reference frames. This approach produces accurate bidirectional frame interpolations that reduce the latent space mismatch between the semantic encoder code of the true frame and the code of the interpolated proxy, lowering the residual entropy and improving rate-distortion performance, particularly for videos with low to medium temporal information content where motion is structured.

The total expected distortion for a GOP under bidirectional prediction is

$$D_{\text{total}} = \mathbb{E} \left[\sum_{k=0}^{N-1} \|F_{m+k} - F'_{m+k}\|_2^2 \right], \quad (3.17)$$

and the rate-distortion objective minimised by the encoder is

$$J = D_{\text{total}} + \lambda (R_{\text{key}} + R_{\text{latent}} + R_{\text{residual}}), \quad (3.18)$$

where λ controls the trade-off between bitrate and distortion, and R_{key} , R_{latent} , and R_{residual} denote the bitrates of the key frame, latent vector, and residual components, respectively. This Lagrangian formulation is the standard operational approach connecting theoretical rate-distortion analysis to practical encoding decisions, and it underpins rate-distortion optimisation in conventional codecs, including VVC.

The ablation study of Section 3.3.3 shows that bidirectional hierarchical prediction is not only helpful but necessary: unidirectional prediction using the semantic autoencoder alone produces a BD-rate of +4.68% relative to VTM 17.0, indicating worse

performance than the reference codec, because semantic prediction errors accumulate across sequential frames without bidirectional context to constrain the prediction. The bidirectional approach achieves -21.1% BD-rate, confirming that the bounded temporal window provided by past and future anchor frames is necessary for stable semantic representations.

Training Procedure

The training procedure is designed to ensure synchronisation between the source and destination semantic decoders while adapting to the content of each individual GOP. At the start of each GOP, both the source and destination autoencoders are initialised with the Glorot uniform initialiser [166] using an identical fixed random seed, guaranteeing

$$\boldsymbol{\theta}_{e,\text{src}}^{(0)} = \boldsymbol{\theta}_{e,\text{dst}}^{(0)}. \quad (3.19)$$

Parameter updates for optimisation step p follow

$$\boldsymbol{\theta}^{(p+1)} = \boldsymbol{\theta}^{(p)} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{MSE}}^{(p)}, \quad (3.20)$$

such that $\boldsymbol{\theta}_{e,\text{src}}^{(p)} = \boldsymbol{\theta}_{e,\text{dst}}^{(p)}$ and $\boldsymbol{\theta}_{d,\text{src}}^{(p)} = \boldsymbol{\theta}_{d,\text{dst}}^{(p)}$ hold for all epochs under deterministic optimisation. Early stopping [178] is employed to prevent the autoencoder from fitting to pixel-level noise or encoding artefacts, halting training after 10 consecutive stagnant validation epochs and restoring the best weights.

At the source, the semantic encoder is trained using all frames in the current GOP, intentionally overfitting to the local spatio-temporal statistics of the scene. The source semantic decoder is trained using only the intra-decoded key frames, which serve as both the training signal and the shared context received at the destination. At the destination, an identical decoder is trained using the received latent vectors and the VVC-decoded key frame, without requiring a feedback channel or pre-transmission of the trained model. To reproduce the same stopping point under early stopping, the number of epochs to convergence, E^* , is transmitted as side information. This asynchronous remote training strategy is mathematically justified by the fact that, when

both decoders minimise the same MSE loss with identical initialisations, deterministic update rules, and the same stopping epoch E^* , their weight trajectories converge to the same solution. Table 3.11 confirms this synchronisation empirically, showing that the PSNR and SSIM of the source decoder and the remotely trained destination decoder are identical to sub-decimal-place precision across GOP sizes of 8, 16, and 32.

Table 3.11: Average rate–distortion performance comparison between the source and destination semantic decoders, confirming synchronisation under remote training [22].

GOP Size	Source		Destination	
	PSNR (dB)	SSIM	PSNR (dB)	SSIM
8	31.720	0.9152	31.720	0.9152
16	30.962	0.9085	30.962	0.9085
32	29.322	0.8634	29.322	0.8634

Another property of the adaptive per-GOP training procedure is that the number of epochs required to reach convergence differs between the first frame of a new scene and the later frames in the same scene. For a GOP of size 8, the autoencoder is initialised using the first frame, the final frame, and the RIFE-predicted intermediate frame; the weights learned for this initial set are retained as the starting point for predicting later frames in the same scene, resulting in much faster convergence for those frames. Figure 3.14 illustrates training loss curves for the first frame and three representative later frames of a scene, confirming this progressive reduction in the number of epochs required to satisfy the early stopping criterion.

A critical distinction between the residual-assisted and semantic-only operating modes is the importance of synchronisation. When residuals are used, a deviation between source and destination decoder predictions causes the residual computed at the source to be applied to a different base prediction at the destination, introducing reconstruction errors that compound across the GOP. When operating in semantic-only mode (without residuals), small deviations produce only slightly different semantic reconstructions without catastrophic failure, since the decoded key frames provide sufficient context for meaningful reconstruction. Each GOP boundary resets any accumulated synchronisation drift by re-initialising the decoders from the shared Glorot

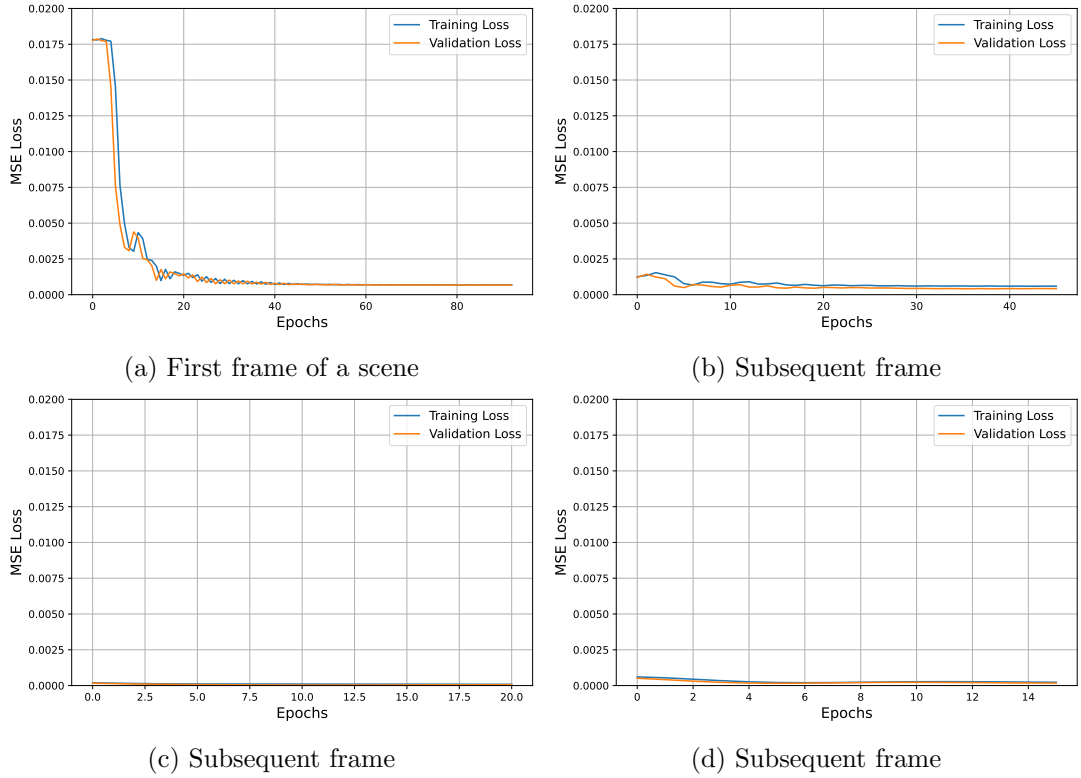


Figure 3.14: Training loss convergence curves for the first frame of a scene and three representative subsequent frames, averaged over a set of video scenes. Weight transfer from the first-frame training to subsequent frames reduces the number of epochs required to reach the early stopping criterion [22].

seed, limiting the temporal scope of any semantic noise.

Latent Entropy Coding

To further reduce the semantic channel bitrate, a residual latent coding strategy combined with DeepCABAC entropy coding is introduced. Rather than entropy-coding each latent vector independently, the framework computes the residual latent

$$\Delta l = l_{\text{sem}} - l_{\text{RIFE}}, \quad (3.21)$$

where l_{sem} is the latent vector produced by the semantic autoencoder for the frame to be predicted, and l_{RIFE} is the latent representation of the RIFE-interpolated proxy frame, predicted from the two reference key frames at the source without requiring additional transmission. Because RIFE produces a temporally coherent proxy that is semantically similar to the target frame, the mutual information $I(l_{\text{sem}}; l_{\text{RIFE}})$ is high, and the entropy $H(\Delta l) \approx H(l_{\text{sem}}) - I(l_{\text{sem}}; l_{\text{RIFE}})$ is substantially lower than that of the original latent vector.

The residual latents Δl are entropy-coded using DeepCABAC with a quantisation interval of 0.1, a step size of 0.1, and a rate parameter $\lambda = 0$. DeepCABAC employs a learned context model that adaptively estimates the probability distribution of quantised latent symbols from previously encoded symbols, enabling efficient arithmetic coding. This entropy coding stage reduces the latent bitrate from 4.096 kbit/s (without entropy coding) to 0.19 kbit/s for 720p video at GOP size 32, contributing an additional 2.06 percentage points of BD-rate improvement relative to VTM 17.0. Table 3.12 quantifies this contribution, confirming that the residual latent coding combined with DeepCABAC entropy coding is an effective and low-overhead mechanism for minimising semantic channel bitrate.

Table 3.12: BD-rate gains attributable to DeepCABAC-based latent entropy coding relative to VTM 17.0 [22].

Configuration	Latent Bitrate (kbit/s)	BD-rate vs VTM 17.0
No entropy coding	4.096	−19.04%
DeepCABAC (proposed)	0.19	−21.1%

Figure 3.15 shows the cross-entropy between the semantic latent l_{sem} and the RIFE-predicted latent l_{RIFE} , evaluated as the mean over selected elements. The high mutual information between the two latent vectors confirms that the RIFE-predicted proxy provides an effective approximation of the semantic latent, justifying the residual latent coding strategy and validating the entropy reduction achieved by DeepCABAC.

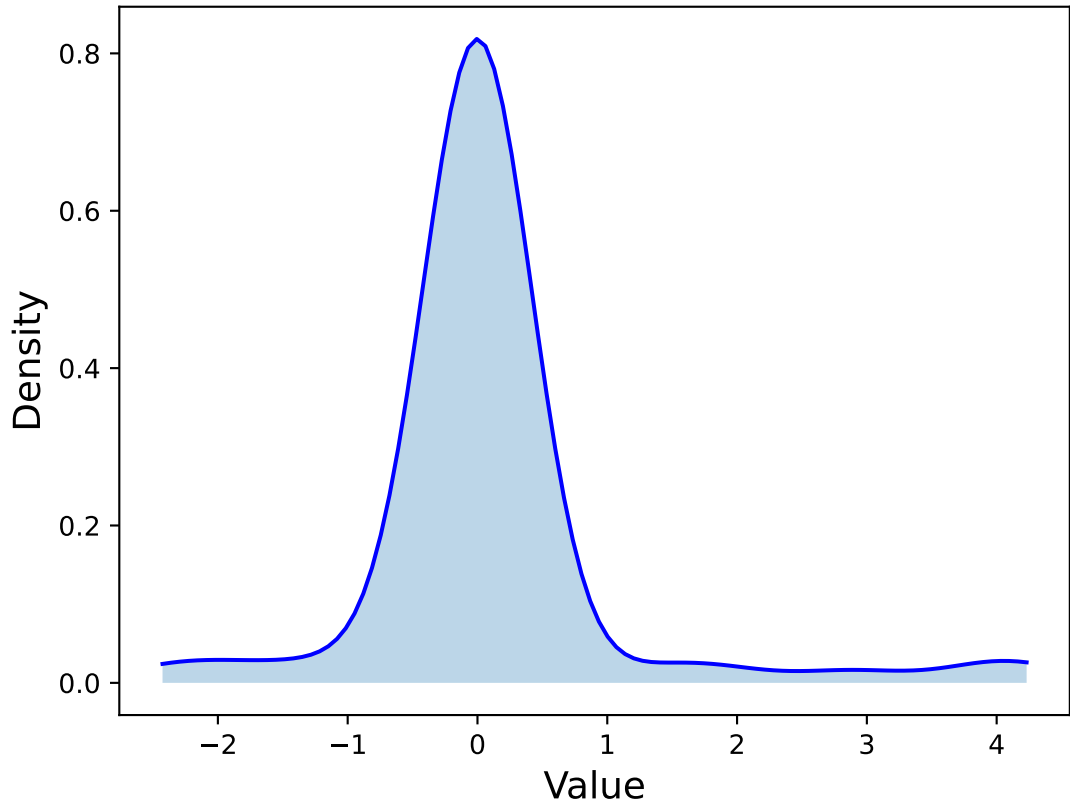


Figure 3.15: Cross-entropy between the semantic latent l_{sem} and the RIFE-predicted latent l_{RIFE} , evaluated as the mean over selected elements, confirming high mutual information between the two representations [22].

Block-Based High-Resolution Extension

The original framework was restricted to a spatial resolution of 320×180 pixels due to the memory requirements of the fully connected autoencoder. For an input of dimension 86,400 (corresponding to a 320×180 YUV420 block), the first dense layer alone requires $86,400 \times 1,024 = 88.5$ million parameters, corresponding to approximately 354 MB at

single-precision (FP32). Including optimiser states, peak memory per block reaches approximately 3.85 GB, which is within the capacity of a single NVIDIA A100 GPU (40 GB) and allows approximately ten blocks to be processed in parallel. Scaling to 360p (640×360) blocks would require approximately 15.4 GB per block, exceeding practical capacity for batch training.

The enhanced framework supports arbitrarily high spatial resolutions by partitioning each input frame into non-overlapping 320×180 blocks, processing each block independently through the semantic encoder and decoder, and reassembling the predictions into a full frame prior to residual computation. The residual R_{m+k} is then computed for the full reassembled frame as

$$R_{m+k} = Q_i(F_{m+k}) - \hat{F}_{m+k}^{\text{full}}, \quad (3.22)$$

where $\hat{F}_{m+k}^{\text{full}}$ denotes the semantic prediction assembled from the individual block predictions. The VVC residual encoder processes the full-resolution residual without alignment to the 320×180 semantic block boundaries, so any block-boundary discontinuities introduced by the independent processing of adjacent semantic blocks are captured in the residual and corrected upon decoding:

$$F'_{m+k} = \hat{F}_{m+k}^{\text{full}} + Q_r(R_{m+k}). \quad (3.23)$$

VVC's inherent in-loop deblocking filters additionally suppress any residual discontinuities during residual encoding and decoding, providing a further layer of artefact suppression without requiring explicit post-processing.

The effectiveness of this artefact mitigation strategy is confirmed in Figure 3.16, which presents a visual and quantitative analysis of a 4K video sequence. At semantic block boundaries, the PSNR is 37.84 dB and SSIM is 0.88 for the semantic-only prediction, compared with 39.06 dB and 0.91 in the interior of the block. After VVC residual compensation, the boundary PSNR improves to 38.55 dB and SSIM to 0.89, showing that the residual coding stage is mainly used to correct boundary artefacts. A subjective experiment conducted with 15 video experts following the double-stimulus

methodology of ITU-R BT.500 [179] confirmed that participants did not perceive noticeable blocking artefacts at standard 4K playback, although some perceptual impact was detected in high-motion scenes, likely due to a combined effect of blocking and motion blur.

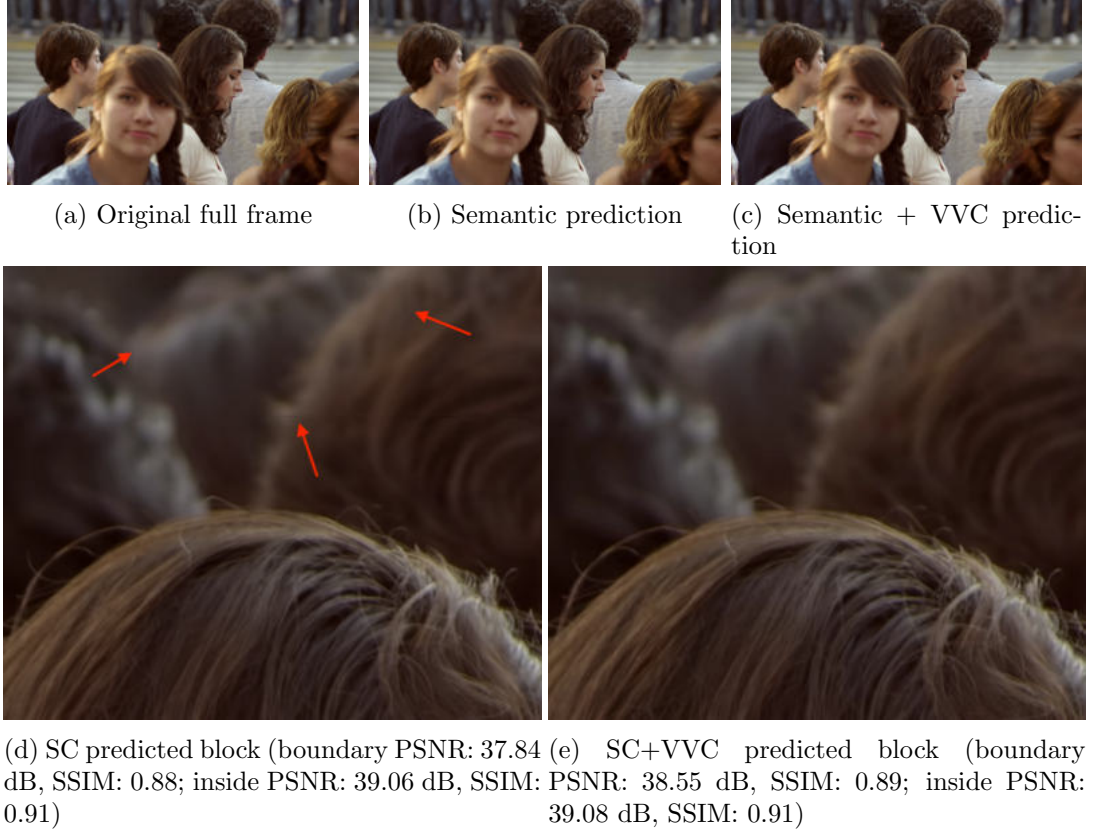


Figure 3.16: Visual and quantitative analysis of block boundary artefacts for a 4K video sequence. Top row: full-frame comparison of original, semantic prediction, and semantic+VVC prediction. Bottom row: zoomed block regions with localised PSNR/SSIM measurements at block boundaries and interior regions [22].

Dataset and Test Videos

The enhanced framework is evaluated on four standard benchmark datasets: the HEVC Class B and Class E test sequences [172], the MCL-JCV dataset [174], and the UVG dataset [173]. All test videos are downsampled to 720p resolution for the primary rate-distortion experiments. For 4K evaluation, five sequences from the Netflix Open Content dataset [180] at Digital Cinema Initiatives (DCI) 4K resolution ($4,096 \times 2,160$) are

also included.

Test videos from all four benchmark datasets are categorised into low, medium, and high complexity classes based on their SI and TI values [181], computed at 720p resolution using the Video Quality Experts Group (VQEG) SI/TI tools [182]. The SI-TI scatter plot in Figure 3.17 shows the distribution of test videos across the three complexity categories. Training of the semantic encoder and decoder is performed on an NVIDIA A100 GPU accessed via Google Colaboratory [183].

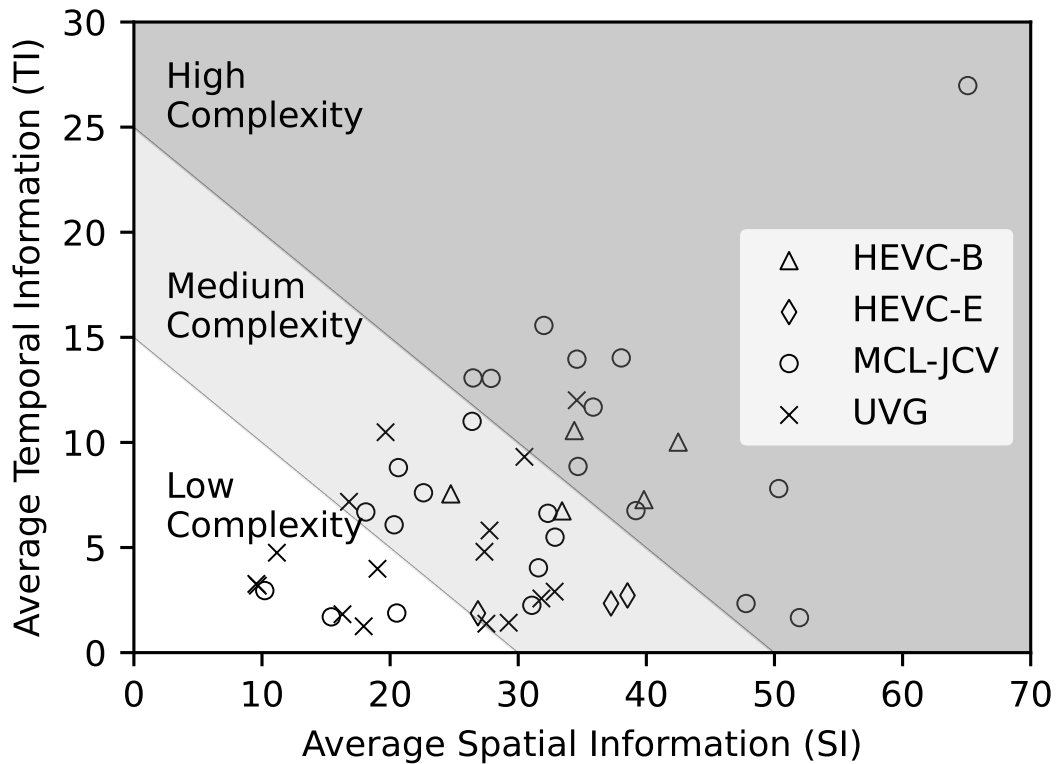


Figure 3.17: Classification of test videos from the HEVC Class B, HEVC Class E, MCL-JCV, and UVG datasets based on spatial information and temporal information, at 720p resolution [22].

Baseline Codecs and Benchmark Configuration

Two reference configurations are used for rate–distortion benchmarking. The primary comparison for standard-resolution test sequences (HEVC Class B through E, UVG, and MCL-JCV) is against VTM 17.0 [184] in random access configuration, using QP

values $\{27, 32, 37, 42\}$ following the JVET Common Test Conditions [185]. For 4K resolution test sequences, the comparison is against VTM 23.0 [184] in the same random access configuration. VTM 17.0 is retained for the standard-resolution experiments because these results reproduce the benchmark settings used by the corresponding published learned-codec references and the earlier hybrid-codec study, both of which report BD-rate relative to VTM 17.0. VTM 23.0 is used for the later 4K experiments because those experiments were run separately with the newer JVET anchor on the Netflix Open Content set. Accordingly, BD-rate results are compared only within the same anchor version and are not interpreted as direct cross-version comparisons. An additional comparison against state-of-the-art learned video codecs, specifically DCVC-FM [52] and DCVC-DC [53], is also conducted using VTM 17.0 [184] as the common reference codec with intra period 32 in order to reproduce the benchmark settings of [52]. Additional comparisons are provided against the HEVC Test Model (HM 16.25) [186], the Enhanced Compression Model (ECM 5.0) [187], and the AOMedia Video Model (AVM) [188] on the same datasets.

In the proposed codec, the intra-frame encoder and the residual encoder are implemented using the expert mode of VVenC 1.12.1 [84]. The intra-frame encoder operates in all-intra configuration at each target QP value to produce key frames. Residuals are buffered and VVC-encoded with a GOP size of 32 and a variable GOP structure optimised for rate-distortion performance.

Evaluation Metrics

Rate-distortion performance is assessed using the same three objective quality metrics introduced in Section 3.2: PSNR [171], SSIM [164], and VMAF [165], with BD-rate [162, 189] as the summary performance measure, computed relative to each reference codec by integrating across the four QP operating points of the rate-distortion curve.

3.3.3 Results and Discussion

Rate-Distortion Performance Comparison

Figures 3.18, 3.19, and 3.20 present the rate–distortion performance of the proposed codec against VTM 17.0 in random access configuration for the low, medium, and high complexity test sets, respectively; Figure 3.21 presents the comparison against VTM 23.0 for the 4K test set. Each figure shows PSNR, SSIM, and VMAF curves as functions of bitrate, together with the corresponding BD-rate.

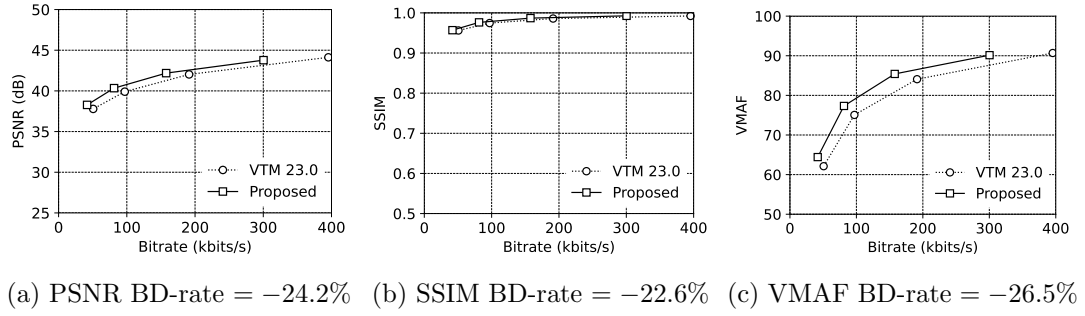


Figure 3.18: Average rate–distortion performance of the proposed codec and VTM 17.0 for low-complexity test videos. Left to right: PSNR, SSIM, VMAF. Note: the y -axes are truncated for clarity [22].

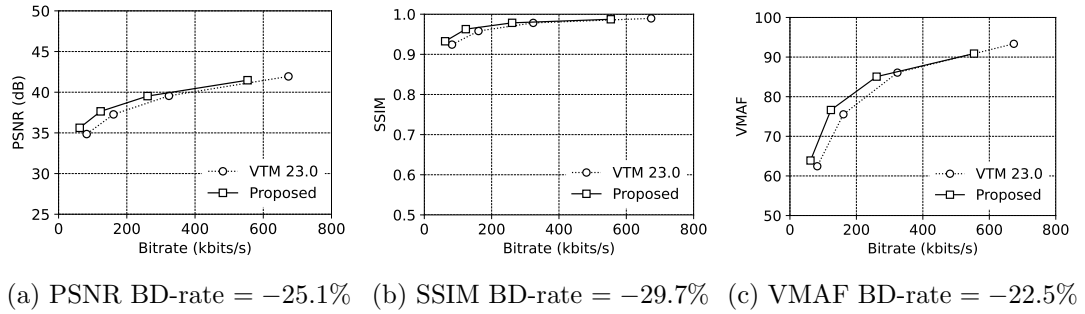


Figure 3.19: Average rate–distortion performance of the proposed codec and VTM 17.0 for medium-complexity test videos. Left to right: PSNR, SSIM, VMAF. Note: the y -axes are truncated for clarity [22].

The proposed codec achieves negative BD-rates against VTM 17.0 across all standard-resolution complexity categories and against VTM 23.0 at 4K resolution, confirming consistent rate–distortion superiority. The gains are largest for low and medium complexity content, with PSNR BD-rates of -24.2% and -25.1% , respectively, and SSIM

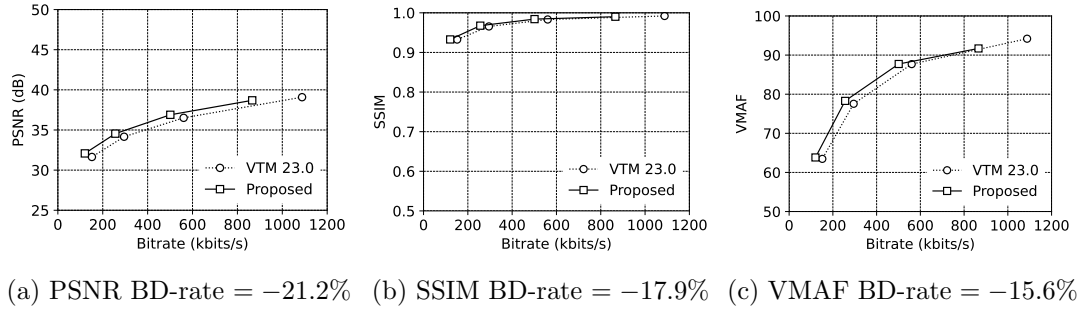


Figure 3.20: Average rate–distortion performance of the proposed codec and VTM 17.0 for high-complexity test videos. Left to right: PSNR, SSIM, VMAF. Note: the y -axes are truncated for clarity [22].

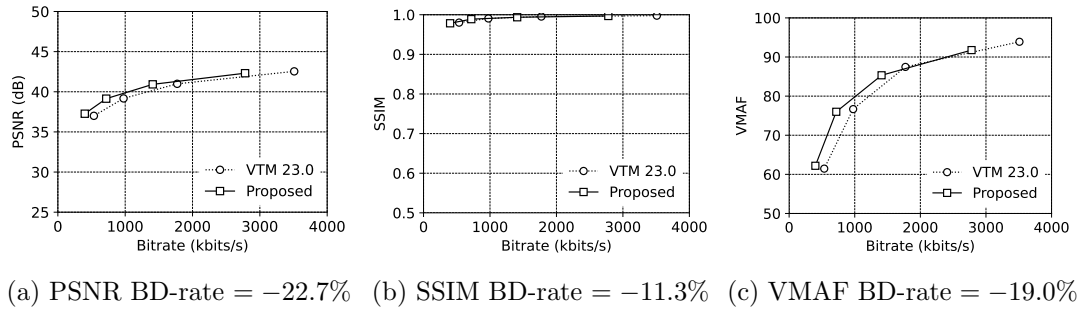


Figure 3.21: Average rate–distortion performance of the proposed codec and VTM 23.0 for 4K resolution test videos from the Netflix Open Content dataset. Left to right: PSNR, SSIM, VMAF. Note: the y -axes are truncated for clarity [22].

BD-rates reaching -22.6% and -29.7% . For high-complexity content, the PSNR BD-rate is -21.2% and the VMAF BD-rate is -15.6% . At 4K resolution, the proposed codec achieves a PSNR BD-rate of -22.7% and a VMAF BD-rate of -19.0% , demonstrating that the block-based partitioning strategy successfully extends the framework’s advantages to high-resolution content.

The variation of gains across content complexity classes can be understood through the high-rate model for transform and entropy coding [89],

$$R_{\text{residual}} \propto \frac{1}{2} \log \left(\frac{\sigma_R^2}{D_q} \right), \quad (3.24)$$

where σ_R^2 is the residual variance and D_q is the effective quantisation distortion. Bidirectional hierarchical prediction reduces σ_R^2 most effectively when inter-frame motion is structured and moderate, as is characteristic of low and medium temporal information content, yielding the largest BD-rate savings. For high temporal information content (rapidly moving scenes) and high-resolution 4K content, the semantic predictor leaves a larger innovation component, so σ_R^2 decreases less, and the BD-rate savings are proportionally smaller. Nevertheless, the semantic bottleneck compresses even high-complexity content to a constant latent bitrate regardless of scene complexity, while VVC’s inter-prediction cost increases with complexity [190], which explains the consistent advantage of the proposed codec across all content types.

Comparison with State-of-the-Art Learned Video Codecs

Table 3.13 compares the PSNR BD-rate of the proposed codec against DCVC-FM [52], DCVC-DC [53], ECM 5.0 [187], HM 16.25 [186], and AVM [188], all measured relative to VTM 17.0 on the HEVC Class B, HEVC Class E, MCL-JCV, and UVG datasets in the YUV420 colour space with intra period 32. The results for DCVC-FM and DCVC-DC are cited from [52] to provide benchmark comparisons under matched settings, using identical datasets, colour space, GOP size, and reference codec.

The proposed codec achieves the best average PSNR BD-rate of -21.1% , surpassing both DCVC-FM (-19.8%) and DCVC-DC (-15.2%). On the HEVC Class B and UVG

Table 3.13: PSNR BD-rate (%) comparison in YUV420 colour space with intra period 32 relative to VTM 17.0. Lower values indicate better rate-distortion performance [22].

Codec	HEVC-B	HEVC-E	MCL-JCV	UVG	Average
VTM 17.0	0.0	0.0	0.0	0.0	0.0
HM 16.25	39.3	37.5	38.0	45.9	40.2
ECM 5.0	−12.7	−11.0	−11.5	−15.0	−12.6
DCVC-DC	−10.8	−20.4	−17.8	−12.0	−15.2
DCVC-FM	−16.3	−30.1	−21.6	−11.4	−19.8
Proposed	−29.3	−11.1	−14.8	−29.2	−21.1

datasets, the proposed codec achieves substantially better performance than DCVC-FM (−29.3% versus −16.3% on HEVC-B, and −29.2% versus −11.4% on UVG). On the HEVC Class E and MCL-JCV datasets, DCVC-FM achieves better BD-rates than the proposed codec (−30.1% and −21.6% versus −11.1% and −14.8%). This dataset-dependent behaviour is attributable to the distribution of temporal information content: HEVC Class E and MCL-JCV contain a comparatively higher proportion of sequences with low average temporal information (less than 5), where the bidirectional semantic predictor risks overcompensation, causing the residual to increase in size with long GOPs. This also explains why DCVC-FM, trained on Vimeo-90k [175] and generalising broadly across content types, performs better on low-motion datasets, while the per-GOP overfit strategy of the proposed codec is more effective on high-motion and high-complexity content.

Importantly, the proposed codec achieves these results without requiring pretraining on any external dataset and without deploying a pretrained model at the receiver, which represents a fundamental practical advantage over DCVC-FM and DCVC-DC. Furthermore, the proposed codec achieves an average encoding time of 14.38 s per frame for 720p video on an NVIDIA A100 GPU, compared with 32.13 s for VTM 17.0 and approximately 0.5 s for DCVC-FM. While the proposed codec is slower than DCVC-FM at inference time, this comparison is not equivalent: DCVC-FM’s inference speed reflects the use of a pretrained model, whereas the proposed codec performs per-GOP adaptation that is absent from DCVC-FM’s inference pipeline.

Bitstream Composition Analysis

Table 3.14 presents the average composition of bits per frame for a 720p video encoded with the proposed codec at GOP size 32 using VVenC-based intra-frame and residual encoders across the four QP values of the JVET Common Test Conditions.

Table 3.14: Average composition of bits per frame for a 720p video from the proposed codec at GOP size 32, showing the relative contributions of the key frame, residual, and latent vector components [22].

QP	Latent (kbits)	Key Frame (kbits)	Residual (kbits)
27	0.27 (0.07%)	357.90 (95.60%)	16.20 (4.32%)
32	0.19 (0.80%)	209.41 (97.06%)	6.15 (2.85%)
37	0.12 (0.10%)	115.20 (97.96%)	2.27 (1.93%)
42	0.08 (0.13%)	59.97 (98.78%)	0.66 (1.08%)

The latent vector accounts for less than 1% of the total bitstream at all QP values, confirming that the per-GOP training strategy achieves genuine semantic compression of temporal structure rather than content memorisation. If the autoencoder were merely memorising frames, the latent representation would require a substantially larger bit allocation to reproduce them accurately. The key frame dominates the bitstream, contributing between 95.60% and 98.78% of all bits, while the residual contributes between 1.08% and 4.32%. The near-negligible latent overhead is consistent with the residual latent entropy relation $H(\Delta l) \approx H(l_{\text{sem}}) - I(l_{\text{sem}}; l_{\text{RIFE}})$: as the quality of RIFE-based semantic prediction improves, the mutual information between the semantic latent and the RIFE-predicted latent increases, further reducing the entropy of the residual latent and the semantic channel bitrate.

Ablation Study

Effect of Residual Encoding Orders. Table 3.15 presents the average PSNR, SSIM, and bitrate for an example 720p video at GOP size 32, showing the progressive effect of including successive orders of residual encoding in the hierarchical prediction structure.

The semantic-only reconstruction (no residuals, key frame only) achieves 26.61 dB

Table 3.15: Ablation study on the effect of residual frame order in the proposed codec for GOP size 32, with per-order marginal contributions [22].

Order of Residuals	PSNR (dB)	SSIM	Bitrate (kbit/s)	Δ PSNR (dB)	Δ Bitrate (kbit/s)	η (dB/Mbit)
None (key frame only)	26.61	0.7508	331.65	—	—	—
+ 1st order	31.91	0.8226	368.96	+5.30	+37.31	142.03
+ 2nd order	35.53	0.8823	431.92	+3.62	+62.96	57.50
+ 3rd order	37.99	0.9539	466.75	+2.46	+34.83	70.63
+ 4th order	39.37	0.9840	545.75	+1.38	+79.00	17.47

PSNR and 0.7508 SSIM at a bitrate of 331.65 kbit/s, demonstrating that the semantic channel alone provides sufficient fidelity for machine vision tasks such as scene understanding and object identification. Adding the first-order residual yields the largest marginal gain of +5.30 dB PSNR at a rate–distortion efficiency of 142.03 dB/Mbit, because first-order frames have the greatest temporal distance from key frames ($N/2 = 16$ frames at GOP size 32) and therefore exhibit the highest prediction error energy. Subsequent orders yield progressively smaller marginal gains as the temporal prediction distance halves at each level. The monotonic improvement from 26.61 dB to 39.37 dB empirically validates the additive distortion model $D_{\text{total}} = D_c + \alpha D_s$: in the absence of channel errors ($D_c = 0$), successive reductions of semantic distortion D_s through residual coding produce predictable perceptual quality improvements.

Figure 3.22 visualises the PSNR and bitrate trade-offs across residual orders. The rate–distortion efficiency η drops below 20 dB/Mbit at the fourth order, suggesting that for bandwidth-constrained scenarios, encoding up to the third order (PSNR 37.99 dB, bitrate 466.75 kbit/s) provides an excellent rate–distortion trade-off, while the fourth order offers diminishing returns.

Bidirectional versus Unidirectional Prediction. Table 3.16 compares unidirectional (sequential) and bidirectional (hierarchical) prediction modes.

Table 3.16: Ablation study comparing unidirectional and bidirectional frame prediction modes relative to VTM 17.0 [22].

Prediction Mode	BD-rate	Outcome
Unidirectional (sequential)	+4.68%	Worse than VTM
Bidirectional (hierarchical)	−21.1%	Best

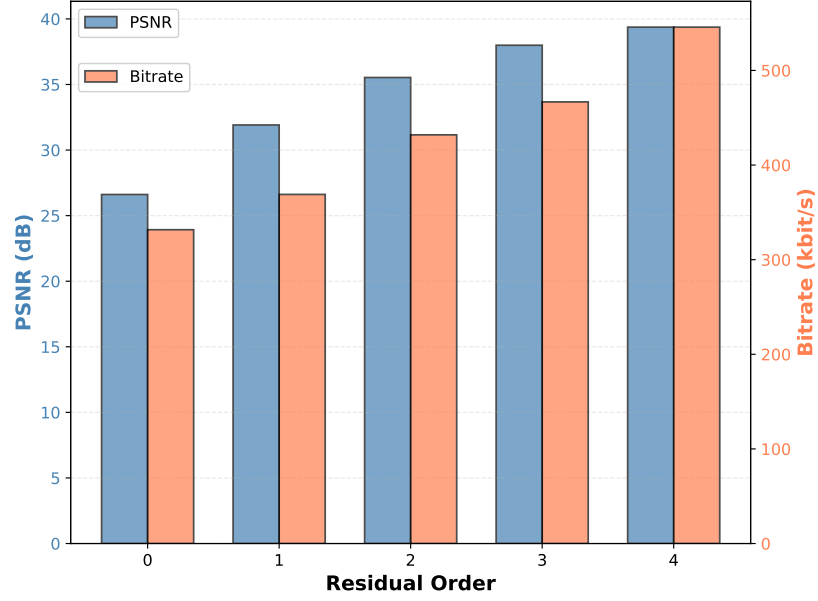


Figure 3.22: PSNR and bitrate trade-offs across residual orders. The monotonic increase in PSNR from 26.61 dB to 39.37 dB with increasing residual orders is accompanied by diminishing marginal PSNR gains: 1st order (+5.30 dB), 2nd order (+3.62 dB), 3rd order (+2.46 dB), 4th order (+1.38 dB) [22].

Unidirectional prediction fails to exploit the temporal structure within a GOP: without access to future reference frames, the semantic autoencoder must extrapolate temporal patterns, leading to increasingly inaccurate predictions as the temporal distance from the last key frame grows. This accumulation of semantic prediction errors results in a BD-rate of +4.68%, indicating a codec that performs worse than VTM 17.0. Bidirectional hierarchical prediction eliminates this error accumulation by constraining every prediction within a bounded temporal window defined by two anchor frames, achieving a BD-rate of −21.1%.

Semantic Encoder Architecture. Tables 3.17, 3.18, and 3.19 present ablation results for the hidden layer dimension d_h , the activation function, and the latent vector dimension L , respectively.

The hidden layer dimension $d_h = 1,024$ provides the optimal balance between reconstruction quality and computational cost: increasing to $d_h = 2,048$ yields only +0.28 dB improvement at 31% greater encoding time and double the parameter count.

Table 3.17: Ablation study on the hidden layer dimension of the semantic encoder [22].

d_h	PSNR (dB)	Encoding Time (s/f)	Parameters (M)
256	25.87	11.2	44.2
512	26.92	12.5	88.5
1024	29.23	14.4	177.0
2048	29.51	18.9	354.0
4096	29.63	29.7	708.0

Table 3.18: Ablation study on the activation function of the semantic encoder [22].

Activation	PSNR (dB)	Convergence Epochs	Training Stability
LeakyReLU ($\alpha = 0.1$)	29.23	42	Stable
ReLU	28.11	45	Stable
Tanh	27.85	63	Unstable
Sigmoid	27.97	72	Unstable

Table 3.19: Ablation study on the latent vector dimension of the semantic encoder [22].

L	PSNR (dB)	Semantic Bitrate (kbit/s)
4	25.42	0.09
8	29.23	0.19
16	29.42	0.38
32	29.64	0.77
64	29.81	1.54

LeakyReLU achieves the highest PSNR and the fastest convergence among the activation functions evaluated, with Tanh and Sigmoid exhibiting unstable training dynamics due to vanishing gradient issues. The latent dimension $L = 8$ is the minimum size that avoids a significant drop in reconstruction quality: increasing to $L = 16$ yields only +0.19 dB at double the semantic bitrate, confirming $L = 8$ as the optimal operating point.

Block Size Selection. Table 3.20 presents an empirical comparison of rate–distortion and computational trade-offs across three semantic block sizes.

Table 3.20: Ablation study on rate–distortion and computational trade-offs at matched bitrate according to semantic block size [22].

Block Size	PSNR (dB)	Encoding Time (s/f)	Memory per Block
90p (160×90)	35.42	3.6 s	0.96 GB
180p (320×180)	37.23	14.4 s	3.85 GB
360p (640×360)	37.35	67.3 s	15.4 GB

The 180p block size provides an optimal trade-off between reconstruction quality and computational cost. The 90p block is 1.81 dB worse in PSNR, because smaller blocks often contain only partial objects, fragmenting semantic understanding. The 360p block achieves only +0.12 dB improvement over 180p at $4.68\times$ greater encoding time and $4\times$ higher memory per block, approaching the memory capacity limit of a single GPU. A 180p block typically contains one to three complete objects or scene elements in natural videos, providing sufficient semantic context for the autoencoder to learn meaningful representations.

YUV Channel Processing Strategy. Table 3.21 compares the joint YUV420 flattening approach against a strategy employing separate autoencoders for the Y, U, and V channels.

Separate Y, U, V autoencoders achieve +0.64 dB improvement over the joint approach at the cost of a 165% increase in encoding time. The joint approach achieves comparable reconstruction quality while maintaining practical encoding latency, making it the preferred choice. The implicit luminance weighting resulting from the 4:1 pixel

Table 3.21: Ablation study comparing joint YUV420 and separate channel processing strategies [22].

Approach	PSNR (dB)	Encoding Time (s/frame)
Separate Y, U, V autoencoders	29.87	38.14
Joint YUV420 (proposed)	29.23	14.38

ratio of Y to U/V in YUV420 is sufficient for the primary objective of demonstrating semantic communication viability.

RIFE and Semantic Autoencoder Complementarity. The ablation further examines the complementary roles of the RIFE intermediate flow estimation module and the semantic autoencoder in the bidirectional prediction framework. The semantic autoencoder alone captures high-level structural information and performs temporal prediction, but typically achieves 1–2 dB less PSNR than RIFE alone for sequences with complex optical flow patterns. The combination of RIFE with the proposed semantic autoencoder achieves a 2–4 dB improvement over either component individually. This superior performance arises from the complementary strengths of the two components: RIFE excels at pixel-level motion alignment and handles structured motion effectively, whereas the semantic autoencoder captures high-level scene context and content-specific patterns that RIFE cannot access. The per-GOP training of the semantic autoencoder allows it to develop specialised representations that, combined with RIFE’s general-purpose optical flow estimation, achieve superior interpolation quality, particularly for video sequences with complex or irregular motion patterns. Figure 3.23 provides a representative example demonstrating the qualitative improvement achieved by this integration: RIFE alone achieves 30.426 dB PSNR and 0.868 SSIM, whereas the proposed combination reaches 32.374 dB PSNR and 0.874 SSIM.

Error Resilience Analysis

An intrinsic advantage of semantics-based bidirectional prediction is increased resilience of the encoded bitstream to transmission errors. Figure 3.24 shows the variation in PSNR, SSIM, and VMAF as functions of BER for low, medium, and high-complexity



Figure 3.23: Prediction quality comparison illustrating the complementary roles of RIFE and the proposed semantic autoencoder. Combining RIFE with the per-GOP trained semantic autoencoder achieves a 1.95 dB PSNR improvement over RIFE alone [22].

test videos, comparing the proposed codec against VTM 17.0.

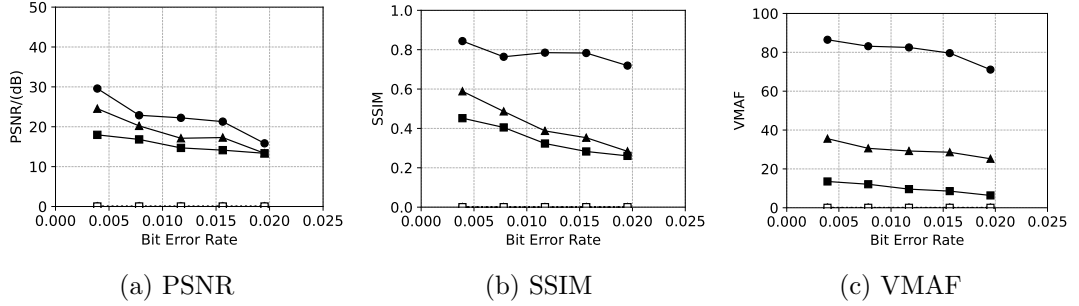


Figure 3.24: Rate-distortion performance of low- ($\circ$), medium- ($\triangle$), and high- ($\square$) complexity videos under bit errors, for the proposed codec (solid lines) and VTM 17.0 (dotted lines) at GOP size 32. Left to right: (a) PSNR, (b) SSIM, (c) VMAF [22].

The proposed codec exhibits a graceful decline in reconstructed video quality as BER increases, while VTM 17.0 fails to produce decodable output at $\text{BER} \geq 10^{-5}$ due to intra-tile CABAC context desynchronisation causing cascading syntax errors that slice-based resynchronisation cannot recover. This failure persists even when VVC is configured with extensive error-resilience settings, including independent tile picture partitioning, explicit tile entry points, disabled cross-boundary loop filtering, and clean instantaneous decoder refresh (IDR) random access points.

The error-resilient behaviour of the proposed codec is understood through the following distortion decomposition [125]. Let bit errors corrupt only residual symbols

with probability p_b . The expected total distortion decomposes as

$$\mathbb{E}[D_{\text{total}}] = D_s + (1 - p_b) D_c + p_b D_c^{\text{err}}, \quad (3.25)$$

where D_s is the semantic distortion from the semantic reconstruction (key frames and latent vectors) which is unaffected by small syntactic corruptions, D_c is the residual channel distortion under error-free conditions, and $D_c^{\text{err}} > D_c$ reflects the increased residual distortion under bit error conditions. Because D_s is often the dominant term at moderate bitrates, small p_b produces only gradual quality loss. When bit errors corrupt the residual bitstream, VVC decoding fails and the system automatically falls back to semantic-only prediction mode, maintaining reconstruction quality in the range of 24 to 30 dB PSNR, which is sufficient for many practical applications. In contrast, the predictive syntax of VTM is highly sensitive to bitstream errors, leading to decoder desynchronisation and catastrophic quality failure.

The interaction between semantic noise and channel noise is modelled following the framework of [125]. The total noise at spatial coordinates (x, y) and temporal coordinate t is expressed as

$$n_{\text{total}}(x, y, t) = n_{\text{sem}}(x, y, t) + n_{\text{tec}}(x, y, t), \quad (3.26)$$

where $n_{\text{sem}} \sim \mathcal{N}(0, \sigma_{\text{sem}}^2)$ and $n_{\text{tec}} \sim \mathcal{N}(0, \sigma_{\text{tec}}^2)$ under a Gaussian approximation. Assuming statistical independence between the two noise sources, the total noise variance is $\sigma_{\text{total}}^2 = \sigma_{\text{sem}}^2 + \sigma_{\text{tec}}^2$.

Figure 3.25 shows the distribution of semantic prediction residual values across the three content complexity categories.

High-complexity content exhibits a wide, relatively flat residual distribution with $\sigma_{\text{sem}} \approx 16.73$, indicating that the semantic predictor struggles to capture complex scene dynamics. Medium-complexity content shows a narrower distribution with $\sigma_{\text{sem}} \approx 9.49$, and low-complexity content demonstrates a highly concentrated distribution around zero with $\sigma_{\text{sem}} \approx 3.35$. These content-dependent σ_{sem} values directly correlate with the BD-rate variations observed across content types: low-complexity sequences achieve

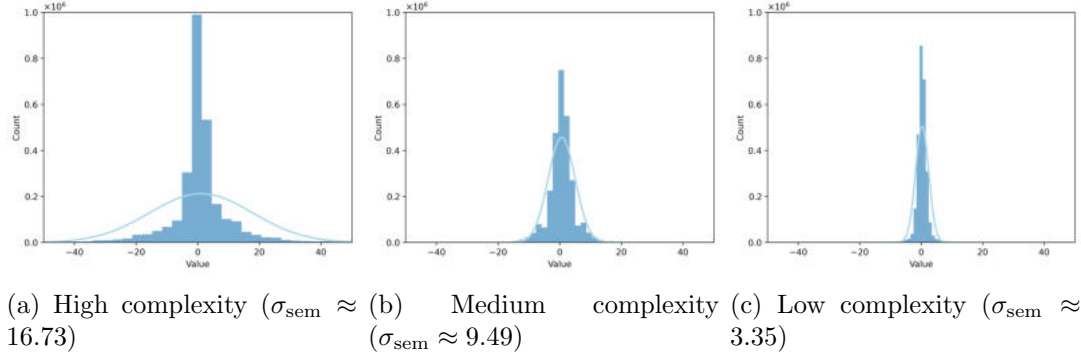


Figure 3.25: Distribution of semantic prediction residual values for different video content complexities, with overlaid fitted Gaussian curves. High-complexity content exhibits a wide, flat distribution indicating large semantic distortion; low-complexity content demonstrates a highly concentrated distribution around zero indicating minimal semantic distortion [22].

the largest BD-rate savings (up to -29.7% in SSIM) because the semantic predictor alone provides near-transparent reconstruction and leaves minimal residual energy for VVC to encode, while high-complexity sequences require proportionally more residual bits to maintain perceptual quality.

Computational Complexity

Table 3.22 presents a comprehensive runtime and complexity comparison between the proposed codec, VTM 17.0, and DCVC-FM for 720p YUV420 video.

Table 3.22: Comprehensive runtime and complexity comparison for 720p YUV420 video encoding and decoding [22].

Metric	Proposed Codec	VTM 17.0	DCVC-FM
Encoding time (s/frame)	~ 14.38	~ 32.13	~ 0.5
Decoding time (s/frame)	~ 3.85	~ 0.04	~ 0.3
Processor utilisation	GPU (A100)	CPU	GPU
Model transmission required	No	No	Yes
Pre-training required	No	No	Yes (Vimeo-90k)

The proposed codec requires an average encoding time of 14.38 s/frame and a decoding time of 3.85 s/frame on an NVIDIA A100 GPU for 720p video. The encoding process comprises approximately 6.3 s for inter-frame semantic encoding, 3.2 s for resid-

ual coding, and 4.4 s for intra-frame key frame coding. VTM 17.0 requires 32.13 s/frame for encoding at matched quality but only 0.04 s/frame for decoding (CPU). Notably, the residual coding time for the proposed codec (6.07 s/frame) is substantially lower than encoding the same video directly with VVenC (24.14 s/frame at QP 27), demonstrating that the semantic prediction reduces the residual entropy sufficiently to lower the per-frame VVC encoding workload.

The semantic encoder and decoder together require approximately 176,963,584 MAC operations per 180p block, corresponding to 353,927,168 FLOPs. This complexity remains significantly below that of VVC inter-coding because the per-GOP training of the semantic autoencoder captures inter-frame temporal correlations, enabling the residual VVC coding to operate predominantly in intra-prediction mode with reduced motion estimation complexity. The proposed framework does not require model transmission or pre-training, which represents a further practical advantage over learned video codecs such as DCVC-FM that require the deployment of a large trained model at the decoder.

3.3.4 Conclusion

This section has presented an enhanced hybrid semantic–VVC video codec that builds on the basic system of Section 3.2 by adding bidirectional hierarchical frame prediction, block-based high-resolution processing, and residual latent entropy coding [22]. The key findings are summarised below.

The enhanced system achieves an average PSNR BD-rate of -21.1% relative to VTM 17.0 across the standard-resolution test sequences (HEVC Class B, HEVC Class E, MCL-JCV, and UVG), which is the best result among all systems tested, including DCVC-FM (-19.8%) and DCVC-DC (-15.2%). The gains are largest on high-complexity sequences: HEVC Class B sequences give an average BD-rate of -29.3% and the UVG dataset gives -29.2% , both well above what learned codecs achieve on the same content. On HEVC Class C through E sequences and the MCL-JCV dataset the gains are smaller, which agrees with the finding in Section 3.2 that medium-complexity content with structured motion is where conventional motion compensation works best.

The bidirectional hierarchical prediction structure is the main contributor to the rate–distortion gain, as shown by the ablation study. Removing the RIFE interpolation module and switching to unidirectional prediction raises the PSNR BD-rate relative to VTM 17.0 to +4.68%, a drop of 25.8 percentage points, confirming that bidirectional prediction with optical flow interpolation is needed for the system to work well. The block-based high-resolution extension scales the framework from 320×180 to 4K content with little loss in compression performance, confirming that the resolution flexibility of the basic system extends to higher, practical resolutions.

The bitstream composition analysis confirms that the semantic latent channel accounts for less than 1% of the total bitstream at typical operating points, with most bits used by VVC residual coding. This means that better temporal prediction will give the largest bitrate savings in future work. The DeepCABAC latent entropy coding stage gives a small additional saving on the latent component, showing that entropy coding of the small latent vectors is a useful addition to the main coding stages.

The error resilience analysis shows a clear difference between the proposed system and VTM 17.0 under channel error conditions. At BER of 10^{-5} and above, VTM 17.0 fails completely because decoding errors spread through the coding chain, while the proposed system loses quality gradually because each GOP is coded independently with no inter-GOP prediction. This robustness to channel errors is a built-in property of the per-GOP structure and does not need additional error protection, which is a practical advantage for wireless video transmission.

Encoding time on an NVIDIA A100 GPU averages 14.38 s/frame for 720p content, made up of approximately 6.3 s for inter-frame semantic encoding, 4.4 s for key frame intra coding, and 3.2 s for residual coding. The encoding time is higher than the 32.13 s/frame of VTM 17.0 software encoding on CPU, but the decoding time of 3.85 s/frame is near real-time and can be reduced further with dedicated hardware. The system does not need a pretrained model and does not need to send model weights over the channel, which are practical advantages over learned codecs in resource-limited settings.

This section addresses Gap 1 from Chapter 2 with better performance than the

basic system of Section 3.2, showing that bidirectional hierarchical prediction in the hybrid framework matches or outperforms the state of the art in both conventional and learned video coding across a wide range of test conditions.

3.4 Chapter Summary and Conclusion

This chapter has presented two original contributions that together show that hybrid semantic and conventional video coding is a practical and effective alternative to both fully conventional and fully learned video coding. The two contributions address Objective 1 and Objective 2 of the thesis and together resolve Gap 1 identified in Chapter 2: the absence of a hybrid system combining semantic inter-frame prediction with the proven context and residual coding capabilities of a standard conventional codec. Section 3.2 presented the foundational hybrid codec, whose main innovation is the remote decoder training mechanism [1]. By initialising the semantic autoencoder at both transmitter and receiver from the same Glorot uniform seed [166], the system allows the receiver to independently recreate the transmitter’s decoder using only the decoded key frames and received latent vectors, without sending any network parameters. The average PSNR difference between locally trained and remotely trained decoder outputs across seven test sequences was negligible, confirming that both decoders produce the same result. Rate-distortion evaluation across twenty test sequences covering low, medium, and high complexity content showed consistent BD-rate improvements over AVC and HEVC of 19–58% in PSNR, and performance at or above VVC for low and high complexity content. Resolution invariance from 256×144 to 1920×1080 and a $2.65\times$ reduction in VVenC per-frame encoding time through semantic pre-processing were also confirmed.

Section 3.3 presented the enhanced codec, which fixes the two main limitations of the basic system by adding bidirectional hierarchical frame prediction with RIFE [72], block-based high-resolution processing, and DeepCABAC residual latent entropy coding [22, 148]. The enhanced system achieved an average PSNR BD-rate of -21.1% relative to VTM 17.0 across the standard-resolution test sequences, which is better than all evaluated learned codec references including DCVC-FM (-19.8%) and DCVC-DC

(−15.2%). The ablation study showed that RIFE-based bidirectional prediction was the main source of this gain, with removing the interpolation module degrading performance to +4.68% BD-rate relative to VTM 17.0. The system was also shown to lose quality gradually under channel bit errors at rates where VTM 17.0 fails completely, confirming a robustness advantage built into the per-GOP structure of the codec.

Together, the two contributions show a clear research progression: from establishing the hybrid architecture and remote training concept, to a system that matches or outperforms the state of the art across standard test datasets at practical resolutions. The main open limitation identified by both contributions is that VVC residual coding dominates the total bitrate, which means that improving semantic temporal prediction quality will give the largest gains in future work. This motivates the research in Chapter 4, which reduces the per-GOP retraining cost by using pretrained context-aware semantic models and content-adaptive parameter specialisation, lowering the computational cost of the semantic coding stage while also improving temporal prediction quality.

Chapter 4

Context-Aware Semantic Video Coding

4.1 Chapter Introduction

The hybrid semantic video codecs developed in Chapter 3 showed that combining autoencoder-based semantic inter-frame prediction with VVC residual coding is effective, with average BD-rate reductions greater than -20% relative to VTM 17.0 [22]. The main limitation of both systems is the need for per-GOP training from scratch. The semantic autoencoder must be retrained from random initialisation for every GOP, which makes encoding sequential, impractical for long videos, and increasingly expensive as video duration grows. Solving this problem defines Gap 2, context-conditioned semantic video coding with pretrained models, and Gap 3, content-adaptive per-GOP parameter overfitting, identified in Chapter 2. These are the main challenges addressed in this chapter.

Section 4.2 presents a context-aware semantic video coding framework (CASVC) that addresses Gap 2 by replacing per-GOP retraining with a single shared pretrained autoencoder used without per-GOP parameter updates. The framework has three key parts: a motion-adaptive model-selection strategy that applies a dedicated fine-tuned variant to sequences with TI above 15, signalled by one flag; hierarchical bidirectional prediction through the RIFE interpolation network [72], which generates motion-

aligned context frames for the semantic decoder; and differential PCA–DeepCABAC [147, 148] entropy coding, which uses the strong correlation between target-frame and context-frame latents to keep the semantic sub-stream below 0.4% of total bitrate. The system is evaluated on four standard 1080p datasets, UVG [173], MCL-JCV [174], HEVC Class B and Class E [172], and a 4K dataset from the Netflix Open Content library [180], all under JVET random-access common test conditions [184].

Section 4.3 presents a content-adaptive parameter overfitting mechanism (CAPO) that addresses Gap 3 by bringing per-GOP content specialisation back into the CASVC framework at much lower cost than full retraining. The main idea is that the fixed global decoder of CASVC represents an average over the Vimeo-90k training data [175]. Any mismatch between this average model and the local content of the current GOP appears as structured prediction error that the VVC residual path must code. CAPO reduces this mismatch by fine-tuning only a small set of decoder parameters, namely per-channel scale multipliers and transposed-convolution biases, totalling about 643 values, while keeping all convolutional kernels, normalisation parameters, and the full encoder fixed. A tapered $32 \rightarrow 64 \rightarrow 96 \rightarrow 128$ encoder with a mirrored decoder and soft-quantisation bottleneck reduces per-block processing complexity by about $4.9\times$ compared with CASVC. A unified PCA–DeepCABAC pipeline transmits the per-GOP parameter update together with the differential latents and keeps this extra sub-stream below 1% of total bitrate.

The method in Section 4.2 was reported in the first-author publication “Context-Aware Semantic Video Coding Using Pretrained Motion-Adaptive Models and Bidirectional Prediction” [73], while Section 4.3 incorporates the subsequent study “Context-Aware Semantic Video Coding via Content-Adaptive Parameter Overfitting.” Their progression moves the thesis from per-GOP training to pretrained coding and then restores content specialisation through lightweight decoder adaptation. Section 4.4 summarises these contributions and links them to the research objectives of the thesis.

4.2 Context-Aware Coding with Pretrained Motion-Adaptive Models and Bidirectional Prediction

Publication Notice

The research presented in this section has been published in the following peer-reviewed publication:

P. Samarathunga, Y. Ganearachchi, T. Fernando, A. Jayasingam, T. Sivalingham, N. Rajatheva, and A. Fernando, “Context-Aware Semantic Video Coding Using Pretrained Motion-Adaptive Models and Bidirectional Prediction,” *IEEE Access*, pp. 1–1, 2026, doi: 10.1109/ACCESS.2026.3686329.

This material is reproduced and adapted under the terms of the IEEE Copyright policy for use in a doctoral thesis. The content has been revised and extended for inclusion and integration in this doctoral thesis, including additional explanations, expanded discussion, and alignment with the overarching research narrative.

4.2.1 Introduction

The motivation and principal components of CASVC are summarised in the Chapter Introduction. This section gives the full method for replacing the per-GOP training used by the codecs in Chapter 3 with a pretrained motion-adaptive model, bidirectional context conditioning, and differential latent coding.

The full methodology, including the autoencoder architecture, context-conditioning mechanism, motion-adaptive model selection, and entropy-coding pipeline, is presented in Section 4.2.2. Experimental results against VTM 17.0, ECM-11.0, and state-of-the-art learned codecs on four 1080p benchmarks and one 4K dataset, together with ablation studies and complexity analysis, are presented in Section 4.2.3.

4.2.2 Methodology

System Overview

The proposed framework is a hybrid semantic video codec that combines a context-conditioned semantic autoencoder with VVC residual coding and RIFE-based bidirec-

tional frame interpolation [72]. As shown in Fig. 4.1, the transmitter splits each input frame into blocks of 144×256 pixels, encodes periodic key frames as VVC intra frames, and processes each intermediate frame through three parallel paths: (i) RIFE-based bidirectional interpolation to generate a motion-aligned context frame C_t , (ii) semantic encoding through a pretrained autoencoder to produce a compact latent vector h_t , and (iii) differential latent encoding with DeepCABAC to compress the difference between h_t and the latent derived from C_t . VVenC [84] then encodes the residual between the original frame and the joint semantic-plus-context prediction, and the receiver adds the decoded residual to the semantic reconstruction to recover the full-quality video signal. The total bitrate is the sum of three components:

$$R_{\text{total}} = R_{\text{key}} + R_{\text{latent}} + R_{\text{residual}}, \quad (4.1)$$

where R_{key} , R_{latent} , and R_{residual} denote the bitrates contributed by intra-coded key frames, differential semantic latents, and VVC-coded residuals, respectively. A key design goal is to keep R_{latent} very small, so that the semantic channel adds only negligible overhead compared with R_{key} and R_{residual} .

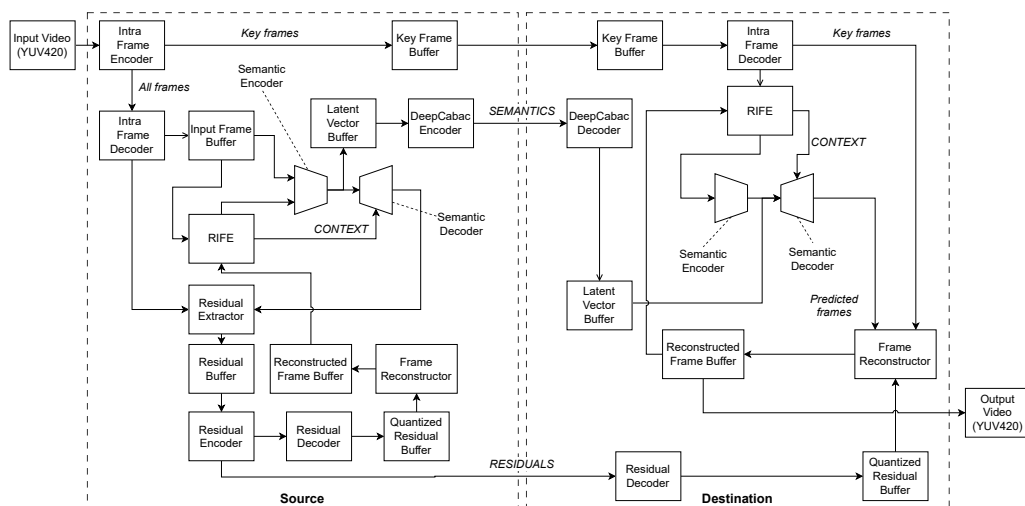


Figure 4.1: End-to-end architecture of the proposed context-conditioned hybrid video codec.

Semantic Autoencoder Architecture

The semantic autoencoder consists of a convolutional encoder ϕ_{θ_e} and a context-conditioned decoder ψ_{θ_d} . The encoder maps a quantised red-green-blue (RGB) input frame $Q_i(F_t) \in \mathbb{R}^{144 \times 256 \times 3}$ to a compact latent representation $h_t \in \mathbb{R}^{9 \times 16 \times 32}$, giving a spatial compression ratio of 256:1 with 4,608 latent values per block:

$$h_t = \phi_{\theta_e}(Q_i(F_t)). \quad (4.2)$$

The encoder has five sequential stages. The first four use stride-2 3×3 convolutions with 128 filters [47], followed by batch normalisation [191], LeakyReLU activation ($\alpha = 0.01$), and a residual block similar to the one used in [51]. Together, these four downsampling stages reduce the spatial resolution from 144×256 to 9×16 . The fifth stage uses a stride-1 convolution with 32 filters [147] to produce the latent bottleneck. This design places most of the computation in the early stages where the spatial resolution is highest, with the first encoder stage accounting for about 66% of the total encoder floating-point operations.

The decoder ψ_{θ_d} reverses this process through four transposed-convolution upsampling stages with 128 filters, stride 2, and 3×3 kernels. Each stage is followed by batch normalisation, LeakyReLU activation, and a residual block. The final decoder stage adds *context conditioning*: the RIFE-interpolated context frame C_t is concatenated with the upsampled feature map before a residual block with 64 filters, a Dropout layer with rate $p = 0.2$, and a final transposed convolution to three output channels. The conditional reconstruction is expressed as

$$\hat{F}_t = \psi_{\theta_d}(h_t, C_t). \quad (4.3)$$

This conditioning mechanism uses the spatial alignment between C_t and the target frame F_t to provide explicit scene-level guidance to the decoder, which substantially reduces residual prediction error. The encoder requires 8.19 GFLOPs (4.10 GMACs) per 144×256 -pixel block, and the decoder requires 52.22 GFLOPs (26.11 GMACs), reflecting the common encoder–decoder asymmetry of learned compression systems.

The network architecture is illustrated in Fig. 4.2.

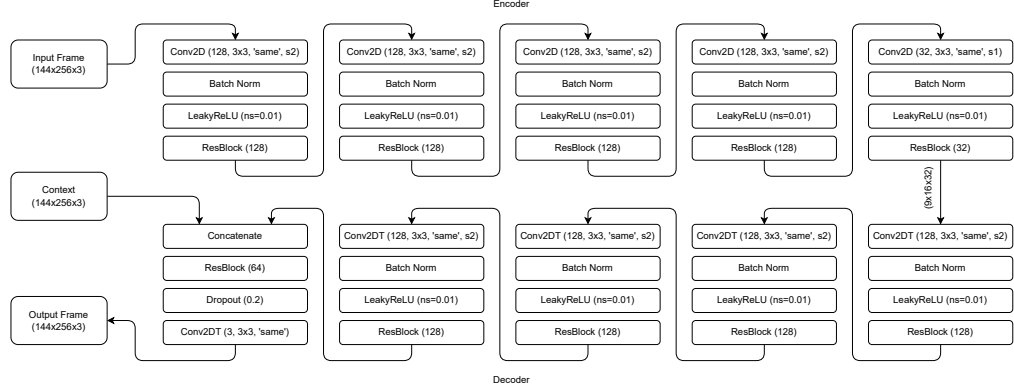


Figure 4.2: Network architecture of the proposed semantic autoencoder.

Design Rationale

The architectural choices described in Section 4.2.2 are motivated by established principles from learned image and video compression [47, 51, 147] applied to the specific requirements of hybrid semantic video coding.

The latent dimensionality of $9 \times 16 \times 32 = 4,608$ values represents a deliberate balance between compression efficiency and reconstruction fidelity. Retaining a spatial structure of 9×16 , rather than flattening the latent to a one-dimensional vector, preserves spatial locality so that each latent position corresponds to a specific 16×16 -pixel patch of the original block. This spatial organisation provides the decoder with implicit positional guidance that substantially aids in the reconstruction of spatially coherent outputs. The channel depth of 32 is empirically validated in the ablation study (Section 4.2.3): increasing from 16 to 32 channels yields a +6.6 dB gain in semantic PSNR, while further increasing to 64 channels yields only +0.4 dB at twice the latent bitrate. The 32-channel representation is therefore the operating point at which the marginal gain in reconstruction quality per unit of additional latent dimensionality falls below a practical threshold. The total latent size of 4,608 values per block is sufficiently compact to permit negligible bitrate overhead through differential DeepCABAC entropy coding, as confirmed by the bitrate composition analysis showing that semantic latents

account for only 0.3 to 0.4% of total bitrate across the full QP range.

The incorporation of the RIFE-interpolated context frame C_t in the decoder is the defining characteristic that situates the proposed approach within the semantic communication paradigm. Conventional autoencoders reconstruct outputs solely from their latent codes, requiring the latent to carry the full informational burden of the target frame. By contrast, the context-conditioned decoder $\psi_{\theta_d}(h_t, C_t)$ leverages two complementary information sources: the compact learned features encoded in h_t , which capture scene-level semantic structure, and the explicit spatiotemporal reference provided by C_t , which conveys motion dynamics, scene continuity, and fine-grained spatial detail derived from already-decoded reference frames. This design embodies the core semantic communication principle: instead of transmitting exhaustive pixel-level syntactic information, the encoder transmits a compact semantic token that the decoder contextualises using shared knowledge, namely, the RIFE prediction computed identically at both encoder and decoder. The context conditioning mechanism resolves reconstruction ambiguities that would otherwise require a substantially larger latent representation, enabling the system to achieve 28–35 dB semantic PSNR from the compressed latent alone before any VVC residual correction.

The systematic use of residual blocks (ResBlocks) [51] throughout both the encoder and decoder serves two purposes. First, skip connections within each residual block ensure unimpeded gradient flow during backpropagation, preventing the vanishing gradient phenomenon that degrades training of deep convolutional networks [192]. Second, the residual formulation encourages each processing stage to learn an incremental refinement of the feature representation rather than a complete transformation, enabling the network to build a hierarchical abstraction of the video frame from pixel values at the input to compact semantic features at the latent bottleneck. This property is particularly important for semantic video coding, where the encoder must learn a highly nonlinear mapping between pixel space and a compact semantic space while retaining the information necessary for accurate reconstruction.

The architecture is deliberately constrained to efficient convolutional operations with uniform 3×3 kernels throughout, avoiding attention mechanisms, deformable

convolutions, or dense fully connected layers. This constraint serves two objectives: (i) inference efficiency, as convolutional operations are highly optimised on contemporary GPU hardware and facilitate hardware acceleration, and (ii) inductive bias, as translation-equivariant convolutional operations are well matched to the local spatial structure of natural video frames. The uniform kernel size also simplifies hardware implementation relative to multi-scale or depthwise separable architectures, supporting the near-real-time decoding performance of 19.2 fps reported in Section 4.2.3.

Batch normalisation is applied after each convolutional layer to stabilise training by normalising internal feature distributions, accelerating convergence and reducing sensitivity to learning rate selection. Dropout (rate $p = 0.2$) is applied exclusively in the final decoder reconstruction stage to provide regularisation during training on the Vimeo-90k dataset, preventing the decoder from overfitting to specific texture patterns or motion statistics of the training sequences. At inference time, Dropout is disabled so that the full decoder capacity is available for reconstruction.

Hierarchical Bidirectional Prediction

Inter-frame prediction is organised into a hierarchical bidirectional structure that extends across a GOP of size $N = 2^n$, enabling the codec to exploit temporal correlations at multiple timescales simultaneously. For a GOP spanning frames F_m through F_{m+N} , the midpoint frame $F_{m+N/2}$ is predicted first using both boundary key frames as references:

$$\hat{F}_{m+N/2} = f_{\text{sem}}(F_m, F_{m+N}; \boldsymbol{\theta}_f), \quad (4.4)$$

where $f_{\text{sem}}(\cdot)$ denotes the combined RIFE interpolation and semantic decoding pipeline parameterised by $\boldsymbol{\theta}_f$. After the midpoint is fully reconstructed (semantic prediction plus VVC residual), the quarter-point frames $F_{m+N/4}$ and $F_{m+3N/4}$ are predicted using the nearest already-reconstructed frames as bidirectional references. This process recurses until all intra-GOP frames have been processed. For a $\text{GOP} = 8$, the processing order is: $F_4 \rightarrow \{F_2, F_6\} \rightarrow \{F_1, F_3, F_5, F_7\}$.

At each level, residual formation and reconstruction follow the hybrid relations already established in Eqns. (3.15) and (3.16): VVC codes the difference between the

quantised target and semantic prediction, and the decoded residual is added to that prediction. Here the same relations are applied recursively, with $\hat{F}_{m+N/2}$ supplied by the pretrained context-conditioned predictor rather than the per-GOP-trained predictor of Chapter 3. The use of fully reconstructed frames (semantic prediction plus residual) as references at each level propagates maximum fidelity information to subsequent prediction stages, progressively reducing residual energy across the hierarchy. This prediction order is the same hierarchical GOP structure introduced in Chapter 3 and illustrated in Fig. 3.13. The difference here is the predictor: CASVC uses the shared pretrained context-conditioned model rather than the per-GOP-trained semantic model used in Section 3.3.

Context Frame Generation with RIFE

The context frame C_t required by the decoder is generated at both encoder and decoder using RIFE [72], a learned interpolation network. At each level of the hierarchy, F_t is the temporal midpoint between the two already reconstructed reference frames $F_{t-\tau}$ and $F_{t+\tau}$. RIFE therefore performs $2\times$ interpolation with a target interpolation fraction of 0.5:

$$C_t = f_{\text{RIFE}}(F_{t-\tau}, F_{t+\tau}). \quad (4.5)$$

Because C_t is generated from the same decoded reference frames, no per-frame motion field or context frame is transmitted: both encoder and decoder run the same fixed RIFE model on the same inputs. A codec implementation must standardise the model version and numerical inference procedure to keep both reconstructions consistent. This design removes per-frame motion and context signalling, unlike learned video codecs that entropy-code motion latents [49], but it adds RIFE model storage and inference complexity at both ends.

All autoencoder and RIFE operations are performed in the RGB colour space to avoid the precision losses introduced by repeated YUV420 conversions. Conversions to YUV420, required for VVC residual coding, are applied only at the input to VVenC and inverted immediately after VVC decoding, using FFmpeg [170] for colour space management.

Differential Latent Compression with DeepCABAC

A main efficiency mechanism of the proposed framework is the differential encoding of semantic latents. Let $l_t = \phi_{\theta_e}(Q_i(F_t))$ denote the latent of the target frame and $l_t^{(C)} = \phi_{\theta_e}(C_t)$ the latent derived from the RIFE context frame. The differential latent is defined as

$$\Delta l_t = l_t - l_t^{(C)}. \quad (4.6)$$

Because C_t is a high-quality approximation of F_t , the latent representations l_t and $l_t^{(C)}$ are strongly correlated. The differential Δl_t therefore has most of its values close to zero, as shown by the kernel density estimate in Fig. 4.3, yielding an entropy upper bound of

$$H(\Delta l_t) \leq H(l_t) - I(l_t; l_t^{(C)}), \quad (4.7)$$

where $I(\cdot; \cdot)$ is mutual information. This inequality shows the mutual-information gain obtained by transmitting Δl_t instead of l_t directly.

Before entropy coding, PCA [147] is applied to Δl_t to decorrelate its components and concentrate variance in a small number of principal components. The PCA-transformed differential latent is then entropy-coded with DeepCABAC [193], a neural-network-enhanced context-adaptive binary arithmetic coder from the Neural Network Compression and Representation (NNR) standard [193], using a uniform quantisation step size of 0.1. The decoded receiver reconstructs l_t by inverting the PCA transform and adding $l_t^{(C)}$.

Context-Conditioned Semantic Rate-Distortion

The semantic rate-distortion trade-off used by the proposed framework can be stated formally as follows. Given an input frame F_t and a side-information context frame C_t available at the decoder, the context-conditioned semantic rate-distortion function is

$$R_s(D_s) = \min_{\substack{p(\hat{F}_t|F_t,C_t): \\ \mathbb{E}[d_s(F_t,\hat{F}_t)] \leq D_s}} I(F_t; \hat{F}_t | C_t), \quad (4.8)$$

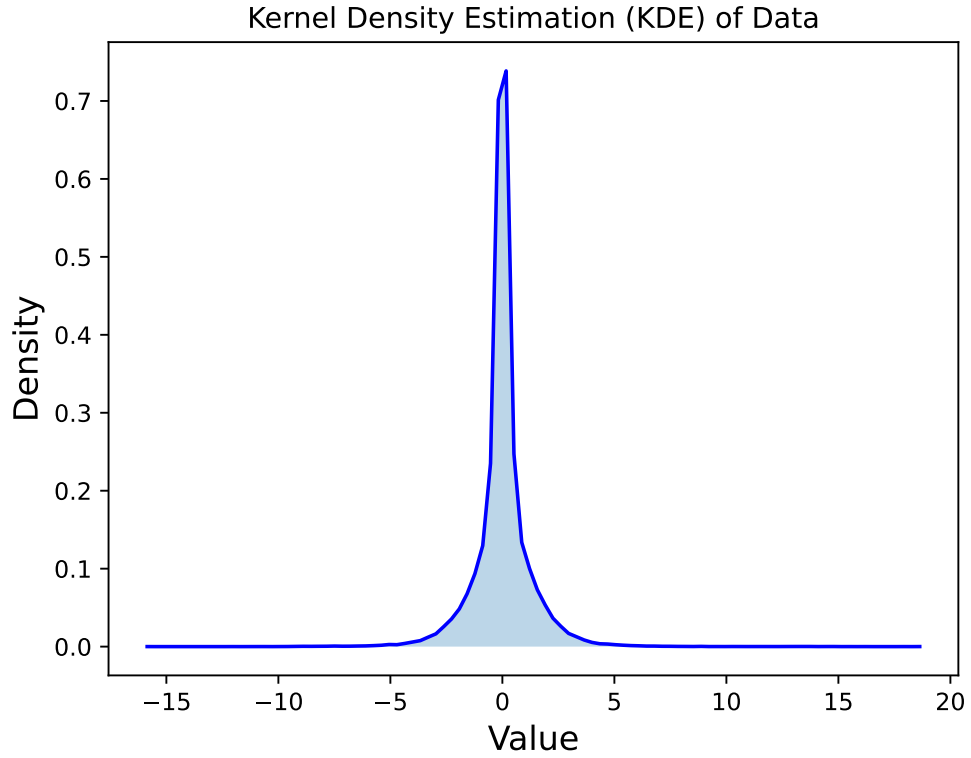


Figure 4.3: Kernel density estimate of the element-wise difference $\Delta l_t = l_t - l_t^{(C)}$ between the target-frame latent and the RIFE-predicted context-frame latent. The sharply peaked distribution centred near zero confirms the strong correlation between the two latent representations, motivating the differential encoding strategy.

where d_s is the per-frame semantic distortion measure and $I(\cdot; \cdot | \cdot)$ denotes conditional mutual information. By conditioning on C_t , the required mutual information $I(F_t; \hat{F}_t | C_t)$ is strictly smaller than the unconditional $I(F_t; \hat{F}_t)$ for any fixed distortion level, since C_t already provides substantial information about F_t . This theoretical view shows that the context conditioning mechanism has a clear information-theoretic basis, and the differential encoding of latents (Eq. 4.6) is a practical way to achieve this reduction. The end-to-end training objective minimises MSE over a batch of GOP sequences:

$$\mathcal{L} = \frac{1}{N} \sum_{k=0}^{N-1} (Q_i(F_{m+k}) - \hat{F}_{m+k})^2. \quad (4.9)$$

Training Procedure and Motion-Adaptive Model Selection

The autoencoder is trained on randomly sampled 144×256 -pixel RGB crops from the Vimeo-90k dataset [175] with a batch size of 8, using the Adam optimiser on an NVIDIA A100 GPU (Google Colab [183]). Training employs a three-phase learning-rate schedule: 1×10^{-4} for 75,000 iterations, 5×10^{-5} for a further 15,000 iterations, and 1×10^{-5} for 20,000 iterations, producing a base model at 110,000 total iterations.

To handle video content with different motion characteristics, a motion-adaptive model selection strategy is used. The TI of each sequence is computed over the first GOP using the standardised VQEG SITI tools [182]. A single base model handles low and medium-motion content ($\text{TI} \leq 15$), while a fine-tuned variant is obtained by continuing training for an additional 30,000 iterations on Vimeo-90k clips with higher frame-to-frame energy. At encoding time, the model is selected according to the rule:

$$\text{Model}(F) = \begin{cases} \text{fine-tuned model} & \text{if } \text{TI}(F) > 15, \\ \text{base model} & \text{otherwise.} \end{cases} \quad (4.10)$$

This binary selection adds no meaningful bitrate overhead, since the model identity can be conveyed as a single flag in the bitstream. Because only the decoder parameters are updated during fine-tuning and both models are pre-shared with the receiver, no model transmission is required at encoding time.

Block-Based Processing for Arbitrary Resolutions

The autoencoder is designed for 144×256 -pixel (144p) inputs. To support arbitrary resolutions including 1080p and 4K, input frames are partitioned into non-overlapping 144×256 -pixel blocks; the autoencoder processes each block independently, and the decoded blocks are then reassembled to form the full-resolution output. VVC in-loop deblocking filters applied during residual decoding suppress boundary artefacts at inter-block seams. The number of blocks per 1080p frame is approximately $(1920/256) \times (1080/144) \approx 56$, and for 4K the count increases to approximately 240 blocks; GPU parallelism reduces the effect of this increase on per-frame processing time.

4.2.3 Results and Discussion

Experimental Setup

Rate-distortion performance is evaluated on five standard benchmark datasets. Four 1080p datasets are assessed with respect to VTM 17.0 [184] as the anchor: the UVG dataset [173] (7 sequences), the MCL-JCV dataset [174] (30 sequences), HEVC Class B [172] (5 sequences, high-motion, diverse spatial complexity), and HEVC Class E [172] (4 sequences, video-conferencing content with low spatial information). A fifth evaluation uses Netflix Open Content 4K sequences [180], assessed with respect to VTM 23.0. VTM 17.0 is used for the four 1080p datasets to keep the anchor identical to the reproduced learned-codec baselines and the original CASVC study. VTM 23.0 is used for the later 4K evaluation because that experiment was run with the newer JVET anchor on the Netflix Open Content benchmark. The 1080p and 4K BD-rate results are therefore interpreted within their own anchor version rather than compared directly across VTM releases. The SI and TI profiles of these four 1080p datasets were introduced in Chapter 3 and are shown in Fig. 3.17. CASVC retains the same three complexity groups so that its content-dependent results can be interpreted consistently with the enhanced hybrid codec.

All sequences are encoded at four QP values $\{27, 32, 37, 42\}$ in random-access mode with intra period = -1 and a GOP size of 32. VVC residual encoding is performed by

VVenC [84] in expert mode. RIFE $2\times$ temporal interpolation is applied at each hierarchical prediction level. Rate-distortion performance is reported in terms of PSNR [171], SSIM [164], and VMAF [165]. Compression efficiency is quantified using the BD-rate metric [162, 189], computed across the four QP operating points. Negative BD-rate values indicate bitrate savings relative to the anchor at equivalent quality.

Comparative baselines include: VTM 17.0 (0%, anchor), HM-16.25 [186], ECM-11.0 [187] (a next-generation conventional codec), DCVC-DC [53, 99], DCVC-FM [52] and its half-precision variant (DCVC-FM fp16), and DCVC-RT (fp16) [26]. BD-rate figures for learned codecs are drawn from the reported values in the DCVC-RT publication [26].

Rate-Distortion Performance

Rate-distortion curves for all four 1080p datasets and the 4K evaluation are presented in Figs. 4.4–4.7.

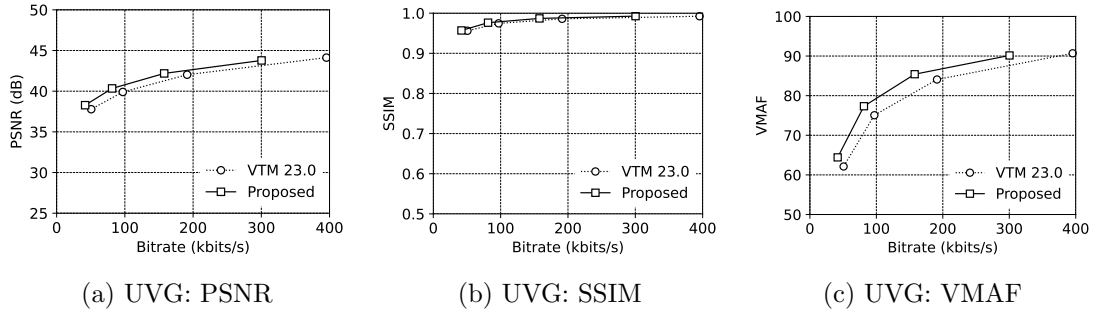


Figure 4.4: Average rate-distortion performance for the UVG dataset. The horizontal axis is bitrate (kbps) and the vertical axis is quality (PSNR in dB, SSIM, or VMAF). Each curve connects four operating points at $QP \in \{27, 32, 37, 42\}$.

The proposed framework achieves large and consistent PSNR BD-rate reductions across all evaluated datasets relative to VTM 17.0. For the UVG dataset, BD-rate reductions of -26.7% , -20.9% , and -19.3% are obtained for PSNR, SSIM, and VMAF, respectively. This strong performance is mainly due to the predictable, low-complexity motion in the UVG sequences, for which RIFE-based interpolation generates accurate context frames and semantic prediction leaves very little residual energy for VVC. For MCL-JCV, BD-rates of -19.7% , -17.5% , and -11.8% are achieved, reflecting

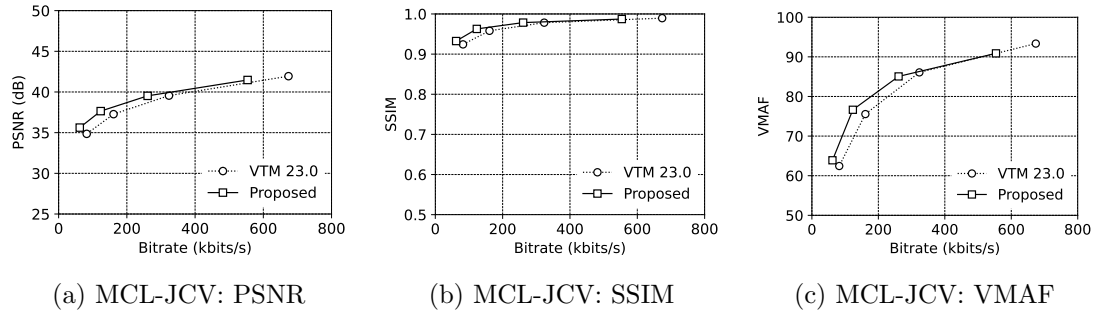


Figure 4.5: Average rate-distortion performance for the MCL-JCV dataset.

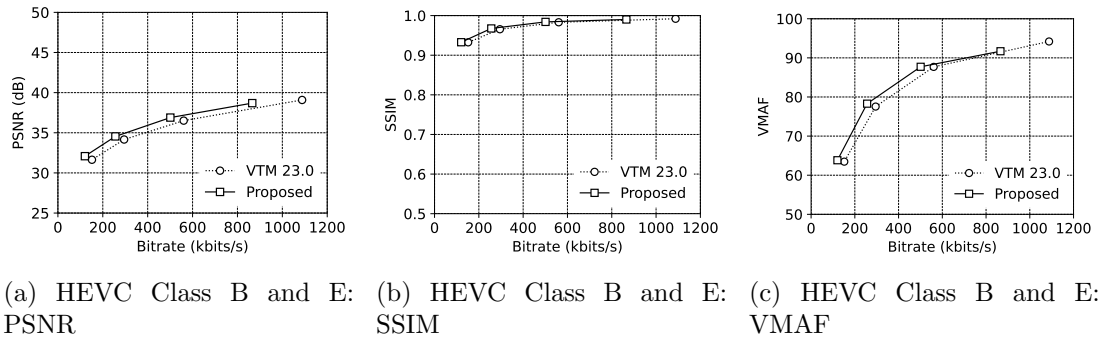


Figure 4.6: Average rate-distortion performance for the HEVC Class B and Class E datasets.

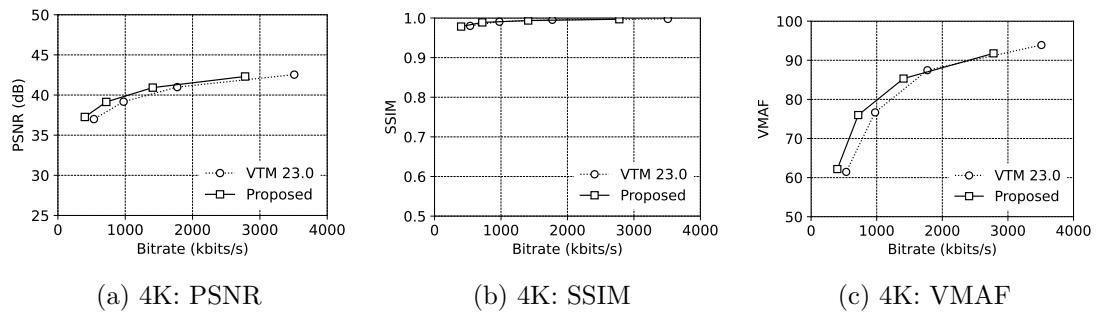


Figure 4.7: Average rate-distortion performance for 4K content from the Netflix Open Content dataset, assessed against VTM 23.0.

the greater content diversity of the 30-sequence dataset and the different difficulties in preserving pixel-level fidelity and perceptual quality. HEVC Class B yields BD-rates of -25.5% , -22.2% , and -21.6% , demonstrating robustness to spatially complex high-motion content. HEVC Class E, characterised by low spatial complexity but varied motion patterns, achieves -15.9% , -12.5% , and -10.9% ; the lower gains for this dataset reflect the relative advantage of VTM’s own temporal prediction tools for the videoconferencing-style content in this class. Averaged across the four 1080p datasets, the proposed framework achieves a PSNR BD-rate reduction of -22.03% relative to VTM 17.0. For 4K content evaluated against VTM 23.0, BD-rates of -22.6% , -21.1% , and -19.3% for PSNR, SSIM, and VMAF confirm that the block-based 144p processing approach scales well to ultra-high-definition resolutions.

Per-sequence variability within each dataset is reported in Table 4.1. Importantly, every individual test sequence achieves a negative PSNR BD-rate, confirming that the reported averages are not caused by aggregation effects and that the framework generalises across diverse content within each dataset.

Table 4.1: Summary statistics of per-sequence PSNR BD-rates (%) within each 1080p dataset, relative to VTM 17.0. Every sequence achieves a negative BD-rate.

Dataset	Mean (%)	Std. dev. (%)	Range (%)
UVG	-26.7	3.4	$[-29.2, -21.8]$
MCL-JCV	-19.7	5.6	$[-28.9, -9.1]$
HEVC Class B	-25.5	4.1	$[-29.1, -19.4]$
HEVC Class E	-15.9	3.7	$[-19.3, -12.1]$

Rate-distortion gains are most pronounced at medium-to-low quality operating points ($QP \in \{37, 42\}$), where the semantic prediction channel is the dominant contributor to the bitstream and VVC residual energy is small. At the highest quality point ($QP = 27$), the performance gap narrows as the VVC residual pathway dominates the total bitrate and the additional coding gain from semantic prediction diminishes relative to fine-grained pixel-level corrections. This QP-dependent behaviour supports the main design principle of the hybrid architecture: semantic communication provides its largest compression advantage in the perceptually relevant medium-quality range,

while VVC residual coding guarantees robustness across the full quality range.

Comparative Analysis with State-of-the-Art Codecs

Table 4.2 presents the full comparative BD-rate analysis with respect to all evaluated baselines.

Table 4.2: PSNR BD-rate (%) comparison in YUV420 colour space for 1080p datasets. Reference: VTM 17.0 (0%). Negative values indicate bitrate savings (lower is better). BD-rate figures for DCVC-DC, DCVC-FM, DCVC-FM fp16, and DCVC-RT fp16 are reproduced from [26]. Enc. fps and Dec. fps denote encoding and decoding throughput in frames per second on an NVIDIA A100 GPU.

Method	UVG	MCL-JCV	HEVC B	HEVC E	Average	Enc. fps	Dec. fps
VTM-17.0	0.0	0.0	0.0	0.0	0.0	0.01	23.6
HM-16.25	40.1	46.6	41.0	42.8	43.1	0.05	39.6
ECM-11.0	-20.0	-22.1	-22.2	-17.2	-20.3	0.002	3.4
DCVC-DC	6.5	-4.4	13.1	90.2	26.4	3.3	4.3
DCVC-FM	-17.6	-8.4	-15.7	-23.0	-16.1	3.4	4.2
DCVC-FM (fp16)	-16.8	-8.0	-15.4	-20.2	-15.1	5.0	8.9
DCVC-RT (fp16)	-24.0	-14.8	-16.6	-22.4	-19.4	125.2	112.8
Proposed	-26.7	-19.7	-25.5	-15.9	-22.0	1.23	19.2

The proposed framework achieves the lowest average PSNR BD-rate (-22.0%) among all evaluated codecs. Against the state-of-the-art learned codec DCVC-FM, the proposed system improves BD-rate performance by approximately 6 percentage points (-22.0% versus -16.1%). Even against DCVC-RT (fp16), which is specifically designed for real-time encoding and achieves -19.4% average BD-rate through architectural simplifications and half-precision arithmetic, the proposed framework maintains a 2.6-percentage-point advantage. DCVC-DC exhibits highly inconsistent performance, with positive BD-rates of $+6.5\%$ and $+13.1\%$ on the UVG and HEVC Class B datasets, respectively, and a severe degradation to $+90.2\%$ on HEVC Class E. This inconsistency highlights the challenges that purely learned codecs face on diverse, high-complexity content. By contrast, the proposed hybrid architecture maintains negative BD-rates on every evaluated dataset and every individual sequence, showing much more consistent generalisation across content types.

ECM-11.0 achieves an average BD-rate of -20.3% , representing the current capa-

bility frontier of conventional hybrid coding through exhaustive algorithmic refinement. The proposed framework surpasses ECM-11.0 by 1.7 percentage points on average despite operating under fundamentally different principles, and does so at substantially greater encoding speed (1.23 fps versus ECM-11.0’s 0.002 fps). This comparison is particularly significant because it demonstrates that integrating semantic communication principles with hierarchical bidirectional prediction yields tangible rate–distortion advantages beyond what conventional coding tool optimisation alone can achieve.

On HEVC Class E, the proposed system achieves -15.9% , which is less favourable than DCVC-FM (-23.0%) and DCVC-RT (-22.4%). This observation is attributed to the compact, low-texture nature of HEVC Class E videoconferencing content, for which VTM’s own intra and inter prediction tools are already highly efficient, leaving less margin for the semantic prediction channel to contribute. The same characteristic content also explains why the base model rather than the fine-tuned model is selected for HEVC Class E sequences, which uniformly satisfy $TI \leq 15$.

Computational Complexity

The proposed system achieves an encoding throughput of 1.23 fps and a decoding throughput of 19.2 fps for 1080p content on an NVIDIA A100 GPU. While encoding is not real-time, this represents a 123-fold improvement over VTM 17.0 (0.01 fps), mainly due to the elimination of VVC’s exhaustive rate–distortion mode-decision search for most frames: since semantic prediction greatly reduces residual energy, VVC converges more rapidly when coding residuals than when coding raw source frames. The encoding complexity can be decomposed as: autoencoder inference (60.41 GFLOPs per 144p block; 8.19 GFLOPs encoder, 52.22 GFLOPs decoder), RIFE interpolation, differential latent PCA and DeepCABAC encoding, and VVenC residual coding. For a 1080p frame partitioned into approximately 56 blocks, semantic inference alone requires approximately 3,383 GFLOPs per frame.

The decoding throughput of 19.2 fps approaches near-real-time playback and is substantially faster than all evaluated learned codecs: DCVC-DC (4.3 fps), DCVC-FM (4.2 fps), and DCVC-FM fp16 (8.9 fps) by factors of $4.5\times$, $4.6\times$, and $2.2\times$, respec-

tively. Decoding is faster than encoding because it executes a deterministic sequence of operations (DeepCABAC decoding, PCA inverse transform, autoencoder inference, VVC residual decoding) without mode-decision search or rate-distortion optimisation. ECM-11.0, despite achieving competitive BD-rate performance, decodes at only 3.4 fps, which is $5.6\times$ slower than the proposed system. The encoder-decoder complexity ratio of approximately 1:6.4 reflects the fundamental asymmetry of learned compression: the encoder prioritises compact latent extraction while the decoder emphasises high-fidelity reconstruction from compressed representations.

It is important to place the comparison with DCVC-RT in its proper hardware context. DCVC-RT achieves 125.2 fps encoding through a combination of half-precision (fp16) inference and cuDNN-optimised kernel fusion running on dedicated high-performance GPU hardware. All experiments in this thesis were conducted on an NVIDIA A100 GPU accessed through Google Colaboratory [183], a shared cloud research environment that does not provide the same dedicated GPU resources, persistent memory allocation, or cuDNN kernel optimisations used in the DCVC-RT benchmarking setup. On equivalent dedicated hardware with the same cuDNN fp16 optimisations, the proposed framework would be expected to achieve substantially higher encoding throughput, as its per-frame computational graph consists of standard feed-forward convolutional inference without the recurrent optical-flow estimation and iterative motion refinement steps that dominate DCVC-RT’s encoding pipeline.

Ablation Study

Component-wise contribution. Table 4.3 isolates the contribution of RIFE-based context generation to the overall framework performance. Without RIFE, the system achieves an average PSNR BD-rate of -13.04% relative to VTM 17.0, confirming that learned semantic encoding alone provides substantial compression gains over conventional inter prediction. Incorporating RIFE improves the average BD-rate to -22.03% , representing an additional gain of approximately 9 percentage points. This synergistic interaction demonstrates that RIFE’s pixel-accurate motion alignment enables the semantic decoder to generate substantially more accurate intermediate-frame recon-

structions, reducing the residual energy that VVC must encode.

Table 4.3: Component-wise ablation: average PSNR BD-rate (%) relative to VTM 17.0. Negative values indicate bitrate savings.

Configuration	Average BD-rate (%)
VTM 17.0 (anchor)	0.0
Semantic + VVC residual (no RIFE)	−13.04
Proposed (RIFE + Semantic + VVC residual)	−22.03

Bitrate composition. Table 4.4 decomposes the total bitrate across QP values. Semantic latents account for only 0.3 to 0.4% of the total bitrate throughout the full QP range, validating the efficiency of the differential encoding strategy. Keyframes dominate the bitrate budget, comprising 65.3% at QP 27 and increasing to 94.1% at QP 42 as residual energy is progressively suppressed by more aggressive quantisation. Residuals contribute 34.3% at QP 27 and decline to 5.6% at QP 42. This composition confirms that the semantic channel overhead is negligible and that the hybrid architecture’s compression gains are achieved without imposing meaningful additional bitrate from the semantic pathway.

Table 4.4: Bitrate composition: percentage of total bitrate contributed by each coding component at different QP values.

QP	Total (kbps)	Key (%)	Latents (%)	Residuals (%)
27	933.5	65.3	0.4	34.3
32	389.3	85.4	0.4	14.2
37	198.2	90.0	0.3	9.7
42	119.0	94.1	0.3	5.6

Motion-adaptive model selection. Table 4.5 presents a sensitivity analysis for the TI threshold used in model selection. For sequences with $TI \leq 15$, the fine-tuned model provides negligible benefit (average -0.3 dB relative to the base model), confirming that the base model is optimal for low- and medium-motion content. For sequences with $TI \in (15, 20]$, the fine-tuned model yields a gain of $+3.9$ dB in semantic PSNR, and for $TI > 20$, the gain increases to $+4.7$ dB. The threshold $TI = 15$ therefore identifies the

point at which fine-tuning changes from slightly harmful to clearly beneficial, providing an effective balance between gains on high-motion content and preservation of quality on low-motion content. Selecting a lower threshold (e.g., $\text{TI} > 10$) would apply the fine-tuned model to medium-motion content, incurring a -0.3 dB penalty without compensating benefit. Selecting a higher threshold (e.g., $\text{TI} > 20$) would forgo the $+3.9$ dB gain on sequences in the $\text{TI} \in (15, 20]$ range.

Table 4.5: Sensitivity analysis of the TI threshold for motion-adaptive model selection. Values are semantic PSNR (dB). Bold entries indicate the selected model for each content category.

TI Category	TI Range	Base (dB)	Fine-tuned (dB)	Δ (dB)
Low	≤ 10	34.2	33.9	-0.3
Medium	$(10, 15]$	33.8	33.5	-0.3
Medium-High	$(15, 20]$	28.5	32.4	$+3.9$
High	> 20	26.8	31.5	$+4.7$

Architectural ablations. Table 4.6 evaluates the sensitivity of semantic PSNR to the two primary architectural choices: latent channel dimensionality and spatial block resolution. Increasing the latent from 16 to 32 channels yields a gain of $+6.6$ dB (from 27.2 dB to 33.8 dB), confirming the importance of sufficient representational capacity. A further increase to 64 channels provides only $+0.4$ dB additional gain (34.2 dB) while doubling the latent bitrate, yielding strongly diminishing returns; the 32-channel configuration therefore provides the optimal capacity-efficiency trade-off. For spatial block resolution, 96p blocks achieve 29.1 dB, insufficient spatial context for reliable scene modelling. 288p blocks achieve 34.5 dB, a gain of $+0.7$ dB over the 144p configuration but at substantially higher training memory, per-iteration computational cost, and reduced spatial adaptability; the 144p block size therefore represents the optimal balance.

Semantic-only versus hybrid reconstruction. Semantic decoding without VVC residual correction achieves 28–35 dB PSNR and SSIM 0.82–0.91 across test sequences, sufficient for many perceptual applications. However, achieving the 38–42 dB PSNR

Table 4.6: Architecture ablation: semantic PSNR (dB) as a function of latent dimensionality and spatial block resolution. Bold entries indicate the proposed configuration.

Configuration	Semantic PSNR (dB)
<i>Latent channel dimensionality</i>	
$9 \times 16 \times 16$ channels	27.2
$9 \times 16 \times 32$ channels (proposed)	33.8
$9 \times 16 \times 64$ channels	34.2
<i>Spatial block resolution</i>	
96p (128×96 pixels)	29.1
144p (256×144 pixels, proposed)	33.8
288p (512×288 pixels)	34.5

range associated with broadcast-quality coding would require substantially deeper autoencoders with orders-of-magnitude higher latent dimensionality, incurring semantic bitrate contributions of 15–25% of the total bitrate and negating the compression advantages of the semantic pathway. Furthermore, overfitting to training-set statistics becomes severe at such high reconstruction targets, degrading generalisation to out-of-distribution content. VVC residual coding handles the pixel-level correction task far more efficiently than expanding latent dimensionality, accounting for only 5.6–34.3% of total bitrate across the QP range. The hybrid architecture therefore achieves a clear division of work: the semantic encoder learns compact perceptual representations, while VVC encodes the statistically simpler, spatially localised residual signal with its mature transform and entropy coding pipeline.

Hierarchical residual layer contribution. Table 4.7 presents the incremental PSNR and bitrate contributions of each order of hierarchically predicted residual frames within a GOP of size 32, evaluated at QP 32. Adding first-order residuals yields the highest marginal efficiency, at 87.39 dB/Mbit, increasing PSNR from 28.47 dB (keyframe-only semantic reconstruction) to 32.26 dB at a cost of only 43.37 kbps. Second-order residuals add +3.65 dB at 52.31 dB/Mbit, and third-order residuals add +2.46 dB at 28.22 dB/Mbit. Fourth-order residuals provide only +1.61 dB at 16.78 dB/Mbit, indicating approaching marginal returns. The overall progression from

28.47 dB to 39.98 dB with total bitrate increasing from 389.21 kbps to 685.48 kbps confirms that each hierarchical level contributes positively to rate-distortion efficiency and that the adopted GOP-32 structure with four orders of residuals is well-justified.

Table 4.7: Incremental contribution of hierarchical residual layers in a GOP of size 32, averaged across all 1080p test sequences at QP 32. ΔPSNR and $\Delta\text{Bitrate}$ denote marginal gains from each additional order; $\eta = \Delta\text{PSNR}/\Delta\text{Bitrate}$ is marginal coding efficiency (dB/Mbit).

Residual order	PSNR (dB)	SSIM	Bitrate (kbps)	ΔPSNR (dB)	$\Delta\text{Bitrate}$ (kbps)	η (dB/Mbit)
None (keyframe only)	28.47	0.784	389.21	N/A	N/A	N/A
+ 1st order	32.26	0.829	432.58	+3.79	43.37	87.39
+ 2nd order	35.91	0.883	502.34	+3.65	69.76	52.31
+ 3rd order	38.37	0.950	589.52	+2.46	87.18	28.22
+ 4th order	39.98	0.989	685.48	+1.61	95.96	16.78

Visual Quality Comparison

The progressive quality enhancement achieved through the three-stage hybrid architecture is illustrated in Fig. 4.8 for four representative test frames drawn from the Vimeo-90k evaluation set. Each row presents, from left to right, the RIFE-based bidirectional interpolation result, the context-conditioned semantic prediction output (SC predicted), and the final reconstruction after VVC residual correction.

The RIFE-interpolated predictions achieve PSNR values in the range 20.96 to 25.57 dB and SSIM in the range 0.74 to 0.82 across the four examples. While RIFE provides motion-aligned temporal context that is structurally coherent with the target frame, it operates without access to the compressed latent and cannot recover fine-grained texture, object boundaries, or high-frequency detail beyond what interpolation can estimate from the surrounding reference frames. The variability in RIFE quality across examples (spanning approximately 4.6 dB) reflects the sensitivity of learned interpolation to motion complexity: example 3 (Row 3, 25.57 dB) exhibits moderate translational motion for which the bidirectional optical flow estimation is reliable, whereas examples 1 and 2 (Rows 1 and 2, 21.13 dB and 20.96 dB) involve more complex scene dynamics that introduce interpolation artefacts.

The context-conditioned semantic prediction (SC predicted) achieves PSNR values



Figure 4.8: Visual comparison of progressive reconstruction quality across four representative test frames (rows). Each column shows: (left) the RIFE-based bidirectionally interpolated prediction, (centre) the context-conditioned semantic communication (SC) prediction from the autoencoder, and (right) the final reconstruction after VVC residual correction. PSNR and SSIM values are reported per frame. Quality improves monotonically from left to right across all examples, validating the complementary roles of motion-aligned interpolation, semantic encoding, and residual correction.

of 29.41 to 30.42 dB and SSIM of 0.82 to 0.88, representing an improvement of 4.84 to 9.46 dB over the corresponding RIFE baseline. Critically, the semantic PSNR is substantially more stable across examples (range < 1 dB) than the RIFE PSNR (range ≈ 4.6 dB), confirming that the autoencoder compensates for variable interpolation quality through its learned latent representations. When RIFE produces a low-quality interpolation (e.g., Row 2, 20.96 dB), the autoencoder’s semantic encoding of the target frame provides the additional information necessary to reconstruct a substantially higher-quality output (30.42 dB), using the RIFE prediction only as a context reference rather than a direct prediction signal. The SSIM gains (0.82 to 0.88) confirm that the semantic encoder preserves structural and luminance attributes simultaneously, not solely optimising for MSE-based PSNR.

VVC residual correction elevates reconstruction quality to PSNR values of 38.01 to 43.14 dB and SSIM of 0.98 to 0.99, which represent broadcast-quality reconstruction in all cases. The final quality is both substantially higher and markedly more consistent across examples (range ≈ 5.1 dB, versus 4.6 dB for RIFE and 1.0 dB for SC) than either upstream stage, demonstrating that the VVC residual pathway reliably converges to high fidelity regardless of the quality of the upstream semantic prediction. This convergence behaviour validates the principal design rationale of the hybrid architecture: the semantic channel provides a strong perceptual prediction that drastically reduces residual energy, while the VVC residual channel provides a deterministic correction mechanism that guarantees broadcast-quality fidelity independent of semantic channel uncertainty.

The visual progression is also qualitatively informative. In the RIFE-interpolated column, motion-blur artefacts and ghosting are visible at object boundaries where the interpolated optical flow introduces blending between foreground and background regions. The SC-predicted column eliminates much of this ghosting, producing sharper object boundaries and more coherent texture, attributable to the semantic autoencoder’s learned representation of scene structure. The VVC-corrected column recovers fine-grained high-frequency detail that the semantic predictor smooths over, restoring textures and sharp edges to near-lossless fidelity. These complementary failure modes

of the individual stages confirm that the three components are non-redundant: RIFE provides motion alignment, semantic encoding provides structural coherence, and VVC provides pixel-level precision.

Discussion

The experimental results collectively support the core claim of this contribution: integrating a pretrained, motion-adaptive semantic autoencoder with RIFE-based bidirectional context generation and VVC residual coding within a hierarchical prediction structure achieves rate-distortion performance superior to both state-of-the-art conventional codecs (including the next-generation ECM-11.0) and the leading purely learned video codecs (DCVC-FM, DCVC-RT) on standard 1080p and 4K benchmarks.

Several observations merit emphasis. First, the negligible latent overhead (0.3–0.4% of total bitrate) demonstrates that the differential encoding strategy combined with DeepCABAC effectively exploits the strong correlation between the target-frame latent and the RIFE-predicted context-frame latent, as visualised in the kernel density estimate (Fig. 4.3). Second, the multi-metric consistency of the gains (PSNR, SSIM, and VMAF all improve) confirms that the semantic prediction mechanism preserves luminance, structural, and perceptual quality attributes simultaneously, rather than optimising a single metric at the expense of others. Third, the 4K evaluation confirms that the block-based architecture scales without architectural modification, with deblocking artefacts at block boundaries effectively suppressed by the VVC in-loop filtering pipeline. Fourth, the comparison with DCVC-DC illustrates a practical limitation of purely end-to-end learned codecs: the +90.2% BD-rate on HEVC Class E demonstrates that learned codecs can fail severely on content outside their effective training distribution, whereas the proposed hybrid architecture maintains consistent negative BD-rates across all evaluated content types.

The framework is not without limitations. Encoding throughput (1.23 fps for 1080p) precludes real-time applications such as live streaming and video conferencing; future work targeting real-time deployment would benefit from lightweight semantic encoder architectures, model quantisation, and hardware-accelerated DeepCABAC implemen-

tations. Performance is subject to variation across datasets, with HEVC Class E representing a content class on which the proposed approach is comparatively less advantageous than learned codecs optimised for low-motion scenarios. Content substantially outside the Vimeo-90k training distribution (for example, animation, screen content, or synthetic computer graphics) may exhibit reduced semantic prediction accuracy, and the 144p block-boundary overhead becomes proportionally more significant at 8K and higher resolutions. VVC patent licensing constraints also impose commercial deployment considerations that do not apply to open codec alternatives.

4.2.4 Conclusion

The context-conditioned semantic video coding framework presented in this section shows that per-GOP retraining can be fully removed without losing rate-distortion performance. Tested under the JVET random-access common test conditions relative to VTM 17.0, the framework achieves an average PSNR Bjøntegaard delta-rate of -22.03% across the four standard 1080p benchmarks: UVG (-26.7%), MCL-JCV (-19.7%), HEVC Class B (-25.5%), and HEVC Class E (-15.9%), and -22.6% relative to VTM 23.0 on 4K Netflix Open Content sequences. Every individual test sequence across all four 1080p benchmarks achieves a negative PSNR BD-rate, confirming that the gains are consistent across the full test suite and are not a result of dataset-level averaging. These results are better than the next-generation conventional codec ECM-11.0 (-20.3%) and all evaluated learned video codecs, including DCVC-FM (-16.1%) and DCVC-RT fp16 (-19.4%), on average across the evaluated benchmarks.

The ablation study shows that RIFE-based context conditioning accounts for approximately nine percentage points of the overall BD-rate gain relative to the baseline without motion-aligned context, confirming that motion-aligned decoder conditioning and compact semantic encoding work well together. Differential PCA-DeepCABAC entropy coding keeps the semantic latent sub-stream to 0.3–0.4% of total bitrate across the full operating range, confirming that the semantic channel adds negligible overhead. The motion-adaptive model selection strategy provides a further improvement over the base model for high-TI sequences, confirming that a single fine-tuning step handles

higher temporal variability without per-sequence fine-tuning. Encoding and decoding speeds of 1.23 fps and 19.2 fps on an NVIDIA A100 GPU represent a $123\times$ speedup over VTM 17.0 while approaching near-real-time decoding, confirming that the system is practical for offline-encode, online-decode use. This contribution directly addresses Gap 2 identified in Chapter 2: the lack of a shared pretrained foundation model for context-conditioned semantic video coding.

Despite these results, one limitation of the framework is that the decoder is fixed at pretraining time and represents an average over the Vimeo-90k training data. For GOP content whose statistics differ from this training data, the fixed decoder introduces prediction error that the VVC residual pathway must absorb at additional bitrate cost. This limitation is most noticeable for content types such as CCTV surveillance footage and 3D computer animation, and is the main motivation for the content-adaptive parameter overfitting mechanism presented in Section 4.3.

4.3 Content-Adaptive Parameter Overfitting for Per-GOP Specialisation

4.3.1 Introduction

As established in the Chapter Introduction, the CASVC framework of Section 4.2 removes per-GOP retraining, but its fixed decoder represents an average over the Vimeo-90k training data [175] and therefore produces higher prediction error for any GOP whose content differs from that data, increasing residual bitrate without a matching improvement in semantic quality. This is Gap 3 identified in Chapter 2: the lack of a practical per-GOP content adaptation mechanism within a fully pretrained hybrid semantic codec.

This section presents the CAPO mechanism, which brings back per-GOP content adaptation into the CASVC framework while keeping all convolutional kernels, normalisation parameters, and the entire encoder fixed. The key idea is that updating only per-channel Learnable Scale multipliers and transposed-convolution biases, totalling approximately 643 scalars, is enough to adapt the decoder from the global Vimeo-

90k average to the local content of each GOP, matching the per-GOP adaptivity of full retraining at a much lower cost. For each GOP, these parameters are found by an online Adam optimisation using the fixed RIFE context frames as inputs, and the resulting update is transmitted as a PCA–DeepCABAC parameters sub-stream alongside the differential semantic latents, keeping total parameter overhead below 1% of bitrate. A further architectural contribution is a tapered encoder with channel widths $32 \rightarrow 64 \rightarrow 96 \rightarrow 128$ and a mirrored decoder with a soft-quantisation bottleneck, which reduces per-block processing cost by approximately $4.9\times$ relative to CASVC while maintaining or improving rate–distortion performance.

The methodology, covering the tapered autoencoder architecture, Learnable Scale modulation, per-GOP overfitting, and joint PCA–DeepCABAC coding of latents and parameters, is detailed in Section 4.3.2. Experimental results for the combined framework on the four standard 1080p benchmarks, a 4K dataset, and out-of-distribution content including CCTV surveillance and 3D animation, together with ablation studies and runtime analysis, are presented in Section 4.3.3.

4.3.2 Methodology

System Overview

The proposed framework extends CASVC [73] by reintroducing per-GOP content adaptivity into an otherwise fully deployable pretrained codec. The proposed solution is a *content-adaptive parameter overfitting* mechanism: for every GOP, only a tiny subset of decoder parameters, comprising per-channel Learnable Scale multipliers and transposed-convolution biases and totalling approximately 643 scalars, is overfitted online to the current GOP’s content and transmitted to the receiver as an additional parameter sub-stream. All convolutional kernels, normalisation parameters, the entire encoder, and the RIFE-based context-generation pipeline remain frozen and identical at both endpoints.

As illustrated in Fig. 4.9, the transmitter partitions each input video in YUV420 into 144×256 -pixel blocks, intra-codes GOP boundary frames with VVenC [84], and for each subsequent intermediate frame: (i) generates a motion-aligned context frame

using bidirectional RIFE interpolation [72]; (ii) encodes the frame semantically with the frozen autoencoder; (iii) performs a pre-overfitting prediction pass with $\Delta_g = 0$ to obtain the frozen contexts $\tilde{C}_{m+k}$ used in the overfitting solve; (iv) optimises the 643 Learnable Scale and bias parameters against the locally available reconstruction targets; and (v) entropy-codes the resulting parameter update Δ_g^* alongside the differential latents and VVC residual. The resulting bitstream has four distinct sub-streams:

$$R_{\text{total}} = R_{\text{key}} + R_{\text{latent}} + R_{\text{param}} + R_{\text{residual}}, \quad (4.11)$$

where R_{key} , R_{latent} , R_{param} , and R_{residual} denote the key-frame, differential-latent, parameter-update, and VVC-residual bitrates, respectively, with $R_{\text{param}}, R_{\text{latent}} \ll R_{\text{key}}, R_{\text{residual}}$ by design. At the receiver, the decoder reproduces the frozen contexts $\tilde{C}_{m+k}$ via a pre-overfitting forward pass, applies the decoded Δ_g^* to instantiate the GOP-specialised semantic decoder, and adds the dequantised VVC residual to obtain the full-quality output.

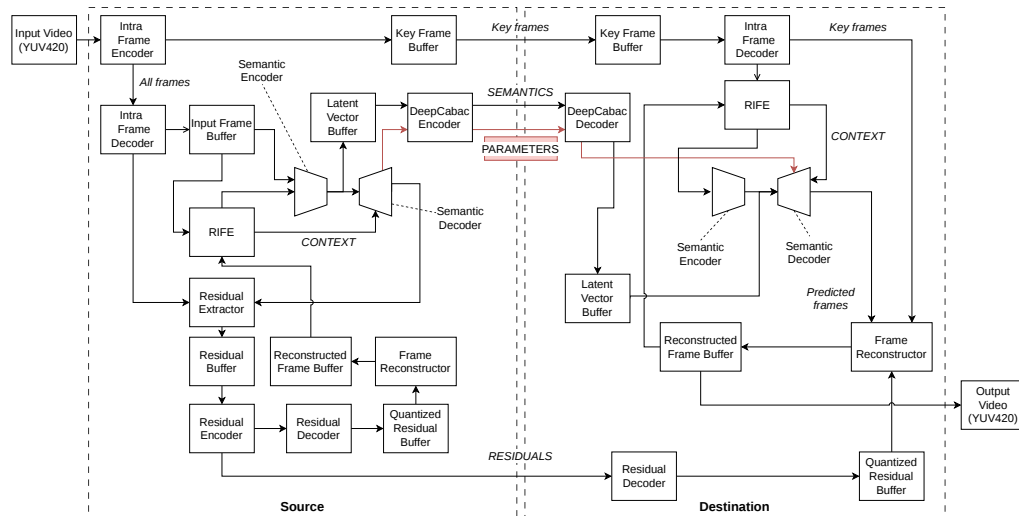


Figure 4.9: End-to-end architecture of the proposed context-aware semantic video codec with content-adaptive parameter overfitting.

Tapered Semantic Autoencoder

The semantic autoencoder is the core computational module, responsible for extracting compact latent representations and reconstructing them conditioned on the RIFE-interpolated context and the per-GOP parameter update. The network architecture is illustrated in Fig. 4.10 and departs from the uniform-width design of CASVC which employed 128 filters at every encoder and decoder stage in two essential respects: the channel widths are tapered, and per-channel Learnable Scale layers are inserted into the decoder upsampling path.

The encoder maps an input RGB block $F_t \in \mathbb{R}^{144 \times 256 \times 3}$ to a compact latent $h_t \in \mathbb{R}^{9 \times 16 \times 32}$ through four successive stride-2 3×3 convolutions with a *tapered* channel configuration of $32 \rightarrow 64 \rightarrow 96 \rightarrow 128$ filters, each followed by batch normalisation [191], LeakyReLU activation ($\alpha = 0.01$), and a residual block [51, 192]. The tapered design progressively widens the channel dimension to reflect the increasing abstraction of semantic features at coarser spatial scales, while keeping the computationally expensive early stages which operate at the full 144×256 resolution narrow. A fifth stride-1 convolution reduces the channel count from 128 to 32, followed by a *soft-quantisation layer* that injects zero-mean uniform noise on $[-0.5, 0.5]$ during training and performs hard rounding at inference; this quantisation-aware design removes the train-test distribution mismatch present in CASVC. The four downsampling stages collectively reduce spatial dimensions from 144×256 to 9×16 , yielding a latent of $9 \times 16 \times 32 = 4,608$ values per block (compression ratio approximately 24:1 relative to the $144 \times 256 \times 3$ input). The encoder is expressed formally as

$$h_t = \phi_{\theta_e}(Q_i(F_t)), \quad (4.12)$$

where ϕ_{θ_e} is the frozen encoder parameterised by weights θ_e and $Q_i(\cdot)$ is the intra-frame quantisation operator.

The decoder inverts this pathway through four transposed-convolution upsampling stages with channel widths that mirror the encoder taper: $128 \rightarrow 96 \rightarrow 64 \rightarrow 32$. At each stage $\ell \in \{1, 2, 3, 4\}$, the 3×3 transposed convolution (Conv2DT) with stride 2

is immediately followed by a per-channel *Learnable Scale* (LS) layer that performs the element-wise affine modulation

$$y = m^{(\ell)} \odot \text{Conv2DT}(x) + b^{(\ell)}, \quad (4.13)$$

where $m^{(\ell)} \in \mathbb{R}^{C^{(\ell)}}$ is the per-channel scale-multiplier vector, $b^{(\ell)} \in \mathbb{R}^{C^{(\ell)}}$ is the transposed-convolution bias, and $C^{(\ell)} \in \{128, 96, 64, 32\}$ is the channel count of stage ℓ . The scale multipliers $m^{(\ell)}$ are initialised to all-ones and the biases to zero, so the Learnable Scale layers behave as identities at the start of training; once trained, each layer provides a compact, per-channel knob through which the decoder can rescale the activations of the corresponding transposed convolution. Batch normalisation and LeakyReLU follow each Learnable Scale layer, then a stage-width residual block. The context-conditioned reconstruction is expressed as

$$\hat{F}_t = \psi_{\theta_d^{\text{frozen}}, \theta_d^{\text{base}} + \Delta_g}(h_t, C_t), \quad (4.14)$$

where θ_d^{frozen} collects all frozen kernels and normalisation parameters, and $\theta_d^{\text{base}} + \Delta_g$ applies the per-GOP update on top of the pretrained Learnable Scale and bias values. The final decoder stage concatenates the upsampled feature map at $144 \times 256 \times 32$ with the RIFE-interpolated context $C_t \in \mathbb{R}^{144 \times 256 \times 3}$, processes the result through a 64-filter residual block, applies Dropout (rate $p = 0.2$), and produces the reconstructed RGB block via a 3×3 output convolution with sigmoid activation.

This taper reduces computation most strongly at the highest-resolution decoder stages and also limits the number of Learnable Scale parameters transmitted per GOP. Section 4.3.3 quantifies the resulting encoder, decoder, and total FLOP reductions against CASVC so that the architecture and complexity results are reported only once.

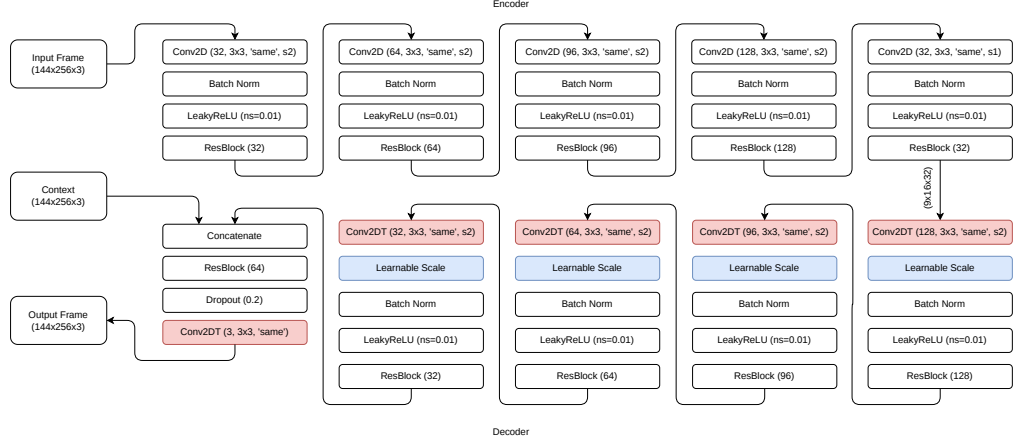


Figure 4.10: Network architecture of the proposed tapered semantic autoencoder. Red-highlighted blocks indicate the overfittable transposed-convolution bias parameters; blue-highlighted blocks indicate the per-channel Learnable Scale multipliers.

Content-Adaptive Parameter Overfitting

For each GOP g starting at frame index m with length N , the encoder solves the per-GOP overfitting objective

$$\Delta_g^* = \arg \min_{\Delta_g} \frac{1}{N} \sum_{k=0}^{N-1} \|Q_i(F_{m+k}) - \hat{F}_{m+k}(\Delta_g; \tilde{C}_{m+k})\|_2^2 + \beta \|\Delta_g\|_2^2, \quad (4.15)$$

where $\hat{F}_{m+k}(\Delta_g; \tilde{C}_{m+k})$ is the context-conditioned reconstruction obtained by patching $\theta_d^{\text{base}} + \Delta_g$ into the frozen decoder, and $\beta = 1 \times 10^{-4}$ is a small ℓ_2 regularisation weight that discourages large parameter excursions and promotes sparsity in the entropy-coded representation.

The overfittable parameter set Δ_g is restricted to two compact families:

1. the per-channel Learnable Scale multipliers $\{m^{(\ell)} \in \mathbb{R}^{C^{(\ell)}}\}_{\ell=1}^4$ inserted after each of the four **Conv2DT** upsampling stages, with $C^{(\ell)} \in \{128, 96, 64, 32\}$; and
2. the bias vectors $\{b^{(\ell)} \in \mathbb{R}^{C^{(\ell)}}\}_{\ell=1}^4$ of the four corresponding **Conv2DT** layers, together with the bias of the final 3×3 output convolution (3 additional scalars).

The total overfittable parameter count per GOP is therefore $2 \times (128 + 96 + 64 + 32) + 3 =$

643 scalars, less than 0.1% of the full decoder and three to four orders of magnitude smaller than the autoencoder weight payload implicitly transmitted by the per-GOP from-scratch retraining of earlier systems [1, 22].

The choice of Learnable Scale multipliers and **Conv2DT** biases is deliberate: together they parameterise the affine modulation of (4.13) on top of an otherwise fixed feature dictionary, acting multiplicatively and additively on already-learned features. Updating full convolutional kernels would require $O(C_{\ell+1} \times C_\ell \times k^2)$ parameters per layer, totalling tens of thousands of values across the four upsampling stages, which would make per-GOP online optimisation impractical and the parameter sub-stream too large to remain a small fraction of the video bitrate. Channel-wise scale multipliers and biases are the smallest affine adapter for each decoder layer: they adjust the per-channel gain and offset of the decoded feature maps without modifying the spatial patterns encoded in the frozen convolutional kernels, which is sufficient to correct the main mismatch between the pretrained general model and the local statistics of each GOP. The two pipelines use *different* contexts depending on which decoder is running. In the base pass ($\Delta_g = 0$), the context is the RIFE-interpolated frame $C_t = f_{\text{RIFE}}(F'_{t-\tau}, F'_{t+\tau})$, identical to CASVC. In the overfitting pass, the context $\tilde{C}_{m+k}$ is the semantic prediction produced by the base decoder (i.e. the pre-overfitting reconstruction with $\Delta_g = 0$), computed once before optimisation begins and then kept frozen throughout. Because $\tilde{C}_{m+k}$ is produced deterministically by the shared pretrained baseline, both encoder and receiver can reproduce it without any additional transmission, guaranteeing perfect synchronisation.

The optimisation of (4.15) is initialised with $\Delta_g = 0$ and solved with Adam at learning rate 1×10^{-3} ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 1×10^{-4}) for approximately 60 to 100 epochs over the N GOP frames, with early stopping triggered when the per-GOP loss ceases to decrease for five consecutive epochs. Because only 643 scalars are updated and the contexts are precomputed and frozen, each epoch reduces to a single forward-backward pass through the decoder, keeping the per-GOP solve cost two orders of magnitude below the from-scratch retraining of [1, 22]. No iterative learning is performed at the receiver: the decoder simply decodes Δ_g from the parameters

sub-stream, patches it into the cached pretrained baseline, reconstructs the GOP, then reverts to the baseline before the next GOP.

Differential PCA–DeepCABAC Entropy Coding

The proposed framework retains the CASVC differential-latent definition in Eqn. (4.6) and its PCA–DeepCABAC pipeline, but extends that pipeline to the per-GOP parameter update Δ_g . Specifically, $l_t^{(C)}$ in Eqn. (4.6) is obtained by encoding the RIFE context F_t^{RIFE} with the same frozen encoder. The resulting Δl is projected by PCA [147] onto a reduced-dimensional subspace, quantised with a step size of 0.1, and entropy coded with DeepCABAC [148].

For the parameter update, the encoder forms the residual $\Delta\theta_g = \Delta_g^*$, which by construction is sharply peaked around zero owing to the ℓ_2 regularisation in (4.15). PCA is applied to $\Delta\theta_g$ using a basis derived offline from the per-GOP parameter residuals of the training sequences, concentrating the update energy and decorrelating the coefficients. The PCA-projected residuals are then quantised with a coarse uniform step selected to keep R_{param} below 1% of the GOP’s total bitrate and entropy coded with DeepCABAC. At the receiver, the inverse PCA transform recovers $\Delta\theta_g$, which is added to the cached pretrained values θ_d^{base} to instantiate the GOP-specialised decoder. This stateless per-GOP update protocol preserves random-access semantics and prevents error drift across GOPs.

The theoretical justification is provided by the conditional rate–distortion bound

$$I(F; \hat{F} \mid C, \Delta_g) \leq I(F; \hat{F} \mid C), \quad (4.16)$$

with strict inequality whenever Δ_g carries GOP-specific information beyond the RIFE context; this tightening of the achievable semantic rate–distortion lower bound is the theoretical motivation for incurring R_{param} .

Training Procedure

The semantic autoencoder is trained offline *once* on the Vimeo-90k dataset [175] using the context-conditioned mean-squared-error objective defined for CASVC in Section 4.2.2. The mini-batch size remains $B = 8$, and each training sample uses its RIFE-interpolated context $C^{(b)}$. Adam is employed with a progressive learning rate schedule: 1×10^{-4} for the first 75,000 iterations, 5×10^{-5} for the next 15,000 iterations, and 1×10^{-5} for the final 20,000 iterations; the soft-quantisation layer is active in noise-injection mode throughout to align latent statistics with the entropy-coded inference regime. Training converges within approximately 110,000 iterations on a single NVIDIA A100 GPU, and the resulting baseline model achieves 28–35 dB semantic-only PSNR, providing the high-quality starting point on which the per-GOP overfitting refinement operates.

At inference time, the only learning operation performed online is the per-GOP overfitting solve of (4.15) run at the encoder. The decoder requires no additional training and simply applies the decoded Δ_g to the shared pretrained baseline, maintaining full deployment parity with CASVC except for the small additional parameters sub-stream.

4.3.3 Results and Discussion

Experimental Setup

Evaluation is performed on six datasets spanning three distinct content classes. The first group comprises the four standard 1080p natural-content benchmarks evaluated against VTM 17.0: UVG (7 sequences) [173], MCL-JCV (30 sequences) [174], HEVC Class B (5 sequences) [172], and HEVC Class E (4 sequences) [172]. The second group consists of two out-of-distribution content classes evaluated against VTM 23.0: a CCTV surveillance set (3 sequences) with a static camera, sparse foreground motion, and stable background statistics, and a computer-generated 3D animation set (5 sequences) with sharp synthetic edges, saturated colour palettes, and rendered noise statistics. These two additional classes are chosen deliberately because they sit at opposite ends of the content-statistics spectrum relative to the Vimeo-90k training distribution, providing

a stringent test of the generalisation capability of the per-GOP overfitting mechanism.

Both the key-frame intra encoder and the residual encoder use VVenC 1.12.1 [84] in expert mode (`vvencFFapp`). The RIFE network [72] is configured for $2\times$ interpolation; the VVC residual is encoded with `IntraPeriod` = -1 (random access, first frame intra coded), a GOP size of 32, and the same QP value applied uniformly to both key frames and residuals. Four operating points $QP \in \{27, 32, 37, 42\}$ are evaluated for all methods. Rate-distortion quality is assessed using the same metrics and BD-rate methodology introduced in Section 4.2.3: PSNR, SSIM, and VMAF in YUV420 colour space, with BD-rates computed across the four QP operating points against the corresponding VTM anchor; negative values indicate bitrate savings at matched quality. All autoencoder and per-GOP overfitting operations are performed on a single NVIDIA A100 GPU. BD-rate figures for DCVC-DC, DCVC-FM, DCVC-FM (fp16), and DCVC-RT (fp16) are reproduced from the unified benchmarking study in [26], which uses an identical experimental protocol (same four datasets, same VTM 17.0 anchor, same QP grid).

Rate-Distortion Performance

The average rate-distortion curves for all six datasets are presented in Figs. 4.11–4.16. On every dataset and on every metric, the proposed RD curve lies strictly above the CASVC curve, which in turn lies above the VTM anchor, confirming the central hypothesis of this work: per-GOP content-adaptive parameter overfitting produces a consistent, measurable rate-distortion improvement on top of CASVC with no observed regression at any operating point.

The per-dataset PSNR BD-rate results are summarised in Table 4.8, SSIM BD-rates in Table 4.9, and VMAF BD-rates in Table 4.10, all relative to VTM 17.0 for the four natural-content datasets. BD-rate figures for the two additional content classes relative to VTM 23.0 are reported in Table 4.11.

Across the four 1080p natural-content datasets, the proposed codec achieves an average PSNR BD-rate of **−23.9%** relative to VTM 17.0, an average SSIM BD-rate of **−22.4%**, and an average VMAF BD-rate of **−21.4%**. The largest gains occur on

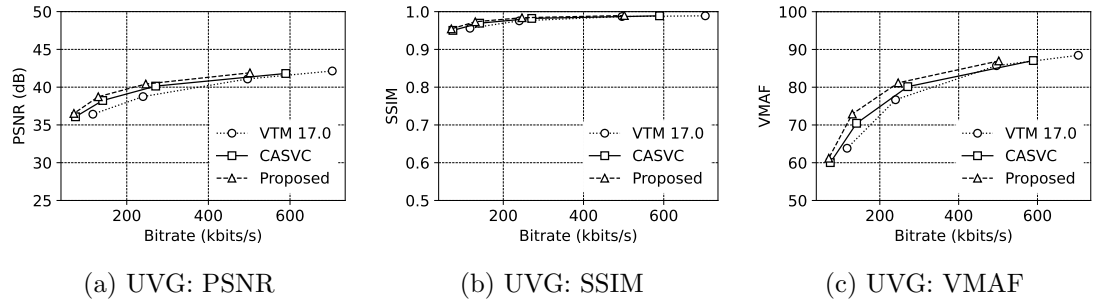


Figure 4.11: Average rate-distortion performance for the UVG dataset. The horizontal axis is bitrate (kbps) and the vertical axis is quality (PSNR in dB, SSIM, or VMAF). Each curve connects four operating points at $QP \in \{27, 32, 37, 42\}$.

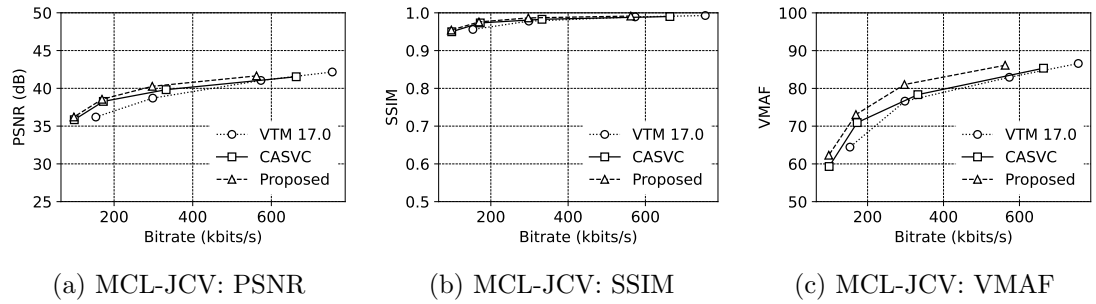


Figure 4.12: Average rate-distortion performance for the MCL-JCV dataset.

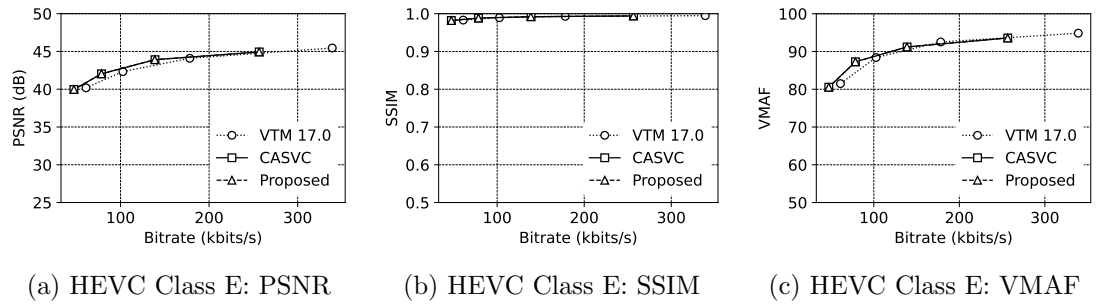


Figure 4.13: Average rate-distortion performance for the HEVC Class E dataset.

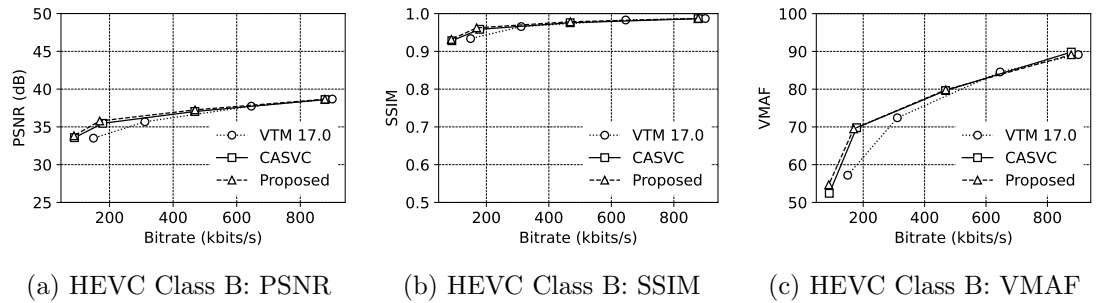


Figure 4.14: Average rate-distortion performance for the HEVC Class B dataset.

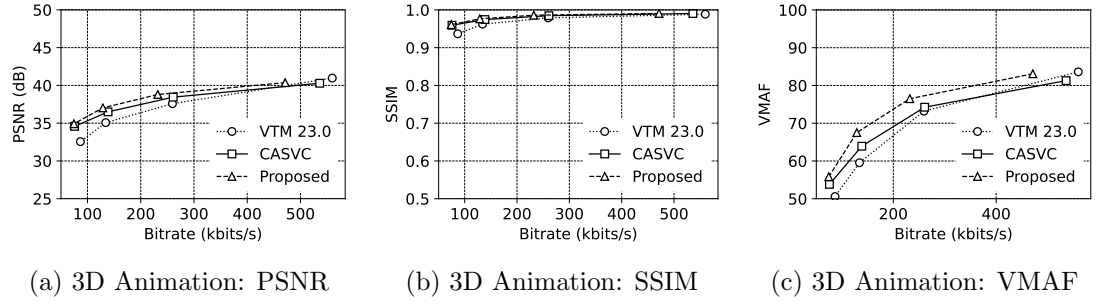


Figure 4.15: Average rate-distortion performance for the 3D animation dataset (average of 5 sequences), evaluated against VTM 23.0.

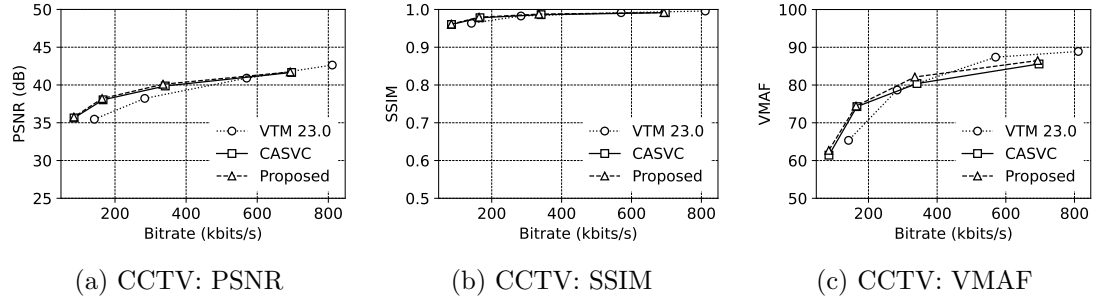


Figure 4.16: Average rate-distortion performance for the CCTV surveillance dataset (average of 3 sequences), evaluated against VTM 23.0.

Table 4.8: PSNR BD-rate (%) comparison in YUV420 colour space for all 1080p datasets. Reference: VTM 17.0 (0%). Encoding and decoding throughput in frames per second (fps) on an NVIDIA A100 GPU. BD-rate figures for DCVC variants reproduced from [26].

Method	UVG	MCL-JCV	HEVC B	HEVC E	Average	Enc. fps	Dec. fps
VTM-17.0	0.0	0.0	0.0	0.0	0.0	0.01	23.6
HM-16.25	40.1	46.6	41.0	42.8	43.1	0.05	39.6
ECM-11.0	-20.0	-22.1	-22.2	-17.2	-20.3	0.002	3.4
DCVC-DC	6.5	-4.4	13.1	90.2	26.4	3.3	4.3
DCVC-FM	-17.6	-8.4	-15.7	-23.0	-16.1	3.4	4.2
DCVC-FM (fp16)	-16.8	-8.0	-15.4	-20.2	-15.1	5.0	8.9
DCVC-RT (fp16)	-24.0	-14.8	-16.6	-22.4	-19.4	125.2	112.8
CASVC [73]	-26.7	-19.7	-25.5	-15.9	-22.0	1.23	19.2
Proposed	-27.9	-22.6	-27.2	-17.9	-23.9	0.18	16.7

Table 4.9: SSIM BD-rate (%) comparison in YUV420 colour space for all 1080p datasets. Reference: VTM 17.0 (0%).

Method	UVG	MCL-JCV	HEVC B	HEVC E	Average
VTM-17.0	0.0	0.0	0.0	0.0	0.0
CASVC [73]	−20.9	−17.5	−22.2	−12.5	−18.2
Proposed	−26.6	−25.5	−24.4	−13.3	−22.4

Table 4.10: VMAF BD-rate (%) comparison in YUV420 colour space for all 1080p datasets. Reference: VTM 17.0 (0%).

Method	UVG	MCL-JCV	HEVC B	HEVC E	Average
VTM-17.0	0.0	0.0	0.0	0.0	0.0
CASVC [73]	−19.3	−11.8	−21.6	−10.9	−15.9
Proposed	−22.9	−24.8	−23.4	−14.6	−21.4

Table 4.11: BD-rate (%) comparison against VTM 23.0 for CCTV (average of 3 sequences) and 3D animation (average of 5 sequences) content. All sequences encoded at $QP \in \{27, 32, 37, 42\}$ with intra period = −1 (random access).

Method	CCTV			3D Animation		
	PSNR	SSIM	VMAF	PSNR	SSIM	VMAF
VTM-23.0	0.0	0.0	0.0	0.0	0.0	0.0
CASVC [73]	−17.0	−15.5	−13.3	−15.0	−14.5	−13.6
Proposed	−18.6	−16.4	−18.8	−19.1	−16.4	−17.8

UVG (−27.9% PSNR, −26.6% SSIM, −22.9% VMAF) and HEVC Class B (−27.2%, −24.4%, −23.4%), where the broadcast-style motion characteristics make the RIFE-interpolated context a highly accurate prior, leaving the per-GOP update Δ_g free to specialise the decoder to fine-grained local statistics. MCL-JCV yields −22.6% PSNR, −25.5% SSIM, and −24.8% VMAF despite its greater content diversity, while HEVC Class E returns −17.9% PSNR, −13.3% SSIM, and −14.6% VMAF. The comparatively smaller HEVC E gains arise from its quasi-static videoconferencing scenes, which contain mainly stationary backgrounds and slow head or upper-body motion. This strong temporal redundancy is already represented efficiently by the block-based inter-prediction and adaptive partitioning tools of VVC [38], while the global pretrained prior also closely matches the local content statistics; consequently, the motion-aligned semantic context provides less additional information and the optimisation of (4.15) converges to a near-zero Δ_g^* . VVC may therefore remain preferable for prolonged static or slow-motion sequences with little scene change, particularly at high-quality operating points where the residual bitstream dominates and the semantic and adaptation overhead cannot be offset by a sufficiently large reduction in prediction error.

Directly compared with CASVC, the proposed framework consistently improves all three metrics on every dataset. In PSNR BD-rate, the improvements are 1.2 pp on UVG, 2.9 pp on MCL-JCV, 1.7 pp on HEVC B, and 2.0 pp on HEVC E, for an overall average improvement of 1.87 pp (−23.9% vs −22.0%). The SSIM and VMAF gains over CASVC are even larger in absolute terms (−22.4% vs −18.2% in SSIM and −21.4% vs −15.9% in VMAF on average), reflecting the fact that the per-GOP update primarily corrects low-frequency, content-correlated biases in the semantic prediction exactly the structured prediction errors that VMAF’s learned perceptual model is most sensitive to. The MCL-JCV VMAF jump from −11.8% (CASVC) to −24.8% (proposed) is particularly striking: diverse, perceptually challenging content benefits disproportionately from a decoder that adapts its reconstruction prior to each individual scene.

The QP dependence follows the mechanism already established for CASVC in Section 4.2.2: gains are largest when residual quantisation makes semantic prediction accuracy most influential, whereas the residual sub-stream dominates at QP 27. CAPO

preserves this trend but improves every operating point by reducing the content-specific prediction error before residual coding.

Out-of-Distribution Generalisation

The results on the two additional content classes in Table 4.11 and Figs. 4.15–4.16 confirm that the rate–distortion advantage of the proposed framework generalises well beyond standard natural-video benchmarks. On CCTV, the proposed codec achieves -18.6% PSNR, -16.4% SSIM, and -18.8% VMAF BD-rate against VTM 23.0, improving over CASVC by 1.67 pp, 0.90 pp, and 5.50 pp, respectively. On 3D animation the gains are larger: -19.1% PSNR, -16.4% SSIM, and -17.8% VMAF against VTM 23.0, improving over CASVC by 4.16 pp, 1.91 pp, and 4.27 pp, respectively.

The 3D animation result, a 4.16 pp CASVC-to-proposed PSNR gain and the largest such gain across all six datasets, directly illustrates the main value of per-GOP overfitting when the content departs from the pretraining distribution. This behaviour is important because the pretrained CASVC decoder carries the visual prior of Vimeo-90k, which mainly contains natural camera-captured video with sensor noise, motion blur, depth-of-field effects, and broad texture variation. The two additional content classes depart from that prior in different ways: CCTV clips contain mostly static backgrounds, smooth low-frequency regions, and limited motion, whereas 3D animation contains piecewise-smooth shading, very sharp synthetic edges, stable highlights, and narrow saturated colour palettes. When a frozen decoder trained on natural video is applied to such out-of-distribution classes, the mismatch appears as a structured, content-correlated, low-frequency prediction error. VVC must then spend residual bits correcting an error pattern that comes from the decoder prior rather than from genuinely unpredictable detail.

The proposed per-GOP update Δ_g^* is designed to remove exactly this mismatch. Because it is solved online from the actual GOP frames, it re-targets the Learnable Scale parameters of the decoder upsampling stages to the local content statistics before residual coding takes place. For CCTV sequences, this update suppresses the texture-generating tendency of the natural-image prior, keeping flat surveillance backgrounds

flat and explaining why the largest improvement is observed in VMAF. For 3D animation, the same mechanism rebalances the decoder so that smooth shading and sharp synthetic edges are reproduced without the soft natural-image texture bias, producing the largest PSNR gain observed in the study. These results support a general conclusion: the benefit of the 643-scalar update increases as the gap between the current GOP statistics and the Vimeo-90k pretraining distribution becomes larger.

Bitstream Composition Ablation

To quantify *why* the proposed codec improves over CASVC, Table 4.12 reports the relative change of each bitstream component when switching from CASVC to the proposed codec, computed as $(\text{Proposed} - \text{CASVC})/\text{CASVC} \times 100\%$ at each of the four QP operating points, averaged over two sequences from each dataset. The proposed bitstream has four distinct sub-streams: (i) intra-coded *key frames* (R_{key}), identical in both codecs since both use the same VVenC intra configuration; (ii) VVC-coded inter-frame *residuals* (R_{residual}), which correct the error between the semantic prediction and the original frame; (iii) the new *parameters* sub-stream (R_{param}), which carries the per-GOP decoder update Δ_g and is unique to the proposed codec (CASVC has no such stream); and (iv) DeepCABAC-coded differential *latents* (R_{latent}), which transmit the compact semantic representation of each frame after inter-GOP differential coding. The rate budget identity is $R_{\text{total}} = R_{\text{key}} + R_{\text{residual}} + R_{\text{param}} + R_{\text{latent}}$ as defined in (4.11). Because CASVC has no parameters sub-stream, R_{param} is reported as a percentage of the proposed codec’s own total bitrate.

Four observations follow from Table 4.12.

First, $\Delta R_{\text{residual}} = -9.01\%$ on average, with reductions of 8–16% at the three medium-to-low quality operating points. This large drop in residual bits confirms that the per-GOP update absorbs the structured, content-correlated prediction error that the CASVC decoder averaged over the Vimeo-90k distribution could not correct, leaving VVC to encode only the genuinely high-frequency, non-predictable texture detail.

Second, the parameter overhead is negligible and scales gracefully: $R_{\text{param}}/R_{\text{total}}$ ranges from only 0.14% at QP = 27 to 1.42% at QP = 42, averaging 0.69% across all

Table 4.12: Bitstream composition ablation: per-component percentage change of the proposed codec relative to CASVC at each QP operating point (average of 2 sequences from each dataset). The $R_{\text{param}}/R_{\text{total}}$ column reports the parameter sub-stream as a fraction of the proposed codec’s total bitrate (CASVC has no parameter sub-stream). Quality columns report percentage change relative to CASVC; positive values indicate improvement.

QP	ΔR_{key}	$\Delta R_{\text{residual}}$	$R_{\text{param}}/R_{\text{total}}$	ΔR_{latent}	ΔR_{total}	ΔPSNR	ΔSSIM	ΔVMAF
27	0.00%	−0.11%	+0.14%	0.00%	−0.34%	+0.08%	+0.01%	+0.85%
32	0.00%	−8.34%	+0.38%	0.00%	−2.30%	+0.48%	+0.29%	+2.18%
37	0.00%	−11.16%	+0.80%	0.00%	−3.30%	+0.84%	+0.29%	+7.55%
42	0.00%	−16.43%	+1.42%	0.00%	−3.98%	+1.21%	+0.50%	+7.47%
Avg	0.00%	−9.01%	+0.69%	0.00%	−2.48%	+0.65%	+0.27%	+4.51%

operating points. Even at the worst-case QP = 42, the 1.42% parameter overhead is more than $11\times$ smaller than the 16.43% residual saving it enables at the same operating point, confirming that the exchange is highly favourable throughout the entire rate–distortion range.

Third, $\Delta R_{\text{latent}} = 0$ at every QP without exception. The encoder ϕ_{θ_e} and the DeepCABAC differential-latent pipeline are both frozen and identical in CASVC and the proposed codec; the per-GOP update Δ_g touches only the decoder side. This column therefore serves as a control variable confirming that the two codecs produce bit-for-bit identical latent streams.

Fourth, the net saving in total bitrate (−2.48% average) is accompanied by simultaneous improvements in all three quality metrics: PSNR by +0.65%, SSIM by +0.27%, and VMAF by +4.51% on average, peaking at +1.21% PSNR and +7.55% VMAF at the most practically relevant operating points. The disproportionately large VMAF gains (+4.51% versus +0.65% PSNR) reflect the nature of the corrected error: the Vimeo-90k prior introduces a low-frequency, content-correlated texture bias whose perceptual saliency is amplified by VMAF’s learned metric while having a more moderate effect on pixel-level PSNR.

Comparative Analysis with State-of-the-Art Codecs

The PSNR BD-rate comparison in Table 4.8 situates the proposed codec in the broader landscape. Against the conventional reference HM-16.25 (+43.1% BD-rate), the proposed codec saves approximately 67 pp. Against ECM-11.0, the most competitive conventional baseline at -20.3% , the proposed codec improves by 3.6 pp (-23.9% vs -20.3%), a meaningful margin given that ECM-11.0 represents the strongest signal-domain VVC extension currently available.

Among the learned baselines, DCVC-DC exhibits instability on HEVC E (+90.2%), yielding a positive average BD-rate of +26.4%. DCVC-FM and DCVC-FM (fp16) achieve -16.1% and -15.1% on average; the proposed codec improves over them by 7.8 pp and 8.8 pp, respectively. The real-time DCVC-RT (fp16) achieves -19.4% average BD-rate at 112.8 fps decoding; the proposed codec improves its BD-rate by 4.5 pp while decoding at 16.7 fps. The proposed codec is the only system in Table 4.8 that simultaneously: (i) achieves better than -20% PSNR BD-rate on average; (ii) achieves a negative BD-rate on every individual dataset, including HEVC E where DCVC-DC and DCVC-FM struggle; and (iii) exceeds both the strongest conventional baseline (ECM-11.0) and the strongest learned codec in terms of average BD-rate.

Runtime and Computational Complexity

Table 4.8 also reports encoding and decoding throughput on an NVIDIA A100 GPU. The proposed codec encodes 1080p content at 0.18 fps and decodes at 16.7 fps. The encoding throughput is lower than CASVC (1.23 fps) because of the per-GOP overfitting solve in (4.15): running 60–100 Adam epochs over the N GOP frames with frozen contexts introduces an unavoidable additional cost relative to CASVC’s fully frozen-decoder pipeline. Despite this overhead, the proposed encoder remains $18\times$ faster than VTM 17.0 (0.01 fps) and is well suited to offline-encode/online-decode applications such as video-on-demand and archival compression. No iterative learning is performed at the receiver: the per-GOP overfitting overhead is entirely an encoder-side cost.

Decoding throughput drops modestly from CASVC’s 19.2 fps to 16.7 fps. The three additional decoder-side operations, entropy-decoding Δ_g , applying the inverse

PCA transform, and running one pre-overfitting forward pass to reproduce the frozen contexts $\tilde{C}_{m+k}$, account for the entire slowdown; the tapered autoencoder itself is faster than CASVC’s because of the $\approx 4.9\times$ FLOP reduction. The decoder at 16.7 fps remains faster than all DCVC variants: DCVC-FM (4.2 fps), DCVC-FM (fp16) (8.9 fps), DCVC-DC (4.3 fps), and ECM-11.0 (3.4 fps); only DCVC-RT (fp16) at 112.8 fps decodes faster, at the cost of 4.5 pp worse average BD-rate. As noted in Section 4.2.3, DCVC-RT’s high throughput relies on cuDNN-optimised fp16 inference on dedicated GPU hardware, a deployment environment not available within the Google Colaboratory [183] cloud platform used for all experiments in this thesis; on equivalent dedicated hardware, the proposed framework would achieve substantially higher throughput.

In terms of inference complexity, the proposed tapered autoencoder requires 1.17 GFLOPs per 144×256 block at the encoder (versus 8.19 GFLOPs for CASVC, a $7.0\times$ reduction) and 11.26 GFLOPs at the decoder (versus 52.22 GFLOPs for CASVC, a $4.6\times$ reduction), for a net per-block reduction of $\approx 4.9\times$ ($60.41 \rightarrow 12.43$ GFLOPs). The savings concentrate at the highest-resolution decoder stages: replacing the uniform 128-channel feature maps of CASVC with the $96 \rightarrow 64 \rightarrow 32$ taper eliminates the bulk of the CASVC `ResBlk(128)` cost at 144×256 , which alone exceeds 20 GFLOPs in CASVC. The Learnable Scale layers add only approximately 2×10^6 FLOPs per block ($< 0.05\%$ of decoder cost) and are therefore negligible relative to the FLOP savings delivered by the channel taper.

Visual Quality Comparison

Fig. 4.17 illustrates the progressive quality enhancement achieved by the four stages of the proposed pipeline: RIFE-only interpolation, CASVC semantic prediction, proposed parameter-updated semantic prediction, and final hybrid reconstruction with the VVC residual.

Three observations emerge consistently across all six rows. *First*, RIFE-only prediction (column 1) yields modest fidelity (PSNR 21–26 dB, SSIM 0.74–0.88): it correctly reproduces global motion but loses texture and fine structure, as expected from a flow-based interpolator without access to the current-frame latent. The CCTV row (Row 5)

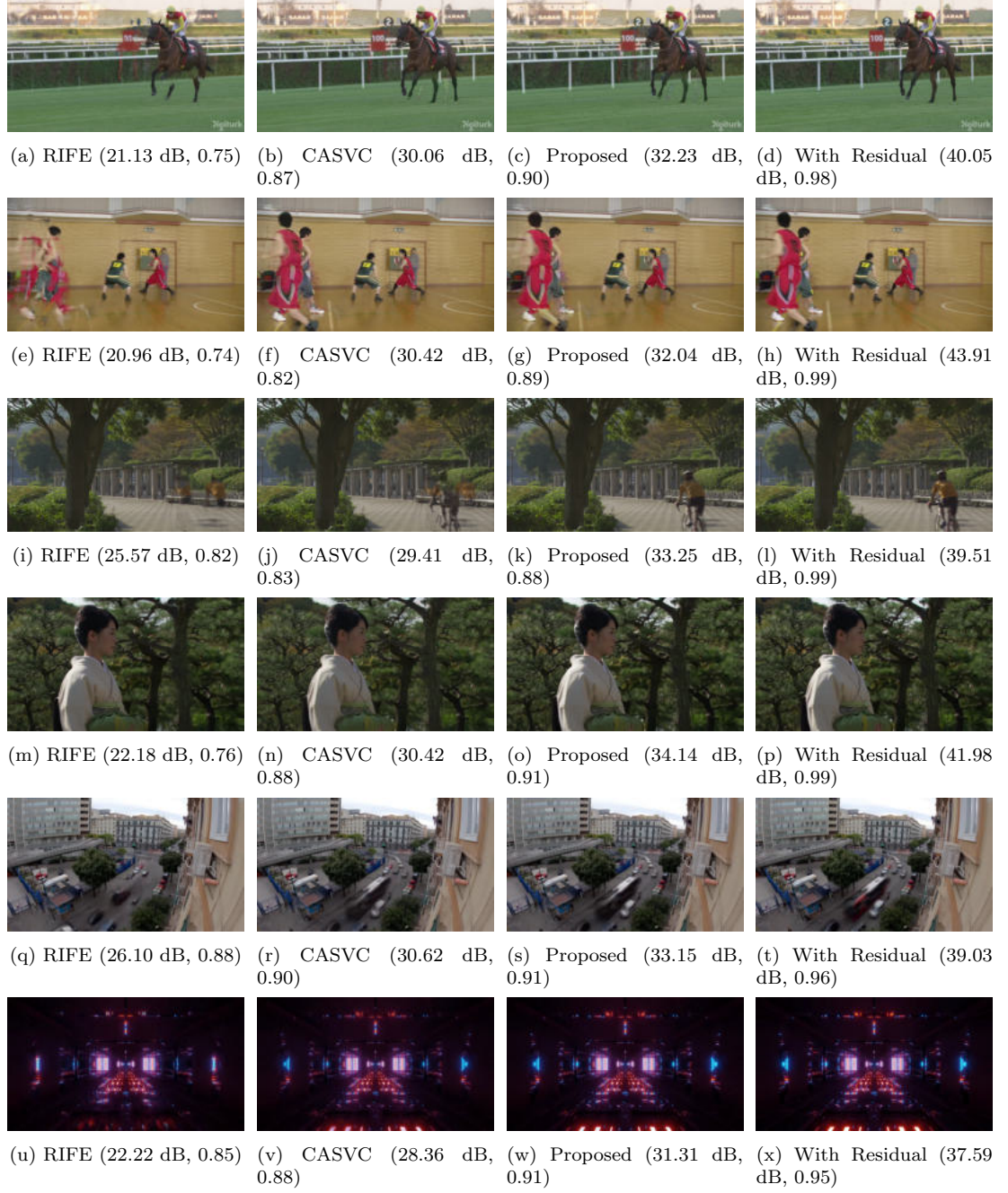


Figure 4.17: Visual quality comparison across six representative frames spanning all three evaluated content classes. Each row shows a frame at four stages: (1) RIFE-only interpolation, (2) CASVC semantic prediction, (3) proposed parameter-updated semantic prediction, and (4) final hybrid reconstruction with the VVC residual. Captions report (PSNR in dB, SSIM). Rows 1–4 are natural-video content; Row 5 is a CCTV surveillance frame; Row 6 is a 3D animation frame. The per-GOP parameter update Δ_g consistently raises semantic prediction quality by 1.6–3.8 dB PSNR and up to 0.06 SSIM over CASVC before the VVC residual is added.

starts from a higher SSIM (0.88) owing to the largely static background, while the 3D animation row (Row 6) similarly benefits from the piecewise-smooth synthetic regions that RIFE handles well. *Second*, conditioning the CASVC semantic decoder on the RIFE context (column 2) lifts PSNR by 4–10 dB and SSIM by 0.01–0.13, quantitatively confirming the value of the learned semantic latent on top of the motion-aligned context. *Third*, replacing the frozen CASVC decoder with the per-GOP parameter-updated decoder (column 3) provides a further 1.6–3.8 dB PSNR gain and an SSIM improvement of up to 0.06 on the same frames, with no additional residual transmitted at this stage. This isolated semantic improvement is visible across all three content classes and is particularly pronounced on the 3D animation row (Row 6), where the proposed decoder raises the semantic prediction from 28.36 dB / 0.88 (CASVC) to 31.31 dB / 0.91, mirroring the largest single-dataset CASVC-to-proposed PSNR BD-rate gap (4.16 pp on 3D animation) reported in Table 4.11. The visual improvement from the parameter update manifests primarily as sharper edges and reduced low-frequency colour drift, consistent with Δ_g rescaling the decoder upsampling feature maps to match local content statistics rather than the Vimeo-90k global average. Finally, after VVC residual correction (column 4), all six rows reach 37.6–43.9 dB PSNR with SSIM 0.95–0.99, confirming that the hybrid architecture achieves broadcast-grade fidelity across natural, CCTV, and 3D animation content alike.

Discussion and Limitations

The proposed framework establishes that per-GOP content adaptivity can be re-introduced into a fully deployable pretrained hybrid semantic codec at a cost of only 643 additional scalars per GOP (averaging 0.69% of the total bitrate), yielding consistent BD-rate improvements across six diverse content classes with no observed regression at any operating point. The three key design decisions restricting the adaptive set to Learnable Scale multipliers and Conv2DT biases, transmitting the update differentially via PCA–DeepCABAC, and precomputing the frozen contexts before the overfitting solve collectively ensure that the encoder-side computational overhead is manageable, the decoder requires no retraining, and the protocol supports random access at GOP

boundaries.

Several limitations remain. The per-GOP overfitting solve reduces encoder throughput from 1.23 fps (CASVC) to 0.18 fps, making the proposed codec better suited to offline encoding workflows (archival compression, video-on-demand) than to real-time streaming. Performance on HEVC Class E (low-motion videoconferencing) is comparatively modest because the global pretrained prior already closely matches quasi-static content, leaving limited room for per-GOP improvement; gains grow with the gap between GOP content statistics and the Vimeo-90k training distribution. The 144×256 block-processing architecture introduces tile-boundary artefacts at high quality operating points, partially mitigated by bilateral filtering at block seams. VVC patent licensing obligations also apply to commercial deployment, as in all VVC-based systems.

4.3.4 Conclusion

The content-adaptive parameter overfitting mechanism presented in this section shows that per-GOP content adaptation can be added back into a fully pretrained hybrid semantic codec at very low overhead. Updating only per-channel Learnable Scale multipliers and transposed-convolution biases (643 scalars per GOP, transmitted at an average of 0.69% of total bitrate via PCA–DeepCABAC compression) produces consistent BD-rate improvements over CASVC [73] across all evaluated content classes without any regression at any operating point. Relative to VTM 17.0, the combined framework achieves average Bjøntegaard delta-rates of -23.9% PSNR, -22.4% SSIM, and -21.4% VMAF across the four standard 1080p benchmarks, improving over CASVC by an average of 1.87 percentage points in PSNR. The bitstream composition ablation confirms that residual bitrate falls by 9.01% on average relative to CASVC, while the parameter sub-stream adds only 0.69% overhead, showing that the mechanism works by reducing content-related prediction error rather than moving distortion between rate components.

The out-of-distribution evaluation clearly shows the scope of the mechanism: PSNR BD-rates of -18.62% and -19.12% relative to VTM 23.0 on CCTV surveillance and 3D

animation content, respectively, confirm that the combined framework maintains strong performance on content types that are far from the Vimeo-90k training distribution. The improvement over CASVC is 1.67 percentage points on CCTV surveillance and 4.16 percentage points on 3D animation, with the latter being the largest single-dataset gain observed in the entire experimental study. This confirms that the benefit of the per-GOP overfitting step grows with how different the GOP content is from the Vimeo-90k training distribution, which is exactly the case where the fixed global decoder of CASVC is least well suited. The combined framework achieves competitive performance against all evaluated state-of-the-art learned codecs on the standard benchmarks, with gains over DCVC-FM by 7.8 percentage points, DCVC-RT by 4.5 percentage points, and ECM-11.0 by 3.6 percentage points in average PSNR BD-rate.

The tapered autoencoder reduces per-block processing from 60.41 GFLOPs (CASVC) to 12.43 GFLOPs, a $4.9\times$ reduction that partially offsets the encoding cost of the per-GOP solve; the encoder nonetheless runs at 0.18 fps, which is $18\times$ faster than VTM 17.0 (0.01 fps), while the decoder runs at 16.7 fps, making the system suitable for offline-encode, online-decode use cases such as video-on-demand and archival compression. This contribution directly addresses Gap 3 identified in Chapter 2: the lack of a practical mechanism for per-GOP content adaptation in a fully pretrained hybrid semantic codec. Together with the context-conditioned semantic codec of Section 4.2, the combined framework provides a two-stage approach to pretrained hybrid semantic video coding: a shared global model that removes per-GOP training overhead, combined with a compact per-GOP decoder update that restores content-specific adaptation at low signalling cost.

4.4 Chapter Summary and Conclusion

This chapter has presented two contributions to context-aware hybrid semantic video coding, moving from the per-GOP training approach of Chapter 3 towards a fully pretrained architecture with content-adaptive decoding.

Section 4.2 introduced a context-conditioned semantic video coding framework that addresses Gap 2 identified in Chapter 2 by replacing per-GOP from-scratch training

with a shared pretrained autoencoder. The framework uses RIFE-based hierarchical bidirectional context generation [72] for motion-aligned decoder conditioning, differential PCA-DeepCABAC [147, 148] entropy coding that keeps semantic latent overhead below 0.4% of total bitrate, and a TI-based model selection strategy that assigns high-motion content to a dedicated fine-tuned model. Evaluated under the JVET random-access common test conditions, the framework achieves an average PSNR BD-rate of -22.03% relative to VTM 17.0 across UVG, MCL-JCV, HEVC Class B, and HEVC Class E at 1080p, and -22.6% relative to VTM 23.0 at 4K, outperforming ECM-11.0 (-20.3%), DCVC-FM (-16.1%), and DCVC-RT fp16 (-19.4%) while running at 1.23 fps encoding and 19.2 fps decoding on an NVIDIA A100 GPU. This contribution fulfils Objective 6 of the thesis and shows that per-GOP retraining can be fully removed within a competitive hybrid semantic video codec.

Section 4.3 introduced a content-adaptive parameter overfitting mechanism that addresses Gap 3 identified in Chapter 2 by bringing back per-GOP content adaptation into the CASVC framework at the cost of 643 scalars per GOP. Updating only per-channel Learnable Scale multipliers and transposed-convolution biases allows the decoder to adapt from the global Vimeo-90k average to the local content of each GOP, reducing residual bitrate by 9.01% on average relative to CASVC while the parameter sub-stream adds only 0.69% of total bitrate. A tapered $32 \rightarrow 64 \rightarrow 96 \rightarrow 128$ encoder reduces per-block processing by $4.9\times$ relative to CASVC. The combined framework achieves average BD-rates of -23.9% PSNR, -22.4% SSIM, and -21.4% VMAF relative to VTM 17.0, improving over CASVC by 1.87 percentage points on the standard benchmarks and by 4.16 percentage points on 3D animation, the content class furthest from the training distribution. This contribution fulfils Objective 7 of the thesis.

Together, the two contributions of this chapter show that a compact, $\mathcal{O}(\text{channels})$ -sized affine decoder update, transmitted via an entropy-coded parameter sub-stream, is enough to recover the per-GOP content adaptivity of full retraining at a much lower cost. The results on CCTV surveillance and 3D animation content show that the approach is not limited to the Vimeo-90k training distribution, making content-adaptive parameter overfitting a practical mechanism for hybrid semantic video compression.

The main remaining limitations are encoder throughput (0.18 fps), which prevents real-time encoding, and that the semantic autoencoder is trained only on SDR Vimeo-90k content. These limitations motivate Chapter 5, which extends the CASVC-CAPO framework to linear HDR EXR video through an automatic, per-frame, fully invertible tone-mapping step.

Chapter 5

Semantic Video Coding for High Dynamic Range Content

5.1 Chapter Introduction

The growing use of BT.2020/PQ HDR and WCG content in broadcast and streaming has raised the dynamic range and perceptual quality expected from modern video codecs [67, 68]. Standard codecs such as H.266/VVC [38] support HDR/WCG content through ten-bit profiles, luma mapping with chroma scaling (LMCS), and supplemental enhancement information messages [69]. However, their rate-distortion and entropy-coding tools were largely designed and tested on SDR content, which limits performance on HDR material. The context-aware semantic video codecs developed in Chapter 4, namely CASVC in Section 4.2 and its CAPO extension in Section 4.3, have the same limitation because both were trained and evaluated only on ten-bit SDR data. If a linear HDR EXR sequence is fed directly into either system, the semantic autoencoder sees content far from its Vimeo-90k training distribution, which increases semantic reconstruction error and residual bitrate and removes the BD-rate advantage observed on SDR content. This is the limitation defined as Gap 6 in Chapter 2: how to extend a semantic video codec to HDR sources without retraining its learned components specifically for HDR data.

The framework in this chapter addresses Gap 6 by adding per-frame invertible tone

mapping as the smallest HDR-specific step needed to bridge the gap between linear HDR EXR source material and an SDR-trained semantic codec. The automatic photographic tone-mapping operator of Reinhard et al. [150] is chosen because it satisfies three requirements at once. It produces ten-bit SDR output whose statistics closely match the training distribution used by the CASVC–CAPO semantic autoencoder, it is fully invertible in closed form, and it is automatic because both required parameters can be estimated for each frame from the log-average scene luminance without manual tuning. This choice is important because the HDR–SDR stage is not used only to compress dynamic range for display; it must also transform the HDR input into the SDR-domain signal statistics expected by the semantic codec. A Reinhard-only compression step controls highlight roll-off, but it does not by itself guarantee that mid-tones will fall at the code values expected in BT.709/sRGB-like SDR content. For this reason, the proposed bridge combines per-frame exposure normalisation, Reinhard compression, and a per-frame gamma adjustment, so that both dynamic range and mid-tone placement are aligned with the SDR domain while remaining analytically invertible. The Hable Filmic operator uses a configurable toe, linear section, shoulder, and white point to simulate a film-like response [194]. This gives greater control over shadow contrast and highlight roll-off than the global Reinhard curve, but also introduces more curve choices. A fixed, unclipped Filmic curve can be inverted [195], so invertibility alone does not exclude this operator; Reinhard is preferred here because its simpler rational form provides a direct closed-form inverse and requires only the exposure and gamma parameters already estimated for each frame. Learned tone-mapping methods can optimise brightness, contrast, and temporal consistency [196], but they require training and neural-network inference and are not generally constrained to provide a one-to-one mapping with a closed-form inverse. A separate learned inverse model [197] estimates an HDR reconstruction rather than algebraically reversing the forward mapping. Bilateral tone mapping [198] preserves local detail through an image-dependent spatial decomposition, but this spatially varying operation cannot be reversed from two global parameters. Reinhard is therefore selected for its automatic operation, low complexity, compact metadata, and simple inverse, with the limitation that its global curve

provides less control over local contrast and display appearance.

The established CASVC–CAPO pipeline is reused as defined in Section 4.3, so this chapter describes only the HDR-specific processing and evaluation changes. The semantic autoencoder is retrained once offline on tone-mapped clips from the HDRVD2K HDR video dataset [23], so that its training data matches the SDR statistics produced at inference time. To the best of the author’s knowledge, this is the first hybrid semantic video codec evaluated end to end on the JVET HDR/WCG common test conditions [23].

The main contributions of this chapter are as follows.

- *Per-Frame Invertible Tone-Mapping Bridge.* An automatic per-frame Reinhard pre-process estimates an exposure multiplier e_t and a gamma exponent γ_t in closed form from the per-frame log-average luminance, mapping each linear HDR EXR frame to a ten-bit SDR frame whose statistics match the same perceptual grey level as the CASVC–CAPO training distribution. The two scalar parameters per frame are entropy-coded as a compact CABAC-coded HDR metadata sidestream, and the receiver recovers the linear HDR float32 signal through algebraic inversion.
- *HDR Extension with No Architectural Change.* All four components of CASVC–CAPO are reused without modification: the bidirectional hierarchical prediction structure, the RIFE-based motion-aligned context generation [72], the tapered quantisation-aware semantic autoencoder, and the PCA–DeepCABAC differential entropy coder [147, 148, 193]. The autoencoder is retrained once offline on tone-mapped clips from the HDRVD2K dataset [23]; no per-sequence or per-GOP HDR-specific retraining is required.
- *HDR Evaluation.* The framework is evaluated on the HDM-HDRV cinematic dataset and the JVET HDR/WCG common test condition sequences [23] against VTM 23.0 at four random-access quantisation points. Five metrics are reported: SDR-domain PSNR, SSIM [164], and VMAF on the tone-mapped output, and HDR-domain PU-PSNR and PU-SSIM computed via PU21 perceptual encod-

ing [27].

- *Ablation of Per-Frame Tone-Mapping.* A direct comparison against a fixed-parameter Reinhard baseline on representative HDM-HDRV frames shows that per-frame estimation of the two scalar parameters is needed to maintain consistent SDR statistics across frames with widely varying HDR scene luminance.

The first-author study “Context-Aware Semantic Video Coding via Content-Adaptive Parameter Overfitting for High Dynamic Range Videos” is integrated across Section 5.2, which presents its method and ablation studies, and Section 5.2.2, which reports evaluation against VTM 23.0 on the HDM-HDRV and JVET-HDRV benchmark datasets [23]. This work extends the pretrained and content-adaptive framework of Chapter 4 to HDR and WCG sources. Section 5.3 summarises the contribution and relates it to the thesis objectives.

5.2 Content-Adaptive Parameter Overfitting for HDR Video Coding

5.2.1 Methodology

System Overview

As explained in the Chapter Introduction, the standard CASVC–CAPO model of Chapter 4 is trained on 8-bit SDR clips from the Vimeo-90k dataset [175] and operates fully in the 8-bit SDR domain, so sending linear HDR EXR content directly into that codec exposes the semantic autoencoder to a distribution far from its training set. This increases both semantic reconstruction error and residual bitrate. The proposed framework addresses this by training a separate HDR model instance that shares the same architecture as CASVC–CAPO but is trained on 10-bit SDR clips derived from the HDRVD2K HDR dataset via the tone-mapping bridge described below, so that its training distribution matches the 10-bit SDR statistics produced by tone-mapping at inference time. This separation means the original Vimeo-90k-trained CASVC–CAPO model and its SDR performance are not affected. The framework extends CASVC–

CAPO (Section 4.3) to linear HDR EXR sources by adding an automatic, per-frame, fully invertible photographic tone-mapping pre-process [150]. This transforms each linear HDR EXR frame into a 10-bit SDR frame whose statistics match the training distribution of the HDR model. The two tone-mapping parameters for each frame, the exposure multiplier e_t and the gamma exponent γ_t , are entropy coded as a compact HDR metadata side stream, which lets the receiver recover the linear HDR float32 signal through a closed-form inverse. All four parts of CASVC-CAPO are then applied without modification to the tone-mapped SDR stream. The tone-mapping stage, the end-to-end pipeline, and the per-GOP overfitting procedure are described in Sections 5.2.1–5.2.1. The full architecture is shown in Fig. 5.1.

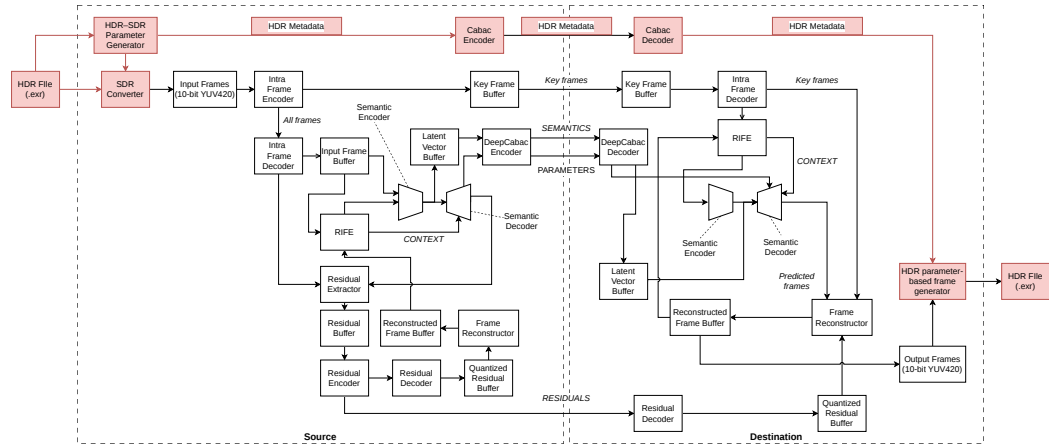


Figure 5.1: End-to-end architecture of the proposed context-aware semantic HDR video codec.

HDR-to-SDR Tone Mapping

Let an HDR video sequence be denoted $\mathcal{F} = \{F_t^{\text{HDR}}\}_{t=0}^T$, where each frame $F_t^{\text{HDR}} \in \mathbb{R}^{H \times W \times 3}$ is a linearised RGB image read directly from the linear HDR EXR source, with values representing scene-linear light. The automatic tone-mapping operator maps each frame to a 10-bit SDR frame $F_t^{\text{SDR}} \in [0, 1023]^{H \times W \times 3}$ through three steps.

Step 1: Automatic Exposure. The per-frame log-average luminance is computed as

$$L_{w,t} = \exp\left(\frac{1}{HW} \sum_p \ln(\varepsilon + L_p)\right),$$

$$L_p = 0.2126 r_p + 0.7152 g_p + 0.0722 b_p, \quad (5.1)$$

where (r_p, g_p, b_p) are the linearised RGB values of pixel p , $\varepsilon = 10^{-4}$ guards against $\ln(0)$, and the coefficients $(0.2126, 0.7152, 0.0722)$ are the ITU-R BT.709 relative luminance coefficients [199]. BT.709 coefficients are used because the tone-mapped signal is intended to serve as an SDR intermediate representation, so its luminance statistics should remain consistent with the SDR-domain semantic codec rather than the original HDR BT.2020/PQ source convention. The per-frame exposure multiplier is set to

$$e_t = \frac{0.18}{L_{w,t}}, \quad (5.2)$$

which maps the geometric mean luminance of each frame to the photographic middle-grey key value of 0.18 [150]. The constant 0.18 is used because 18% reflectance is a conventional photographic middle-grey reference [200], giving a neutral exposure target that usually places average scene brightness at a visually balanced mid-tone before the later SDR-domain gamma adjustment. Estimating e_t independently per frame enables the tone-mapping to track scene-cut luminance changes and intra-sequence dynamic range variation, ensuring that every frame is individually exposed to a consistent perceptual level before entering the CASVC–CAPO pipeline.

Step 2: Reinhard Mapping and Automatic Gamma. Each RGB channel is scaled by e_t and passed through the Reinhard curve $R(x) = x/(1 + x)$, which maps $[0, \infty)$ to $[0, 1)$ while preserving relative luminance order [150]. Reinhard compression is used to bring the HDR luminance range into the SDR interval with a simple order-preserving mapping, but this step alone does not guarantee that the resulting SDR frame places mid-tones at the code values expected by the SDR-domain semantic codec. A per-frame gamma exponent γ_t is then estimated automatically from the mean tone-

mapped luminance $\bar{m}_t$ of the Reinhard-mapped frame as

$$\gamma_t = \frac{\ln(\bar{m}_t)}{\ln(0.45)}, \quad \gamma_t \in [1.8, 3.2], \quad (5.3)$$

which ensures that the mean brightness is placed at the perceptually appropriate mid-tone of 0.45, between the approximately 41% code-value placement of 18%-grey under the BT.709 opto-electronic transfer function (OETF) ($V = 1.099 \times 0.18^{0.45} - 0.099 \approx 0.41$) and the approximately 46% placement under the sRGB transfer function ($C = 1.055 \times 0.18^{1/2.4} - 0.055 \approx 0.46$). The gamma step is therefore used as a distribution-alignment stage rather than a purely visual adjustment: it shifts the post-Reinhard mid-tone level toward the SDR statistics seen during autoencoder training, while remaining exactly invertible once γ_t is transmitted as metadata. The full per-frame tone-mapping is applied as

$$F_t^{\text{SDR}} = \text{clip}(\text{R}(e_t \cdot F_t^{\text{HDR}}), 0, 1)^{1/\gamma_t}, \quad (5.4)$$

where $\text{clip}(x, 0, 1) = \max(0, \min(x, 1))$ is applied elementwise, the exponentiation is applied channel-wise, and the output is subsequently quantised to 10-bit precision.

Step 3: 10-Bit Quantisation and YUV420 Conversion. The tone-mapped RGB output $F_t^{\text{SDR}} \in [0, 1]^{H \times W \times 3}$ is quantised to 10-bit precision and converted to YUV420 using BT.709 matrix coefficients, producing the SDR stream $\{F_t^{\text{SDR}}\}$ that enters the CASVC-CAPO pipeline. This representation includes one luminance component and two chrominance components, so the framework codes both luminance and colour information rather than operating on luminance alone. However, BT.2020 defines a wider colour gamut than BT.709, and converting wide colour gamut content to the BT.709 intermediate can map colours that lie outside the BT.709 gamut [68, 201]. In addition, YUV420 stores each chrominance component at half the horizontal and vertical resolution of the luminance component. The inverse tone-mapping can restore the compressed dynamic range, but it cannot recover colour-gamut or chrominance detail removed by these two operations. The present framework therefore handles luminance and chrominance jointly, but it is not a native gamut-preserving BT.2020 codec.

Invertibility. A critical property of the proposed tone-mapping is that it admits a closed-form inverse. Given the decoded 10-bit SDR reconstruction and the transmitted metadata scalars (e_t, γ_t) , the receiver recovers the linear HDR float32 signal through three steps: (i) inverse gamma correction, which raises each channel to the power γ_t ; (ii) inverse Reinhard mapping $R^{-1}(x) = x/(1 - x)$; and (iii) division by the exposure multiplier e_t . Local or bilateral operators [198] and display-referred operators such as the Academy Color Encoding System (ACES) film transform do not admit algebraic inverses of this form and would preclude lossless HDR recovery at the decoder with only two transmitted scalars per frame.

End-to-End Pipeline

After tone mapping, the 10-bit SDR stream $\{F_t^{\text{SDR}}\}$ follows the CASVC–CAPO processing sequence defined in Sections 4.2.2–4.3.2, including VVenC key frames and residuals, RIFE context generation, differential latent coding, and the 643-scalar per-GOP decoder update. The only additional operation is the coding of the tone-mapping metadata, so the four-part CAPO rate budget in Eqn. (4.11) becomes

$$R_{\text{total}} = R_{\text{key}} + R_{\text{latent}} + R_{\text{param}} + R_{\text{residual}} + R_{\text{meta}}, \quad (5.5)$$

where R_{meta} is the bitrate of the CABAC-coded HDR metadata sidestream $\{(e_t, \gamma_t)\}_{t=0}^T$, which contributes negligible bitrate at typical frame rates and video bitrates. The VVenC key-frame and residual sub-streams are standard SDR VVC bitstreams; no HDR-specific Supplemental Enhancement Information (SEI) messages are required, since HDR recovery at the receiver is achieved entirely through the algebraic inverse tone-mapping driven by the metadata sidestream.

Semantic Autoencoder and Content-Adaptive Parameter Overfitting

The HDR model retains the tapered autoencoder in Fig. 4.10 and the parameter set defined in Section 4.3.2 without architectural changes. For each GOP, it applies the objective in Eqn. (4.15) to the tone-mapped frames F_{m+k}^{SDR} . The same 643 Learnable

Scale and `Conv2DT` bias parameters, regularisation weight, optimiser settings, and early-stopping rule are used; only the training domain differs.

Training Procedure and Entropy Coding

The HDR model is a separate trained instance from the CASVC–CAPO model of Chapter 4. The standard CASVC–CAPO model is trained on 8-bit SDR clips from the Vimeo-90k dataset [175], which represents the statistics of conventional 8-bit SDR video. The HDR model shares the same architecture, training objective, and hyperparameters, but is trained from scratch on 10-bit SDR clips derived from the HDRVD2K dataset [23] by applying the same automatic Reinhard tone-mapping of Section 5.2.1 to the raw linear HDR EXR frames. This produces training data whose bit depth and statistical distribution match exactly the 10-bit SDR stream seen by the codec at inference time, eliminating any residual domain mismatch that would arise from applying the Vimeo-90k-trained model to tone-mapped HDR content. Because the two models are independent instances, retraining for HDR does not affect the performance of the Vimeo-90k-trained CASVC–CAPO model; all SDR evaluations in Chapter 4 use the original Vimeo-90k model without modification. Training uses the context-aware objective referenced in Section 4.3.2, with mini-batch size $B = 8$, Adam optimisation, and a progressive learning-rate schedule: 1×10^{-4} for the first 40,000 iterations, 5×10^{-5} for the next 15,000 iterations, and 1×10^{-5} for the final 20,000 iterations, with early stopping on the HDRVD2K validation split.

Three entropy-coding paths are employed for the bitstream sub-streams. The VVC syntax elements in the key-frame and residual sub-streams are handled by VVenC’s standard CABAC engine [84]. The per-frame tone-mapping parameters (e_t, γ_t) are entropy-coded as the HDR metadata sidestream via a conventional CABAC encoder, contributing negligible bitrate at typical frame rates. The learned components, specifically the differential semantic latents and the per-GOP parameter update Δ_g^* , are handled by the unified differential PCA–DeepCABAC pipeline [147, 148]: the semantic latent of each frame is coded as the difference between the latent of the original SDR frame and the latent of the RIFE-interpolated frame, and the per-GOP parameter

update is projected onto its dominant principal components, quantised with a coarse uniform step, and entropy-coded with DeepCABAC [148].

5.2.2 Results and Discussion

Experimental Setup

The proposed framework is evaluated on two standard 1080p HDR benchmark sets. The HDM-HDRV cinematic dataset provides naturally captured HDR content with a WCG and high peak luminance, and is widely used to benchmark HDR/WCG codecs in the broadcast research community. The JVET HDR/WCG common test condition sequences are the reference material used by the JVET to validate HDR coding tools in HEVC and VVC, and represent the most operationally relevant HDR benchmark for a VVC-anchored evaluation. Both datasets are described in [23]. All source sequences are provided as scene-linear HDR EXR frames at 1080p resolution. VTM 23.0 is used here, rather than VTM 17.0, because the HDR study was carried out later and follows the then-current JVET HDR common test conditions and reference software. As in the 4K evaluation, the BD-rate results in this chapter are interpreted only against the same VTM 23.0 anchor.

The proposed system processes each source sequence as follows. The linear HDR EXR frames are read directly as float32 RGB, requiring no transfer function inversion. The automatic Reinhard tone-mapping of Section 5.2.1 then converts the linear HDR RGB to a 10-bit SDR frame. The resulting 10-bit SDR stream is encoded by the CASVC-CAPO pipeline (Section 4.3) with intra period = -1 (random access), GOP size = 32, and $QP \in \{27, 32, 37, 42\}$ applied uniformly. The reference anchor is VTM 23.0 configured at the same four QP operating points; for VTM evaluation, the EXR frames are first converted to 10-bit BT.2020/PQ YUV420 format following the JVET HDR common test conditions [67]. All semantic encoder and decoder and per-GOP overfitting operations are performed on a single NVIDIA A100 GPU.

Quality Metrics. Two complementary families of metrics are reported. SDR-domain quality of the tone-mapped 10-bit SDR reconstruction is assessed by PSNR (Y component), SSIM [164] (Y component), and VMAF. HDR-domain perceptual quality

is reported through PU-PSNR and PU-SSIM, both computed in the linear HDR domain against the original EXR ground-truth frames using PU21 perceptual encoding [27]. For the proposed method, the reconstructed 10-bit SDR frame is first converted back to a linear HDR EXR by applying the exact closed-form inverse of the tone-mapping (inverse gamma, inverse Reinhard $R^{-1}(y) = y/(1-y)$, and division by the transmitted exposure multiplier e_t); the recovered linear HDR EXR and the original EXR reference are then both encoded with PU21, and PSNR and SSIM are evaluated on the PU21-encoded values. For VTM 23.0, the decoded 10-bit BT.2020/PQ YUV420 output is converted to a linear HDR EXR through the full inverse pipeline: code-value dequantisation, bilinear 4:2:0 chroma upsampling, inverse BT.2020 Y'CbCr matrix ($K_r=0.2627$, $K_b=0.0593$), and inverse PQ EOTF (ST.2084) [149]; the resulting linear EXR is compared against the same original EXR reference using the same PU21 encoding. This ensures that PU-PSNR and PU-SSIM measure distortion in an identical perceptual linear-light domain for both systems, providing a directly comparable evaluation. For the VTM 23.0 anchor, conventional PSNR, SSIM, and VMAF are additionally computed in the PQ-coded domain following the standard JVET anchor evaluation practice. BD-rates are computed in the Bjøntegaard sense [162] against VTM 23.0 at the four QP operating points; negative values denote bitrate savings at matched quality.

Rate–Distortion Performance

The average rate–distortion curves for both datasets are presented in Figs. 5.2 and 5.3. On every dataset and for every metric, the proposed method’s rate–distortion (RD) curve lies strictly above the VTM 23.0 anchor, confirming consistent bitrate savings at matched quality across both SDR-domain and HDR-domain perceptual evaluation criteria.

The per-dataset BD-rate figures are summarised in Tables 5.1 and 5.2, and the average across both datasets is reported in Table 5.3.

Averaged across both datasets, the proposed codec achieves BD-rate savings of -16.42% in PSNR, -14.14% in SSIM, -10.27% in VMAF, -15.31% in PU-PSNR, and -11.12% in PU-SSIM against VTM 23.0. The HDM-HDRV dataset yields large

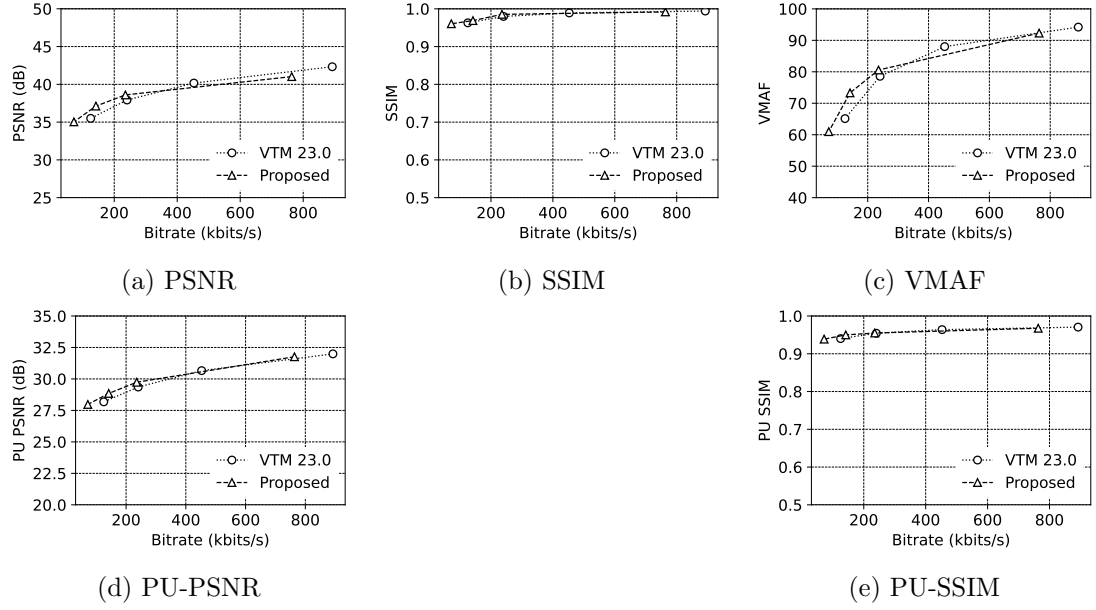


Figure 5.2: Average rate-distortion performance of the proposed method and VTM 23.0 on the HDM-HDRV dataset [23]. Subfigures (a)–(c) show SDR-domain metrics (PSNR, SSIM, VMAF) computed on the tone-mapped 10-bit SDR output; subfigures (d)–(e) show PU21-encoded HDR perceptual metrics (PU-PSNR, PU-SSIM) computed on the recovered linear HDR signal. Each curve connects four operating points at $QP \in \{27, 32, 37, 42\}$.

Table 5.1: BD-rate (%) of the proposed method against VTM 23.0 on the HDM-HDRV dataset [23]. PSNR, SSIM, and VMAF are computed on the tone-mapped 10-bit SDR output; PU-PSNR and PU-SSIM are computed on the recovered HDR signal via PU21 encoding [27]. Negative values denote bitrate savings (lower is better).

Method	PSNR	SSIM	VMAF	PU-PSNR	PU-SSIM
VTM 23.0	0	0	0	0	0
Proposed	−16.20%	−14.52%	−13.35%	−16.60%	−12.56%

Table 5.2: BD-rate (%) of the proposed method against VTM 23.0 on the JVET-HDRV dataset [23]. PSNR, SSIM, and VMAF are computed on the tone-mapped 10-bit SDR output; PU-PSNR and PU-SSIM are computed on the recovered HDR signal via PU21 encoding [27]. Negative values denote bitrate savings (lower is better).

Method	PSNR	SSIM	VMAF	PU-PSNR	PU-SSIM
VTM 23.0	0	0	0	0	0
Proposed	−16.65%	−13.76%	−7.18%	−14.02%	−9.68%

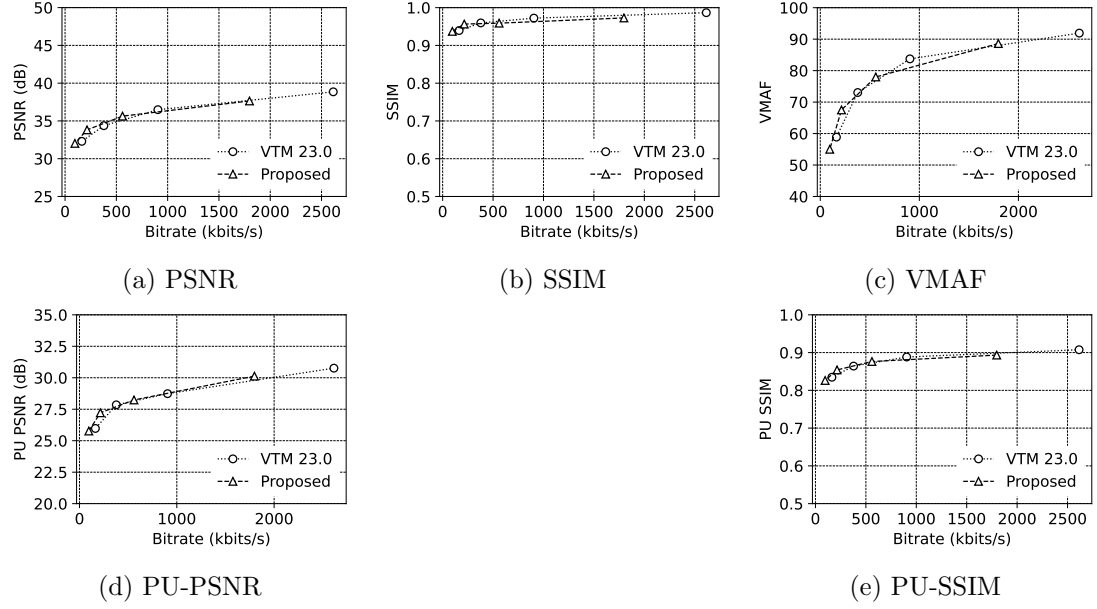


Figure 5.3: Average rate-distortion performance of the proposed method and VTM 23.0 on the JVET-HDRV dataset [23]. Subfigures (a)–(c) show SDR-domain metrics computed on the tone-mapped 10-bit SDR output; subfigures (d)–(e) show PU21-encoded HDR perceptual metrics. Each curve connects four operating points at $QP \in \{27, 32, 37, 42\}$.

Table 5.3: Average BD-rate (%) across HDM-HDRV and JVET-HDRV datasets against VTM 23.0. Negative values denote bitrate savings (lower is better).

Method	PSNR	SSIM	VMAF	PU-PSNR	PU-SSIM
VTM 23.0	0	0	0	0	0
Proposed	−16.42%	−14.14%	−10.27%	−15.31%	−11.12%

PSNR and PU-PSNR gains (-16.20% and -16.60% respectively), while the JVET-HDRV sequences yield comparable PSNR gains (-16.65%) but a more modest VMAF gain (-7.18%). This is consistent with the JVET-HDRV sequences containing a higher proportion of quasi-static studio-style content where the VMAF perceptual model is less sensitive to the structured prediction error that the per-GOP overfitting mechanism eliminates. The consistency of the savings across both SDR-domain metrics (PSNR, SSIM, VMAF) and HDR-domain perceptual metrics (PU-PSNR, PU-SSIM) confirms that the closed-form inverse tone-mapping faithfully recovers the linear HDR signal at the receiver and that the quality gains are not an artefact of SDR-domain evaluation.

Ablation: Per-Frame versus Fixed Tone-Mapping Parameters

A central design choice of the proposed framework is that the exposure multiplier e_t and the gamma exponent γ_t are estimated independently for every frame rather than set to fixed values for the entire sequence. To validate this choice, Fig. 5.4 compares the automatic per-frame estimator against a fixed baseline in which a single pair $(e, \gamma) = (500, 2.0)$ is applied uniformly to every frame of an HDM-HDRV test sequence. Four representative frames are shown: frame 1 (a high-luminance scene) and frames 21, 41, and 61 (low-luminance scenes).

Frame 1 has high log-average luminance, and the automatic estimator correctly sets a small multiplier ($e_1 = 6.40$) to prevent clipping of the highlights. The fixed multiplier of $e = 500$ is approximately $78\times$ larger than this automatic value, saturating frame 1 entirely to white as visible in the top row of Fig. 5.4. Frames 21, 41, and 61 contain dark content with low log-average luminance, requiring large exposure multipliers ($e_t \approx 950\text{--}1010$) to lift the signal into the mid-tone range. The fixed multiplier of 500 is approximately half the required value for these frames, leaving them severely underexposed and compressing the shadow detail that the semantic encoder relies upon to distinguish foreground from background. The automatic per-frame estimator resolves both failure modes simultaneously: by anchoring every frame at the same perceptual grey level via $e_t = 0.18/L_{w,t}$ and computing γ_t in closed form to place the mean tone-mapped luminance at 0.45, the operator eliminates the systematic exposure mismatch

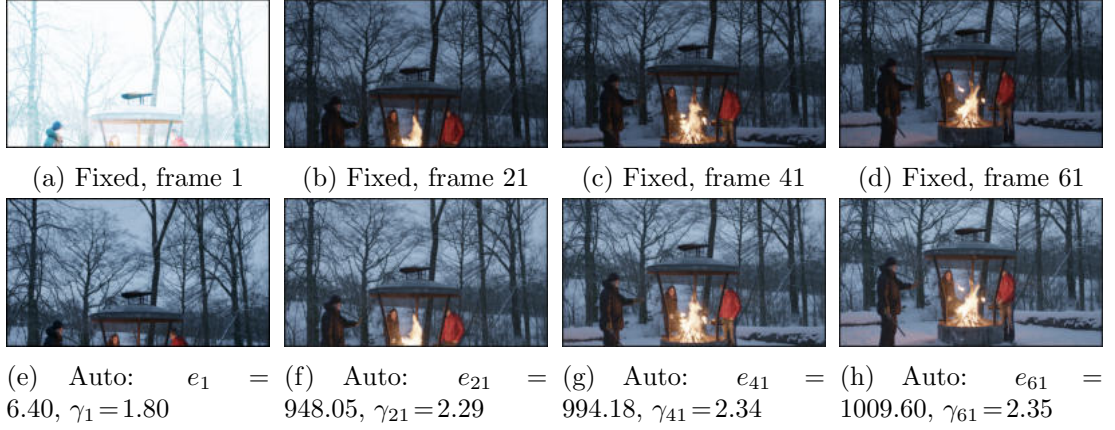


Figure 5.4: Ablation of per-frame versus fixed tone-mapping parameters on four representative frames of an HDM-HDRV sequence. **Top row:** Fixed baseline ($e = 500, \gamma = 2.0$) applied uniformly to all frames. Frame 1 (high-luminance scene) is saturated to white because the fixed multiplier is approximately $78\times$ larger than the automatic estimate ($e_1 = 6.40$). Frames 21, 41, and 61 (low-luminance scenes) are underexposed because the fixed multiplier is roughly half the required value ($e_t \approx 950\text{--}1010$). **Bottom row:** Automatic per-frame estimator. Each frame is individually normalised to the correct mid-tone level, preserving highlight detail in the bright scene and shadow detail in the dark scenes.

and keeps every SDR output frame inside the training distribution of the semantic autoencoder, regardless of absolute HDR scene luminance.

Ablation: Luminance and Chrominance Reconstruction

To determine whether the framework reconstructs colour information as well as luminance, Table 5.4 reports dataset-average peak signal-to-noise ratio (PSNR) separately for the Y, U, and V components of the reconstructed 10-bit YUV420 SDR intermediate. The comparison uses the same four quantisation parameter (QP) operating points as the main evaluation.

PSNR decreases smoothly for all three components as QP increases, showing that the expected quantisation response is present in both luminance and chrominance. Across the four operating points, average U-PSNR and V-PSNR are 41.0696 dB and 41.8835 dB, respectively, compared with 35.7704 dB for Y-PSNR. Thus, the codec does not obtain its overall reconstruction quality by preserving luminance while neglecting chrominance. The higher chrominance PSNR values should not be interpreted

Table 5.4: Component-wise reconstruction quality on the evaluated HDR dataset in 10-bit YUV420 format. Higher PSNR is better.

QP	Y-PSNR (dB)	U-PSNR (dB)	V-PSNR (dB)
27	40.4683	43.9230	44.8457
32	37.1067	41.9552	42.7730
37	34.1142	39.9840	40.7013
42	31.3922	38.4160	39.2141
Average	35.7704	41.0696	41.8835

as greater perceptual colour accuracy because U and V are spatially subsampled in YUV420 and usually contain less high-frequency energy than Y. The ablation establishes good reconstruction of the chrominance samples that remain in the BT.709 intermediate, but it does not demonstrate exact preservation of the original BT.2020 gamut. A native BT.2020 path with explicit gamut mapping and comparison against 4:4:4 coding is therefore required to assess wide-gamut colour preservation independently of chroma subsampling [68, 201].

Computational Complexity

The computational complexity of the proposed framework is dominated by the per-GOP overfitting solve and the RIFE context generation, both of which are inherited unchanged from CASVC-CAPO (Section 4.3). The automatic Reinhard tone-mapping pre-process, comprising per-frame log-average luminance estimation, exposure computation, and gamma estimation, consists entirely of closed-form algebraic operations applied independently to each frame and contributes less than 0.1% of the total encoder computational budget on an NVIDIA A100 GPU.

The proposed framework achieves an encoder throughput of 0.17 fps and a decoder throughput of 16.4 fps on 1080p HDR sequences. The encoder throughput is marginally lower than CASVC-CAPO (0.18 fps on SDR) because the tone-mapping overhead is absorbed within the per-GOP scheduling. The decoder recovers each frame through a single forward pass of the patched semantic decoder, one RIFE interpolation, dequantisation and addition of the VVenC residual, and the three-step closed-form inverse tone-mapping; together these steps sustain the 16.4 fps decoder rate. The heavy op-

timisation burden of the per-GOP overfitting solve is placed entirely at the encoder, which operates offline, while the lightweight decoder is well-suited to near real-time playback applications.

Discussion and Limitations

The rate–distortion and ablation results support three principal conclusions.

Per-frame invertible tone-mapping is sufficient to bridge an HDR source to an SDR-trained hybrid semantic codec. The proposed framework reuses every component of CASVC–CAPO (Section 4.3) without modification and achieves consistent BD-rate savings of -16.42% , -14.14% , -10.27% , -15.31% , and -11.12% in PSNR, SSIM, VMAF, PU-PSNR, and PU-SSIM respectively against VTM 23.0. The gains are consistent across both SDR-domain and HDR-domain metrics, confirming that the closed-form inverse tone-mapping preserves the linear HDR signal at broadcast fidelity.

Per-frame estimation of the tone-mapping parameters is necessary, not optional. The ablation demonstrates that a fixed pair $(e, \gamma) = (500, 2.0)$, chosen at the centre of a typical HDR luminance range, fails on heterogeneous HDR content: the bright frame is saturated to white and the dark frames are severely underexposed. The automatic per-frame estimator, by anchoring every frame at the same perceptual grey level, eliminates this systematic mismatch and keeps the SDR output statistics inside the autoencoder’s training distribution at all times.

The HDR metadata sidestream and the standard 10-bit SDR VVC bitstream are jointly sufficient for closed-form HDR recovery. The proposed framework carries only two scalars (e_t, γ_t) per frame in a CABAC-coded sidestream of negligible bitrate. The inverse Reinhard curve, the inverse gamma correction, and the inverse exposure scaling recover the linear HDR float32 signal from the decoded SDR reconstruction, with the only lossy step being the 10-bit SDR quantisation applied at the encoder.

Several limitations remain. The proposed framework inherits the encoder-side computational overhead of CASVC–CAPO, making the system better suited to offline encoding workflows such as archival compression and video-on-demand than to real-time HDR streaming. The Reinhard operator is a global per-channel curve and may under-

utilise the local-contrast headroom of bright HDR scenes; piecewise-invertible local operators with closed-form inverses and a small number of per-frame parameters represent a natural direction for future investigation. Direct comparison against learned HDR codecs and against AV2-based HDR pipelines [85] is left to future work.

5.3 Chapter Summary and Conclusion

This chapter has presented a contribution that extends the semantic video coding systems developed in Chapter 4 to the HDR domain, fulfilling Objective 8 of the thesis. The main finding is that the gap between linear HDR EXR source material and the SDR training distribution of the CASVC–CAPO semantic autoencoder can be bridged by a simple, non-architectural step: an automatic, per-frame, fully invertible photographic tone-mapping that maps each linear HDR EXR frame to a ten-bit SDR frame whose statistics are anchored at the same perceptual grey level, regardless of absolute HDR scene luminance. The two scalar tone-mapping parameters per frame are transmitted as a compact CABAC-coded metadata sidestream at negligible bitrate, and the receiver recovers the linear HDR float32 signal through closed-form inversion of the Reinhard curve and the gamma correction. All components of CASVC–CAPO are reused without modification, and the semantic autoencoder is retrained once offline on tone-mapped clips from the HDRVD2K HDR video dataset [23], removing any per-sequence or per-GOP HDR-specific training cost.

The framework was evaluated on the HDM-HDRV cinematic dataset [23] and the JVET HDR/WCG common test condition sequences [23] against VTM 23.0 under the JVET random-access common test conditions, achieving average Bjøntegaard delta-rate savings of -16.42% in PSNR, -14.14% in SSIM, -10.27% in VMAF, -15.31% in PU-PSNR, and -11.12% in PU-SSIM. The gains are consistent across both SDR-domain metrics (PSNR, SSIM, VMAF on the tone-mapped ten-bit SDR output) and HDR-domain metrics (PU-PSNR and PU-SSIM on the linear HDR signal recovered through closed-form inverse tone-mapping), confirming that the system achieves broadcast-quality fidelity in both domains. The ablation study confirms that per-frame estimation of the tone-mapping parameters is necessary: a fixed-parameter baseline fails on HDR

content with varying luminance by saturating bright frames and underexposing dark frames, whereas the automatic per-frame estimator keeps stable SDR statistics across all luminance conditions and keeps every tone-mapped frame within the autoencoder’s training distribution.

The main limitation of the framework is the encoder-side computational cost inherited from the CASVC–CAPO per-GOP overfitting mechanism, which limits the system to offline encoding workflows. Future work will look at piecewise-invertible local tone-mapping operators with closed-form inverses and a small number of per-frame parameters, joint learning of the tone-mapping parameters and the autoencoder, and encoder-side speed improvements via warm-started optimisation and early-exit scheduling. Comparison against learned HDR codecs [23] and against royalty-free HDR pipelines based on AV2 [85] is a further priority for future work.

Chapter 6 addresses a related aspect of robust semantic communication by introducing path diversity through multiple description coding, targeting the single-path failure modes of the hybrid semantic video coding systems developed in the previous contribution chapters.

Chapter 6

Robust Semantic Communication via Multiple Description Coding

6.1 Chapter Introduction

The rapid growth of video and image traffic over wireless networks, with projected peak data rates above 1 Tbps for sixth-generation (6G) systems [29–31], has increased the need for robust visual communication systems that maintain perceptual quality under changing wireless channel conditions. Conventional compression standards such as H.264/AVC [36], H.265/HEVC [37], and H.266/VVC [177], as well as learned image compression frameworks, achieve high rate–distortion efficiency under the assumption of error-free transmission. Their tightly coupled entropy-coded bitstreams, however, degrade badly when channel noise or packet loss passes a critical level. Semantic communication, realised through Deep JSCC systems [59,60,62,63], addresses this problem by learning latent representations that are robust to noise and degrade smoothly as channel conditions worsen. Yet all existing single-path Deep JSCC systems share the same structural weakness identified in Chapter 1: they send one latent representation over one wireless path. If that path suffers a deep fade, shadowing, or persistent narrowband interference, the entire semantic representation is lost regardless of its noise robustness, so a severe failure mode returns at the path level.

MDC provides the theoretical basis for handling such path-level failures. Introduced

by El Gamal and Cover [71] and surveyed by Goyal [70], MDC encodes a source into several independent but correlated descriptions so that any received subset can still be decoded, with quality improving as more descriptions arrive. Classical MDC based on multiple description scalar quantisation [152] and later video extensions [155] showed the value of this idea for lossy packet networks. Neural image MDC introduced learned feature and latent description generation [156, 157, 202, 203]. These methods establish strong source-coding and missing-description baselines, but do not jointly evaluate semantic complementarity, explicit channel coding, and physical fading across image and video transmission. The gap addressed here is therefore a channel-aware end-to-end framework for complementary semantic descriptions over path-diverse wireless links.

The proposed framework differs from the closest work in three main ways. First, compared with single-path Deep JSCC systems [62, 63], it adds structural path diversity. Second, compared with learned image MDC [156, 157, 202, 203], it distributes features from a shared semantic latent across three descriptions, trains one fusion decoder for every non-empty received subset, and extends the same design from images to context-conditioned video. Third, compared with classical MDC video systems [155], the proposed framework jointly trains semantic description generation and subset-aware fusion rather than adding error-resilient source coding to a fixed conventional codec. The distinctive contribution is the combination of these properties with separately entropy-coded and polar-protected descriptions within one physical-channel evaluation framework.

This chapter presents two contributions that fill this gap. The first, SMDC for Image Transmission (Section 6.2), proposes an end-to-end convolutional autoencoder with an interleaved channel-wise splitting strategy that generates three semantically complementary latent descriptions from a shared $9 \times 16 \times 24$ bottleneck tensor. A two-stage training procedure ensures that each description alone carries enough semantic information for a coherent reconstruction, while all three together achieve near-optimal quality. The framework is evaluated against JPEG [48], JPEG2000 [74], HEIF [75], and CompressAI [76] baselines under the same three-path transmission conditions across AWGN, Rayleigh, and Rician fading channels. The key contributions are:

1. *Learned semantic multiple description coding*: A deep learning architecture that generates three semantically complementary descriptions of an image, directly optimising the trade-off between central and side distortion in the semantic latent space.
2. *Interleaved channel-wise splitting*: A deterministic method that assigns every third latent channel to each description, so that each description samples feature maps uniformly across all spatial positions and frequency sub-bands, giving global semantic coverage from any single received description.
3. *Semantic redundancy*: Use of semantic redundancy as an alternative to bit-level redundancy, ensuring that key semantic features are present in every description with minimal bitrate overhead.
4. *Guaranteed performance floor*: A solution to the single point of failure in Deep JSCC, providing a guaranteed PSNR floor above 29 dB from any single received description across all evaluated channel conditions.
5. *Fusion decoder*: A neural network decoder that uses learned inter-description correlations to reconstruct the image from any received subset under severe and uneven channel degradation across three independent paths.

The second contribution, Semantic MDC for Robust Video Transmission (Section 6.3), extends the SMDC image framework to video sequences using a Group-of-Pictures structure of size 8. A dedicated MDC Intra-Frame Semantic Encoder handles boundary key frames using the image SMDC architecture, while a context-conditioned MDC Inter-Frame Semantic Encoder encodes only the difference between each intermediate frame’s latent and its RIFE-interpolated temporal context [72], which greatly reduces transmitted data by using inter-frame temporal correlation. Generating all context frames in parallel from the fixed boundary frames removes sequential inter-frame dependencies, allowing fully parallel encoding within each GOP. Building on the context-conditioned semantic video coding framework of Chapter 4, this contribution adds structural path diversity that removes the single-path weakness of the semantic

video coding systems developed in earlier chapters. The key contributions are:

1. *Semantic MDC Video framework*: A complete end-to-end SMDC framework for video transmission that combines intra-frame and inter-frame SMDC modules within a GOP-8 structure.
2. *Context-conditioned inter-frame MDC*: A differential latent encoding strategy that uses RIFE-based temporal interpolation to generate motion-aligned context frames, transmitting only the difference between the inter-frame latent and its context to achieve much lower bitrate than encoding raw inter-frame latents.
3. *Parallel context generation*: An interpolation scheme where all seven inter-frames in a GOP derive their contexts from the same pair of decoded boundary intra-frames, enabling fully parallel inter-frame encoding without sequential dependencies.
4. *Three-path evaluation with combined SNR analysis*: A multi-channel evaluation framework with asymmetric per-path SNR assignment and a Shannon-capacity-based combined SNR metric, giving a clear basis for comparing multi-description and single-path operation.
5. *Robustness gains*: Experimental results on 20 test video sequences show 4–5 dB combined SNR robustness gains over VVenC [84] under the same three-path transmission conditions across AWGN, Rayleigh, and Rician fading channels.

The image framework in Section 6.2, including its complementary descriptions and subset-aware reconstruction, was reported in the first-author publication “Semantic Communication Based Multiple Description Coding for Robust Image Transmission over Wireless Channels” [204]. Section 6.3 incorporates the follow-on study “Semantic Communication Based Multiple Description Coding for Robust Video Transmission over Wireless Channels,” which extends the approach through temporal context and differential latent coding. Together, the studies add path diversity and resilience to the thesis research narrative. Section 6.4 summarises their findings, relates them to the broader thesis objectives, and identifies open research directions.

6.2 Semantic Multiple Description Coding for Robust Image Transmission

Publication Notice

The research presented in this section has been published in the following peer-reviewed publication:

P. Samarathunga, U. Jayasinghe, Y. Ganearachchi, T. Fernando, T. Sivalingam, N. Rajatheva, and A. Fernando, “Semantic Communication Based Multiple Description Coding for Robust Image Transmission over Wireless Channels,” *IEEE Access*, pp. 1–1, 2026, doi: 10.1109/ACCESS.2026.3716070.

This material is reproduced and adapted under the terms of the IEEE Copyright policy for use in a doctoral thesis. The content has been revised and extended for inclusion and integration in this doctoral thesis, including additional explanations, expanded discussion, and alignment with the overarching research narrative.

6.2.1 Introduction

The Chapter Introduction establishes the path-level failure of single-path Deep JSCC, while Section 2.15 reviews the relevant classical and learned MDC methods. Building directly on that foundation, this section addresses the remaining gap: existing methods optimise reconstruction from missing descriptions, whereas SMDC jointly considers semantic latent complementarity, subset-aware decoding, channel coding, and fading-channel robustness.

The proposed SMDC framework addresses this limitation by training a convolutional autoencoder end to end to generate three semantically complementary descriptions from a shared $9 \times 16 \times 24$ latent bottleneck tensor. An interleaved channel-wise splitting strategy ensures that each description captures complementary features across all frequency sub-bands and spatial positions instead of grouping information into one local channel subset. A fusion decoder learns the correlations between descriptions so that it can reconstruct the image from any available subset under uneven channel degradation. Table 6.1 compares the proposed approach with earlier work and identi-

fies SMDC as the first framework to combine semantic representation, path diversity, end-to-end learning, and channel awareness in one jointly trained system.

Table 6.1: Comparison of approaches for robust image transmission over wireless channels.

Approach	Semantic Repr.	Path Diversity	End-to-End (E2E) Learned	Channel Aware
Classical MDC [70, 152]	×	✓	×	×
Deep JSCC [62, 63]	✓	×	✓	✓
Neural MDC [156, 157, 202, 203]	Partial	✓	✓	×
SMDC (Proposed)	✓	✓	✓	✓

The five contributions of this section, as listed in the Chapter Introduction, are the learned SMDC architecture, interleaved channel-wise splitting strategy, semantic redundancy formulation, guaranteed performance floor, and fusion decoder; their implementation is detailed in the following subsections.

6.2.2 Methodology

System Overview

The proposed SMDC framework combines semantic communication with path diversity through end-to-end neural network training to solve the single point of failure found in existing Deep JSCC systems [62, 63]. It does this by generating three semantically complementary descriptions of the input image and sending each description over a separate wireless path, so that even one surviving description can still produce a semantically coherent reconstruction.

The overall architecture is shown in Fig. 6.1. An input image $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ with $H = 144$, $W = 256$, and $C = 3$ is processed by the SMDC semantic encoder, which maps it to a latent tensor of size $9 \times 16 \times 24$ and then splits it deterministically through interleaved channel-wise splitting into three descriptions $\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3 \in \mathbb{R}^{9 \times 16 \times 8}$. Each description is entropy coded with DeepCABAC [148], protected with polar codes [205], and modulated before being sent over an independent wireless channel. At the receiver, a fusion decoder reconstructs the image from the available subset $\mathcal{A} \subseteq \{1, 2, 3\}$ of

received descriptions using learned inter-description correlations.

Importantly, the SMDC pipeline follows the classical separation principle. Semantic source coding through the autoencoder, entropy coding through DeepCABAC, channel coding through polar codes, and modulation through Binary Phase Shift Keying (BPSK) or Differential Binary Phase Shift Keying (DBPSK) are applied as separate, sequential stages. This is different from JSCC systems such as DeepJSCC [62], which map image pixels directly to channel symbols through a single neural network, with no separate entropy coding, channel coding, or modulation stage. Because SMDC produces discrete bitstreams and JSCC produces continuous channel symbols, the two approaches use different rate measures: bits per pixel for SMDC and channel uses per pixel for JSCC. A direct comparison at the same operating point would require conversion assumptions that have no clear theoretical basis. For this reason, Deep JSCC and similar JSCC systems are not included as baselines; the evaluation instead compares against separation-based methods (JPEG, JPEG2000, HEIF, and CompressAI) that operate in the same bitstream-based framework as SMDC.

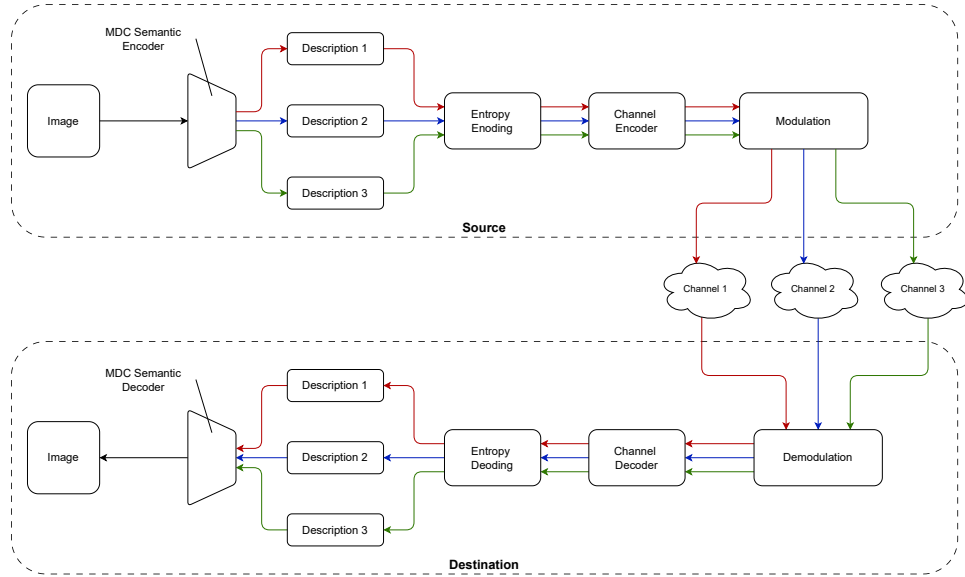


Figure 6.1: End-to-end SMDC pipeline. The three colours represent the three independent channel paths for transmitting the three descriptions.

SMDC Autoencoder and Interleaved Channel-Wise Splitting

The SMDC autoencoder is a CNN inspired by learned compression and variational transforms [47, 51]. The encoder consists of four downsampling blocks, each comprising a strided convolution ($\times 2$ spatial reduction), batch normalisation, and LeakyReLU activation with negative slope 0.01, followed by residual connections to stabilise training and improve gradient flow [51]. The four downsampling stages reduce the spatial dimensions from 144×256 to 72×128 , 36×64 , 18×32 , and 9×16 while maintaining 128 feature channels. A bottleneck stage maps the features to 24 channels, producing the latent tensor $\mathbf{Z} \in \mathbb{R}^{9 \times 16 \times 24}$.

The semantic encoder output is formally defined as

$$\{\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3\} = \mathcal{E}_{\text{sem}}(\mathbf{X}), \quad (6.1)$$

where $\mathbf{D}_i \in \mathbb{R}^{9 \times 16 \times 8}$ for $i \in \{1, 2, 3\}$ are the three latent descriptions, and the brace notation denotes the ordered collection indexed by $i = 1, 2, 3$. The three descriptions are generated by an interleaved channel-wise splitting of the 24-channel bottleneck tensor: Description 1 receives channel indices $\{0, 3, 6, 9, 12, 15, 18, 21\}$, Description 2 receives $\{1, 4, 7, 10, 13, 16, 19, 22\}$, and Description 3 receives $\{2, 5, 8, 11, 14, 17, 20, 23\}$. This interleaved assignment ensures that every description samples feature maps uniformly across all 9×16 spatial positions and all frequency sub-bands, providing each description with global semantic coverage of the full image extent rather than a localised subset of channels.

The decoder accepts any non-empty subset $\mathcal{A} \subseteq \{1, 2, 3\}$ of the received descriptions, zero-padding missing descriptions to form a $9 \times 16 \times 24$ input tensor, and applies four transposed-convolution upsampling blocks mirroring the encoder to produce the reconstruction $\hat{\mathbf{X}} \in \mathbb{R}^{144 \times 256 \times 3}$. A final 1×1 convolution with Sigmoid activation yields the output RGB image. For larger input resolutions, the framework partitions the image into non-overlapping $144 \times 256 \times 3$ blocks, encodes each independently, and applies a bilateral post-filter [206] ($d = 9$, $\sigma_{\text{color}} = 75$, $\sigma_{\text{space}} = 75$) along a four-pixel boundary band at every block seam to suppress reconstruction discontinuities. This

lightweight post-filter is sufficient because the objective is to restore semantic continuity across neighbouring blocks rather than enforce pixel-perfect seam consistency.

The architecture is illustrated in Fig. 6.2.

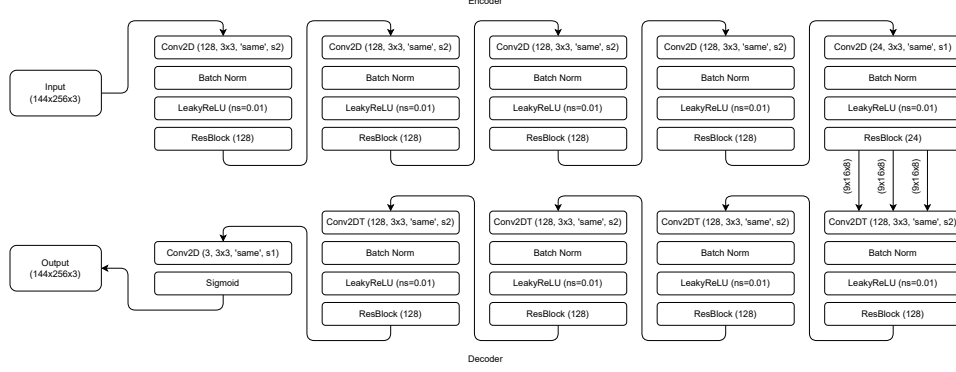


Figure 6.2: SMDC autoencoder architecture with interleaved channel-wise split into three descriptions. The encoder downsamples the input to a $9 \times 16 \times 24$ latent tensor; the decoder upsamples a concatenation (or zero-padded subset) of the three $9 \times 16 \times 8$ descriptions.

Two-Stage Training Strategy

The SMDC autoencoder is trained on the 800-image training split (IDs 0001–0800) of the DIV2K dataset [207] and evaluated on the disjoint 100-image validation split (IDs 0801–0900). Input images are cropped into uniform $144 \times 256 \times 3$ blocks to match the network input dimensions and ensure consistent processing throughout training.

Stage 1: Main Autoencoder Training. The entire autoencoder is jointly optimised using a rate–distortion loss

$$\mathcal{L}_{\text{AE}} = \mathbb{E}_{\mathbf{X}} \left[\|\mathbf{X} - \hat{\mathbf{X}}\|^2 \right] + \lambda R, \quad (6.2)$$

where $\mathbb{E}_{\mathbf{X}}[\cdot]$ denotes the expectation over training images, λ is the rate–distortion trade-off parameter, and R is the estimated bitrate in bits per pixel. The network is trained to reach a target reconstruction quality in the 30–35 dB PSNR range. PSNR follows the thesis-wide definition in Eqn. (2.2); for the unit-normalised inputs used here, its peak

signal value is $L = 1$. Training employs a batch size of 8 with shuffled mini-batches on a Google Colab A100 GPU to improve generalisation and reduce overfitting, and uses a progressive learning-rate schedule: 1×10^{-4} for the first 25,000 iterations, 5×10^{-5} for a further 15,000, and 1×10^{-5} for the final 10,000 to refine the learned representation and stabilise convergence.

Stage 2: Decoder Refinement with Quantisation. Following Stage 1, the decoder is fine-tuned to handle quantised encoder outputs. The continuous bottleneck tensor $\mathbf{Z} \in \mathbb{R}^{9 \times 16 \times 24}$ is quantised channel-wise as $\mathbf{Z}_q = \lfloor \mathbf{Z} + 0.5 \rfloor$, where the floor operation is applied elementwise, yielding three $9 \times 16 \times 8$ descriptions. The decoder training loss is defined over all seven non-empty subsets of descriptions,

$$\mathcal{L}_{\text{Dec}} = \sum_{\mathcal{A} \in \mathcal{P}(\{1,2,3\}) \setminus \{\emptyset\}} w_{\mathcal{A}} \|\mathbf{X} - \mathcal{D}_{\text{sem}}(\{\mathbf{D}_i\}_{i \in \mathcal{A}})\|^2, \quad (6.3)$$

where $\mathcal{P}(\cdot)$ denotes the power set, $\mathcal{D}_{\text{sem}}(\cdot)$ denotes the SMDC semantic decoder, and $w_{\mathcal{A}}$ are non-negative weights reflecting the probability of receiving subset $\mathcal{A}$. This training explicitly covers all reception cases: a single description produces a coarse but meaningful reconstruction, any pair improves fidelity, and all three descriptions recover the highest quality output. This subset-aware training compels the encoder, through the shared rate-distortion objective, to distribute complementary semantic content across all three descriptions: each description must independently carry sufficient spatial and semantic coverage for meaningful reconstruction, since the loss penalises failure under every reception scenario simultaneously.

Robustness Augmentation. To enhance resilience against channel impairments, each description is segmented into 16×1 blocks and random block erasures are simulated during training with probability p_{loss} :

$$\tilde{\mathbf{D}}_i = \mathbf{D}_i \odot \mathbf{M}_i, \quad \mathbf{M}_i \sim \text{Bernoulli}(1 - p_{\text{loss}}), \quad (6.4)$$

where $\odot$ denotes element-wise multiplication and $\mathbf{M}_i \in \{0, 1\}^{9 \times 16 \times 8}$ is a binary block mask whose entries are sampled independently at the block level and then applied uniformly within each 16×1 block. This augmentation explicitly conditions the decoder to reconstruct missing spatial blocks using semantic redundancies from the remaining descriptions. Overall, the two-stage training strategy ensures that the SMDC autoencoder is jointly compression-efficient under normal operation and resilient when only partial or corrupted descriptions are received.

Entropy Coding, Channel Coding, and Modulation

Each description is independently entropy-coded with DeepCABAC [148], using a block partitioning of 16×1 to localise errors and limit burst impact. The DeepCABAC configuration uses step size 1.0 and interval 0.1. Although this introduces modest overhead per block, it improves robustness to residual channel errors by limiting the propagation of local decoding failures. The encoding and decoding pipeline for description i is

$$\mathbf{B}_i = \mathcal{H}(\mathbf{Q}(\mathbf{D}_i)), \quad (6.5)$$

$$\hat{\mathbf{D}}_i = \mathcal{H}^{-1}(\hat{\mathbf{B}}_i), \quad (6.6)$$

where $\mathcal{H}(\cdot)$ denotes the DeepCABAC entropy encoder, $\mathbf{Q}(\cdot)$ the quantisation operator, and $\mathbf{B}_i$ the compressed bitstream for description i .

Polar codes [205] with information length $K = 164$ bits and codeword length $E = 512$ bits (code rate $R \approx 0.32$) are applied to each description's bitstream for forward error correction. The choice $K = 164$ is determined by the DeepCABAC output frame size, and $E = 512$ is the smallest power-of-two codeword length accommodating K at a rate below $\frac{1}{3}$, providing strong error correction for deep-fading conditions at the cost of moderate bandwidth overhead. Polar forward error correction (FEC) is applied only for the fading-channel experiments; for AWGN experiments it is disabled to isolate the inherent robustness of the semantic multi-description architecture.

BPSK modulation is used for AWGN channels, where the channel introduces no phase rotation and coherent detection achieves the theoretical minimum BER without

differential encoding overhead. DBPSK modulation is used for Rayleigh and Rician fading channels, where rapid random phase rotations make reliable carrier-phase recovery for coherent BPSK impractical; differential encoding encodes each bit as the phase change between successive symbols, naturally cancelling the common rotation without requiring explicit channel estimation.

Channel Models and Combined SNR

Three channel models are evaluated: AWGN, Rayleigh fading (normalised path gains, Jakes Doppler spectrum, maximum Doppler shift 0.001 Hz), and Rician fading (K-factor = 3, normalised path gains, Jakes Doppler spectrum, maximum Doppler shift 0.001 Hz) [208].

The three descriptions are transmitted over independent channels with asymmetric SNRs: Channel 1 operates at high SNR (25 dB to −5 dB), Channel 2 at medium SNR (15 dB to −15 dB), and Channel 3 at low SNR (10 dB to −20 dB). This asymmetric configuration reflects practical wireless deployments where multiple paths experience different propagation losses, interference levels, and antenna gains.

To enable fair comparison with single-path baselines, a capacity-equivalent combined SNR is derived from the Shannon–Hartley theorem [209]. For a three-path system with equal bandwidth allocation $B/3$ per path, the total capacity equals that of a single channel at the combined SNR when

$$\left(\frac{S}{N}\right)_{\text{combined}} = \left[\prod_{i=1}^3 \left(1 + \left(\frac{S}{N}\right)_i\right) \right]^{1/3} - 1, \quad (6.7)$$

or equivalently in decibels,

$$\text{SNR}_{\text{combined}} = 10 \log_{10} \left(\left[\prod_{i=1}^3 \left(1 + 10^{\text{SNR}_i/10}\right) \right]^{1/3} - 1 \right). \quad (6.8)$$

Here, $(S/N)_i$ denotes the linear-scale SNR of path i , and $\text{SNR}_i = 10 \log_{10}((S/N)_i)$ denotes the corresponding value in decibels. This derivation assumes that the three transmission paths are statistically independent, that the total bandwidth is divided

equally across the three paths without cross-path interference, and that the reported SNRs for fading channels represent long-term average received SNRs so that the resulting expression can be interpreted as an ergodic-capacity-equivalent measure. Based on the asymmetric per-channel SNR ranges, the resulting combined SNR spans from 16.76 dB to -9.55 dB.

These three models were selected because together they span the most important range of practical wireless channel conditions [158], and AWGN and Rayleigh fading channels are widely adopted as standard benchmarks in the wireless communications research literature [158]. The AWGN model provides an ideal reference with no multipath effects, establishing an upper performance bound. The Rayleigh model represents the most challenging non-line-of-sight (NLOS) scenario, such as dense urban environments with many independent scatterers and no dominant path, and produces the deepest fades. The Rician model with K-factor 3 represents a partial line-of-sight condition, such as suburban or indoor settings with a dominant path alongside scattered components, and produces fades of intermediate severity. Evaluating across all three models therefore reveals how the SMDC framework behaves across the full spectrum from ideal to worst-case fading.

Frequency-selective fading channels, where the channel coherence bandwidth is smaller than the signal bandwidth and different frequency components fade independently, were not explicitly modelled in this evaluation. This is because frequency-selective fading is strongly dependent on the deployment-specific multipath power-delay profile, which varies across environments and is not captured by a single general model. More importantly, frequency-selective fading is the standard motivation for using Orthogonal Frequency-Division Multiplexing (OFDM) modulation, which is employed in all modern cellular standards from 4G Long-Term Evolution (LTE) through to 5G New Radio [30, 66]. In an OFDM system, the wideband frequency-selective channel is divided into many narrow subcarriers. When the subcarrier spacing is sufficiently narrow relative to the coherence bandwidth of the channel, each subcarrier experiences approximately flat fading and the frequency-selective behaviour of the wideband channel is handled at the modulation layer before the SMDC framework operates [158].

Semantic image transmission over OFDM channels has been demonstrated in related work on Deep Joint Source–Channel Coding [167], where the same flat-fading assumption applies per subcarrier. In an OFDM deployment of the SMDC framework, each description would be mapped onto a subset of OFDM resource blocks, and each resource block would experience flat fading characterised by either Rayleigh or Rician statistics depending on the propagation environment. The channel models evaluated in this thesis therefore remain directly applicable to each subcarrier, and the results generalise to OFDM systems through this equivalence.

Baseline Methods

Four baselines are evaluated under the same three-path architecture: JPEG [48], JPEG2000 [74], HEIF [75], and CompressAI (Cheng2020 attention model) [76]. To match the three-path SMDC transmission setting fairly, the input image $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ is first decomposed into three sub-images through interleaved pixel-wise splitting,

$$\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\} = \mathcal{S}(\mathbf{X}), \quad (6.9)$$

where each sub-image selects every third pixel in raster-scan order,

$$\mathbf{X}_i(p) = \mathbf{X}(3p + i - 1), \quad i \in \{1, 2, 3\}. \quad (6.10)$$

Here, $p \in \{0, 1, \dots, HW/3 - 1\}$ indexes the pixels of each sub-image in raster-scan order. Each sub-image is then encoded independently with the corresponding baseline codec,

$$\mathbf{B}_i = \mathcal{E}_{\text{base}}(\mathbf{X}_i), \quad i \in \{1, 2, 3\}, \quad (6.11)$$

and the resulting bitstreams are transmitted over the same independent channels with identical polar coding and modulation. At the receiver, the reconstruction depends on the received subset $\mathcal{A}$,

$$\hat{\mathbf{X}} = \mathcal{F}(\{\hat{\mathbf{X}}_i\}_{i \in \mathcal{A}}), \quad (6.12)$$

where $\hat{\mathbf{X}}_i = \mathcal{E}_{\text{base}}^{-1}(\hat{\mathbf{B}}_i)$ and $\mathcal{F}(\cdot)$ denotes the fusion rule that performs direct interleaving when $|\mathcal{A}| = 3$, bilinear interpolation when $|\mathcal{A}| = 2$, and bicubic interpolation when $|\mathcal{A}| = 1$. If all three sub-images are received, they are reassembled by direct interleaving; if only two are available, missing pixels are estimated by bilinear interpolation; if only one is available, bicubic interpolation [210] is applied to reconstruct the missing two-thirds of the image. All methods are configured to operate at a matched bitrate of approximately 0.33 bits per pixel. This adaptive reconstruction mechanism gives the baselines the same structural opportunity for graceful degradation as SMDC, so that the comparison isolates the benefit of semantic-domain description generation over pixel-domain splitting rather than any asymmetry in the transmission framework. Conventional MDC image codecs were not included as executable baselines because the established scalar-quantisation, lattice-quantisation, and redundant-frame methods define coding constructions rather than a maintained full-colour reference codec that supports the same three-description bitstream, polar-code, modulation, and fading-channel test chain [152–154]. A new implementation would require design choices for colour conversion, entropy coding, three-description index assignment, and rate allocation that are not fixed by those works, so its result would depend strongly on the reimplementation rather than on the underlying MDC principle. The selected baselines are reproducible deployed or open implementations and are converted into independently decodable descriptions using the same interleaved split and channel protection as SMDC. Consequently, the reported experiment compares semantic-domain splitting with a matched pixel-domain MDC construction; it does not claim a direct rate–distortion comparison with every classical MDC algorithm. Implementing a matched-rate conventional MDC codec and the recent LMDC and INR-MDSQC approaches under the same physical-channel conditions remains useful future work.

6.2.3 Results and Discussion

Experimental Setup

Results are averaged over the 100 images of the DIV2K validation split (IDs 0801–0900) [207], which is fully disjoint from the training set. Reconstruction quality is

assessed by PSNR (dB) and SSIM [164], reported as mean $\pm$ standard deviation over the validation set. The three-path MDC results are presented in Fig. 6.3, and the single-channel baseline results are presented in Fig. 6.4 for direct comparison of the multi-description versus single-description operating modes. For the single-channel experiments, each baseline codec encodes the full image and transmits it over a single path using the conventional separation-based pipeline illustrated in Fig. 6.5.

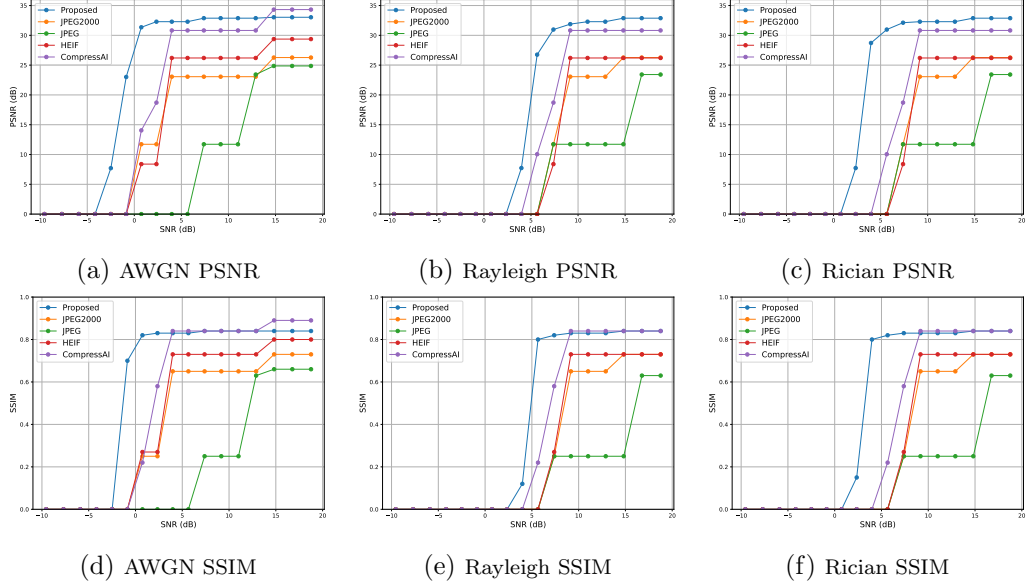


Figure 6.3: Performance comparison of the proposed SMDC framework against baseline methods across three channel models under three-path MDC transmission. (a)–(c) PSNR versus combined SNR for AWGN, Rayleigh, and Rician channels. (d)–(f) SSIM versus combined SNR.

AWGN Channel Results

Figs. 6.3a and 6.3d present PSNR and SSIM performance under AWGN conditions with polar coding disabled, isolating the inherent robustness of the multi-description architecture. At high combined SNR (above 10 dB), learned compression baselines such as CompressAI achieve slightly higher PSNR due to their sophisticated hyperprior entropy models optimised for error-free channels. However, this advantage is strictly confined to the noiseless regime.

Below 5 dB combined SNR, all four baselines fail catastrophically: SSIM collapses

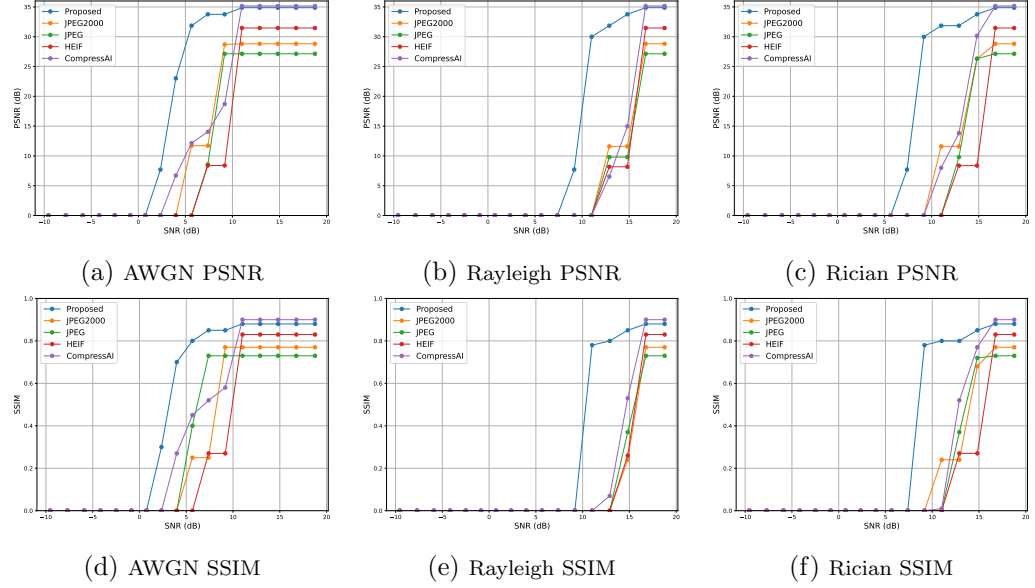


Figure 6.4: Performance comparison across single-channel (non-MDC) configurations. (a)–(c) PSNR versus SNR for AWGN, Rayleigh, and Rician channels. (d)–(f) SSIM versus SNR.



Figure 6.5: Conventional pipeline used for the single-channel (non-MDC) baseline comparison. The entire image is source-coded, channel-coded with polar codes, and transmitted over a single path; the receiver applies the same codec to decode the received bitstream directly without any multi-description fusion.

below 0.2, driven by desynchronisation of their arithmetic decoders from residual bit errors. SMDC, by contrast, maintains PSNR above 30 dB and SSIM above 0.80 down to 1 dB combined SNR, providing a 4 dB robustness advantage over multi-channel baselines and a 10 dB advantage over single-channel baselines. The graceful degradation behaviour is clear throughout: SMDC exhibits smooth, monotonic quality reduction without any abrupt collapse, a property that emerges from the learned semantic redundancy across independent paths and the subset-aware decoder trained over all seven non-empty reception subsets.

Rayleigh and Rician Fading Results

Under Rayleigh fading (Figs. 6.3b and 6.3e), baseline methods fail catastrophically above 12 dB SNR in the multi-channel configuration. SMDC maintains PSNR above 30 dB and SSIM above 0.80 down to 7 dB combined SNR, a 5 dB robustness advantage over multi-channel baselines. The incorporation of polar FEC extends the operational envelope by 3–5 dB relative to the no-FEC AWGN scenario, with SMDC achieving $\text{PSNR} = 31$ dB and $\text{SSIM} \approx 0.82$ at 7 dB combined SNR.

Rician fading results (Figs. 6.3c and 6.3f) exhibit comparable trends with marginally improved performance due to the dominant line-of-sight component (K -factor = 3). The 5 dB combined SNR threshold for robust operation is maintained across both fading scenarios.

An important observation across both fading environments is the SMDC decoder’s ability to use asymmetric channel conditions adaptively. When one or two channels are severely degraded, the learned fusion decoder down-weights corrupted descriptions and relies mainly on higher-quality paths, effectively performing maximum-ratio combining in the semantic domain. This adaptive capability, absent in all single-stream systems, establishes a guaranteed performance floor at approximately 28–30 dB PSNR even when two of three paths fail simultaneously.

Comparing Fig. 6.3 with Fig. 6.4 quantifies the path-diversity gain directly. For AWGN, single-channel SMDC degrades below 30 dB PSNR below 5 dB SNR, while three-path SMDC maintains it down to 1 dB, a 4 dB gain attributable to channel di-

versity. For Rayleigh fading, the corresponding thresholds are 11 dB (single-channel SMDC) versus 7 dB (three-path SMDC). The physical interpretation is that the probability of all three independent channels simultaneously experiencing deep fades is the product of their individual fade probabilities, which is far lower than the single-channel fade probability.

Consolidated Performance Table

Table 6.2 consolidates the minimum combined SNR at which each configuration maintains PSNR above 30 dB and SSIM above 0.80 across all three channel models. In every condition, multi-channel SMDC achieves the lowest threshold, and the gap between multi-channel SMDC and multi-channel baselines (4 dB for AWGN, 5 dB for Rayleigh and Rician) quantifies the benefit of learned semantic redundancy over pixel-domain splitting under identical transmission conditions.

Table 6.2: Minimum combined SNR (dB) to maintain PSNR > 30 dB and SSIM > 0.80 . Lower values indicate greater robustness.

Channel Model	Multi-Channel		Single Channel	
	SMDC	Baseline	SMDC	Baseline
AWGN	1	5	5	11
Rayleigh	7	12	11	16
Rician	5	10	9	14

Ablation Studies

To isolate the contribution of the main architectural and transmission decisions, the ablation analysis addresses five questions: how much robustness is gained from multi-description transmission relative to a single path, what role polar FEC plays beyond the semantic architecture itself, why interleaved splitting is superior to contiguous splitting, whether three descriptions is the best operating point, and why asymmetric channel SNRs are used in the experiments.

Multi-Description versus Single-Description Coding. The comparison between Fig. 6.3 and Fig. 6.4 directly quantifies the benefit of distributing semantic information

across multiple independent paths rather than concentrating it in a single transmission stream. Under AWGN, single-channel SMDC falls below 30 dB PSNR at 5 dB SNR, whereas the three-path configuration maintains the same threshold down to 1 dB combined SNR, yielding a 4 dB gain. The corresponding thresholds are 11 dB versus 7 dB under Rayleigh fading and 9 dB versus 5 dB under Rician fading. This improvement cannot be attributed to bandwidth asymmetry, because the combined-SNR formulation equalises the total transmission capacity. Instead, it arises from path diversity and semantic redundancy: when one path is heavily corrupted, the remaining descriptions still preserve enough global semantic information for the decoder to produce a meaningful reconstruction. The key difference from single-stream transmission is structural. In a single path, a deep fade corrupts one tightly coupled bitstream and causes global reconstruction failure; in the three-path configuration, the probability that all three channels simultaneously undergo severe degradation is the product of their individual fade probabilities, which is substantially smaller. This converts catastrophic failure into graceful degradation and explains why the three-description architecture preserves a stable quality floor even in low-SNR regimes.

Role of Polar Coding. The contrast between the AWGN experiments without polar coding and the Rayleigh/Rician experiments with polar coding clarifies the complementary roles of semantic redundancy and bit-level protection. Under AWGN, SMDC already maintains PSNR above 30 dB and SSIM above 0.80 down to 1 dB combined SNR, showing that the main robustness mechanism is the semantic multi-description architecture itself rather than the channel code. However, fading channels introduce burst errors, deep nulls, and rapid phase changes that are more damaging to the entropy-coded descriptions. In this regime, polar FEC extends the operational envelope by roughly 3–5 dB, allowing the receiver to recover higher-integrity bitstreams before entropy decoding. The two mechanisms therefore operate at different levels: semantic MDC provides coarse-grained structural robustness by ensuring that any surviving description remains meaningful, whereas polar codes provide fine-grained protection within each description by correcting residual bit errors. The observed performance

gains in fading channels depend on both layers working together; MDC alone cannot suppress all residual entropy-decoder failures, and FEC alone cannot prevent outage when an entire path is lost.

Interleaved versus Contiguous Channel-Wise Splitting. To validate the interleaved splitting strategy, performance is compared against a contiguous alternative in which channels $\{0-7\}$, $\{8-15\}$, and $\{16-23\}$ are assigned to the three descriptions. When all three descriptions are received, both strategies achieve comparable performance (34.84 dB vs. 34.81 dB PSNR), confirming equivalent rate-distortion efficiency at full reception. Under partial reception, the interleaved strategy demonstrates substantial advantages: with any two descriptions received, interleaved splitting maintains 32.76 dB PSNR (2.11 dB degradation from full reception) versus 28.53 dB for contiguous splitting (6.28 dB degradation). With a single description, the gap widens further: 29.86 dB (5.01 dB degradation) for interleaved versus 24.19 dB (10.62 dB degradation) for contiguous, a 5.67 dB advantage for the proposed strategy.

The performance difference arises from spatial-semantic coverage: with interleaved splitting, each description samples feature maps uniformly across all 9×16 spatial positions, providing the decoder with global context about the full image extent from any single received path. Contiguous splitting concentrates each description’s information in a localised subset of channels, yielding incomplete spatial coverage when only one or two descriptions are available. During Stage 2 training, the decoder explicitly learns to infer missing channels from spatially complete but channel-sparse inputs, which is precisely the condition created by interleaving. That inference problem is much harder under contiguous splitting because the missing information is not only channel-sparse but also less spatially representative, forcing the decoder to hallucinate larger portions of the latent structure. Table 6.3 summarises this comparison.

Number of Descriptions. A comparison of two-, three-, and four-description configurations at constant total bitrate (0.33 bpp) reveals the optimal operating point. With two descriptions (12 channels each), single-description PSNR is highest (31.15 dB) but path redundancy is minimal: any single link failure constitutes a near-worst-case sce-

Table 6.3: Performance comparison: interleaved versus contiguous splitting (AWGN). PSNR and SSIM reported as mean $\pm$ standard deviation over 100 test images.

Splitting Strategy	Desc. Recv.	PSNR (dB)	SSIM	Δ PSNR (dB)
Interleaved	All 3	34.84 ± 1.21	0.91 ± 0.04	–
Interleaved	Any 2	32.76 ± 1.89	0.86 ± 0.05	–2.11
Interleaved	Any 1	29.86 ± 2.54	0.80 ± 0.06	–5.01
Contiguous	All 3	34.81 ± 1.18	0.90 ± 0.04	–
Contiguous	Any 2	28.53 ± 2.42	0.77 ± 0.07	–6.28
Contiguous	Any 1	24.19 ± 2.67	0.64 ± 0.09	–10.62

nario. With three descriptions (8 channels each), single-description PSNR is 29.86 dB (5.01 dB degradation from full reception) but the system can now tolerate the complete loss of any two paths while maintaining perceptually acceptable reconstruction. With four descriptions (6 channels each), single-description PSNR collapses to 25.83 dB (9.04 dB degradation), as the sparse 6-channel allocation fails to capture adequate spatial-frequency content, and the $2^4 - 1 = 15$ subset combinations required for Stage 2 training introduce optimisation instability and substantially increased training cost. Expanding the latent dimensionality to preserve eight channels per description in the four-description case would increase bitrate and transmission overhead, violating the matched-rate design used throughout the evaluation. Moreover, each additional description adds a separate entropy-coded and channel-coded stream, increasing protocol overhead and receiver complexity. Table 6.4 summarises this trade-off; the three-description configuration captures the most significant robustness gain, namely that one surviving path is sufficient for perceptually acceptable reconstruction, while maintaining a viable single-description quality floor and manageable training complexity.

Asymmetric Channel SNRs. The use of asymmetric SNRs across the three transmission paths is intentional rather than incidental. In practical wireless deployments, parallel paths rarely experience identical link quality because of unequal propagation loss, antenna gains, interference levels, or local shadowing. The chosen configuration therefore tests whether the fusion decoder can exploit heterogeneous path reliability instead of assuming balanced channels. At the same time, each experiment keeps the channel type homogeneous across the three paths, that is, all AWGN, all Rayleigh, or

all Rician, so that the effect of SNR heterogeneity is isolated from changes in fading statistics. This makes the comparison more rigorous: performance differences can be attributed to adaptive semantic fusion under unequal link quality, rather than to mixing fundamentally different channel distributions within one experiment. The combined-SNR formulation then ensures that the multi-path system is compared fairly with single-path baselines under equivalent total-capacity conditions.

Table 6.4: Impact of number of descriptions on performance (AWGN). PSNR and SSIM reported as mean $\pm$ standard deviation.

Num. Desc.	Channels per Desc.	Desc. Recv.	PSNR (dB)	SSIM
2	12	Both	34.87 ± 1.19	0.91 ± 0.04
		Any 1	31.15 ± 1.73	0.82 ± 0.05
3	8	All 3	34.84 ± 1.21	0.91 ± 0.04
		Any 1	29.86 ± 2.54	0.80 ± 0.06
4	6	All 4	34.80 ± 1.17	0.91 ± 0.04
		Any 1	25.83 ± 2.91	0.71 ± 0.08

Computational Complexity

The SMDC autoencoder requires approximately 16.38 GFLOPs per forward pass for a $144 \times 256 \times 3$ input block, corresponding to 8.62 million trainable parameters (3.18M encoder, 5.44M decoder). Benchmarked on the same NVIDIA A100 GPU using identical $144 \times 256 \times 3$ input blocks averaged over the 100-image DIV2K validation set, the SMDC autoencoder completes a full encode-decode pass in 290.23 ms per block compared to 330.18 ms for CompressAI [76], making SMDC approximately 12% faster despite generating three descriptions. This efficiency arises because the interleaved channel-wise split is a zero-cost indexing operation on the $9 \times 16 \times 24$ bottleneck tensor, and the decoder processes all descriptions through a single unified network rather than three separate decoders.

Discussion and Limitations

The experimental evidence supports three principal conclusions.

Learned semantic redundancy fundamentally outperforms pixel-domain splitting under partial reception. The 4–10 dB SNR robustness advantage of SMDC over pixel-domain baselines (Table 6.2) arises because interleaved semantic feature distribution ensures that each description independently carries sufficient spatial and semantic coverage for coherent reconstruction. Pixel-domain interpolation is structurally unable to reconstruct high-frequency edges and texture from spatially sparse sub-samples.

Semantic MDC and polar FEC operate in complementary regimes. Semantic MDC provides coarse-grained structural robustness through a guaranteed quality floor from any single surviving description, handling complete path failures and deep fades. Polar FEC provides fine-grained bit-level protection within each description, correcting residual errors to maximise the fidelity of successfully received descriptions. Neither mechanism alone achieves the performance of their combination: MDC without FEC is vulnerable to bit corruption during entropy decoding, and polar codes alone cannot protect against catastrophic link outages.

The guaranteed performance floor is conditioned on the input distribution matching the training corpus. The 29.86 dB single-description PSNR floor corresponds to semantically coherent, perceptually acceptable reconstructions of natural image content within the DIV2K distribution. However, for out-of-distribution domains such as medical imagery, satellite imagery, or synthetic content, the achievable floor may be lower, and domain-specific fine-tuning would be required to restore the robustness guarantees.

Several limitations remain. The per-block processing architecture can introduce boundary artefacts at block seams in full-resolution images, partially mitigated by the bilateral post-filter; joint processing of larger receptive fields would eliminate this limitation. The fixed three-description, 0.33 bpp configuration lacks adaptability to varying channel conditions or bandwidth constraints; a variable-rate, variable-redundancy extension driven by channel state feedback is a natural direction for future work. The current framework operates on still images; the extension to video transmission through temporal prediction and inter-frame semantic diversity is addressed in Section 6.3.

6.2.4 Conclusion

The proposed SMDC framework shows that end-to-end learning of semantic redundancy across multiple independent wireless paths can change the failure behaviour of wireless image transmission from sudden collapse to smooth degradation. Three key findings emerge from the experimental evaluation.

Semantic multiple description coding removes path-level single points of failure. Under AWGN conditions, SMDC maintains PSNR above 30 dB down to a combined SNR of 1 dB (Table 6.2), compared with failure of all four baselines below 5 dB. Under Rayleigh and Rician fading with polar FEC enabled, the 30 dB PSNR threshold is maintained down to 7 dB and 5 dB combined SNR respectively, versus 12 dB and 10 dB for multi-channel baselines. These 4–5 dB robustness gains come from the learned semantic redundancy distributed across three independent paths, supported by the 5.67 dB PSNR advantage of interleaved over contiguous splitting under single-description reception (Table 6.3).

Each description independently preserves semantic content. The three-description configuration achieves a single-description PSNR floor of 29.86 dB, maintaining acceptable reconstruction quality even when two of three paths fail completely. Fig. 6.6 shows this smooth degradation, illustrating how quality reduces progressively from all three descriptions received down to a single description under increasing BERs, rather than failing suddenly as seen in all baselines.



Figure 6.6: Visual comparison of SMDC reconstruction quality under different reception scenarios. From left to right: (a) all three descriptions received; (b) any two descriptions received; (c) single description without bit errors; (d) single description at $\text{BER} = 0.001$; (e) single description at $\text{BER} = 0.01$.

The two-stage training procedure prepares the decoder for all reception scenarios. The Stage 2 decoder refinement trained over all seven non-empty description subsets

and stochastic block erasure augmentation is important for the observed performance floor. Without subset-aware training, the decoder cannot produce useful reconstructions from partial description sets under quantised inputs. The 12% reduction in per-block encode-decode time compared with CompressAI (290.23 ms vs. 330.18 ms on an A100 GPU) shows that the interleaved channel-wise split adds no computational cost at inference time.

The fixed three-description, 0.33 bpp operating point cannot adapt to varying channel conditions or bandwidth requirements, and the training data is limited to natural photographic content from the DIV2K dataset [207]. The per-block processing also introduces boundary artefacts at block seams in full-resolution images, partially reduced by the bilateral post-filter [206].

6.3 Semantic Multiple Description Coding for Robust Video Transmission

6.3.1 Introduction

The channel-robustness motivation, limitations of single-path systems, and relationship to classical and learned MDC are established in the Chapter Introduction and Section 6.2.1. This section extends the image framework to video by combining its three-description representation with the RIFE-based temporal context and differential latent coding defined in Chapter 4. The new GOP structure and inter-frame processing are specified below and evaluated against VVenC 1.12.1 [84] under the same three-path channel conditions.

6.3.2 Methodology

System Overview

The SMDC-V framework, illustrated in Fig. 6.7, is the direct synthesis of two prior contributions in this thesis: the image-domain Semantic MDC framework of Section 6.2, which established the interleaved three-description latent representation and multi-

path polar-coded delivery, and the context-conditioned semantic video coding systems of Chapter 4 (Sections 4.2), which introduced RIFE-based temporal context generation and differential latent encoding for inter-frame compression. Building on these foundations, the SMDC-V system extends semantic MDC to video sequences by exploiting temporal redundancy between frames to reduce the bitrate allocated to intermediate frames while preserving the graceful-degradation properties of the multi-description architecture. The resulting system decomposes each GOP into two complementary subsets: boundary intra-frames processed by the MDC Intra-Frame Semantic Encoder inherited from Section 6.2, and intermediate inter-frames processed by a newly introduced MDC Inter-Frame Semantic Encoder that encodes only the deviation of each inter-frame’s latent representation from a RIFE-interpolated [72] temporal context.

A video sequence $\mathcal{F} = \{F_t\}$ is partitioned into non-overlapping GOPs of size $N = 8$. Within each GOP, the two boundary frames F_m and F_{m+8} are encoded as intra-frames via the MDC Intra-Frame Semantic Encoder, while the seven intermediate frames $\{F_{m+k}\}_{k=1}^7$ are encoded as inter-frames conditioned on context interpolated from the two boundary frames. Each frame yields three independently transmitted descriptions; the receiver reconstructs any accessible frame from the available subset of descriptions, with the GOP structure ensuring that boundary-frame errors propagate no further than the current GOP.

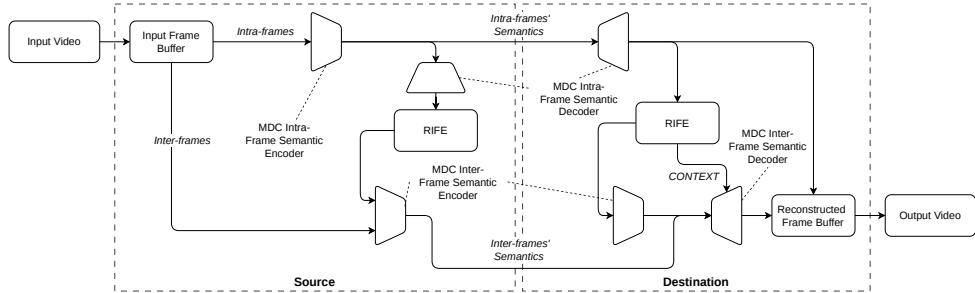


Figure 6.7: End-to-end SMDC-V pipeline for video transmission over three independent wireless channels. Boundary frames (F_m , F_{m+8}) are encoded by the MDC Intra-Frame Semantic Encoder; the seven intermediate frames are encoded by the MDC Inter-Frame Semantic Encoder conditioned on RIFE-interpolated context. Three descriptions per frame are independently entropy-coded, channel-coded, and transmitted.

Video Decomposition and GOP Structure

Given a GOP of size $N = 8$ with boundary frames F_m and F_{m+8} , a temporal context frame is generated for each intermediate position $k \in \{1, \dots, 7\}$ by the RIFE optical-flow-based frame interpolation network [72]:

$$C_{m+k} = f_{\text{RIFE}}(F_m, F_{m+8}; \tau_k), \quad \tau_k = \frac{k}{8}, \quad (6.13)$$

where $\tau_k \in (0, 1)$ specifies the temporal interpolation fraction and f_{RIFE} denotes the RIFE model. All seven context frames are generated in parallel from the same pair of decoded boundary frames, unlike hierarchical inter-frame coding schemes [51], which propagate errors from near-boundary frames to further positions. This parallel generation strategy ensures that inter-frame reconstruction quality degrades no further for the middle frame ($k = 4$, worst-case temporal distance) than for adjacent frames ($k = 1$ or $k = 7$); the temporal estimation error $\|C_{m+k} - F_{m+k}\|$ is bounded by the RIFE worst-case interpolation accuracy at $\tau = 0.5$ regardless of the frame index, since all context frames originate from the same decoded source pair.

The GOP size $N = 8$ is selected to balance three competing requirements: temporal distance, bounded by $k/8 \leq 7/8$, constrains the RIFE interpolation accuracy; error propagation is limited to a single GOP since each intra-frame is encoded independently; and the ratio of intra-frames (2 per GOP) to inter-frames (7 per GOP) determines the bitrate overhead from intra-frame coding. Experiments with larger GOPs ($N = 16$, $N = 32$) confirm degraded RIFE interpolation quality at distant intermediate positions and increased sensitivity to intra-frame channel errors, confirming that $N = 8$ achieves the best trade-off between temporal prediction accuracy and error isolation [73].

MDC Intra-Frame Semantic Encoder

Boundary frames use the image SMDC encoder, interleaved description split, and subset-aware decoder defined in Sections 6.2.2 and 6.2.2. The network is initialised from the image model and fine-tuned on video frames; this fine-tuning is the only change to the intra-frame module.

Because the intra-frame structure is unchanged, its encoder, interleaved three-description split, and subset-aware decoder are not redrawn here. They are shown in the image SMDC pipeline in Fig. 6.1 and in the detailed autoencoder diagram in Fig. 6.2; only the video-domain fine-tuning described above distinguishes this module.

MDC Inter-Frame Semantic Encoder

The MDC Inter-Frame Semantic Encoder introduces the key technical innovation of the proposed framework: rather than encoding each intermediate frame F_{m+k} from scratch, it encodes only the deviation of F_{m+k} 's latent representation from the corresponding RIFE context C_{m+k} , exploiting the temporal correlation structure captured by the context interpolation.

Differential Latent Encoding. Let $\phi_{\theta_e}(\cdot)$ denote the shared encoder backbone (structurally identical to the intra-frame encoder: four strided-convolution blocks, 128 channels, $9 \times 16 \times 24$ output). The latent representations of the source frame and its context are computed as

$$l_{m+k} = \phi_{\theta_e}(F_{m+k}), \quad (6.14)$$

$$l_{m+k}^C = \phi_{\theta_e}(C_{m+k}). \quad (6.15)$$

The differential latent is defined as

$$\Delta l_{m+k} = l_{m+k} - l_{m+k}^C. \quad (6.16)$$

Its entropy advantage follows the mutual-information argument established for differential context latents in Eqn. (4.7). When the interpolated context is a useful predictor, Δl_{m+k} is more compressible than the original latent l_{m+k} .

Multiple Description Generation. The differential latent $\Delta l_{m+k} \in \mathbb{R}^{9 \times 16 \times 24}$ is split into three descriptions by the same interleaved channel-wise strategy as the intra-

frame module:

$$\{\mathbf{D}_1^{\text{inter}}, \mathbf{D}_2^{\text{inter}}, \mathbf{D}_3^{\text{inter}}\} = \mathcal{S}(\Delta l_{m+k}), \quad (6.17)$$

where $\mathcal{S}(\cdot)$ denotes the interleaved channel-wise splitting operator.

Context-Conditioned Decoder. At the receiver, the received descriptions are assembled to form $\widehat{\Delta l}_{m+k}$ and the estimated latent is recovered as

$$\hat{l}_{m+k} = \widehat{\Delta l}_{m+k} + l_{m+k}^C, \quad (6.18)$$

where l_{m+k}^C is computed locally at the receiver from the two decoded boundary frames $\hat{F}_m$ and $\hat{F}_{m+8}$ via the same RIFE interpolation. The reconstruction of the inter-frame is then

$$\hat{F}_{m+k} = \psi_{\theta_d}(\hat{l}_{m+k}, C_{m+k}), \quad (6.19)$$

where ψ_{θ_d} denotes the context-conditioned decoder. The decoder concatenates the recovered latent $\hat{l}_{m+k}$ with the interpolated context C_{m+k} at the final decoding stage via a residual block (64 filters, dropout probability $p = 0.2$, followed by a transposed-convolution upsampling block), enabling the network to correct systematic interpolation errors arising from occlusion or non-linear motion. This context conditioning accounts for approximately 62% of the total decoder FLOPs, for a decoder total of 52.22 GFLOPs.

The inter-frame architecture is illustrated in Figs. 6.8 and 6.9.

Transmission Pipeline

Both intra-frame and inter-frame descriptions use the quantisation, DeepCABAC, polar-code, modulation, and three-path channel pipeline defined in Sections 6.2.2 and 6.2.2. The coding parameters, asymmetric per-path SNR ranges, combined SNR definition in Eqns. (6.7)–(6.8), and use of polar coding only for the fading experiments are unchanged.

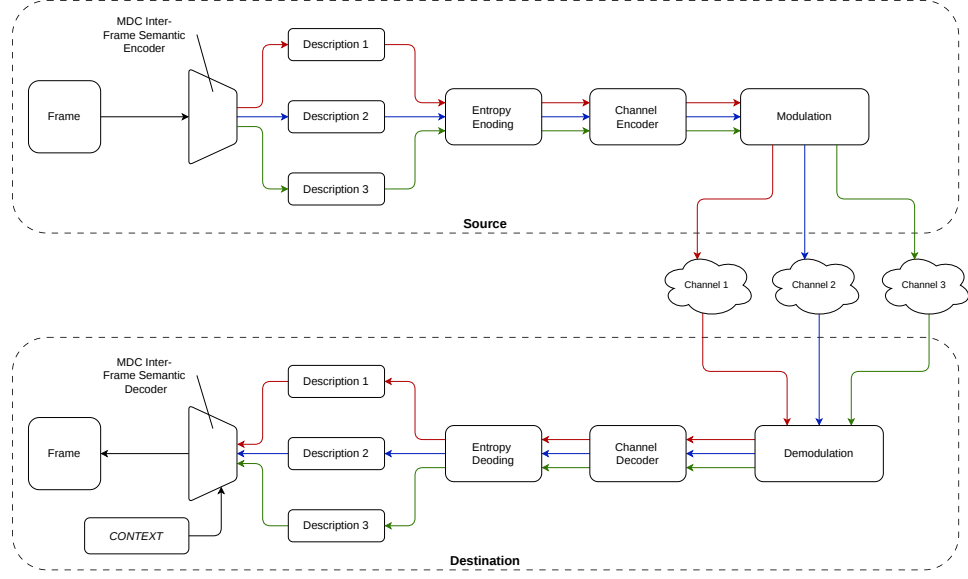


Figure 6.8: MDC Inter-Frame Semantic Encoder and Decoder. The encoder computes the differential latent $\Delta l_{m+k} = l_{m+k} - l_{m+k}^C$ and splits it into three descriptions via interleaved channel-wise assignment. The context-conditioned decoder recovers $\hat{l}_{m+k}$ from the received descriptions and concatenates it with the locally recomputed context C_{m+k} to reconstruct $\hat{F}_{m+k}$.

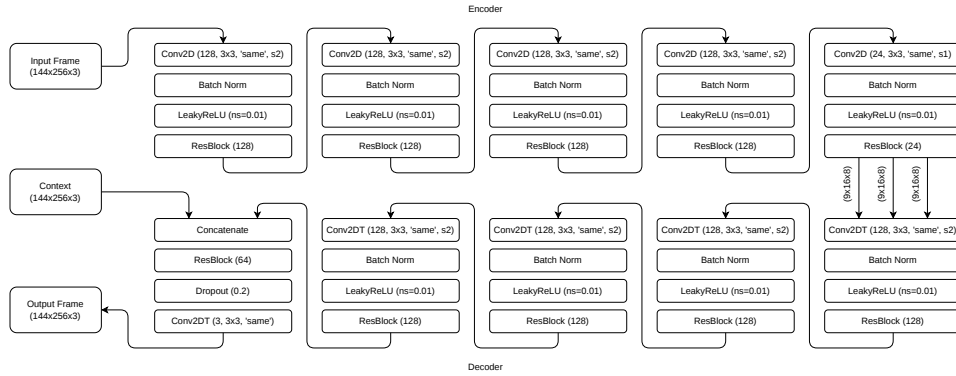


Figure 6.9: Detailed network architecture of the MDC Inter-Frame Semantic Encoder and context-conditioned decoder. The context-conditioning residual block at the final upsampling stage contributes approximately 62% of the total decoder FLOPs, for a decoder total of 52.22 GFLOPs.

Training Strategy

Both the intra-frame and inter-frame modules are trained on the Vimeo-90k dataset [175], which provides diverse natural video content for learning both spatial and temporal semantic representations. Frames are cropped into $144 \times 256 \times 3$ RGB blocks to match the network input dimensions. Training is performed with the Adam optimiser, batch size 8, on a single NVIDIA A100 GPU.

Intra-Frame Training. The intra-frame model uses the two-stage optimisation, subset sampling, learning-rate schedule, and block-erasure augmentation defined in Section 6.2.2; it is fine-tuned on video frames rather than retrained with a different objective.

Inter-Frame Training. The inter-frame module is trained on the Vimeo-90k septuplet split, where each training sample provides the target inter-frames $\{F_{m+k}\}_{k=1}^7$ together with context interpolated from the boundary reference frames F_m and F_{m+8} . In the first training stage, the module minimises the averaged reconstruction loss over all seven intermediate positions:

$$\mathcal{L}_{\text{AE}}^{\text{inter}} = \frac{1}{7} \sum_{k=1}^7 \left[\|F_{m+k} - \hat{F}_{m+k}\|^2 + \lambda R_k \right], \quad (6.20)$$

where R_k is the estimated bitrate for position k . This stage learns to transmit only the residual semantic information that is not already captured by the interpolated context, which preserves bitrate efficiency while improving temporal consistency. The same progressive learning-rate schedule is applied for 110,000 total iterations, targeting a reconstruction-quality range of 28–35 dB PSNR to maintain rate compatibility with the intra-frame module. After this, a second refinement stage applies the same subset-aware decoder training and random block erasure augmentation used for the intra-frame model, but now on the differential latent descriptions and with the context-conditioned decoder. This additional refinement explicitly trains the inter-frame decoder to recover from any non-empty subset of received descriptions and from missing latent blocks

caused by channel impairments.

Baseline Method

The baseline is the VVenC 1.12.1 encoder [84] configured for multiple description video transmission. Rather than applying VVenC directly to the full video, the baseline adopts the same three-path interleaved splitting strategy employed by the image SMDC baselines: the video is decomposed into three sub-videos via pixel-wise interleaved sampling,

$$F_t^{(i)}(p) = F_t(3p + i - 1), \quad i \in \{1, 2, 3\}, \quad (6.21)$$

and each sub-video is independently encoded by VVenC with a GOP size of $N = 8$. Error resilience is enabled through independent tile partitioning, explicit tile entry points, disabled cross-slice loop filter boundaries, clean IDR refresh, and Message-Digest Algorithm 5 (MD5) payload integrity checks. The same polar codes ($K = 164$, $E = 512$) and BPSK/DBPSK modulation are applied to the VVenC bitstream for direct comparison. At the receiver, adaptive reconstruction is applied depending on the number of received sub-videos: three sub-videos are directly interleaved, two sub-videos are merged by bilinear interpolation for missing pixels, and one sub-video is upsampled by bicubic interpolation [210]. VVenC is configured to match the SMDC-V target bitrate of 0.13 bits per pixel. In addition to this three-path MDC configuration, a single-path VVenC baseline is also evaluated, in which the full video (without interleaved splitting) is encoded by VVenC with the same error-resilience settings and channel coding, and transmitted over a single path. This allows the benefit of path diversity to be quantified independently of the semantic coding architecture.

6.3.3 Results and Discussion

Experimental Setup

Evaluation is conducted on 20 test video sequences from the semantic video communication benchmark established in [1], each comprising 65 frames at 144×256 pixels. These sequences represent diverse temporal motion patterns including fast action, slow

panning, and static backgrounds, providing a comprehensive assessment of the proposed framework’s robustness across different video content types. Reconstruction quality is assessed by PSNR (dB), SSIM [164], and VMAF [165], averaged over all 20 sequences and over all 65 frames per sequence. The combined SNR axis follows Eqns. (6.7)–(6.8), spanning from 16.76 dB to -9.55 dB. Both three-path MDC and single-path configurations are evaluated to quantify the benefit of path diversity.

Multi-Channel AWGN Results

Figs. 6.10a and 6.10d present PSNR and SSIM performance under AWGN conditions with polar coding disabled, isolating the semantic multi-description architecture from channel-coding artefacts. The proposed SMDC-V framework maintains PSNR above 30 dB and SSIM above 0.80 down to a combined SNR of 1 dB, whereas the VVenC MDC baseline requires a combined SNR of at least 4 dB to achieve the same threshold, giving a 3 dB robustness advantage. At high combined SNR (above 10 dB), VVenC achieves a peak PSNR of approximately 36.5 dB compared with SMDC-V’s plateau of approximately 33 dB, a difference attributable to VVenC’s sophisticated inter-frame prediction and entropy coding optimised for error-free transmission conditions. Below 0 dB combined SNR, VVenC MDC fails catastrophically as the arithmetic decoder desynchronises from residual bit errors, while SMDC-V maintains a quality floor above 28 dB PSNR, sustaining the graceful-degradation property established in the image SMDC framework. VMAF results (Fig. 6.10g) confirm these trends perceptually: SMDC-V achieves VMAF above 70 down to a combined SNR of approximately 1 dB, while VVenC MDC achieves VMAF above 70 only at approximately 4 dB combined SNR, reflecting the perceptual impact of pixel-domain reconstruction artefacts from the interleaved sub-video splitting strategy.

Multi-Channel Rayleigh and Rician Fading Results

Under Rayleigh fading (Figs. 6.10b and 6.10e), SMDC-V maintains PSNR above 30 dB and SSIM above 0.80 down to a combined SNR of 7 dB; VVenC MDC requires 10 dB combined SNR for the same threshold, a 3 dB robustness advantage. The wider mar-

gin under Rayleigh fading compared with AWGN reflects the greater vulnerability of conventional entropy decoders to the bursty bit errors introduced by deep Rayleigh fades, while the SMDC-V semantic decoder is inherently tolerant to description-level impairments through its subset-aware training procedure.

Rician fading results (Figs. 6.10c and 6.10f) exhibit comparable trends. SMDC-V achieves the 30 dB PSNR threshold at a combined SNR of 5 dB, compared with 9 dB for VVenC MDC, a 4 dB robustness advantage. The stronger line-of-sight component (Rician K-factor of 3) reduces the variance of deep fading events relative to the Rayleigh case, and both methods recover quality more rapidly at higher combined SNR, though the relative 5 dB advantage of SMDC-V is preserved across the full SNR range. VMAF results (Figs. 6.10h and 6.10i) confirm the perceptual robustness advantage: SMDC-V achieves VMAF above 70 down to approximately 7 dB combined SNR under Rayleigh fading and approximately 6 dB under Rician fading, while VVenC MDC achieves VMAF above 70 only at approximately 9 dB under both fading conditions, a 2–3 dB perceptual robustness advantage for SMDC-V.

Single-Channel Versus Multi-Channel Analysis

Single-channel results are presented in Fig. 6.11 for both SMDC-V and VVenC operating over a single transmission path without multiple description coding. Under AWGN (Figs. 6.11a and 6.11d), single-channel SMDC-V achieves the 30 dB PSNR threshold at 6 dB SNR, compared with 10 dB for single-channel VVenC, a 4 dB advantage. The single-channel VVenC performance curve exhibits the characteristic cliff effect: quality remains near zero until approximately 10 dB SNR, then rises abruptly to approximately 37 dB within a 2 dB range, because VVenC’s arithmetic entropy decoder can only produce valid output when the bit-error rate falls below a decodability threshold. SMDC-V exhibits smooth monotonic degradation throughout, a direct consequence of the semantic quantisation tolerance and the subset-aware decoder training.

Comparing the multi-channel and single-channel configurations quantifies the path-diversity gain directly. For AWGN, multi-channel SMDC-V achieves the 30 dB threshold at 1 dB combined SNR versus 6 dB for single-channel SMDC-V, a 5 dB diversity

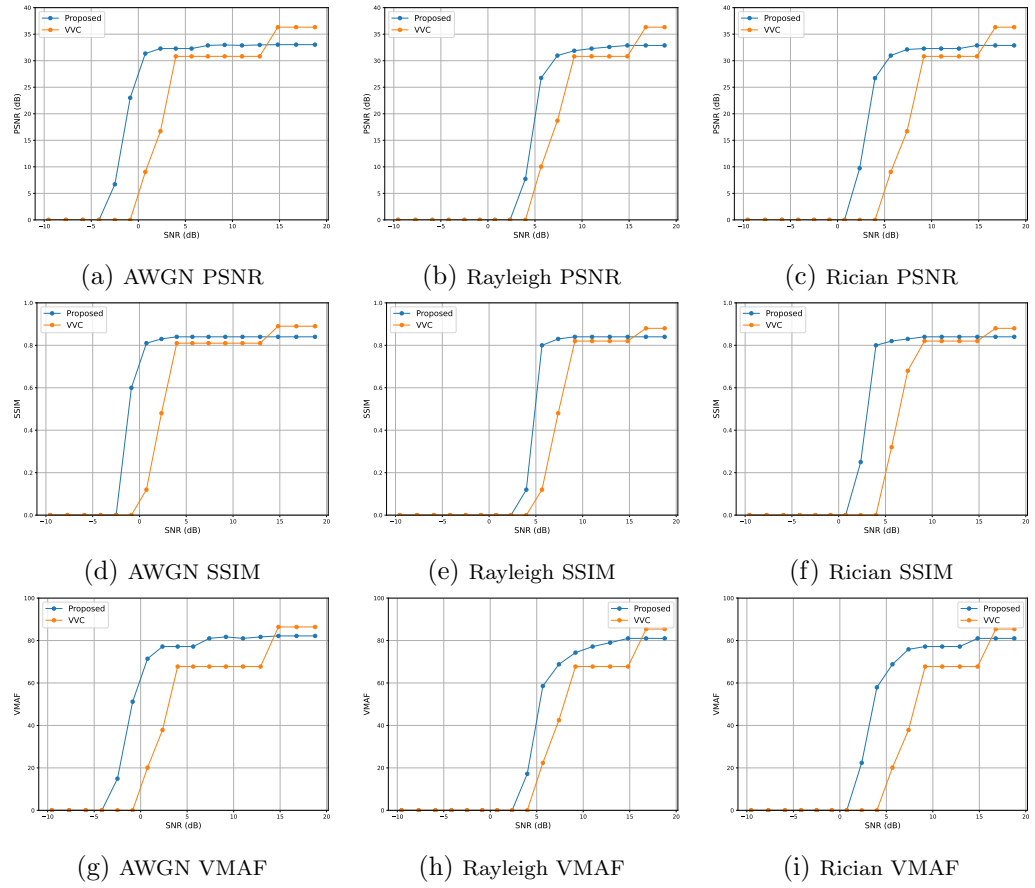


Figure 6.10: Multi-channel MDC performance of SMDC-V against VVenC MDC across three channel models. (a)–(c) PSNR versus combined SNR. (d)–(f) SSIM versus combined SNR. (g)–(i) VMAF versus combined SNR.

gain. For Rayleigh fading, the corresponding thresholds are 7 dB (multi-channel) versus 11 dB (single-channel), again a 4 dB diversity gain. The physical interpretation is that the probability of all three statistically independent paths simultaneously experiencing a deep fade is the product of their individual fade probabilities, which is orders of magnitude lower than the single-path fade probability.

Under Rayleigh and Rician fading (Figs. 6.11b–6.11f), single-channel SMDC-V achieves the 30 dB PSNR threshold at 11 dB and 10 dB SNR respectively, compared with 16 dB and 15 dB for single-channel VVenC, preserving the 4–5 dB advantage observed in the multi-channel configuration. In terms of VMAF (Figs. 6.11g–6.11i), single-channel SMDC-V achieves VMAF above 70 at approximately 6 dB (AWGN), 11 dB (Rayleigh), and 11 dB (Rician), compared with approximately 11 dB, 16 dB, and 16 dB for single-channel VVenC respectively.

Consolidated Performance Table

Table 6.5 consolidates the minimum SNR at which each configuration achieves PSNR above 30 dB, SSIM above 0.80, and VMAF above 70 across all three channel models for both multi-channel MDC and single-channel configurations. Multi-channel SMDC-V achieves the lowest threshold in every condition, with VMAF thresholds 2–3 dB lower than VVenC MDC across all channel models.

Table 6.5: Minimum SNR (dB) to achieve each quality threshold for SMDC-V and VVenC under multi-channel MDC and single-channel configurations. Lower values indicate greater channel robustness. Combined SNR is used for multi-channel; single-path SNR is used for single-channel. Values are approximate to the nearest dB.

Channel	Metric	Multi-Channel		Single Channel	
		SMDC-V	VVenC	SMDC-V	VVenC
AWGN	PSNR > 30 dB	1	4	6	10
	SSIM > 0.80	1	4	6	10
	VMAF > 70	1	4	6	11
Rayleigh	PSNR > 30 dB	7	10	11	16
	SSIM > 0.80	7	10	11	16
	VMAF > 70	7	9	11	16
Rician	PSNR > 30 dB	5	9	10	15
	SSIM > 0.80	5	9	10	15
	VMAF > 70	6	9	11	16

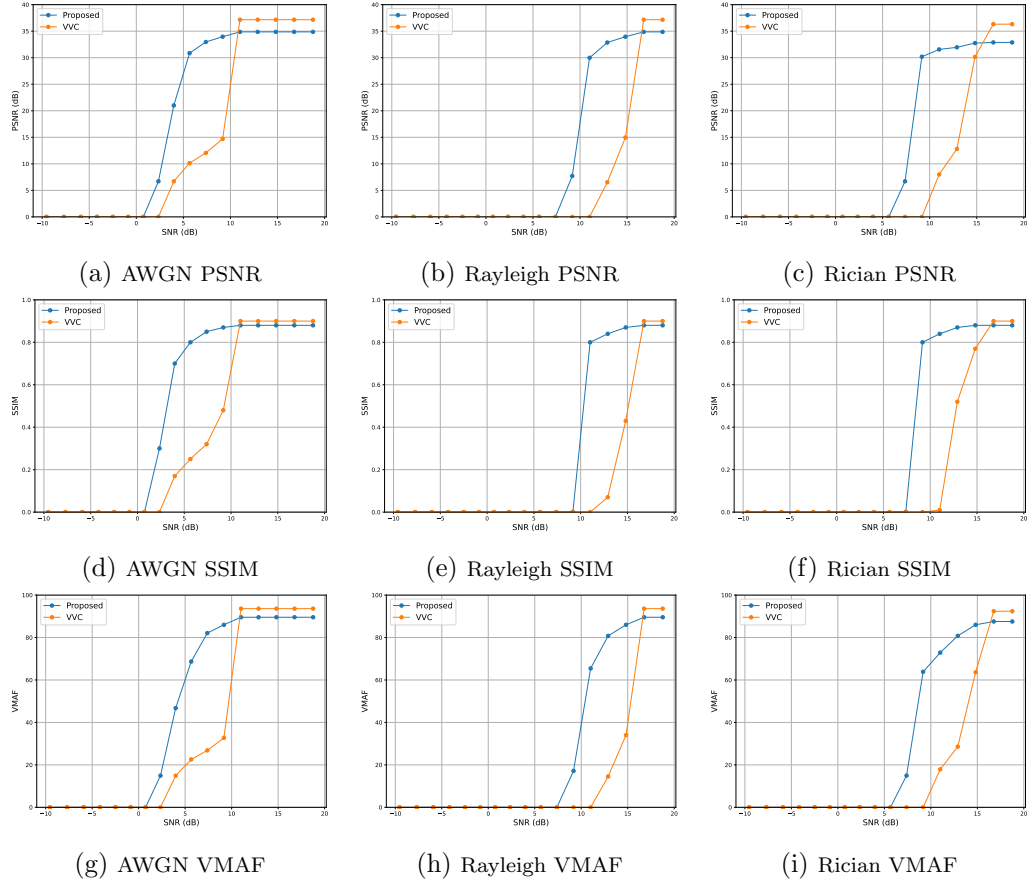


Figure 6.11: Single-channel (non-MDC) performance of SMDC-V against VVenC across three channel models. (a)–(c) PSNR versus SNR. (d)–(f) SSIM versus SNR. (g)–(i) VMAF versus SNR.

Computational Complexity

All complexity measurements are performed on an NVIDIA A100 GPU for $144 \times 256 \times 3$ frames. The MDC Intra-Frame Semantic Encoder requires 16.38 GFLOPs per frame with 8.62 million parameters (3.18M encoder, 5.44M decoder), identical to the image SMDC autoencoder of Section 6.2. The MDC Inter-Frame Semantic Encoder requires 60.41 GFLOPs per frame (8.19 GFLOPs encoder, 52.22 GFLOPs decoder); the decoder’s substantially higher cost is dominated by the context-conditioning residual block at the final upsampling stage, which accounts for approximately 62% of the total decoder FLOPs. The per-GOP complexity is $16.38 + 7 \times 60.41 \approx 439$ GFLOPs for two intra-frames and seven inter-frames.

The SMDC-V encoder throughput of 15.2 fps represents a $69\times$ speedup over VVenC’s 0.22 fps, arising because the proposed framework requires no rate–distortion optimisation search, no motion estimation, and no iterative entropy coding; all computation consists of deterministic forward passes through the encoder and RIFE networks. The decoder throughput of 15.7 fps is lower than VVenC’s 102.04 fps because the RIFE context generation and context-conditioned decoder perform substantially more computation than VVenC’s entropy and motion-compensation decoding; this asymmetry reflects the computational burden being shifted from the discrete arithmetic decoder to the learned neural network.

Discussion and Limitations

The experimental results support three principal conclusions.

Differential latent encoding effectively extends semantic MDC robustness to video.

By encoding the deviation $\Delta l_{m+k} = l_{m+k} - l_{m+k}^C$ rather than the full latent representation, the inter-frame encoder achieves a 4–5 dB robustness advantage over the VVenC MDC baseline consistent with that of the image SMDC framework, confirming that the graceful-degradation property of the three-description architecture is preserved across the temporal prediction chain.

Parallel context generation from fixed boundary frames isolates error propagation.

Unlike hierarchical video coding schemes where a corrupted frame infects all subsequent

frames in the prediction chain, the SMDC-V framework regenerates the RIFE context independently for each GOP at both encoder and decoder, so that any error in an intra-frame is strictly contained within its GOP and does not cascade to subsequent GOPs. This bounded error propagation is a fundamental reliability property absent from standard predictive coding.

A peak quality gap at high SNR is the cost of structural robustness. At combined SNRs above 10 dB, VVenC MDC achieves approximately 3.5 dB higher peak PSNR than SMDC-V (36.5 dB vs. 33 dB), reflecting VVenC’s superior rate–distortion performance under error-free transmission at the target 0.13 bpp bitrate. The two objectives are in direct tension: rate–distortion optimisation for error-free channels conflicts with the semantic redundancy required for graceful degradation under channel impairment, and the 0.13 bpp budget does not accommodate both simultaneously.

Several limitations constrain the scope of the proposed framework. The 28–35 dB PSNR operating range is appropriate for standard and high definition content at moderate bitrates but may be insufficient for broadcast-quality applications. The RIFE context generation depends on decoded boundary frame quality: under severe channel conditions, degraded intra-frame reconstructions reduce context accuracy, weakening the compression efficiency of the inter-frame differential encoding and lowering the achievable PSNR floor. Efficient joint rate allocation between intra-frame and inter-frame descriptions to maximise combined PSNR under a global GOP bitrate constraint represents a natural direction for future investigation.

6.3.4 Conclusion

The Semantic MDC Video framework shows that the smooth-degradation properties of the image SMDC framework extend to video transmission when differential latent encoding of inter-frames is combined with parallel context generation in a GOP-8 structure. Three main findings emerge from the experimental evaluation.

Differential latent encoding keeps semantic MDC robustness across the temporal prediction chain. By encoding the difference $\Delta l_{m+k} = l_{m+k} - l_{m+k}^C$ rather than the full inter-frame latent, the inter-frame encoder achieves 4–5 dB robustness gains over

VVenC MDC across all channel models (Table 6.5), in line with the gains of the image SMDC framework. The framework maintains PSNR above 30 dB and VMAF above 70 down to approximately 1 dB combined SNR (AWGN), 7 dB (Rayleigh), and 5–6 dB (Rician) in the multi-channel configuration, with VMAF thresholds 2–3 dB lower than VVenC MDC across all channel models. The 5 dB diversity gain for AWGN (1 dB multi-channel versus 6 dB single-channel SMDC-V) and 4 dB for fading channels confirm that all three paths contribute to the overall robustness.

Parallel context generation limits error propagation to individual GOPs. Since each GOP’s inter-frame contexts come only from the two decoded boundary intra-frames via an independent RIFE pass, errors within a GOP are contained and do not spread to later GOPs, a key reliability property absent from standard inter-frame predictive coding.

Encoding speed is greatly improved over VVC-based coding. The framework achieves 15.2 fps encoding speed, a 69× improvement over VVenC at 0.22 fps, by replacing exhaustive rate–distortion optimisation and motion estimation with deterministic neural network inference. The per-GOP computational cost of approximately 439 GFLOPs is dominated by the context-conditioned inter-frame encoder at 60.41 GFLOPs per inter-frame.

The framework is limited to a fixed 0.13 bpp bitrate with a three-description GOP-8 structure, without adaptive rate allocation between intra-frame and inter-frame descriptions. The 28–35 dB PSNR operating range is suitable for standard and high definition content at moderate bitrates. The RIFE context generation depends on the quality of decoded boundary frames: under severe channel conditions, degraded intra-frame reconstructions reduce context accuracy, weakening the inter-frame differential encoding efficiency. Extension to variable bitrates, adaptive GOP structures, and joint rate optimisation between intra-frame and inter-frame descriptions is a natural direction for future work.

6.4 Chapter Summary and Conclusion

This chapter has addressed the single-point-of-failure weakness in single-path semantic communication systems through two SMDC frameworks for image and video transmission. Together, these contributions show that end-to-end learning of semantic redundancy distributed across independent wireless paths can improve the robustness and efficiency of visual wireless communication.

The image and video findings are reported in detail in the respective section conclusions (Sections 6.2.4 and 6.3.4) and consolidated numerically in Tables 6.2 and 6.5. Taken together, they show that the same three-description principle operates in both domains: image SMDC learns complementary spatial descriptions, while SMDC-V adds differential temporal latents and RIFE context without changing the underlying path-diversity mechanism. The consistent robustness gains across AWGN, Rayleigh, and Rician channels demonstrate that the image-domain design generalises to video rather than requiring a separate channel-robustness principle.

Together, these contributions address the research gap identified in Chapter 1: the lack of an end-to-end framework that jointly optimises semantic source coding, complementary description generation, and channel-aware training for path-diverse wireless transmission. The work shows that semantic communication systems can be designed to provide both robustness to channel noise, from the learned representations, and robustness to path failure, from the multiple description architecture, giving reliability that neither mechanism can achieve on its own.

The performance floors established by both frameworks make them suitable for applications where minimum quality must be maintained under the worst channel conditions, including autonomous vehicle telemetry, remote surgical systems, and disaster-response communications. The connection between semantic communication system design, DNN-based representation efficiency, and network robustness is continued in Chapter 7, which addresses the related challenge of efficiently transmitting and compressing the neural network parameters that underpin semantic video coding systems.

Chapter 7

DNN Parameter Efficiency for Semantic Video Coding

7.1 Chapter Introduction

The semantic video coding systems developed in Chapters 3–6 achieve good rate–distortion performance by training autoencoder-based semantic encoder and decoder networks to represent video content in a compact latent space. A common strategy in these systems is per-scene or per-GOP overfitting, where the encoder and decoder are trained directly on the video content to produce near-lossless reconstruction, and only the latent vectors and decoder parameters are sent to the receiver. This uses the spatial and temporal redundancy inside a scene to achieve compression ratios that general-purpose autoencoders cannot match. The main problem is that transmitting the decoder parameters from transmitter to receiver accounts for a large part of the total bitrate. In a standard autoencoder for a 320×180 frame, the decoder may contain tens of thousands of floating-point parameters. Even when these are compressed with the NNC standard [193], implemented by the open-source NNCodec toolkit [159], the resulting bitrate is still much higher than that of conventional codecs such as VVC [38].

This chapter addresses the parameter-transmission problem in two ways. The first, presented in Section 7.2, observes that the decoder weights and biases show structured spatial correlation when arranged as two-dimensional arrays. This structure is exploited

by treating the per-GOP decoder weight and bias arrays as frames in a video sequence and encoding them with VVC through the VVenC open encoder [84]. The second, presented in Section 7.3, observes that consecutive video frames in the same scene produce very similar decoder parameters after per-frame overfitting because the visual content changes only slightly from one frame to the next. This temporal redundancy is exploited by a displacement-vector-based prediction method that moves parameter blocks from one frame to the next and sends only the NNCodec-compressed residual.

Both methods are evaluated on the same ten-video test set at 320×180 resolution, covering a range of SI and TI levels, so that their rate-distortion behaviour can be analysed as a function of content complexity. Together, they show that DNN parameter space has usable structure that can reduce transmission overhead in semantic video coding, and they provide a starting point for future parameter-compression schemes that combine spatial and temporal redundancy.

A broader research gap also motivates this chapter. The hybrid frameworks developed in Chapters 3–5 all rely on VVC as the residual coding component. A fully independent semantic video coding pipeline would remove this dependency by overfitting an autoencoder to each scene and sending the learned weights and biases to the receiver as the compressed representation. The receiver would then reconstruct each frame by running the transmitted decoder with the received latent vectors. The contributions of this chapter are the author’s first step in that direction. In particular, Section 7.2 does not use VVC as a residual-coding stage; instead, it repurposes VVC to transmit decoder weights and biases, making it an intermediate step toward removing the VVC residual path altogether. By showing that decoder weights contain useful spatial and temporal structure and that this structure can be compressed to bitrates competitive with NNCodec, the chapter provides the basis for a future pipeline in which the decoder itself, rather than a VVC residual bitstream, is the only transmitted item. This direction is identified as a priority for future work and is discussed further in Section 8.6.

It is important to note that the contributions of this chapter work under a clearly different approach from the hybrid semantic-VVC systems of Chapters 3–5. In those

systems, the semantic decoder is pre-shared between the transmitter and receiver: it is trained once on a general dataset and shared when the system is set up, so no decoder parameters need to be sent per video. Per-GOP content adaptation in the CASVC–CAPO framework (Contribution 4, Section 4.3) only requires sending a small set of approximately 643 numbers, less than 0.1% of the full decoder, together with the semantic latents and VVC residual bitstream. The *transmitted-decoder* approach studied in this chapter applies to a different scenario: the decoder is not pre-shared but is instead trained directly on few frames, so the decoder parameters themselves become the main compressed item that must be sent to the receiver. This approach is more closely related to the long-term goal of a fully self-contained semantic video codec that does not need a VVC residual path, at the cost of higher parameter transmission overhead per scene. The pre-shared decoder approach is suitable for general use where low per-video overhead and compatibility are important; the transmitted-decoder approach is better suited to cases where training the decoder on each scene gives very high reconstruction quality and the cost of sending the decoder can be spread over enough frames. The contributions of this chapter should therefore be seen as early research steps toward a future fully self-contained semantic video codec, not as direct extensions of the CASVC–CAPO system. Section 7.2 presents the spatial parameter-coding method published in “Deep Neural Network Compression for Semantic Video Communications” [24], while Section 7.3 presents the temporal parameter-prediction method published in “Semantic Communication Based Video Coding Using Temporal Prediction of Deep Neural Network Parameters” [25]. Together, these publications examine how the neural models supporting semantic video coding can themselves be transmitted efficiently.

7.2 Deep Neural Network Compression for Semantic Video Communications

Publication Notice

The research presented in this section has been published in the following peer-reviewed publication:

P. Samarathunga, I. Alahapperuma, T. Fernando, Y. Ganearachchi, and A. Fernando, “Deep Neural Network Compression for Semantic Video Communications,” in *Proc. IEEE Int. Conf. Consumer Electron. (ICCE)*, Las Vegas, USA, 2025, pp. 1–4, doi: 10.1109/ICCE63647.2025.10930011.

This material is reproduced and adapted under the terms of the IEEE Copyright policy for use in a doctoral thesis. The content has been revised and extended for inclusion and integration in this doctoral thesis, including additional explanations, expanded discussion, and alignment with the overarching research narrative.

7.2.1 Introduction

As established in the Chapter Introduction, per-GOP-overfitting semantic video coding systems require the decoder parameters to be sent to the receiver for every Group of Pictures before any frame in that GOP can be decoded. Existing approaches based on the NNC standard [193], implemented through DeepCABAC [148], compress the decoder parameters by applying entropy coding to a one-dimensional parameter vector, without using the structured spatial layout of the underlying weight matrices. This section proposes an alternative: the decoder weights and biases are reshaped into two-dimensional matrices, treated as grayscale image frames, and compressed with VVC through the VVenC open-source encoder [84].

The key motivation is that decoder parameters arranged as spatial arrays exhibit smooth gradient patterns arising from the dense layer architecture, because adjacent weights operate on nearby input features and therefore form structured two-dimensional patterns that suit VVC’s block-based prediction tools. Consecutive GOPs trained on slowly changing scene content also produce successive weight frames with strong

temporal correlation that VVC’s random-access inter-prediction can exploit. Neither of these properties is available when NNC treats the parameters as an unstructured scalar sequence.

7.2.2 System Architecture

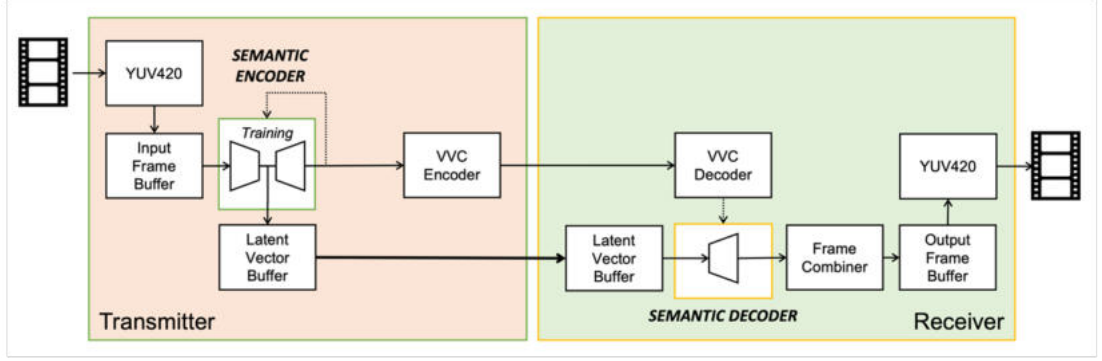


Figure 7.1: System architecture for VVC-based DNN parameter compression. The transmitter trains a two-model semantic autoencoder per GOP, reshapes the decoder weights and biases as image frames, and encodes them using VVenC. The receiver decodes the VVC bitstream to retrieve decoder parameters and reconstructs each video frame from the transmitted latent vectors [24].

The proposed system, illustrated in Fig. 7.1, integrates semantic communication with VVC to compress and transmit decoder DNN parameters in a transmitter-receiver framework. Input video frames are processed in YUV 4:2:0 format, stored in an input frame buffer.

Semantic Autoencoder

The semantic autoencoder comprises two sequential fully connected models. The first model (semantic encoder) flattens the input frame representation into an 86,400-element vector and reduces it to a 24-element latent vector through a dense layer. The second model (semantic decoder) expands through a dense layer and reshapes the output into a $270 \times 320 \times 1$ array, with 86,400 weight values corresponding to the connections in the dense layer. For each GOP of 32 frames, both models are trained jointly using the Adam optimiser with a learning rate of 10^{-3} and MSE loss for up to 500 epochs at a batch size of 1, intentionally overfitting the network to the scene content. Overfitting

ensures near-lossless reconstruction of each frame from its 24-element latent vector at the cost of producing decoder parameters that are specific to each GOP.

Weight and Bias Frame Extraction

After GOP-level training, the transmitter extracts the weights and biases of the semantic decoder. The weight matrix is arranged as a sequence of 24 grayscale frames, each representing the spatial distribution of weights for one latent dimension. The bias values form one additional grayscale frame. This produces $24 + 1 = 25$ parameter frames per 32-frame GOP. Fig. 7.2 illustrates the spatial arrangement of the weight and bias frames: the smooth, structured patterns confirm that the spatial correlation assumption is valid and that block-based prediction tools can capture these patterns efficiently.

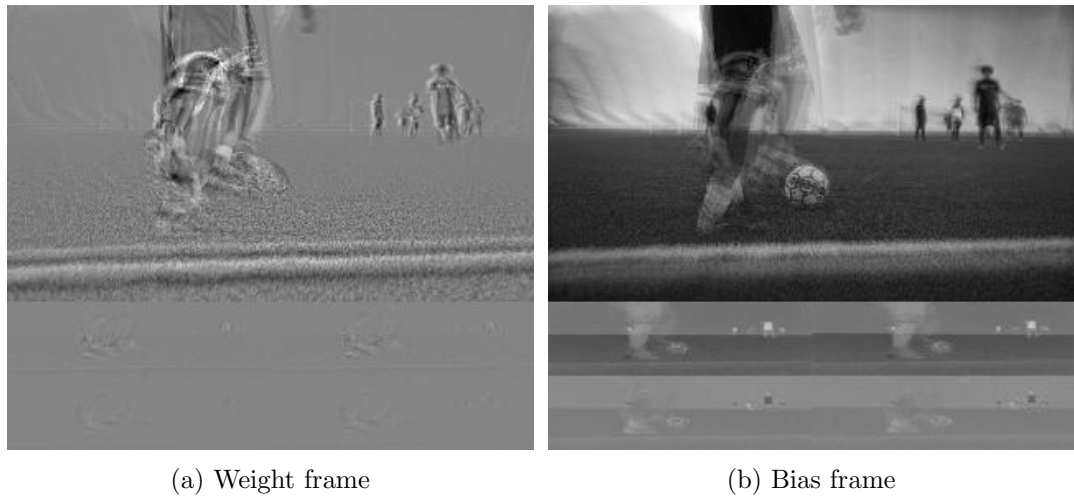


Figure 7.2: Spatial arrangement of decoder (a) weight values and (b) bias values as grayscale image frames for one GOP. The structured spatial patterns confirm that the weight and bias matrices carry spatial correlation that VVC intra- and inter-prediction tools can exploit for efficient compression [24].

VVC Encoding and Decoding

The 25 parameter frames per GOP are encoded using VVenC 1.12.0 in expert mode, configured for the slow random-access profile at a resolution of 320×270 pixels, 20 fps, YUV 4:2:0, 8-bit depth. The QP is swept from 9 to 28 to generate the rate-distortion

operating curve, and the GOP size within VVC is set to 8. The resulting VVC bitstream is transmitted to the receiver together with the 24 latent vectors produced by the encoder for the corresponding 32-frame GOP.

At the receiver, the VVC bitstream is decoded to retrieve the weight and bias frames. These are reshaped to reconstruct the decoder parameter tensors, and the decoder model is reinstantiated with these parameters. The received 24-element latent vectors are passed through the reinstantiated decoder to reconstruct each video frame. The total bitrate is the sum of the VVC parameter bitstream and the latent vector transmission overhead.

7.2.3 Results and Discussion

Experimental Setup

The proposed method is evaluated on ten YUV 4:2:0 videos at 320×180 spatial resolution, covering a range of SI and TI content (see Fig. 3.4 in Section 3.2). The SI metric quantifies the spatial detail in individual frames (high-frequency edge content), while TI captures the magnitude of frame-to-frame luminance change (motion). Performance is assessed using three quality metrics: PSNR, SSIM [164], and VMAF [165]. The baseline is NNCodec [159], the state-of-the-art neural network compression codec implementing the NNC standard. BD-rates [162] with respect to NNCodec summarise the rate-distortion gain for each video and metric.

Rate-Distortion Performance

The average rate-distortion curves over all ten test videos, shown in Fig. 7.3, demonstrate that the proposed approach consistently outperforms NNCodec at all operating points for PSNR, SSIM, and VMAF. Table 7.1 summarises the per-video BD-rate results. All values are negative, confirming bitrate savings at equivalent quality across every video and every metric.

PSNR BD-rate savings range from 38.09% (Video 1) to 78.66% (Video 3). SSIM BD-rate savings span 44.73% to 82.89%, and VMAF savings range from 51.26% to 74.35%.

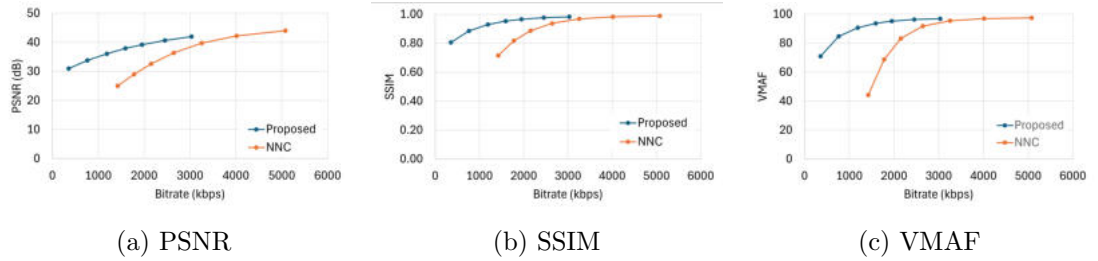


Figure 7.3: Average rate-distortion performance of the proposed VVC-based DNN parameter compression method against NNCodec over the ten test videos for (a) PSNR, (b) SSIM, and (c) VMAF [24].

Table 7.1: BD-rates (%) of the proposed VVC-based DNN parameter compression method versus NNCodec. Negative values indicate bitrate savings at equivalent quality.

Video	PSNR BD (%)	SSIM BD (%)	VMAF BD (%)
Video 1	−38.09	−44.73	−51.26
Video 2	−61.73	−62.58	−71.27
Video 3	−78.66	−82.89	−71.24
Video 4	−54.53	−56.54	−62.80
Video 5	−38.36	−49.88	−51.78
Video 6	−58.97	−70.50	−67.35
Video 7	−59.17	−65.91	−67.49
Video 8	−58.32	−67.60	−74.35
Video 9	−45.49	−55.43	−65.49
Video 10	−58.12	−64.15	−71.64

The largest PSNR BD-rate savings are concentrated among Videos 2, 3, 6, 7, 8, and 10, all of which exhibit moderate-to-high SI and TI. For these videos, NNCodec assigns high bitrate at all quality levels because it cannot adapt its entropy coding model to the spatial structure of the weight matrices. The proposed VVC encoder, in contrast, identifies the smooth gradient patterns in the weight frames as low-entropy regions and allocates bits proportionally, achieving a substantially better rate–distortion operating point.

The smallest savings are observed for Videos 1 and 5, which have lower SI and TI content. For these clips, the NNCodec bitrate is already relatively low, and the VVC encoding overhead begins to represent a larger fraction of the total bitstream, reducing the advantage. Nevertheless, even for these videos, BD-rate savings of 38% for PSNR confirm a consistent advantage of the proposed method.

Discussion and Limitations

The results establish that treating decoder weight and bias matrices as structured two-dimensional image frames enables substantially more efficient compression than the unstructured NNC approach. As motivated in the Section Introduction, the smooth gradient patterns visible in Fig. 7.2 arise from the dense layer architecture, confirming that the spatial correlation assumption holds in practice. VVC’s block-based intra- and inter-prediction tools and rate–distortion-optimised mode selection are inherently well matched to such patterns, explaining the consistent performance advantage.

However, the total bitrates of the proposed system remain much higher than those of conventional VVC applied directly to the source video, because of the fixed per-GOP cost of transmitting 25 parameter frames regardless of content. To put this in context: in the present experiments at 320×180 resolution, the parameter bitstream makes up the largest part of the total bitrate and is much higher than the bitrate of a conventional VVC encoding of the same video at a similar quality level. This confirms that the present contribution is not intended as a short-term replacement for the pre-shared decoder approach of the CASVC–CAPO system, where only 643 numbers are sent per GOP and the parameter overhead is small relative to the video bitstream. Instead, its

value lies in showing that decoder weights have spatial structure that VVC can use, setting a compression baseline for the transmitted-decoder approach and showing that treating parameters as two-dimensional images is much more efficient than treating them as a plain list of numbers in NNC. Reducing the total overhead requires either a smaller decoder (fewer weights per GOP) or longer-range temporal prediction to spread the per-scene cost over more frames. The latter is the subject of the investigation in Section 7.3. A direct comparison against conventional VVC encoding of the source video is not included in the present contribution and is identified as a key target for future work.

7.3 Temporal Prediction of Deep Neural Network Parameters for Efficient Video Coding

Publication Notice

The research presented in this section has been published in the following peer-reviewed publication:

P. Samarathunga, Y. Ganearachchi, T. Fernando, I. Alahapperuma, and A. Fernando, “Semantic Communication Based Video Coding Using Temporal Prediction of Deep Neural Network Parameters,” in *Proc. IEEE Gaming, Entertainment and Media Conf. (GEM)*, Turin, Italy, 2024, pp. 1–6, doi: 10.1109/GEM61861.2024.10585783.

This material is reproduced and adapted under the terms of the IEEE Copyright policy for use in a doctoral thesis. The content has been revised and extended for inclusion and integration in this doctoral thesis, including additional explanations, expanded discussion, and alignment with the overarching research narrative.

7.3.1 Introduction

The per-frame overfitting approach in semantic video coding produces a different decoder state for every video frame within a scene. Sending the full decoder parameters for each frame using NNCodec is too costly: the bitrate overhead grows with the num-

ber of frames, quickly using up most of the total bitstream budget and removing the compression advantage of the semantic encoder. As established in the Chapter Introduction, consecutive video frames within the same scene produce very similar decoder parameter states after per-frame overfitting, with the weight and bias values changing by only small amounts from one frame to the next. This property is similar to the temporal redundancy between consecutive video frames in conventional video coding, where motion-compensated prediction reduces the inter-frame residual to a small fraction of the original frame energy.

This section proposes a displacement-vector-based temporal prediction method for the decoder parameter space. Instead of sending the full decoder parameters for each frame, the system sends only displacement vectors that describe how the parameters have changed from the previous frame, along with the NNCodec-compressed residual of the prediction error. The transmitter and receiver align their decoder states through shared initialisation and key-frame exchange, allowing the receiver to reconstruct each frame’s decoder parameters from the compressed residual alone. Experimental results on ten test videos show PSNR gains of 3 to 25 dB and an average bitrate saving of 54% compared to the NNCodec-only baseline.

7.3.2 System Architecture

The proposed system, illustrated in Fig. 7.4, comprises four principal functional components.

Context Encoder and Decoder

The context encoder identifies scene boundaries using a scene transition detection algorithm [168] and extracts the first frame of each detected scene as the key frame. The key frame is compressed using HEIF and transmitted to the receiver. Since frames within the same scene share strong temporal correlation with the key frame, the key frame provides the shared *context* that anchors the semantic communication paradigm: both transmitter and receiver can use the same key frame as the starting point for their respective decoder training processes.

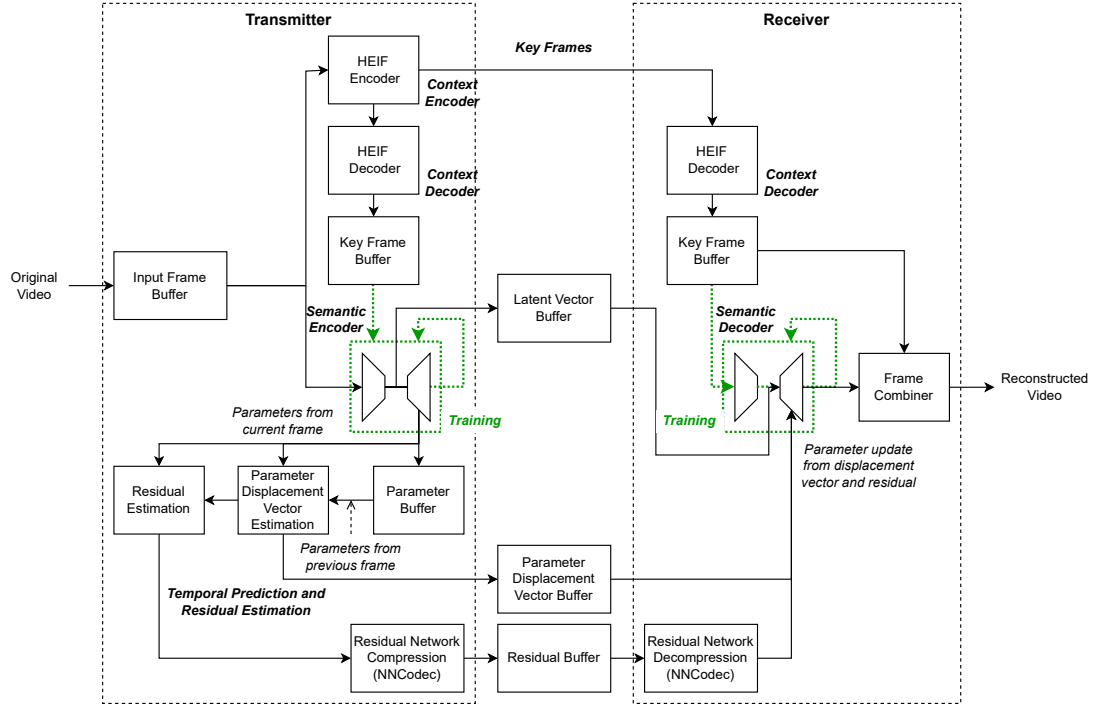


Figure 7.4: Proposed system for temporal prediction of DNN parameters. The system comprises four components: context encoder/decoder (key frame extraction and HEIF compression), semantic encoder/decoder (per-frame overfitting autoencoder), temporal prediction and residual estimation, and NNCodec-based residual compression/decompression [25].

Semantic Encoder and Decoder

The semantic encoder and decoder are implemented as a fully connected autoencoder, illustrated in Fig. 7.5. The encoder compresses each input frame to a 1-element latent vector, and the decoder reconstructs the frame from this latent representation. Both the encoder and decoder are initialized using the Glorot uniform initializer [166] with a fixed seed shared between the transmitter and receiver, ensuring identical starting weights on both sides without any additional signalling overhead.

Model: "sequential_120"		
Layer (type)	Output Shape	Param #
flatten_20 (Flatten)	(None, 86400)	0
dense_40 (Dense)	(None, 1)	86401
Total params: 86401 (337.50 KB)		
Trainable params: 86401 (337.50 KB)		
Non-trainable params: 0 (0.00 Byte)		
Model: "sequential_121"		
Layer (type)	Output Shape	Param #
dense_41 (Dense)	(None, 86400)	172800
reshape_20 (Reshape)	(None, 270, 320)	0
Total params: 172800 (675.00 KB)		
Trainable params: 172800 (675.00 KB)		
Non-trainable params: 0 (0.00 Byte)		

Figure 7.5: Network model for temporal DNN parameter prediction. *sequential_120* denotes the encoder layers and *sequential_121* the decoder layers [25].

At the transmitter, the autoencoder is overfitted to each video frame beginning from the key-frame parameter state: training continues until the per-frame reconstruction PSNR converges to the IEEE 754 single-precision numerical floor of the implementation (typically in excess of 60 dB), at which point the decoder state encodes the frame to within floating-point precision and no further descent reduces the loss. This floor, rather

than mathematical infinity, defines the termination criterion: the one-element latent vector together with the per-frame decoder state jointly carry the frame information, and the convergence threshold is set sufficiently high that the residual error contributes negligibly to the rate-distortion accounting of the subsequent NNCodec stage. The 1-element latent vector and decoder parameters are extracted after each frame. When a new scene is detected, the process restarts with the new key frame.

Temporal Prediction and Residual Estimation

The central innovation of the system is the parameter displacement vector v_k , which quantifies the shift of a block of decoder parameters between two consecutive frames. Let $w_n[k+i-1]$ denote the i -th parameter in block k of the decoder trained on frame n , and let $w_{n-1}[k+i-1-j]$ denote the corresponding parameter from the previous frame under candidate displacement j . The MSE of the block prediction under candidate displacement j is:

$$\text{MSE}_{k,j} = \frac{1}{p} \sum_{i=1}^p (w_n[k+i-1] - w_{n-1}[k+i-1-j])^2, \quad (7.1)$$

where p is the block size, k is the block index, and a is the maximum displacement search radius. The displacement vector v_k is found by minimising $\text{MSE}_{k,j}$ over the integer search range $j \in \{-a, -a+1, \dots, a\}$:

$$v_k = \arg \min_{j \in \{-a, -a+1, \dots, a\}} \text{MSE}_{k,j}. \quad (7.2)$$

The displacement vectors v_k are computed for all parameter blocks and used to predict each current decoder parameter from the previous frame's decoder. A residual is then extracted as the difference between the actual decoder parameters and the displacement-vector prediction. This residual is a vector of parameter values with substantially lower entropy than the full parameter set because the displacement prediction removes the bulk of the inter-frame variation, leaving only the unpredictable component for compression.

Residual Compression and Decompression

The residual parameter vector is compressed using NNCodec [159], with the QP varied to control the quality-bitrate trade-off. The compressed residual, together with the displacement vectors v_k and the 1-element latent vector, is transmitted to the receiver. At the receiver, the decoder parameters are updated by applying the displacement-vector prediction and adding the decompressed residual. This incremental tracking mechanism propagates the transmitter-side decoder state to the receiver at much lower bitrate than transmitting the full parameter set.

7.3.3 Motivating Evidence: Temporal Correlation of DNN Parameters

The fundamental assumption underpinning the temporal prediction is that consecutive-frame decoder parameters are highly correlated. Fig. 7.6 provides direct visual confirmation: the weight and bias distributions for two consecutive frames are nearly identical, with only small structured differences visible in the difference images. This observation supports the hypothesis that a displacement-vector search over a limited range is sufficient to predict the bulk of the parameter shift, reducing the residual to a low-entropy signal that NNCodec can compress far more efficiently than the full parameter vector.

7.3.4 Results and Discussion

Experimental Setup

The system is evaluated on the same ten YUV 4:2:0 videos at 320×180 resolution used in Section 7.2. These clips are drawn from the test set introduced in Fig. 3.3, and their SI/TI characteristics are shown in Fig. 3.4 (Section 3.2); the frames are therefore not reproduced in this chapter. Videos are classified by content complexity according to their SI and TI values: Videos 3, 4, and 5 are low complexity (low SI and TI); Videos 1, 7, and 8 are medium complexity; Videos 2, 6, 9, and 10 are high complexity. The baseline is the same semantic coding system without temporal prediction, in which

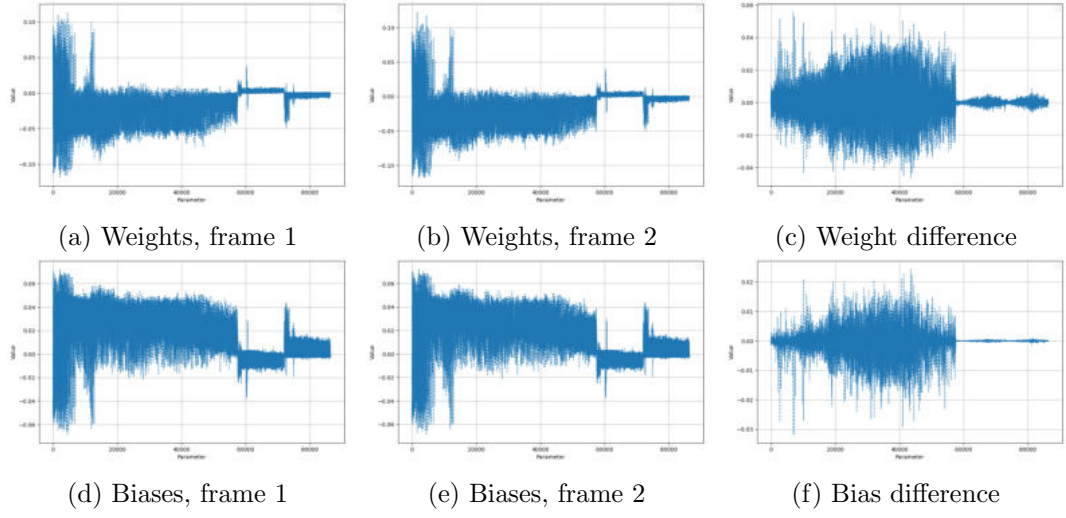


Figure 7.6: Distribution of decoder weight and bias values when trained on two consecutive video frames. The near-identical distributions in (a, b) and (d, e), and the low-amplitude difference signals in (c, f) (plotted at higher resolution), confirm strong inter-frame temporal correlation in the DNN parameter space [25].

the full residual for each frame is transmitted compressed only by NNCodec at varying QP values. Rate-distortion performance is measured as PSNR versus total bitrate.

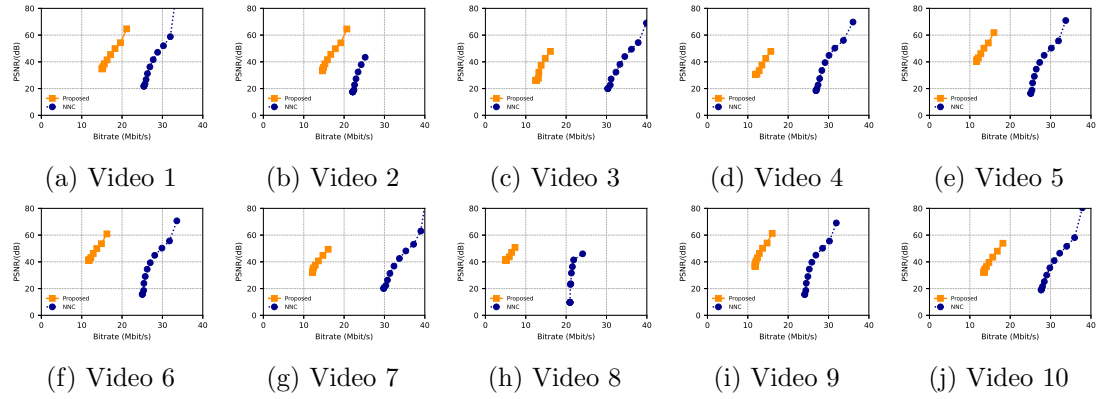


Figure 7.7: Per-video rate-distortion comparison of the proposed temporal DNN parameter prediction method against the NNCodec-only baseline for the ten test videos [25].

Rate-Distortion Gains

The proposed temporal prediction system consistently achieves better rate-distortion performance than the NNCodec-only baseline across all ten test videos, as shown in

Fig. 7.7. The overall PSNR gain ranges from 3 dB to 25 dB depending on content complexity, and the average bitrate saving is approximately 54%. These gains arise because the displacement-vector prediction removes the bulk of the inter-frame parameter variation, leaving a low-entropy residual that NNCodec can compress to a smaller bitstream for the same reconstruction quality.

The largest gains are observed for medium-complexity videos. Video 8 achieves PSNR improvements up to 30 dB at low bitrates, and Videos 1 and 7 also exhibit gains exceeding 20 dB in the low-bitrate regime. This behaviour reflects the interplay between content motion and parameter displacement: medium-complexity videos produce non-trivial inter-frame parameter shifts that are nonetheless sufficiently structured to be well-captured by the displacement vector search within the bounded search range $j \in (-a, a)$.

High-complexity videos (Videos 2, 6, 9, and 10) show more modest gains of 3 to 10 dB. For high SI/TI content, large inter-frame parameter shifts produce larger residuals that are harder to compress efficiently, reducing the advantage of temporal prediction over direct NNCodec compression. Nevertheless, the proposed system consistently outperforms the NNCodec baseline even for the most challenging content, confirming the general applicability of the approach.

Low-complexity videos (Videos 3, 4, and 5) exhibit intermediate gains of 5 to 15 dB. For low-motion content, the parameter displacement vectors are small and the residual magnitude is inherently low, so both the predicted and unpredicted residuals compress to similar sizes under NNCodec. The gain for low-complexity content is therefore primarily attributable to residual compression efficiency rather than to displacement vector prediction accuracy.

Discussion and Limitations

The experimental results confirm that temporal parameter prediction provides a meaningful compression advantage across all video complexity categories. The 54% average bitrate saving demonstrates that inter-frame correlations in the DNN parameter space are consistently exploitable across diverse content types and content complexity levels.

However, as with the spatial compression approach of Section 7.2, the absolute bitrates of both the proposed system and the NNCodec-only baseline remain much higher than those of conventional video codecs such as VVC, AV1, and HEVC, or the hybrid semantic-VVC systems described in Chapters 3 and 4. This gap is a consequence of the very simplified decoder architecture used in this proof-of-concept: the 1-element latent bottleneck severely limits the semantic encoder’s expressive capacity, placing a correspondingly large burden on the decoder parameters to encode all frame information. A more expressive decoder with a larger bottleneck, as in the 24-element design of Section 7.2, would reduce the total number of decoder parameters and bring the parameter transmission bitrate closer to that of conventional codecs.

The displacement vector search is implemented over a one-dimensional parameter sequence with a fixed block size and search range ($\pm a$). Extending the search to two-dimensional parameter matrices, as motivated by the spatial analysis in Section 7.2, may capture additional inter-frame correlation and further reduce the residual entropy. The additional complexity introduced by the displacement vector computation relative to the NNCodec-only baseline has not been quantified in this implementation; complexity characterisation and optimisation are identified as important directions for future investigation.

The temporal correlation assumption holds well within a single scene but can break down at scene changes, where a new scene produces decoder parameters that have no relation to those of the previous scene. The system already handles this through the scene transition detection in the context encoder (Section 7.3.2), which detects scene changes and resets the parameter state using a new key frame. At each scene change, the full parameter set is sent without prediction, at a cost equal to one NNCodec-compressed full parameter set; the exact overhead at scene changes has not been measured in this work but is at most equal to the NNCodec-only baseline bitrate for those frames. Looking ahead, the temporal prediction idea can naturally be extended to the per-GOP level, which is more directly relevant to the CASVC-CAPO framework. In that case, the prediction would work between consecutive GOPs in the same scene, using the small CAPO decoder update (643 numbers per GOP) as the signal to be predicted rather

than per-frame full parameters. This would let the pre-shared decoder approach benefit from the similarity between GOP parameter updates across frames in the same scene, and is a direction for future work.

7.4 Chapter Summary and Conclusion

This chapter has presented two approaches to efficient DNN parameter transmission in semantic video coding.

Section 7.2 showed that the spatial structure of decoder weight and bias matrices can be used by VVC to achieve BD-rate savings of 38% to 83% compared to NNCodec across three quality metrics and ten diverse test videos. The key finding is that weight matrices, when reshaped as two-dimensional image frames, show smooth gradient patterns from the dense layer architecture. VVC’s block-based prediction and transform coding tools capture these patterns more efficiently than the entropy coding applied to the unstructured parameter vector in NNC.

Section 7.3 showed that temporal correlation between consecutive-frame decoder parameters can be used by a displacement-vector-based prediction mechanism, reducing the residual to a low-entropy signal that NNCodec compresses with a 54% average bitrate saving over direct parameter compression. The approach is based on the same principle as motion-compensated prediction in conventional video coding, applied to the DNN parameter space, providing a clear method for inter-frame parameter coding.

Together, the two contributions show that DNN parameter compression benefits from treating the parameter space as a structured signal, whether spatial (as in Section 7.2) or temporal (as in Section 7.3). Both approaches share the same foundation as the per-GOP parameter overfitting strategy first used in the context-aware hybrid systems of Chapter 4, and provide a starting point for future joint spatial-temporal parameter compression systems. The total bitrates of both systems remain much higher than conventional video codecs, which motivates further work on compact decoder architectures, better residual compression, and longer-range temporal prediction. The two contributions in this chapter are early steps toward a future semantic video codec that transmits the decoder itself, not extensions of the pre-shared decoder approach

used in Chapters 3–5. For the near term, the pre-shared decoder with compact CAPO adaptation remains the practical choice, since it sends only 643 numbers per GOP with very little overhead. The work here explores the transmitted-decoder approach, shows which structural properties of the parameter space can be used for compression, and identifies the gap between current bitrates and what would be needed for practical use. These open problems are addressed in the research agenda of Chapter 8.

Chapter 8

Conclusions and Future Work

8.1 Summary of Contributions

This thesis has developed, evaluated, and connected a family of semantic communication systems for video coding, together with related contributions to semantic image transmission. The work covers hybrid coding, content adaptation, HDR extension, and robust wireless transmission, and addresses nine research objectives through nine original contributions that together close the seven research gaps identified in Chapter 2. The contributions progress from hybrid semantic-VVC architectures to context-aware and content-adaptive extensions, HDR domain adaptation, path-diverse robust transmission, and DNN parameter efficiency. The following subsections summarise these contributions by research thread.

Research Thread 1: Hybrid Semantic and VVC-Based Video Coding

Chapter 3 addresses Research Question RQ1 through two contributions that establish the hybrid semantic and VVC paradigm and fulfil Objectives 1 and 2.

The first contribution (Section 3.2) proposes a hybrid video coding system combining autoencoder-based semantic inter-frame prediction with VVC for intra-frame and residual coding [1]; as confirmed by the systematic search described in Section 2.12, this is the first such hybrid system reported in the literature. A remote decoder training mechanism exploits deterministic Glorot uniform initialisation to ensure that both

transmitter and receiver independently reproduce identical per-GOP-trained decoders without parameter transmission, achieving a negligible average PSNR difference between locally and remotely trained outputs. Rate-distortion evaluation across twenty test sequences demonstrates BD-rate savings of 19–58% over AVC and HEVC, and performance at or marginally above H.266/VVC for low- and high-complexity content, with a $2.65\times$ reduction in per-frame VVenC encoding time through semantic pre-processing.

The second contribution (Section 3.3) extends the hybrid codec with bidirectional hierarchical frame prediction via RIFE, block-based high-resolution processing, and DeepCABAC residual latent entropy coding [22]. The enhanced system achieves a PSNR BD-rate of -21.1% relative to VTM 17.0 while also demonstrating graceful degradation under channel bit errors at rates where VTM undergoes catastrophic quality collapse. Ablation analysis establishes that proposed RIFE-based bidirectional prediction is the dominant contributor to the BD-rate gain.

Research Thread 2: Context-Aware and Content-Adaptive Semantic Video Coding

Chapter 4 addresses both Research Questions RQ1 and RQ2 through two contributions. Contribution 3 extends the hybrid semantic-VVC thread by eliminating per-GOP re-training through a shared pretrained model, thereby fulfilling Objective 3 under RQ1. Contribution 4 then recovers per-scene content adaptivity through lightweight decoder specialisation, thereby fulfilling Objective 4 under RQ2.

The third contribution (Section 4.2) introduces a context-conditioned semantic video coding framework that replaces per-GOP from-scratch training with a shared pretrained autoencoder and a motion-adaptive model selection strategy [73]. Proposed RIFE-based hierarchical bidirectional temporal interpolation generates motion-aligned context frames that condition the decoder without transmitting any motion information, and differential PCA-DeepCABAC entropy coding constrains semantic latent overhead below 0.4% of total bitrate. The framework achieves an average PSNR BD-rate of -22.03% relative to VTM 17.0.

The fourth contribution (Section 4.3) introduces a content-adaptive parameter over-

fitting mechanism that reintroduces per-GOP content specialisation at a cost of only 643 scalars per GOP, less than 0.1% of the full decoder [73]. Per-channel learnable scale multipliers and transposed-convolution biases enable the decoder to shift its reconstruction prior from the global Vimeo-90k average to the local statistics of each GOP, reducing residual bitrate by 9.01% relative to the CASVC baseline while a tapered autoencoder architecture reduces per-block inference by $4.9\times$. The combined framework achieves average BD-rates of -23.9% PSNR relative to VTM 17.0, with consistent generalisation to other video types such as CCTV surveillance and 3D animation contents.

Research Thread 3: Semantic Video Coding for High Dynamic Range Content

Chapter 5 addresses the HDR extension aspect of Research Question RQ2, fulfilling Objective 5 by extending the CASVC-CAPO framework to HDR sources through a non-architectural adaptation.

The fifth contribution (Section 5.2) introduces an automatic, per-frame, fully invertible Reinhard tone-mapping pre-process [150] that bridges the domain gap between linear HDR EXR source material and the SDR training distribution of the semantic autoencoder. An exposure multiplier and a gamma exponent are estimated in closed form from log-average luminance, mapping each linear HDR EXR frame to a ten-bit SDR frame whose statistics match the autoencoder’s training distribution, while the two scalar parameters per frame are CABAC-coded as a compact HDR metadata sidestream that enables exact closed-form HDR recovery at the receiver. Evaluation on the HDM-HDRV cinematic dataset and the JVET HDR/WCG common test condition sequences yields average BD-rate savings of -16.42% PSNR, -14.14% SSIM, -10.27% VMAF, -15.31% PU-PSNR, and -11.12% PU-SSIM relative to VTM 23.0, with consistent gains across both SDR-domain and HDR-domain perceptual metrics.

Research Thread 4: Robust Semantic Communication via Multiple Description Coding

Chapter 6 addresses Research Question RQ3 through two contributions that provide path diversity and guaranteed perceptual quality floors for wireless transmission, fulfilling Objectives 6 and 7.

The sixth contribution (Section 6.2) proposes the SMDC framework for robust image transmission. A novel interleaved channel-wise splitting strategy generates three semantically complementary latent descriptions from a shared bottleneck tensor, each independently protected by polar codes and transmitted over a dedicated wireless path. Evaluated on the DIV2K validation set, SMDC maintains PSNR above 30 dB and SSIM above 0.80 down to 1 dB, 7 dB, and 5 dB combined SNR for AWGN, Rayleigh fading, and Rician fading channels respectively, compared with catastrophic failure of conventional baselines (JPEG, JPEG2000, HEIF, and CompressAI) at 5 dB, 12 dB, and 10 dB.

The seventh contribution (Section 6.3) extends SMDC to video transmission under a GOP-8 structure. A context-conditioned inter-frame MDC encoder transmits only the differential latent between each inter-frame and its RIFE-interpolated temporal context, substantially reducing transmission entropy. Evaluated across 20 test video sequences, the framework achieves 4–5 dB combined SNR robustness gains over a VVenC baseline under all three channel models, with SMDC-V reaching VMAF above 70 at approximately 1 dB, 7 dB, and 6 dB combined SNR for AWGN, Rayleigh, and Rician channels against approximately 4 dB, 9 dB, and 9 dB for VVenC MDC, and with a $69\times$ encoding speedup over the VVenC MDC reference (15.2 fps versus 0.22 fps).

Research Thread 5: DNN Parameter Efficiency for Semantic Video Coding

Chapter 7 addresses Research Question RQ4 through two contributions that reduce the overhead of transmitting DNN decoder parameters, fulfilling Objectives 8 and 9.

The eighth contribution (Section 7.3) exploits temporal correlations between DNN parameters trained on consecutive video frames through a displacement-vector-based

prediction mechanism [25]. Parameters are predicted from key frames and displacement vectors, and only the NNCodec-compressed residual is transmitted, achieving PSNR gains of 3–25 dB and an average bitrate saving of 54% relative to direct NNCodec compression across ten diverse test video sequences.

The ninth contribution (Section 7.2) treats DNN weight matrices as two-dimensional signals and compresses them using VVC, exploiting the spatial correlation structure of dense layer weight matrices when reshaped as 2D images [24]. VVC’s block-based prediction and transform coding tools capture the smooth gradient patterns of weight matrices more efficiently than entropy coding of the unstructured parameter vector, achieving BD-rate savings of 38% to 83% relative to the NNCodec baseline across all three evaluated quality metrics and ten diverse test video sequences.

8.2 Key Findings

The nine contributions presented in this thesis lead to several main empirical findings that are important for both visual communication based on semantic communication research and the practical design of future wireless visual transmission systems. These findings are organised according to the four research questions defined in Chapter 1.

RQ1: Hybrid semantic–VVC video coding performance. The results of Chapters 3 and 4 demonstrate that hybrid semantic–conventional architectures achieve performance competitive with or exceeding both pure conventional codecs and state-of-the-art learned codecs. The enhanced hybrid system [22] surpasses DCVC-FM and DCVC-DC on the PSNR BD-rate metric, showing that the hybrid paradigm can exceed fully learned approaches. A key finding from the ablation study in Chapter 3 is that bidirectional hierarchical prediction is the single most valuable component of the temporal modelling pipeline: its removal degrades performance from -21.1% to $+4.68\%$ BD-rate relative to VTM, a swing of approximately 26 percentage points. This finding identifies temporal modelling quality as the primary lever for further rate–distortion improvement in hybrid semantic video codecs.

RQ2: Content-adaptive specialisation and HDR extension. The CAPO mechanism shows that fewer than 0.1% of the decoder parameters are sufficient to recover the per-sequence content adaptivity of full from-scratch retraining. This result has important theoretical implications: it suggests that the residual information required to adapt a pretrained generic model to a specific video sequence is extremely low-dimensional, and that the adaptation can be described by a compact affine transformation of decoder feature channels. The HDR investigation shows the related finding that domain adaptation between SDR and HDR sources can be achieved entirely within the pre-processing and metadata pipeline without any architectural modification of the trained semantic codec. Two scalar tone-mapping parameters per frame are sufficient to bridge the HDR–SDR domain gap, achieving consistent BD-rate savings of 10–16% across multiple HDR-domain perceptual metrics.

RQ3: MDC robustness and guaranteed quality floors. The SMDC results of Chapter 6 provide the first quantitative description of the robustness benefit of semantic-domain multiple description coding versus pixel-domain splitting. The 4 dB combined SNR advantage of SMDC over conventional baselines on AWGN channels directly measures the value of distributing complementary semantic content across descriptions rather than duplicating the same semantic information in multiple copies. The video SMDC results confirm that the architectural principles established for image transmission transfer to the video domain, with the context-conditioned inter-frame module preserving the graceful-degradation properties of the three-description architecture while exploiting temporal correlations. Crucially, SMDC maintains PSNR above 30 dB and SSIM above 0.80 down to 1 dB combined SNR for AWGN channels, eliminating the cliff effect that causes conventional entropy-coded systems to fail catastrophically at low SNR and establishing a guaranteed perceptual quality floor unreachable by separation-based baselines.

RQ4: DNN parameter compression efficiency. The two DNN parameter efficiency contributions of Chapter 7 show that substantial compression gains are available by treating the parameter space as a structured signal. The spatial correlation approach

achieves BD-rate savings of 38–83% over NNCodec by exploiting the smooth gradient patterns of weight matrices when reshaped as 2D images; the temporal prediction approach achieves 54% average bitrate saving by exploiting strong inter-frame correlations in parameters trained on consecutive video frames. A key insight is that both spatial and temporal redundancies exist in the parameter space at the same time, suggesting that a future system combining both approaches within a joint spatial–temporal parameter coding framework could achieve much larger reductions than either approach alone.

8.3 Overall Conclusion

This thesis demonstrates to video communication researchers that semantic communication is a powerful alternative to both signal-processing-based conventional video coding and to fully learned video coding, across standard SDR video, HDR video, and wireless transmission scenarios. Across hybrid architectures, content adaptation, HDR extension, robust wireless transmission, and DNN parameter efficiency, the presented work consistently shows that semantic communication systems are competitive with and often exceed conventional codecs such as H.266/VVC and learned codecs such as DCVC, achieving BD-rate savings of up to 23.9% while eliminating the catastrophic cliff effect that occurs in conventional entropy-coded systems under low-SNR wireless channels. The framework adapts to HDR content using only two scalar parameters per frame without any architectural modification, generalises across diverse content types including cinematic, surveillance, and animation sequences, and maintains perceptual quality floors down to 1 dB combined SNR on AWGN channels.

Beyond video coding performance, this thesis also shows that the same paradigm can be extended in a different direction: using it to support network-compression-based video coding. The key insight from the CAPO mechanism and the DNN parameter compression work is that fewer than 0.1% of decoder parameters are sufficient for content-specific adaptation, and that weight matrices exhibit strong spatial and temporal redundancies. This insight directly enables the compression and transmission of

neural network models themselves over bandwidth-constrained links. It opens a recursive paradigm in which semantic video codecs transmit not only the compressed video signal but also the DNN decoders that reconstruct it, enabling dynamic model deployment, edge adaptation, and distributed semantic inference. The thesis therefore establishes that semantic communication is not merely a video coding tool but a foundational framework for network-aware, model-transmitting visual communication systems, and points toward future self-adapting codecs that compress both the message and the means of its interpretation.

8.4 Implications for Research and Industry

The contributions of this thesis have implications at several levels: for academic research in semantic communication and video compression, for the design of next-generation wireless communication systems, and for practitioners developing intelligent video and image transmission pipelines.

Implications for semantic communication research. This thesis establishes several reusable methodological frameworks. The hybrid semantic-VVC paradigm of Chapters 3 and 4 shows that semantic communication research benefits from integration with, rather than replacement of, proven conventional coding standards, using decades of optimised entropy coding and bitstream tools within a learned semantic front-end. The content-adaptive parameter overfitting mechanism of Chapter 4 provides a general design template for restoring per-instance adaptivity to pretrained foundation models at negligible bitrate cost, with potential application well beyond semantic video coding. The SMDC framework of Chapter 6 provides the first end-to-end framework for joint semantic source coding and path-diverse channel coding, establishing a design template for robust semantic communication under realistic wireless fading conditions. The DNN parameter efficiency contributions of Chapter 7 identify parameter transmission overhead as a critical research challenge for deployable semantic video coding and provide two complementary solutions that motivate future work on jointly optimised spatial-temporal parameter compression.

Implications for sixth-generation wireless networks. The sixth-generation (6G) wireless standard, currently under development [30,31], identifies native intelligence and semantic communication as key distinguishing features relative to fifth-generation (5G) systems. The results of this thesis provide quantitative evidence for the rate–distortion and robustness gains that semantic communication systems can achieve under realistic channel conditions. The SMDC framework’s demonstrated ability to maintain PSNR above 30 dB and SSIM above 0.80 down to 1 dB combined SNR for AWGN channels is of direct relevance to 6G link budget design for mission-critical applications including autonomous vehicle telemetry, remote surgical systems, and disaster-response communications, where minimum quality-of-service guarantees must be maintained even under worst-case channel conditions. The DNN parameter efficiency frameworks of Chapter 7 address a key deployment barrier for the model-distribution component of any 6G semantic communication infrastructure, where the bandwidth required to distribute updated decoder parameters across heterogeneous edge devices must remain a small fraction of the payload channel capacity.

Implications for broadcasting and HDR streaming. The HDR semantic video coding framework of Chapter 5 is of direct relevance to broadcast and streaming services migrating from SDR to HDR content. The demonstration that a compact closed-form invertible tone-mapping pre-process is sufficient to bridge the SDR–HDR domain gap, achieving 10–16% BD-rate reductions relative to VTM 23.0 on standard HDR benchmark sequences, suggests that semantic video codecs trained on existing SDR datasets can be used for HDR production workflows without full retraining. The PU-PSNR and PU-SSIM improvements confirm that the quality gains are perceptually meaningful in the HDR luminance range, validating the suitability of the framework for premium video distribution and HDR archive applications.

8.5 Limitations

Despite the range and depth of the contributions presented in this thesis, several limitations should be acknowledged. These limitations define the boundaries of the reported

results and point to areas that need more study.

Encoder computational complexity. Most contributions involve offline encoding workflows. The CASVC and CAPO frameworks of Chapter 4 operate at approximately 1.23 fps and 0.18 fps encoding throughput respectively on an NVIDIA A100 GPU, which precludes real-time application. The per-GOP parameter overfitting of Section 4.3 requires a small-batch gradient descent procedure for each GOP, adding latency that is incompatible with live streaming or interactive applications. While decoder throughputs are generally higher (19.2 fps for CASVC), practical one-way semantic video transmission remains limited to scenarios where significant encoding latency is acceptable, such as pre-encoded broadcast production and archival transcoding. The hardware qualification and comparison with DCVC-RT are given in Section 4.2.3. Because all experiments used an NVIDIA A100 through the shared Google Colaboratory environment [183], the reported throughput should not be interpreted as a hardware-normalised comparison with dedicated, cuDNN-optimised fp16 implementations; real-time encoder performance remains unproven. The wireless channel models used throughout this thesis (AWGN, Rayleigh fading, and Rician fading) are theoretical approximations that do not capture the full complexity of practical wireless propagation environments. Real-world channels exhibit frequency-selective fading, Doppler shift from relative motion between transmitter and receiver, spatially correlated interference, and time-varying path geometry. The robustness demonstrations of Chapter 6 should therefore be interpreted as performance under idealised conditions; practical deployment will require channel-specific re-evaluation and possibly channel-adaptive system reconfiguration.

Dataset scale and diversity. The semantic video coding systems of Chapter 4 are trained on the Vimeo-90k dataset, which is representative of consumer-grade video content but does not cover specialised domains such as medical imaging, satellite remote sensing, or infrared thermal imagery. Performance on strongly out-of-distribution content may not replicate the benchmark results, and systematic evaluation across diverse content domains remains a direction for future investigation. The HDR evaluation of Chapter 5 is conducted on the HDM-HDRV and JVET HDR/WCG common test

condition sequences; while these are the established benchmarks for HDR codec evaluation, further validation on user-generated HDR content captured under uncontrolled conditions would strengthen the practical generality of the reported gains.

DNN parameter transmission overhead. The absolute DNN parameter transmission overhead of the systems in Chapter 7 remains much higher than the motion vector overhead of standard conventional codecs. While the CAPO mechanism of Chapter 4 limits adaptation to 643 scalars per GOP (less than 0.1% of the full decoder and an average of 0.69% of total bitrate), the full-parameter approaches of Sections 7.2 and 7.3 involve transmitting entire weight and bias matrices that, even after VVC-based or NNCodec-based compression, constitute bitstreams substantially larger than the semantic latent representations. Closing this gap requires compact decoder architectures and joint parameter–latent rate–distortion optimisation.

Absence of a standardised semantic bitstream syntax. None of the systems proposed in this thesis defines a standardised bitstream syntax for the semantic latent representation. Semantic sub-streams are specific to the particular model architecture and training procedure; a model retrained with different hyperparameters produces semantically incompatible bitstreams. This fragmentation prevents interoperability between independently trained systems, complicates long-term storage and retrieval applications, and is a significant barrier to commercial adoption. The development of a standardised semantic codec bitstream format, similar to the bitstream syntaxes of H.264/AVC or H.265/HEVC, is identified as a critical near-term research and standardisation priority.

8.6 Future Work

The contributions and limitations identified in this thesis motivate several concrete research directions for future investigation.

Fully independent semantic video coding via neural network overfitting. All hybrid semantic–VVC frameworks developed in Chapters 3–5 rely on VVC as the residual coding backbone. The author’s initial work in Chapter 7 explores a complementary direction: overfitting a compact autoencoder to each video scene and transmitting the learned decoder weights and biases as the compressed representation, so that the receiver reconstructs every frame purely by running the received decoder. Showing that DNN parameters carry structured spatial and temporal redundancy that can be compressed to competitive bitrates supports the feasibility of this approach [24, 25, 161]. However, achieving rate–distortion performance that is competitive with VVC-backed hybrid codecs across diverse content types and resolutions without a conventional residual coding stage remains an open research challenge. Key directions include joint optimisation of the overfitting autoencoder architecture and the parameter compression pipeline under a unified rate–distortion objective, adaptive precision allocation across weight tensors based on their contribution to reconstruction quality, and the design of lightweight decoder architectures whose parameter count is small enough that the transmission overhead is negligible relative to the semantic latent bitrate. Progress on this challenge would yield a fully standardisation-free semantic video codec whose only transmitted items are neural network parameters and latent vectors.

Joint spatial–temporal DNN parameter compression. Section 7.2 and Section 7.3 independently exploit spatial correlations within weight matrices and temporal correlations between consecutive parameter frames. A natural next step is a joint spatial–temporal parameter coding framework that simultaneously applies VVC-based 2D spatial prediction and displacement-vector-based temporal prediction within a unified rate–distortion optimised pipeline. Such a joint framework would be expected to clearly outperform either individual approach, since it exploits two different dimensions of the parameter redundancy structure that are both present in the DNN parameter space of semantic video decoders.

Joint optimisation of semantic encoding and VVC residual coding. All hybrid semantic–VVC systems developed in Chapters 3–5 are trained in two separate

stages: the semantic autoencoder is trained first under a rate–distortion objective, and VVC is then applied to compress the residual signal left by the semantic decoder. Because these two stages are trained independently, the semantic encoder has no way to adjust its output based on how well VVC can compress the resulting residual, which limits the overall coding efficiency. A jointly trained system would include a differentiable approximation of the VVC coding process, covering the DCT quantisation and CABAC entropy coding steps, directly in the semantic training loss, so that both the semantic codec and the residual compressor are optimised together. This approximation could take the form of a learned rate estimator that predicts the bitrate VVC would need to compress a given residual, and this estimate would be added to the training loss to encourage the semantic decoder to produce residuals that VVC can compress more efficiently. Such a jointly trained system would be expected to reduce the residual bitrate at the same reconstruction quality, giving further BD-rate improvements over the separate approach used in this thesis, without any changes to the VVC standard or its decoder.

Integration with large-scale visual foundation models. The semantic autoencoders used throughout this thesis are compact convolutional architectures trained on moderate-scale datasets. Recent large-scale visual foundation models based on vision transformers and self-supervised representation learning have demonstrated substantially richer and more generalisable semantic representations. Replacing the semantic encoder backbone with a pretrained large-scale foundation model could improve the quality and content generalisability of the semantic latent representation, with the trade-off of increased encoder-side computational complexity. This direction is particularly worth exploring for the CASVC framework of Section 4.2, where the quality of the pretrained semantic representation directly determines the residual bitrate and thus the overall rate–distortion performance.

Real-time encoder acceleration. The CASVC–CAPO framework is too computationally demanding for real-time operation on current GPU hardware. Several speed-up strategies are needed to make the encoder fast enough for live video applications, includ-

ing lower-precision arithmetic, early stopping of the encoding process, initialising each GOP from the previous GOP’s parameters, reducing the size of the semantic autoencoder, and more efficient per-channel scale multiplier updates. Field-Programmable Gate Array (FPGA) and application-specific integrated circuit implementations of the per-GOP overfitting process are also worth exploring for embedded encoder hardware.

Semantic communication under adversarial and non-stationary conditions.

The systems proposed in this thesis are tested under fixed statistical channel models with standard noise. In practice, semantic communication systems may face hostile conditions such as jamming attacks, deliberate changes to the transmitted semantic data, and changing interference patterns. How well semantic encoders hold up against such attacks, and how to detect and respond to them, are important open problems that are especially relevant for safety-critical applications such as autonomous vehicle telemetry and remote surveillance.

Federated learning for distributed semantic communication. The semantic codecs in this thesis are trained on a fixed dataset in a single location. In practice, devices used in the field may see video content that differs from the training data, and collecting such data centrally is often not possible due to privacy concerns. Federated learning allows semantic encoders to be trained across many devices without sharing raw data, so the model can improve from diverse real-world content while keeping each device’s data private. The CAPO framework fits this setting well, since only its 643 small adapter parameters need to be shared with a central server during training, rather than the full model weights.

Evaluation against learned HDR codecs and royalty-free standards. The HDR semantic video coding framework of Chapter 5 is evaluated against VTM 23.0 as the reference. Comparison against dedicated learned HDR video codecs [23] and against the royalty-free AV2 standard [85] would provide a more complete picture of the proposed approach within the broader HDR compression landscape. Such comparisons are identified as a high priority for future evaluation.

Multi-user and broadcast semantic communication. All systems in this thesis are designed for transmission between one sender and one receiver. Many real-world applications, such as broadcast streaming, vehicle communication, and shared edge computing, require sending data from one source to many receivers, or from many sources to one receiver. Designing semantic communication systems for these broadcast and multicast settings is an important open research direction, since a single transmitted semantic representation must be decoded by different receivers that may have different channel conditions, hardware, and quality requirements. The channel-wise splitting approach of the SMDC framework in Chapter 6 is a good starting point for this, as it allows different receivers to decode different subsets of descriptions based on their available bitrate and quality needs.

8.7 Closing Remarks

This thesis set out to show that semantic communication can deliver real and consistent improvements over conventional and learned video codecs, not only in theory but also across the practical challenges that video transmission systems must handle. The nine contributions presented across five research threads provide a clear and positive answer to this question.

For standard dynamic range video coding, the hybrid semantic-VVC framework achieves competitive rate-distortion performance against both conventional codecs and state-of-the-art learned codecs on BD-rate metrics, with consistent gains shown across a range of video content. Content adaptivity is achieved at very low cost by updating only a small fraction of the decoder parameters per GOP, which is enough to match the performance of a fully content-specific model. For HDR content, only two scalar tone-mapping parameters per frame allow an SDR-trained codec to perform well against VTM on both SDR-domain and HDR-domain quality metrics, without any change to the codec design or training. For robust wireless transmission, the SMDC framework removes the cliff effect by keeping good image quality at very low SNR levels over AWGN channels, where conventional systems fail completely. For DNN parameter transmission, temporal prediction and VVC-based compression both reduce the bitrate

needed to send updated decoder parameters, making content-adaptive semantic video coding practical to use.

These results support the main idea of this work: that sending task-relevant meaning is a more efficient and more robust goal than sending an exact copy of every pixel, and that this holds across different video types, channel conditions, and use cases. The contributions also identify the real limits of these results. High encoder complexity means current systems are suited to offline use rather than real-time applications. Training on standard benchmark data means performance on very different content types has not been fully studied. The channel models used are theoretical, not based on real-world measurements. None of the systems defines a standard bitstream format, which limits interoperability. The future research directions in Section 8.6 address each of these issues.

The main finding of this thesis is that semantic communication and conventional coding work best together, not against each other. The strongest results in this work come from combining the compression power of learned semantic encoding with the reliable entropy coding and bitstream tools of VVC. This relationship holds across all four research areas of the thesis and is its central contribution to the field. It suggests that the next generation of video communication systems will not be built by replacing conventional standards, but by combining them with semantic communication. As global video data traffic continues to grow and the demands of emerging applications continue to increase, semantic communication will play an important role in the design of next-generation video communication systems.

Appendix A

Individual Rate–Distortion Performance for Test Videos for GOP Sizes 8, 16, and 32

The individual rate–distortion performance of each test video for GOP sizes of 8, 16, and 32 is shown in Figs. A.1–A.4, Figs. A.5–A.8, and Figs. A.9–A.12, respectively. For each video clip, the PSNR, SSIM, and VMAF performance of the proposed system (orange with the symbol $\circ$) is shown together with that of VVC (blue with the symbol $\diamond$), HEVC (blue with the symbol $\square$), and AVC (blue with the symbol $\triangle$). Individual video clips are identified by the prefix V and two digits to represent the video number, as referred to in Chapter 3.

The figures in the main text show a summarised version of these results by presenting the averages of PSNR, SSIM and VMAF when the 20 test videos are grouped in terms of complexity.

Appendix A. Individual Rate–Distortion Performance for Test Videos for GOP Sizes 8, 16, and 32

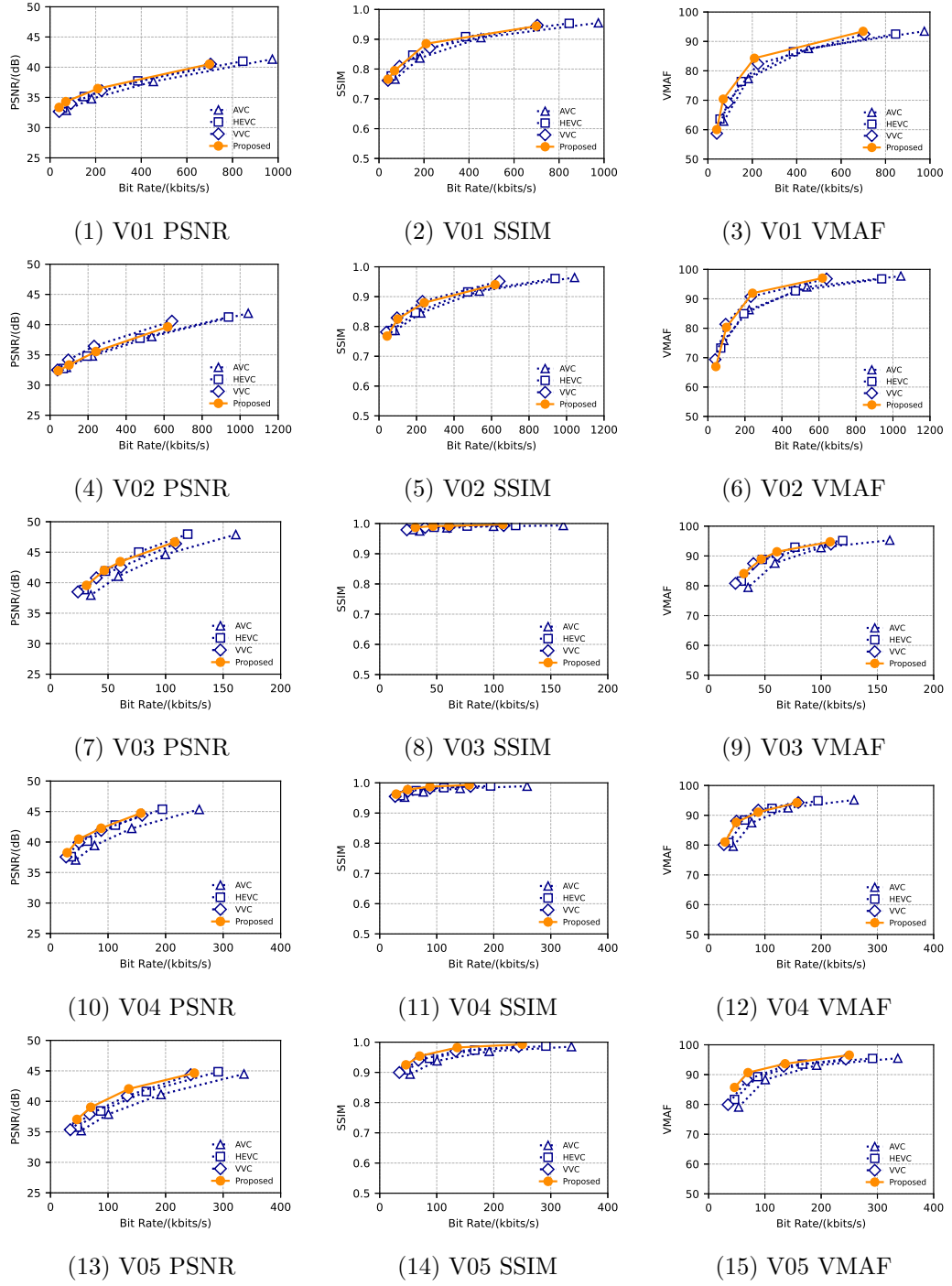


Figure A.1: Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V01–V05, GOP size 8.

Appendix A. Individual Rate–Distortion Performance for Test Videos for GOP Sizes 8, 16, and 32

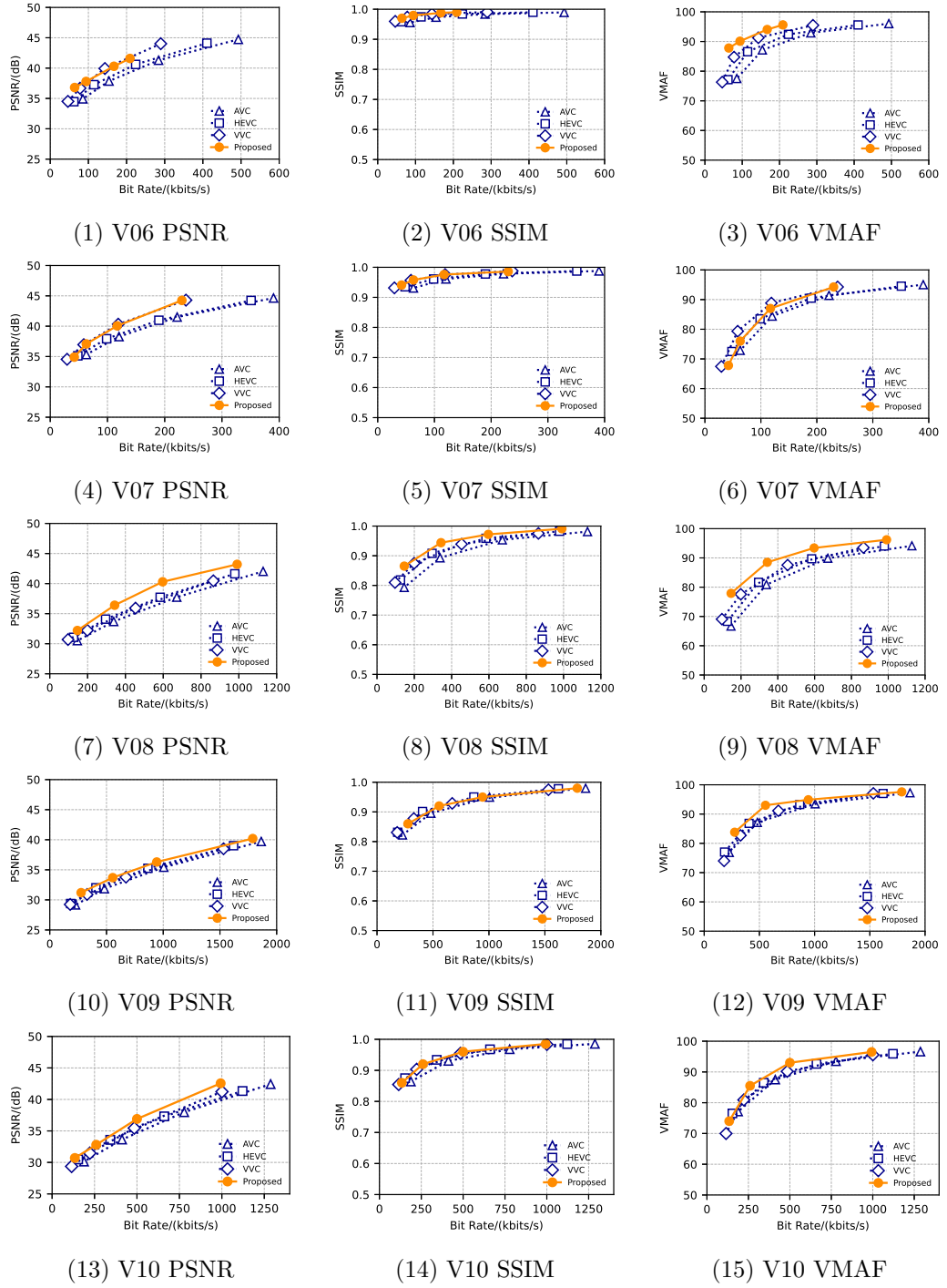


Figure A.2: Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V06–V10, GOP size 8.

Appendix A. Individual Rate–Distortion Performance for Test Videos for GOP Sizes 8, 16, and 32

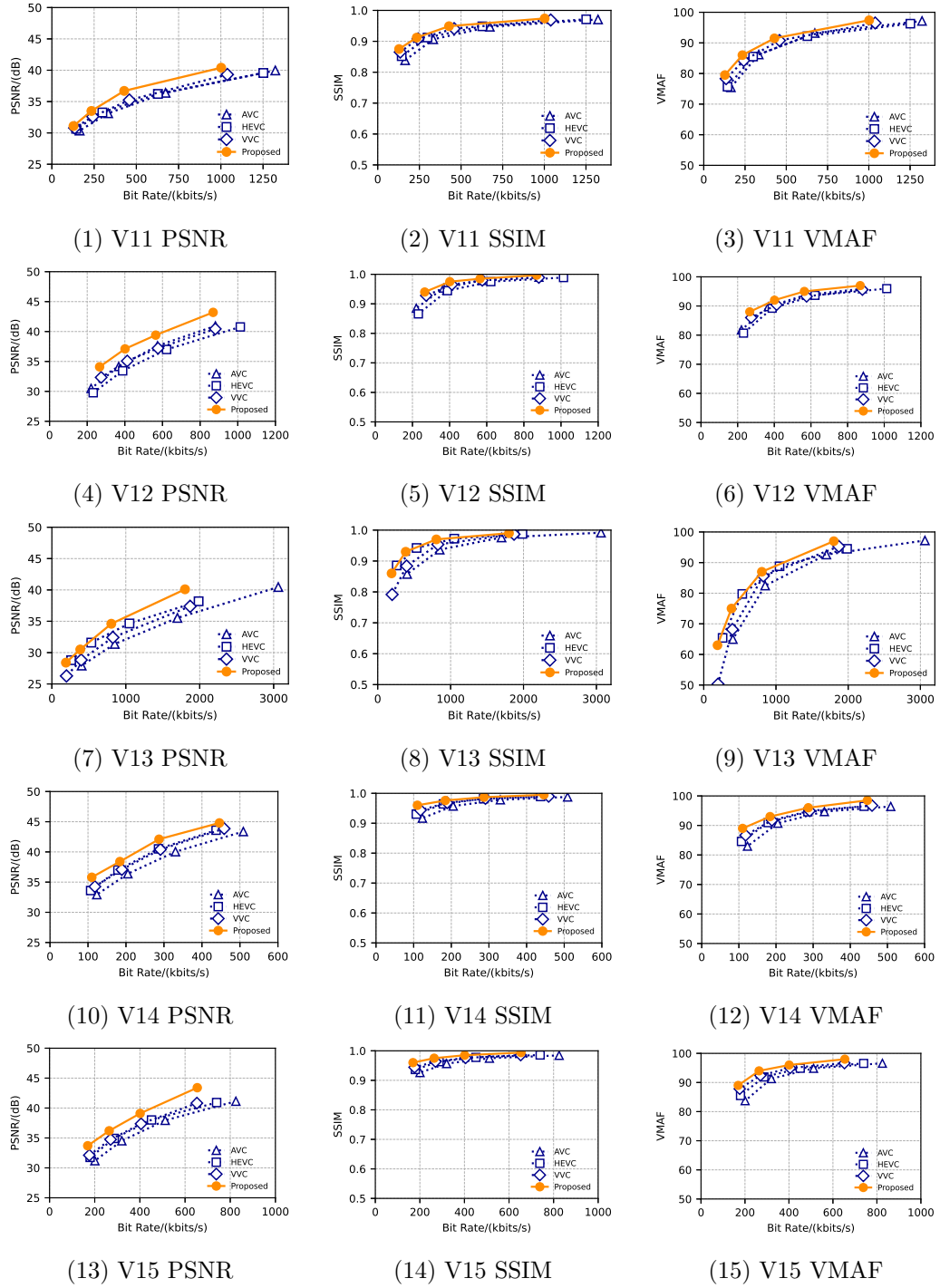


Figure A.3: Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V11–V15, GOP size 8.

Appendix A. Individual Rate–Distortion Performance for Test Videos for GOP Sizes 8, 16, and 32

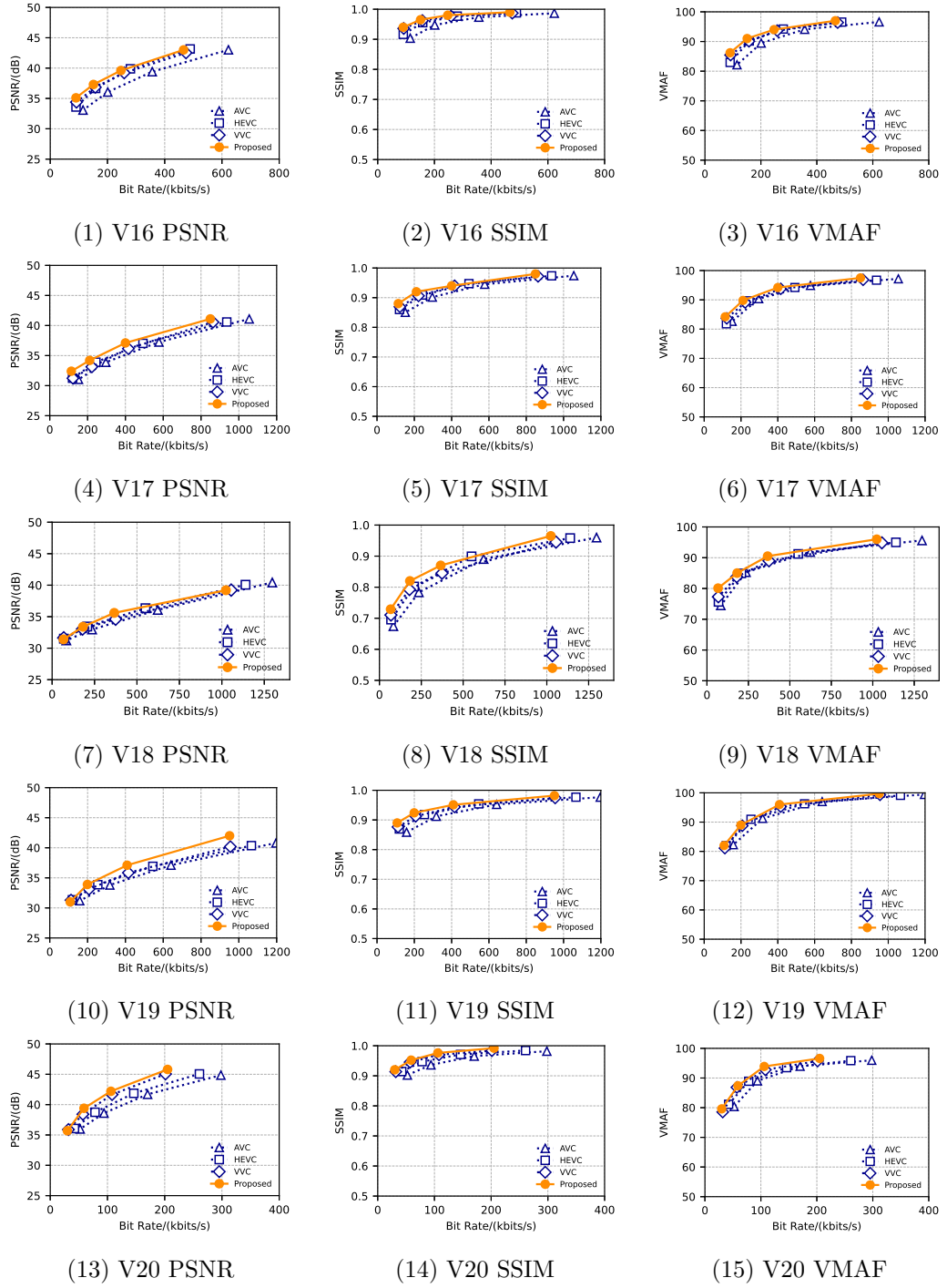


Figure A.4: Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V16–V20, GOP size 8.

Appendix A. Individual Rate–Distortion Performance for Test Videos for GOP Sizes 8, 16, and 32

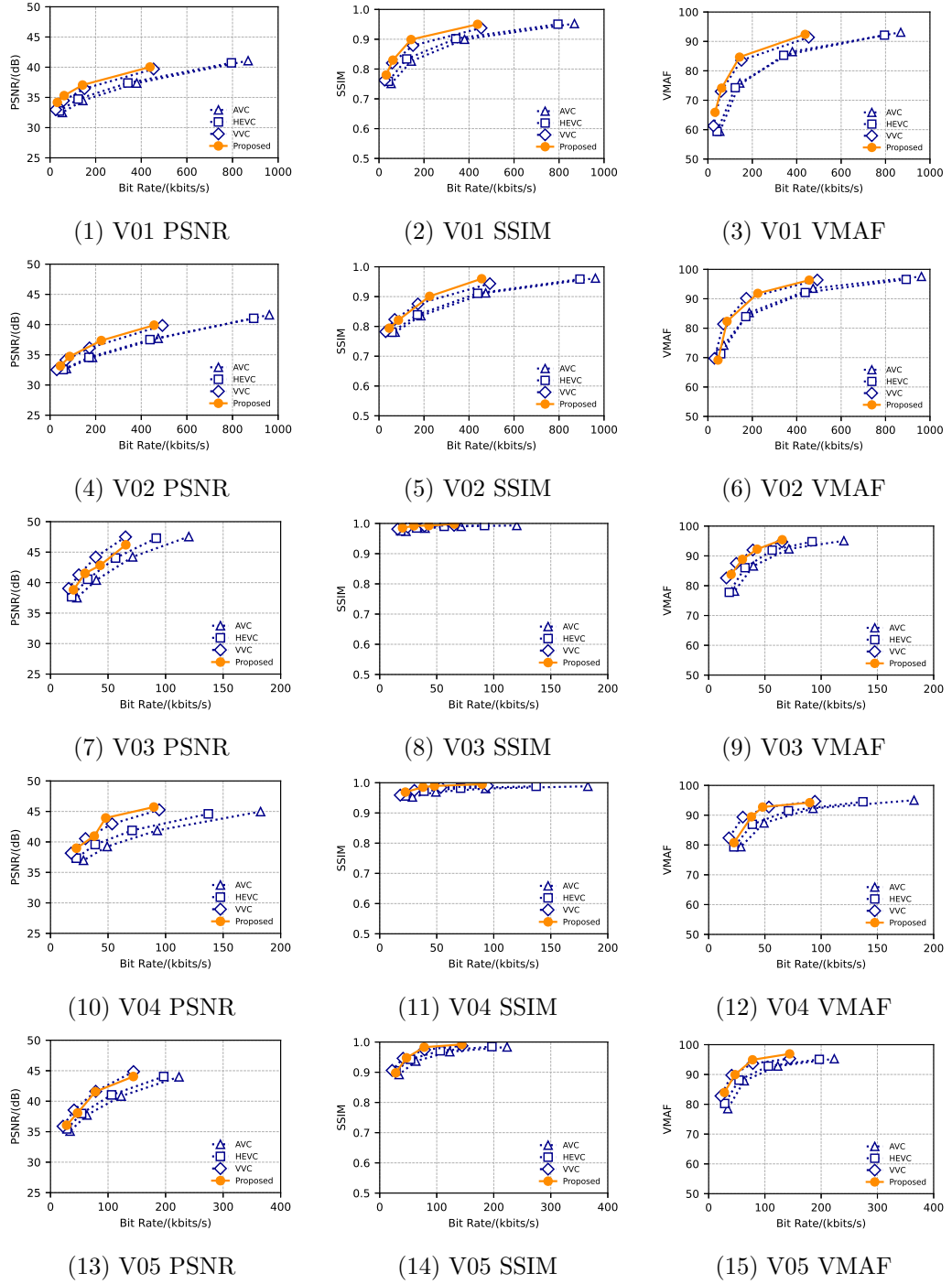


Figure A.5: Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V01–V05, GOP size 16.

Appendix A. Individual Rate–Distortion Performance for Test Videos for GOP Sizes 8, 16, and 32

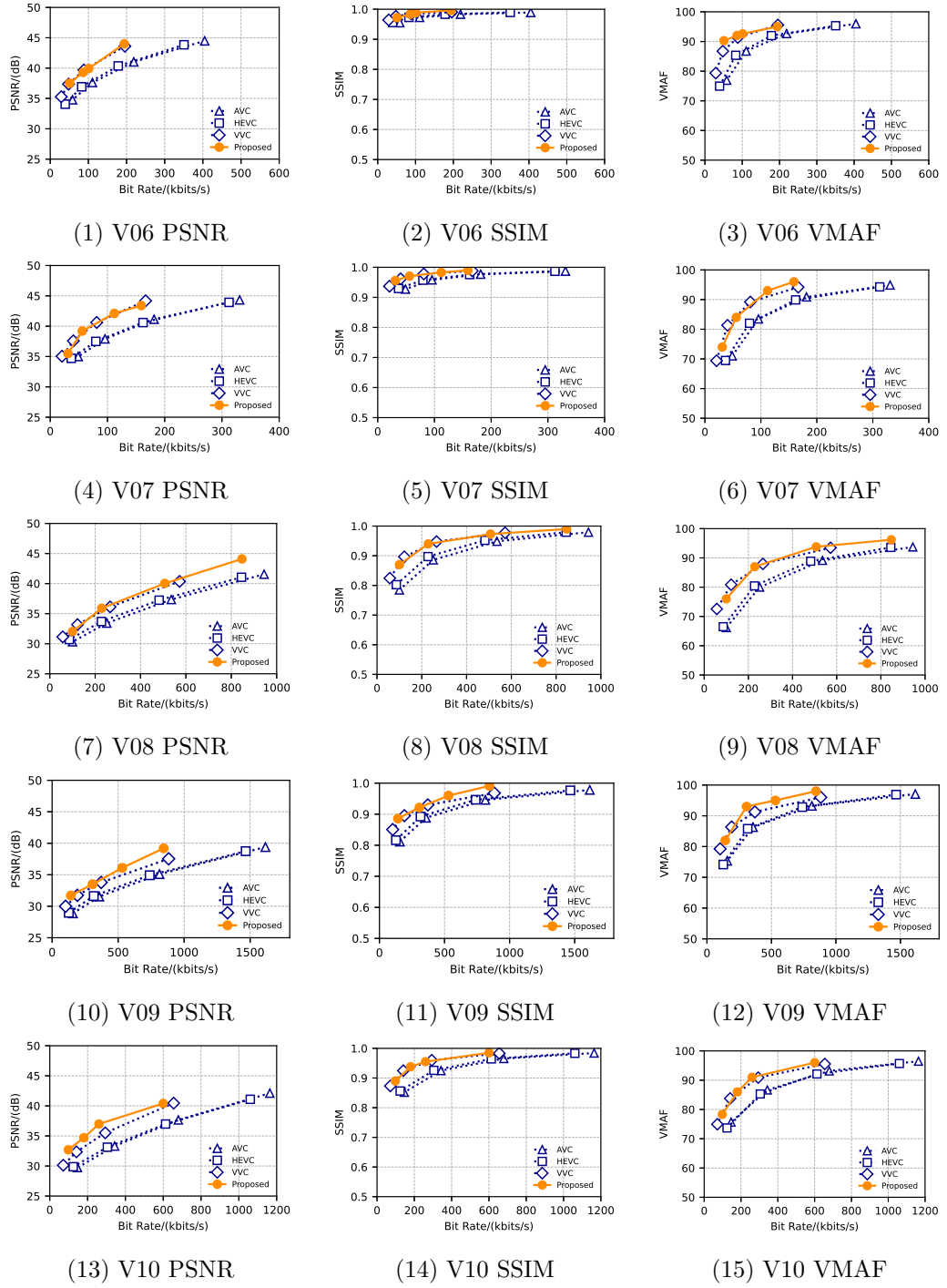


Figure A.6: Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V06–V10, GOP size 16.

Appendix A. Individual Rate–Distortion Performance for Test Videos for GOP Sizes 8, 16, and 32

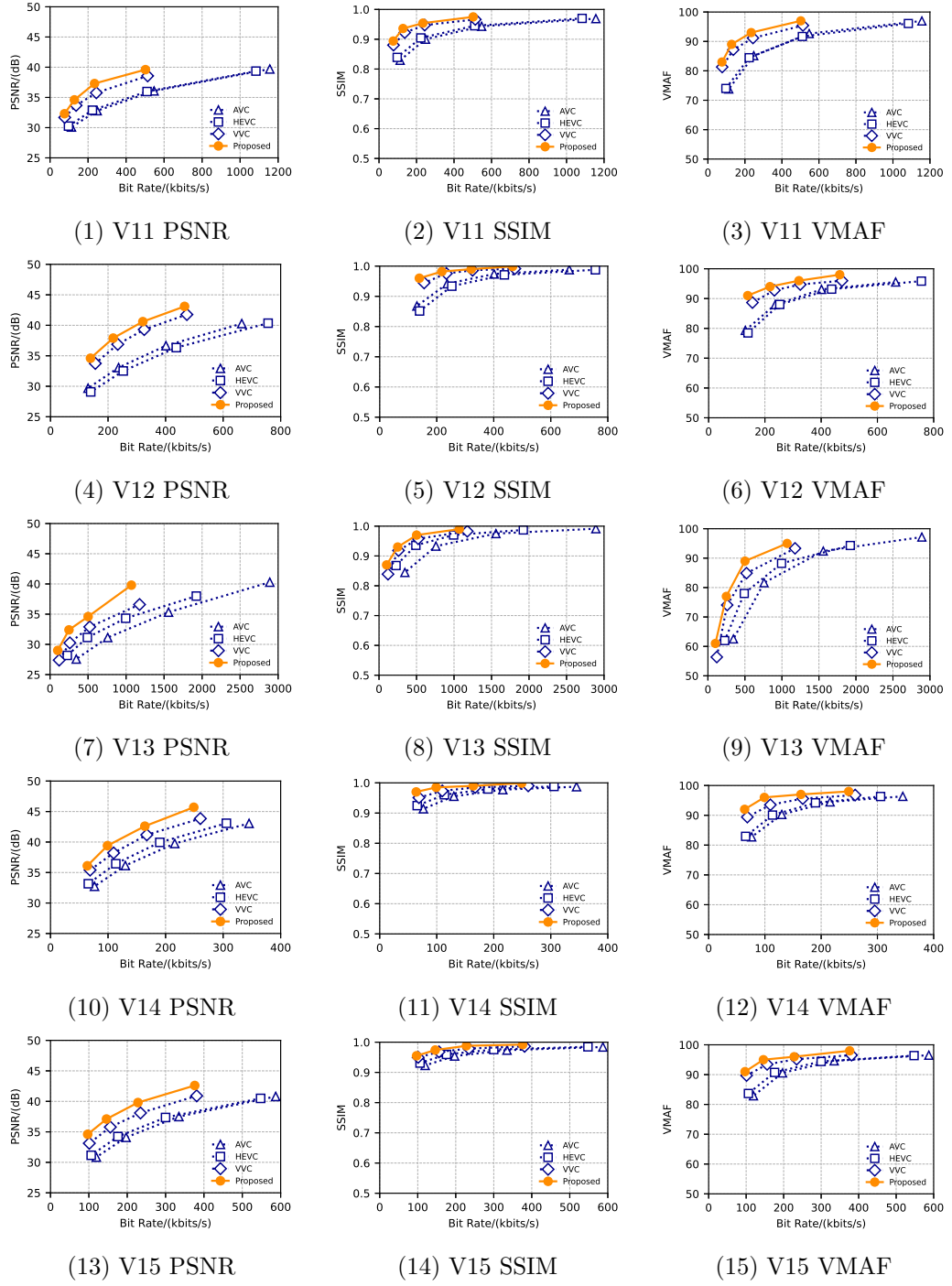


Figure A.7: Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V11–V15, GOP size 16.

Appendix A. Individual Rate–Distortion Performance for Test Videos for GOP Sizes 8, 16, and 32

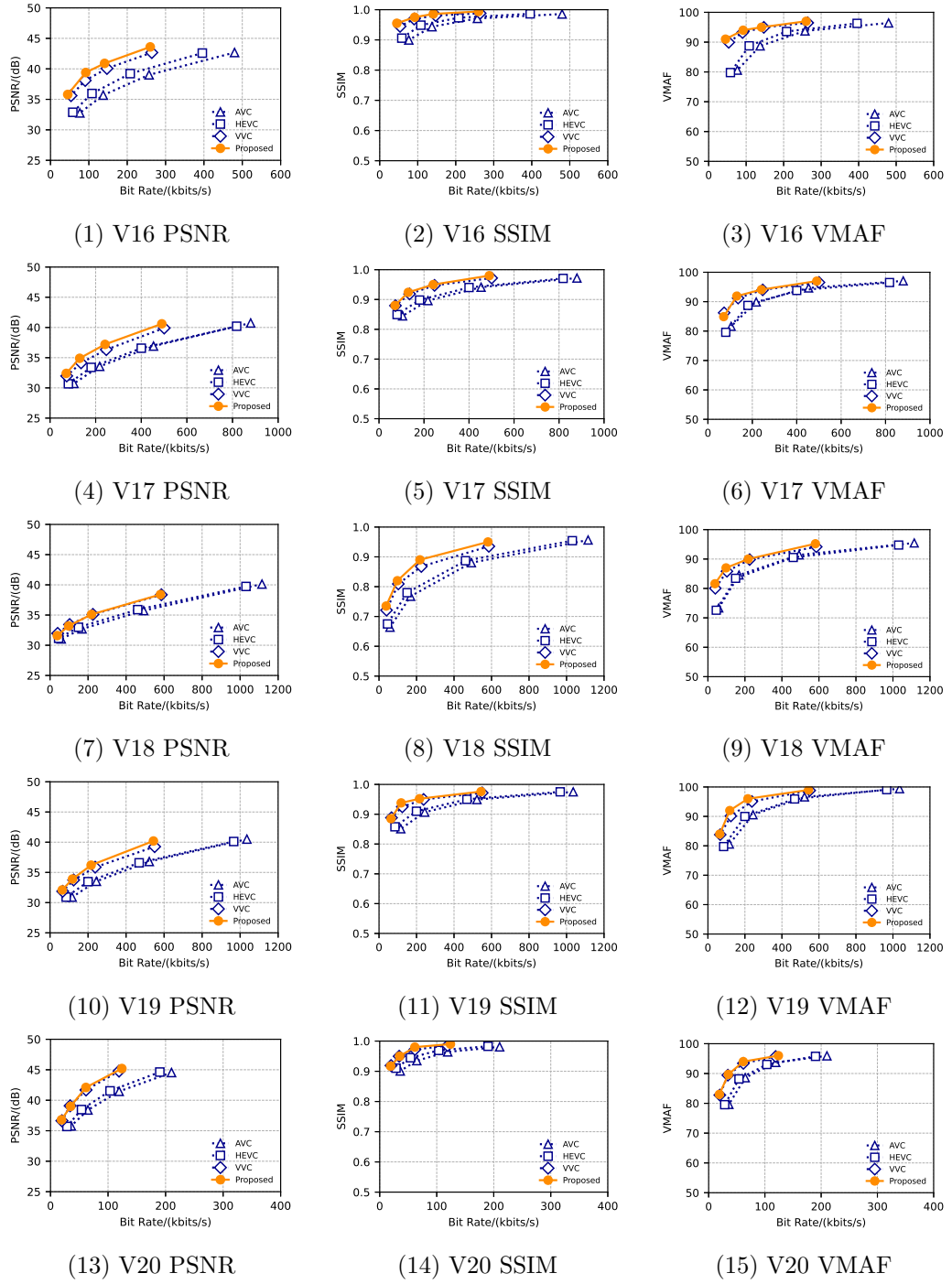


Figure A.8: Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V16–V20, GOP size 16.

Appendix A. Individual Rate–Distortion Performance for Test Videos for GOP Sizes 8, 16, and 32

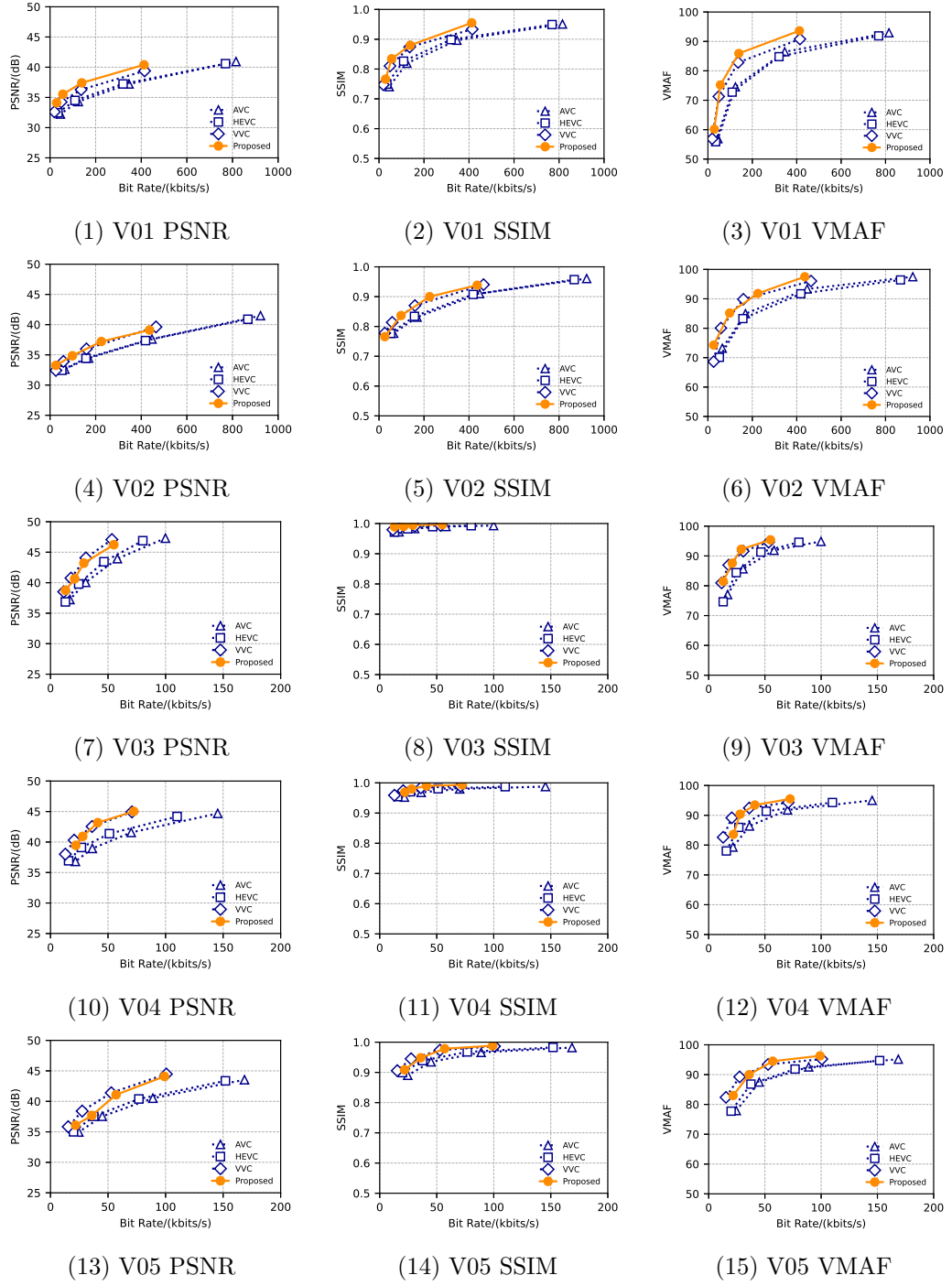


Figure A.9: Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V01–V05, GOP size 32.

Appendix A. Individual Rate–Distortion Performance for Test Videos for GOP Sizes 8, 16, and 32

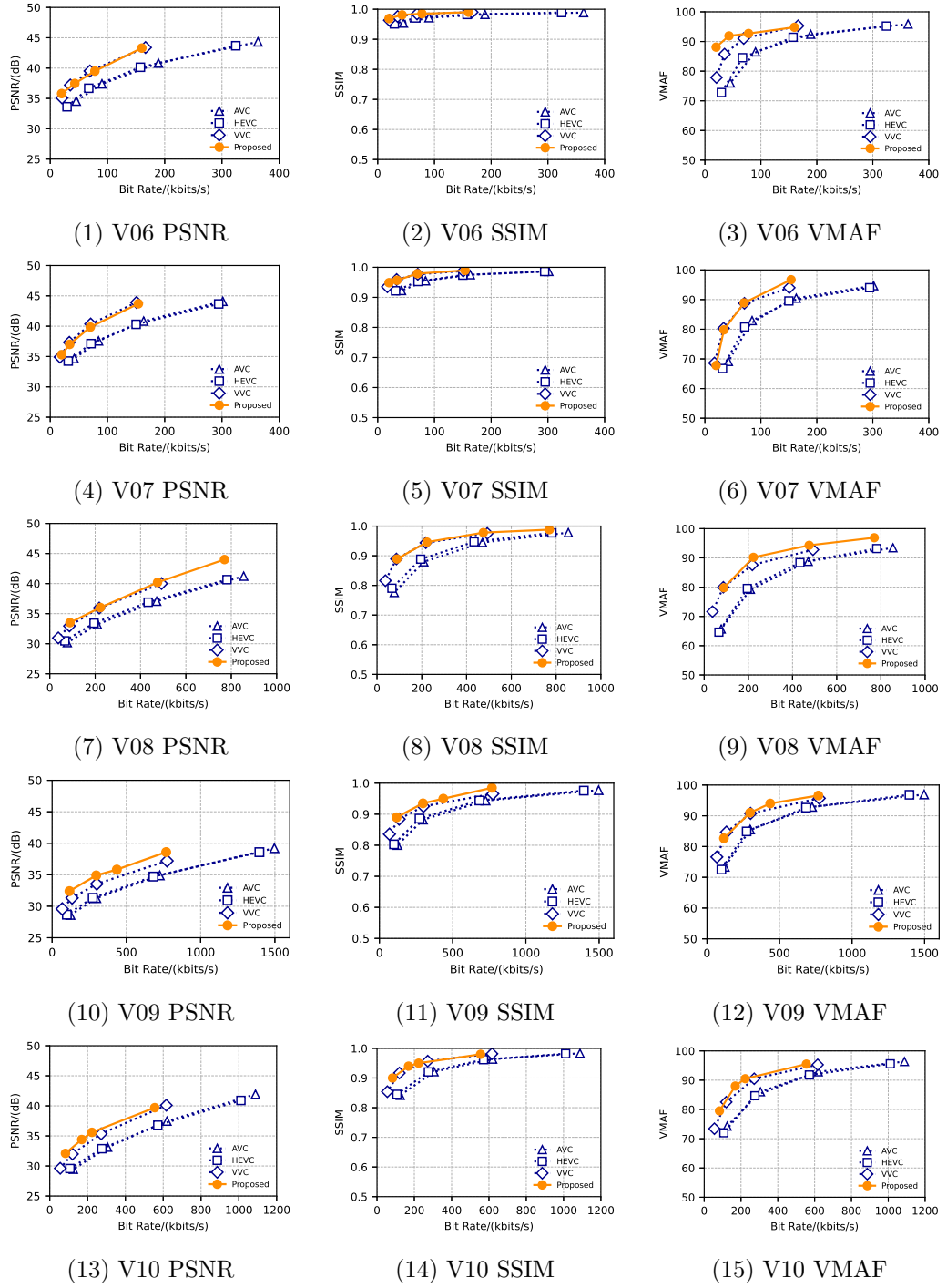


Figure A.10: Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V06–V10, GOP size 32.

Appendix A. Individual Rate–Distortion Performance for Test Videos for GOP Sizes 8, 16, and 32

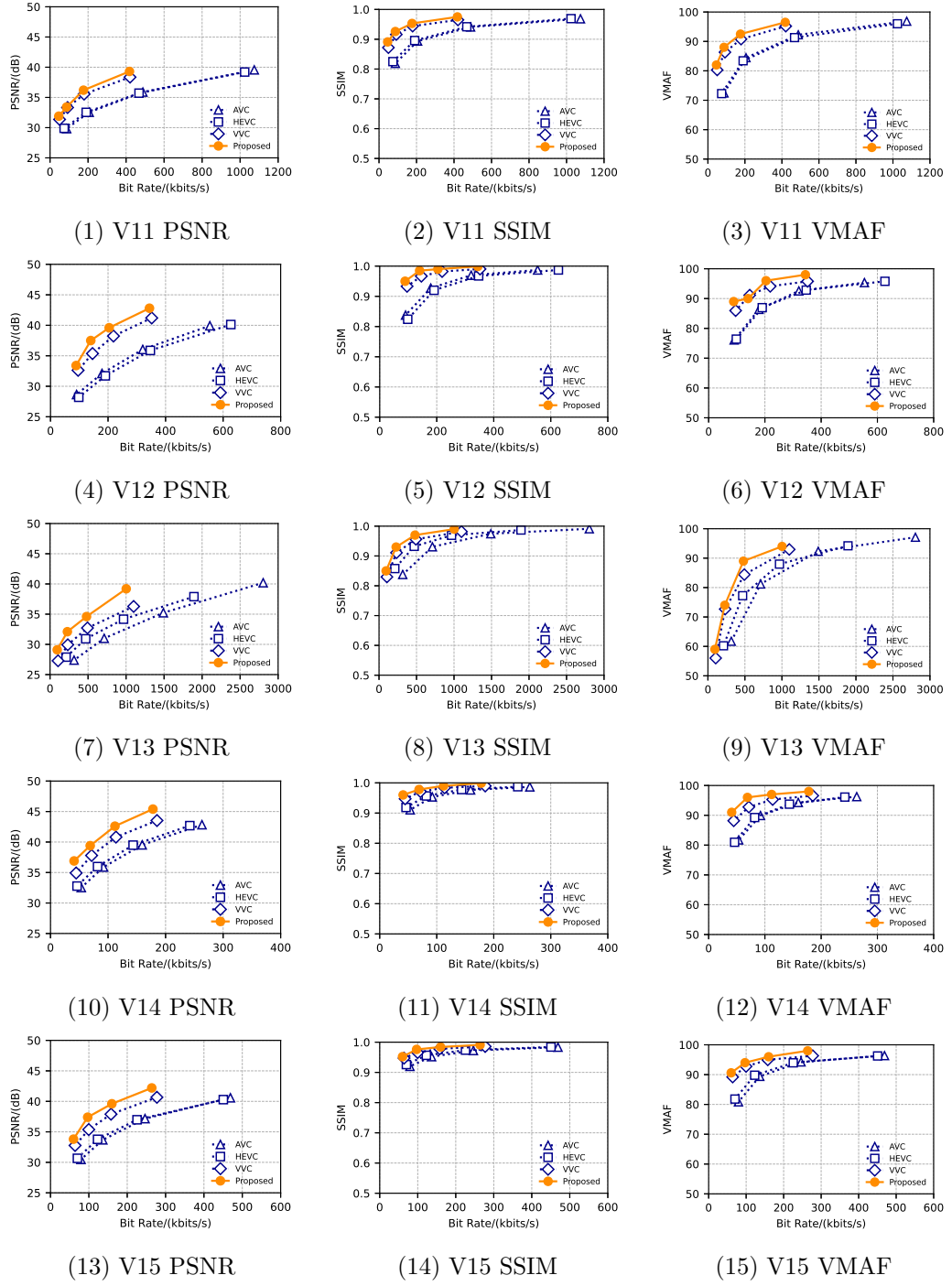


Figure A.11: Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V11–V15, GOP size 32.

Appendix A. Individual Rate–Distortion Performance for Test Videos for GOP Sizes 8, 16, and 32

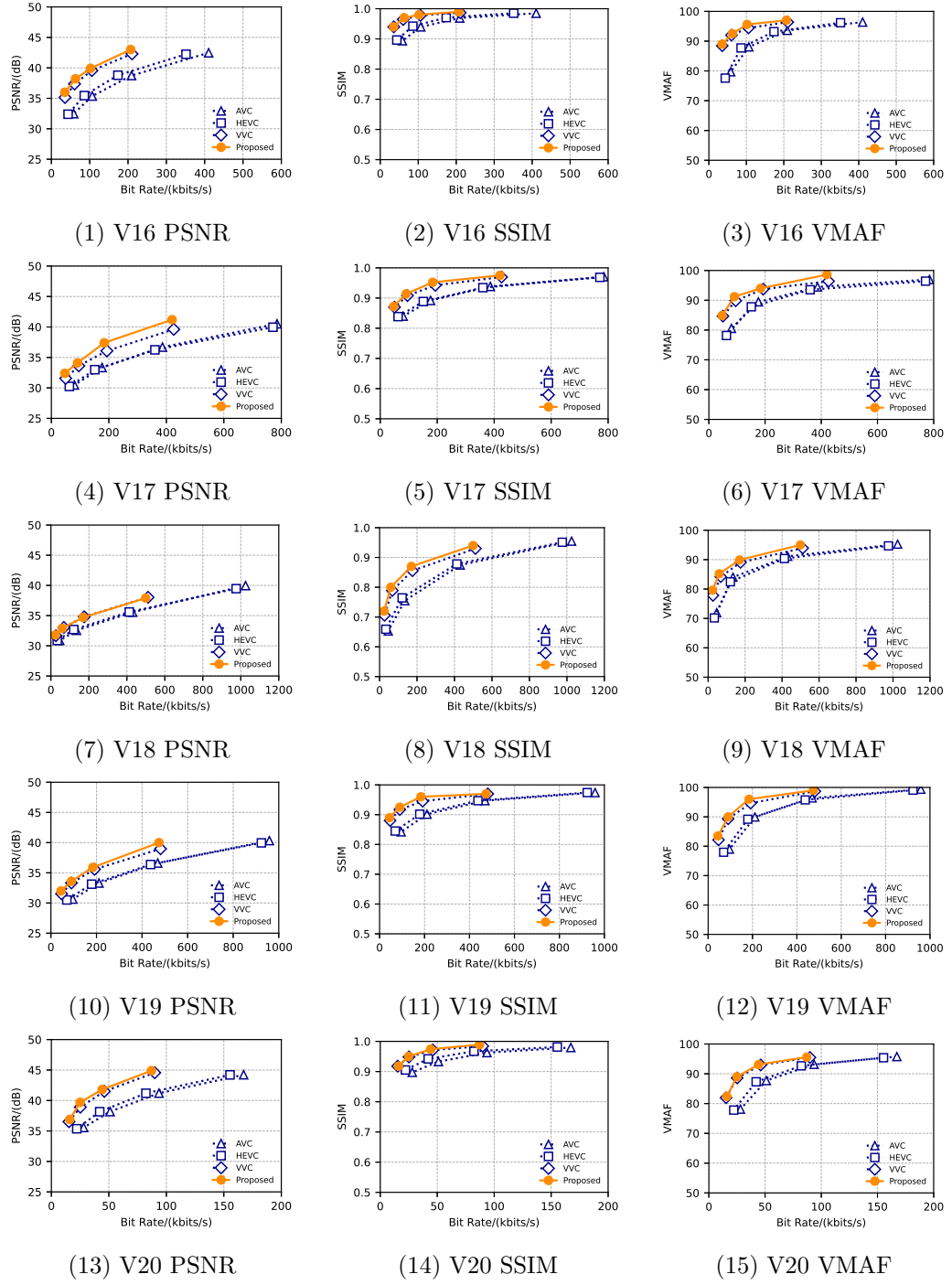


Figure A.12: Rate distortion performance (PSNR, SSIM and VMAF) for test video clips V16–V20, GOP size 32.

Bibliography

- [1] P. Samarathunga, Y. Ganearachchi, T. Fernando, A. Jayasingam, I. Alahapperuma, and A. Fernando, "A semantic communication and vvc based hybrid video coding system," *IEEE Access*, vol. 12, pp. 79 202–79 224, 2024.
- [2] P. Midtrack. People enjoying the day in a beach. [www.pexels.com](https://www.pexels.com/video/people-enjoying-the-day-in-a-beach-3150419/). Accessed: Feb. 12, 2024. [Online]. Available: <https://www.pexels.com/video/people-enjoying-the-day-in-a-beach-3150419/>
- [3] T. Miroshnichenko. People playing soccer. [www.pexels.com](https://www.pexels.com/video/people-playing-soccer-6077718/). Accessed: Feb. 12, 2024. [Online]. Available: <https://www.pexels.com/video/people-playing-soccer-6077718/>
- [4] C. of Couple. A bowl of avocados and vegetables. [www.pexels.com](https://www.pexels.com/video/a-bowl-of-avocados-and-vegetables-7656166/). Accessed: Feb. 12, 2024. [Online]. Available: <https://www.pexels.com/video/a-bowl-of-avocados-and-vegetables-7656166/>
- [5] Pixabay. Blue sky video. [www.pexels.com](https://www.pexels.com/video/blue-sky-video-855005/). Accessed: Feb. 12, 2024. [Online]. Available: <https://www.pexels.com/video/blue-sky-video-855005/>
- [6] T. Elliot. A couple walking towards the launching area of the hot air balloons festival. [www.pexels.com](https://www.pexels.com/video/a-couple-walking-towards-the-launching-area-of-the-hot-air-balloons-festival-3064025/). Accessed: Feb. 12, 2024. [Online]. Available: <https://www.pexels.com/video/a-couple-walking-towards-the-launching-area-of-the-hot-air-balloons-festival-3064025/>
- [7] D. C. Paduret. Culturing a chamomile flower plant. [www.pexels.com](https://www.pexels.com/video/culturing-a-chamomile-flower-plant-3011973/). Accessed: Feb. 12, 2024. [Online]. Available: <https://www.pexels.com/video/culturing-a-chamomile-flower-plant-3011973/>

Bibliography

- [8] S. Garenko. A girl running across a bridge. [www.pexels.com](https://www.pexels.com/video/a-girl-running-across-a-bridge-19805236/). Accessed: Feb. 12, 2024. [Online]. Available: <https://www.pexels.com/video/a-girl-running-across-a-bridge-19805236/>
- [9] V. Singh. Rangdum village in zanskar valley. [www.pexels.com](https://www.pexels.com/video/rangdum-village-in-zanskar-valley-19022224/). Accessed: Feb. 12, 2024. [Online]. Available: <https://www.pexels.com/video/rangdum-village-in-zanskar-valley-19022224/>
- [10] M. Kilinc. Nemrut – bitlis. [www.pexels.com](https://www.pexels.com/video/nemrut-bitlis-18856748/). Accessed: Feb. 12, 2024. [Online]. Available: <https://www.pexels.com/video/nemrut-bitlis-18856748/>
- [11] M. T. Kirkgoz. Cold snow sea dawn. [www.pexels.com](https://www.pexels.com/video/cold-snow-sea-dawn-18051870/). Accessed: Feb. 12, 2024. [Online]. Available: <https://www.pexels.com/video/cold-snow-sea-dawn-18051870/>
- [12] Old town cross. [xiph.org](https://media.xiph.org/video/derf/). Accessed: Apr. 02, 2024. [Online]. Available: <https://media.xiph.org/video/derf/>
- [13] Controlled burn. [xiph.org](https://media.xiph.org/video/derf/). Accessed: Apr. 02, 2024. [Online]. Available: <https://media.xiph.org/video/derf/>
- [14] Ducks take off. [xiph.org](https://media.xiph.org/video/derf/). Accessed: Apr. 02, 2024. [Online]. Available: <https://media.xiph.org/video/derf/>
- [15] Four people. [xiph.org](https://media.xiph.org/video/derf/). Accessed: Apr. 02, 2024. [Online]. Available: <https://media.xiph.org/video/derf/>
- [16] Galleon. [xiph.org](https://media.xiph.org/video/derf/). Accessed: Apr. 02, 2024. [Online]. Available: <https://media.xiph.org/video/derf/>
- [17] L. Photography. A wooden bridge built above water. [www.pexels.com](https://www.pexels.com/video/wood-sea-landscape-nature-3971351/). Accessed: Apr. 02, 2024. [Online]. Available: <https://www.pexels.com/video/wood-sea-landscape-nature-3971351/>
- [18] C. Studio. A person writing on the blueprint. [www.pexels.com](https://www.pexels.com/video/a-person-writing-on-the-blueprint-7484777/). Accessed: Apr. 02, 2024. [Online]. Available: <https://www.pexels.com/video/a-person-writing-on-the-blueprint-7484777/>

Bibliography

- [19] P. Whelen. Man walking on road among pine trees. [www.pexels.com](https://www.pexels.com/video/man-walking-on-road-among-pine-trees-5738706/). Accessed: Apr. 02, 2024. [Online]. Available: <https://www.pexels.com/video/man-walking-on-road-among-pine-trees-5738706/>
- [20] ——. A person doing graffiti art by using a spray paint. [www.pexels.com](https://www.pexels.com/video/a-person-doing-graffiti-art-by-using-a-spray-paint-5621707/). Accessed: Apr. 02, 2024. [Online]. Available: <https://www.pexels.com/video/a-person-doing-graffiti-art-by-using-a-spray-paint-5621707/>
- [21] A. Podrez. A young man executing a kick flip skateboard trick. [www.pexels.com](https://www.pexels.com/video/a-young-man-executing-a-kick-flip-skateboard-trick-4832083/). Accessed: Apr. 02, 2024. [Online]. Available: <https://www.pexels.com/video/a-young-man-executing-a-kick-flip-skateboard-trick-4832083/>
- [22] P. Samarathunga, Y. Ganearachchi, T. Fernando, A. Jayasingam, T. Sivalingam, N. Rajatheva, and A. Fernando, “Enhanced semantic communications and vvc-based hybrid video coding system,” *IEEE Access*, vol. 14, pp. 50 755–50 778, 2026.
- [23] Z. Tian, F. Wang, S. Wang, Z. Zhou, Y. Zhu, and L. Shen, “High Dynamic Range Video Compression: A Large-Scale Benchmark Dataset and A Learned Bit-depth Scalable Compression Algorithm ,” in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2025, pp. 7320–7330. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.00686>
- [24] P. Samarathunga, I. Alahapperuma, T. Fernando, Y. Ganearachchi, and A. Fernando, “Deep neural network compression for semantic video communications,” in *2025 IEEE International Conference on Consumer Electronics (ICCE)*, 2025, pp. 1–4.
- [25] P. Samarathunga, Y. Ganearachchi, T. Fernando, I. Alahapperuma, and A. Fernando, “Semantic communication based video coding using temporal prediction of deep neural network parameters,” in *2024 IEEE Gaming, Entertainment, and Media Conference (GEM)*, 2024, pp. 1–6.

Bibliography

- [26] Z. Jia, B. Li, J. Li, W. Xie, L. Qi, H. Li, and Y. Lu, "Towards practical real-time neural video compression," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 12 543–12 552.
- [27] R. K. Mantiuk and M. Azimi, "Pu21: A novel perceptually uniform encoding for adapting existing quality metrics for hdr," in *2021 Picture Coding Symposium (PCS)*, 2021, pp. 1–5.
- [28] Ericsson, "Video traffic update," Stockholm, 2023. [Online]. Available: <https://www.ericsson.com/en/reports-and-papers/mobility-report/dataforecasts/traffic-by-application>
- [29] Cisco Systems, Inc., "Cisco annual internet report (2018–2023) white paper," Cisco Systems, Inc., White Paper c11-741490, Mar. 2020. [Online]. Available: <https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html>
- [30] S. Dang, O. Amin, B. Shihada, and M.-S. Alouini, "What should 6G be?" *Nature Electronics*, vol. 3, no. 1, pp. 20–29, 2020.
- [31] X. You, C.-X. Wang, J. Huang, X. Gao, Z. Zhang, M. Wang, Y. Huang, C. Zhang, Y. Jiang, J. Wang, M. Zhu, B. Sheng, D. Wang, Z. Pan, P. Liu, Z. Luo, J. Shi, L. Jin, and L. Hanzo, "Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts," *Science China Information Sciences*, vol. 64, no. 11, p. 110301, 2021.
- [32] ITU-T, "Video codec for audiovisual services at p x 64 kbit/s," International Telecommunication Union, Recommendation H.261, 1993. [Online]. Available: <https://www.itu.int/rec/T-REC-H.261/e>
- [33] ISO/IEC JTC1/SC29/WG11, "Coding of moving pictures and associated audio for digital storage media at up to about 1,5 mbit/s," International Organization for Standardization, Standard ISO/IEC 11172-2 (MPEG-1 Video), 1993. [Online]. Available: <https://www.iso.org/standard/22411.html>

Bibliography

- [34] K. Rijkse, “H.263: video coding for low-bit-rate communication,” *IEEE Communications Magazine*, vol. 34, no. 12, pp. 42–45, 1996.
- [35] ISO/IEC JTC1/SC29/WG11, “Generic coding of moving pictures and associated audio information,” International Organization for Standardization, Standard ISO/IEC 13818-2 (MPEG-2 Video), 1995.
- [36] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, “Overview of the H.264/AVC video coding standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.
- [37] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, “Overview of the high efficiency video coding (HEVC) standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.
- [38] B. Bross, J. Chen, J.-R. Ohm, G. J. Sullivan, and Y.-K. Wang, “Developments in international video coding standardization after AVC, with an overview of versatile video coding (VVC),” *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1463–1493, 2021.
- [39] J. Han, B. Li, D. Mukherjee, C.-H. Chiang, A. Grange, C. Chen, H. Su, S. Parker, S. Deng, U. Joshi, Y. Chen, Y. Wang, P. Wilkins, Y. Xu, and J. Bankoski, “A technical overview of AV1,” *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1435–1462, 2021.
- [40] F. Pakdaman, M. A. Adelimanesh, M. Gabbouj, and M. R. Hashemi, “Complexity analysis of next-generation vvc encoding and decoding,” in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 3134–3138.
- [41] N. Ling, C.-C. J. Kuo, G. J. Sullivan, D. Xu, S. Liu, H.-M. Hang, W.-H. Peng, and J. Liu, “The future of video coding,” *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, pp. 1–29, 02 2022. [Online]. Available: <https://doi.org/10.1561/116.000000044>

Bibliography

- [42] F. Bossen, B. Bross, K. Sühning, and D. Flynn, “HEVC complexity and implementation analysis,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1685–1696, 2012.
- [43] J. Ballé, P. A. Chou, D. Minnen, S. Singh, N. Johnston, E. Agustsson, S. J. Hwang, and G. Toderici, “Nonlinear transform coding,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 2, pp. 339–353, 2021.
- [44] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, and S. Wang, “Image and video compression with neural networks: A review,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1683–1698, 2020.
- [45] G. Toderici, S. M. O’Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, and R. Sukthankar, “Variable rate image compression with recurrent neural networks,” 2016. [Online]. Available: <https://arxiv.org/abs/1511.06085>
- [46] J. Ballé, V. Laparra, and E. P. Simoncelli, “End-to-end optimized image compression,” *CoRR*, vol. abs/1611.01704, 2016. [Online]. Available: <http://arxiv.org/abs/1611.01704>
- [47] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, “Variational image compression with a scale hyperprior,” 2018. [Online]. Available: <https://arxiv.org/abs/1802.01436>
- [48] G. K. Wallace, “The JPEG still picture compression standard,” *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.
- [49] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, “DVC: An end-to-end deep video compression framework,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 006–11 015.
- [50] E. Agustsson, D. Minnen, N. Johnston, J. Ballé, S. J. Hwang, and G. Toderici, “Scale-space flow for end-to-end optimized video compression,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8500–8509.

Bibliography

- [51] J. Li, B. Li, and Y. Lu, “Deep contextual video compression,” 2021. [Online]. Available: <https://arxiv.org/abs/2109.15047>
- [52] —, “Neural video compression with feature modulation,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 26 099–26 108.
- [53] —, “Neural video compression with diverse contexts,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 22 616–22 626.
- [54] R. Yang, F. Mentzer, L. Van Gool, and R. Timofte, “Learning for video compression with recurrent auto-encoder and recurrent probability model,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 2, pp. 388–401, 2021.
- [55] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, “Deep learning enabled semantic communication systems,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, 2021.
- [56] C. E. Shannon, “A mathematical theory of communication,” *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [57] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press, 1964.
- [58] M. U. Lokumarambage, V. S. S. Gowrisetty, H. Rezaei, T. Sivalingam, N. Rajatheva, and A. Fernando, “Wireless end-to-end image transmission system using semantic communications,” *IEEE Access*, vol. 11, pp. 37 149–37 163, 2023.
- [59] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Y. Li, “Semantic communications: Principles and challenges,” 2022. [Online]. Available: <https://arxiv.org/abs/2201.01389>
- [60] X. Luo, H.-H. Chen, and Q. Guo, “Semantic communications: Overview, open issues, and future research directions,” *IEEE Wireless Communications*, vol. 29, no. 1, pp. 210–219, 2022.

Bibliography

- [61] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen, and C. Miao, “Semantic communications for future internet: Fundamentals, applications, and challenges,” *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 213–250, 2023.
- [62] E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, “Deep joint source-channel coding for wireless image transmission,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.
- [63] D. B. Kurka and D. Gündüz, “Deep joint source-channel coding of images with feedback,” in *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 5235–5239.
- [64] H. Zhang, S. Shao, M. Tao, X. Bi, and K. B. Letaief, “Deep learning-enabled semantic communication systems with task-unaware transmitter and dynamic data,” *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 170–185, 2023.
- [65] ITU-R, “Requirements related to technical performance for IMT-advanced radio interface(s),” International Telecommunication Union, Report ITU-R M.2134, 2008. [Online]. Available: <https://www.itu.int/pub/R-REP-M.2134>
- [66] —, “IMT vision – framework and overall objectives of the future development of IMT for 2020 and beyond,” International Telecommunication Union, Recommendation ITU-R M.2083-0, 2015. [Online]. Available: <https://www.itu.int/rec/R-REC-M.2083/en>
- [67] —, “Image parameter values for high dynamic range television for use in production and international programme exchange,” International Telecommunication Union, Recommendation BT.2100, 2018.
- [68] —, “Parameter values for ultra-high definition television systems for production and international programme exchange,” International Telecommunication Union, Recommendation BT.2020, 2015. [Online]. Available: <https://www.itu.int/rec/r-rec-bt.2020>

Bibliography

- [69] E. François, C. Fogg, Y. He, X. Li, A. Luthra, and A. Segall, “High dynamic range and wide color gamut video coding in hevc: Status and potential future enhancements,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 63–75, 2016.
- [70] V. K. Goyal, “Multiple description coding: Compression meets the network,” *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 74–93, 2001.
- [71] A. Gamal and T. Cover, “Achievable rates for multiple descriptions,” *IEEE Transactions on Information Theory*, vol. 28, no. 6, pp. 851–857, 1982.
- [72] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou, “Real-time intermediate flow estimation for video frame interpolation,” in *Computer Vision – ECCV 2022*. Springer, 2022, pp. 624–642.
- [73] P. Samarathunga, Y. Ganearachchi, T. Fernando, A. Jayasingam, T. Sivalingam, N. Rajatheva, and A. Fernando, “Context-aware semantic video coding using pretrained motion-adaptive models and bidirectional prediction,” *IEEE Access*, vol. 14, pp. 69 730–69 755, 2026.
- [74] A. Skodras, C. Christopoulos, and T. Ebrahimi, “The jpeg 2000 still image compression standard,” *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 36–58, 2001.
- [75] J. Lainema, M. M. Hannuksela, V. K. M. Vadakital, and E. B. Aksu, “HEVC still image coding and high efficiency image file format,” in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 71–75.
- [76] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, “Learned image compression with discretized Gaussian mixture likelihoods and attention modules,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7939–7948.

Bibliography

- [77] ITU-T, “H.120: Codecs for Videoconferencing Using Primary Digital Group Transmission,” International Telecommunication Union, Tech. Rep., 1984, revised 1988, 1993. [Online]. Available: <https://www.itu.int/rec/T-REC-H.120/en>
- [78] N. Barman, M. G. Martini, and Y. Reznik, “Revisiting Bjontegaard Delta Btrate (BD-BR) Computation for Codec Compression Efficiency Comparison,” in *MHV 2022 - Proceedings of the 1st Mile-High Video Conference*. Association for Computing Machinery, Inc, 3 2022, pp. 113–114.
- [79] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, “Comparison of the coding efficiency of video coding standards—including high efficiency video coding (hevc),” *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 22, no. 12, p. 1669–1684, Dec. 2012. [Online]. Available: <https://doi.org/10.1109/TCSVT.2012.2221192>
- [80] S. Bouaafia, R. Khemiri, and F. E. Sayadi, “Rate-distortion performance comparison: Vvc vs. hevc,” in *2021 18th International Multi-Conference on Systems, Signals & Devices (SSD)*, 2021, pp. 440–444.
- [81] Y.-W. Huang, J. An, H. Huang, X. Li, S.-T. Hsiang, K. Zhang, H. Gao, J. Ma, and O. Chubach, “Block partitioning structure in the vvc standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3818–3833, 2021.
- [82] J. Pfaff, A. Filippov, S. Liu, X. Zhao, J. Chen, S. De-Luxán-Hernández, T. Wiegand, V. Rufitskiy, A. K. Ramasubramonian, and G. Van der Auwera, “Intra prediction and mode coding in VVC,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3834–3847, 2021.
- [83] X. Zhao, S.-H. Kim, Y. Zhao, H. E. Egilmez, M. Koo, S. Liu, J. Lainema, and M. Karczewicz, “Transform coding in the VVC standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3878–3890, 2021.
- [84] A. Wieckowski, J. Brandenburg, T. Nguyen, C. Bartnik, V. George, G. Hege, C. Helmrich, A. Henkel, C. Lehmann, C. Stoffers, I. Zupancic, B. Bross,

Bibliography

- and D. Marpe, “VVenC: An open and optimized VVC encoder implementation,” in *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2021, pp. 1–2.
- [85] A. Norkin, “AV2 video codec architecture,” in *Proceedings of the 17th International Conference on Quality of Multimedia Experience (QoMEX)*. Madrid, Spain: IEEE, 2025. [Online]. Available: https://norkin.org/pdf/QoMEX_2025_AV2_Architecture_slides.pdf
- [86] H. Gao, S. Esenlik, E. Alshina, and E. Steinbach, “Geometric partitioning mode in versatile video coding: Algorithm review and analysis,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 9, pp. 3603–3617, 2021.
- [87] M. Karczewicz, N. Hu, J. Taquet, C.-Y. Chen, K. Misra, K. Andersson, P. Yin, T. Lu, E. François, and J. Chen, “VVC in-loop filters,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3907–3925, 2021.
- [88] S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P. A. Chou, R. A. Cohen, M. Krivokuća, S. Lasserre, Z. Li, J. Llach, K. Mammou, R. Mekuria, O. Nakagami, E. Siahaan, A. Tabatabai, A. M. Tourapis, and V. Zakharchenko, “Emerging mpeg standards for point cloud compression,” *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 133–148, 2019.
- [89] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley-Interscience, 2006. [Online]. Available: <https://onlinelibrary.wiley.com/doi/book/10.1002/047174882X>
- [90] G. W. Cottrell, P. Munro, and D. Zipser, “Image compression by back propagation: An example of extensional programming,” in *Proceedings of the 9th Annual Conference of the Cognitive Science Society*, 1987, pp. 461–473.
- [91] D. Minnen, J. Ballé, and G. D. Toderici, “Joint autoregressive and hierarchical priors for learned image compression,” in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman,

Bibliography

- N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/53edebc543333dfbf7c5933af792c9c4-Paper.pdf
- [92] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. Van Gool, “Generative adversarial networks for extreme learned image compression,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 221–231.
- [93] F. Mentzer, G. Toderici, M. Tschannen, and E. Agustsson, “High-fidelity generative image compression,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS ’20. Red Hook, NY, USA: Curran Associates Inc., 2020.
- [94] M. Lu, P. Guo, H. Shi, C. Cao, and Z. Ma, “Transformer-based image compression,” in *2022 Data Compression Conference (DCC)*, 2022, pp. 469–469.
- [95] R. Zou, C. Song, and Z. Zhang, “The devil is in the details: Window-based attention for image compression,” 2022. [Online]. Available: <https://arxiv.org/abs/2203.08450>
- [96] Z. Hu, D. Xu, G. Lu, W. Jiang, W. Wang, and S. Liu, “Fvc: An end-to-end framework towards deep video compression in feature space,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4569–4585, 2023.
- [97] J. Lin, D. Liu, H. Li, and F. Wu, “M-LVC: Multiple frames prediction for learned video compression,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3543–3551.
- [98] X. Sheng, J. Li, B. Li, L. Li, D. Liu, and Y. Lu, “Temporal context mining for learned video compression,” *IEEE Transactions on Multimedia*, vol. 25, pp. 7311–7322, 2023.

Bibliography

- [99] J. Li, B. Li, and Y. Lu, “Hybrid spatial-temporal entropy modelling for neural video compression,” in *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 2022, pp. 1503–1511.
- [100] A. Wyner and J. Ziv, “The rate-distortion function for source coding with side information at the decoder,” *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 1–10, 1976.
- [101] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006. [Online]. Available: <https://www.science.org/doi/abs/10.1126/science.1127647>
- [102] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
- [103] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- [104] G. Shi, Y. Xiao, Y. Li, and X. Xie, “From semantic communication to semantic-aware networking: Model, architecture, and open problems,” *IEEE Communications Magazine*, vol. 59, no. 8, pp. 44–50, 2021.
- [105] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, “Deep learning based semantic communications: An initial investigation,” in *2020 IEEE Global Communications Conference (GLOBECOM)*, 2020, pp. 1–6.
- [106] A. Li, X. Wei, D. Wu, and L. Zhou, “Cross-modal semantic communications,” *IEEE Wireless Communications*, vol. 29, no. 6, pp. 144–151, 2022.

Bibliography

- [107] Y. Zhang, H. Zhao, J. Wei, J. Zhang, M. F. Flanagan, and J. Xiong, “Context-based semantic communication via dynamic programming,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 3, pp. 1453–1467, 2022.
- [108] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS’13. Red Hook, NY, USA: Curran Associates Inc., 2013, p. 3111–3119.
- [109] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: <https://aclanthology.org/D14-1162/>
- [110] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: <https://aclanthology.org/N18-1202/>
- [111] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2019. [Online]. Available: <https://arxiv.org/abs/1810.04805>
- [112] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” OpenAI, Tech. Rep., 2018.

Bibliography

- [113] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, “Task-oriented multi-user semantic communications,” *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2584–2597, 2022.
- [114] Q. Zhou, R. Li, Z. Zhao, C. Peng, and H. Zhang, “Semantic communication with adaptive universal transformer,” *IEEE Wireless Communications Letters*, vol. 11, no. 3, pp. 453–457, 2022.
- [115] Z. Weng and Z. Qin, “Semantic communication systems for speech transmission,” *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2434–2444, 2021.
- [116] T. Han, Q. Yang, Z. Shi, S. He, and Z. Zhang, “Semantic-preserved communication system for highly efficient speech transmission,” *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 245–259, 2023.
- [117] J. Xu, B. Ai, W. Chen, A. Yang, P. Sun, and M. Rodrigues, “Wireless image transmission using deep source channel coding with attention modules,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2315–2328, 2022.
- [118] S. Wang, J. Dai, Z. Liang, K. Niu, Z. Si, C. Dong, X. Qin, and P. Zhang, “Wireless deep video semantic transmission,” *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 214–229, 2023.
- [119] J. Kang, H. Du, Z. Li, Z. Xiong, S. Ma, D. Niyato, and Y. Li, “Personalized saliency in task-oriented semantic communications: Image transmission and performance analysis,” *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 186–201, 2023.
- [120] P. Samarathunga, V. Gowrisetty, T. Fernando, Y. Ganearachchi, and A. Fernando, “Region of interest scalable image compression using semantic communications,” in *2024 IEEE International Conference on Consumer Electronics (ICCE)*, 2024, pp. 1–5.

Bibliography

- [121] M. Jankowski, D. Gündüz, and K. Mikolajczyk, “Wireless image retrieval at the edge,” *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 89–100, 2021.
- [122] P. Samarathunga, H. Rezaei, M. Lokumarambage, T. Sivalingam, N. Rajatheva, and A. Fernando, “An autoencoder-based task-oriented semantic communication system for m2m communication,” *Algorithms*, vol. 17, no. 11, 2024. [Online]. Available: <https://www.mdpi.com/1999-4893/17/11/492>
- [123] A. Li, X. Liu, G. Wang, and P. Zhang, “Domain knowledge driven semantic communication for image transmission over wireless channels,” *IEEE Wireless Communications Letters*, vol. 12, no. 1, pp. 55–59, 2023.
- [124] E. Erdemir, T.-Y. Tung, P. L. Dragotti, and D. Gündüz, “Generative joint source-channel coding for semantic image transmission,” *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 7, pp. 2030–2045, 2023.
- [125] P. Samarathunga, N. Pollwaththage, T. Fernando, U. Jayasinghe, N. Ihalagamage, Y. Ganearachchi, and A. Fernando, “Semantic and technical noise modeling in semantic image communication over wireless channels,” *IEEE Access*, vol. 13, pp. 150 872–150 900, 2025.
- [126] P. Samarathunga, T. Fernando, V. Gowrisetty, T. Atulugama, and A. Fernando, “Autoencoder based image quality metric for modelling semantic noise in semantic communications,” *Electronics Letters*, vol. 60, no. 4, p. e13115, 2024. [Online]. Available: <https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/ell2.13115>
- [127] P. Samarathunga, Y. Ganearachchi, T. Fernando, L. Thanippulige, and A. Fernando, “Semantic communication based complexity scalable image transmission system for resource constrained devices,” in *2024 IEEE Gaming, Entertainment, and Media Conference (GEM)*, 2024, pp. 1–6.
- [128] P. Samarathunga, T. Fernando, Y. Ganearachchi, N. Pollwaththage, and A. Fernando, “Semantic communication and residual deep learning-driven bitplane cod-

Bibliography

- ing for scalable image compression and communications,” in *2025 IEEE International Conference on Image Processing Workshops (ICIPW)*, 2025, pp. 568–573.
- [129] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, “Wireless semantic communications for video conferencing,” *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 230–244, 2023.
- [130] P. Samarathunga, U. Jayasinghe, Y. Ganearachchi, T. Fernando, N. Senanayake, and A. Fernando, “Semantic communication based lossless distributed video coding for error prone channels,” in *2025 IEEE International Conference on Consumer Electronics (ICCE)*, 2025, pp. 1–4.
- [131] L. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, “Video coding for machines: A paradigm of collaborative compression and intelligent analytics,” *IEEE Transactions on Image Processing*, vol. 29, pp. 8680–8695, 2020.
- [132] MPEG, “MPEG-AI: Video coding for machines (Part 2),” <https://www.mpeg.org/standards/MPEG-AI/2/>, 2025, standard under development.
- [133] J. R. Y. S. H. D. W. Y. C. W. X. W. Y. Z. Z. Z. F. W. S. Cui, “Generative semantic communication: Architectures, technologies, and applications,” *Engineering*, vol. 56, pp. 45–61, 2026. [Online]. Available: <https://www.engineering.org.cn/engi/EN/10.1016/j.eng.2025.07.022>
- [134] M. Chen, Z. Sun, X. He, L. Wang, and A. Al-Dulaimi, “Llm-based semantic communication: The way from task-originated to general,” *IEEE Wireless Communications Letters*, vol. 14, no. 10, pp. 3029–3033, 2025.
- [135] E. Grassucci, S. Barbarossa, and D. Communiello, “Generative semantic communication: Diffusion models beyond bit recovery,” *IEEE Transactions on Cognitive Communications and Networking*, pp. 1–1, 2026.
- [136] H. Tong, Z. Yang, S. Wang, Y. Hu, W. Saad, and C. Yin, “Federated learning based audio semantic communication over wireless networks,” in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 1–6.

Bibliography

- [137] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, “Beyond transmitting bits: Context, semantics, and task-oriented communications,” *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 5–41, 2023.
- [138] P. A. Stavrou and M. Kountouris, “A rate distortion approach to goal-oriented communication,” in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 590–595.
- [139] P. Samarathunga, Y. Ganearachchi, T. Fernando, and A. Fernando, “Error resilient rate-scalable image compression using semantic communications,” in *ICC 2026 - IEEE International Conference on Communications*, 2026, pp. 1–6.
- [140] O. Rippel, A. G. Anderson, K. Tatwawadi, S. Nair, C. Lytle, and L. Bourdev, “ELF-VC: Efficient Learned Flexible-Rate Video Coding,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2021, pp. 14 459–14 468. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01421>
- [141] P. Samarathunga, Y. Ganearachchi, and A. Fernando, “Video compression by chroma prediction using semantic communications,” in *2024 IEEE International Conference on Consumer Electronics (ICCE)*, 2024, pp. 1–4.
- [142] H. Choi and I. V. Bajić, “Deep feature compression for collaborative object detection,” in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 3743–3747.
- [143] A. E. Eshratifar, M. S. Abrishami, and M. Pedram, “Jointdnn: An efficient training and inference engine for intelligent mobile cloud computing services,” *IEEE Transactions on Mobile Computing*, vol. 20, no. 2, pp. 565–576, 2021.
- [144] Z. Teed and J. Deng, “RAFT: Recurrent all-pairs field transforms for optical flow,” in *Computer Vision – ECCV 2020*. Springer, 2020, pp. 402–419.

Bibliography

- [145] Y. Zhang, T. van Rozendaal, J. Brehmer, M. Nagel, and T. Cohen, “Implicit neural video compression,” in *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022. [Online]. Available: <https://openreview.net/forum?id=r4geC2VdP-5>
- [146] Y. Strümpler, J. Postels, R. Yang, L. V. Gool, and F. Tombari, “Implicit neural representations for image compression,” in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 74–91. [Online]. Available: https://doi.org/10.1007/978-3-031-19809-0_5
- [147] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, “Deep convolutional autoencoder-based lossy image compression,” in *2018 Picture Coding Symposium (PCS)*, 2018, pp. 253–257.
- [148] S. Wiedemann, H. Kirchhoffer, S. Matlage, P. Haase, A. Marban, T. Marinč, D. Neumann, T. Nguyen, H. Schwarz, T. Wiegand, D. Marpe, and W. Samek, “Deepcabac: A universal compression algorithm for deep neural networks,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 4, pp. 700–714, 2020.
- [149] S. Miller, M. Nezamabadi, and S. Daly, “Perceptual signal coding for more efficient usage of bit codes,” *SMPTE Motion Imaging Journal*, vol. 122, pp. 52–59, 05 2013.
- [150] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, “Photographic tone reproduction for digital images,” in *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2002, pp. 267–276.
- [151] L. Ozarow, “On a source-coding problem with two channels and three receivers,” *The Bell System Technical Journal*, vol. 59, no. 10, pp. 1909–1921, 1980.
- [152] V. A. Vaishampayan, “Design of multiple description scalar quantizers,” *IEEE Trans. Inf. Theor.*, vol. 39, no. 3, p. 821–834, May 1993. [Online]. Available: <https://doi.org/10.1109/18.256491>

Bibliography

- [153] V. K. Goyal, J. Kovačević, and J. A. Kelner, “Quantized frame expansions with erasures,” *Applied and Computational Harmonic Analysis*, vol. 10, no. 3, pp. 203–233, 2001. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1063520300903403>
- [154] S. Servetto, V. Vaishampayan, and N. Sloane, “Multiple description lattice vector quantization,” in *Proceedings DCC’99 Data Compression Conference (Cat. No. PR00096)*, 1999, pp. 13–22.
- [155] Y. Wang, A. Reibman, and S. Lin, “Multiple description coding for video delivery,” *Proceedings of the IEEE*, vol. 93, no. 1, pp. 57–70, 2005.
- [156] L. Zhao, J. Zhang, H. Bai, A. Wang, and Y. Zhao, “Lmdc: Learning a multiple description codec for deep learning-based image compression,” *Multimedia Tools Appl.*, vol. 81, no. 10, p. 13889–13910, Apr. 2022. [Online]. Available: <https://doi.org/10.1007/s11042-022-12216-4>
- [157] T. H. Le, X. Pic, and M. Antonini, “Inr-mdsqc: Implicit neural representation multiple description scalar quantization for robust image coding,” in *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*, 2023, pp. 1–6.
- [158] J. G. Proakis and M. Salehi, *Digital communications*. McGraw-Hill, 2008.
- [159] D. Becking, P. Haase, H. Kirchhoffer, K. Müller, and W. Samek, “NNCodec: An open source software implementation of the neural network coding ISO/IEC standard,” in *ICML 2023 Workshop Neural Compression: From Information Theory to Applications*, 2023. [Online]. Available: <https://openreview.net/forum?id=5VgMDKUgX0>
- [160] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding,” in *International Conference on Learning Representations (ICLR)*, 2016. [Online]. Available: <https://arxiv.org/abs/1510.00149>

Bibliography

- [161] P. Samarathunga, I. Alahapperuma, Y. Ganearachchi, T. Fernando, and A. Fernando, “Neural network compression for video transmission using semantic communications and vvc,” in *ICC 2026 - IEEE International Conference on Communications*, 2026, pp. 1–6.
- [162] G. Bjontegaard, “Calculation of average PSNR differences between RD-curves,” 2001, vCEG-M33, ITU-T SG16 Q6, Doc. VCEG-M33. Accessed: Feb. 12, 2024. [Online]. Available: https://www.itu.int/wftp3/av-arch/video-site/0104_Aus/VCEG-M33.doc
- [163] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.
- [164] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [165] Netflix, “VMAF: Video multi-method assessment fusion,” <https://github.com/Netflix/vmaf>, 2018.
- [166] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterton, Eds., vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 249–256. [Online]. Available: <https://proceedings.mlr.press/v9/glorot10a.html>
- [167] M. Yang, C. Bian, and H.-S. Kim, “Deep joint source channel coding for wireless image transmission with ofdm,” in *ICC 2021 - IEEE International Conference on Communications*, 2021, pp. 1–6.
- [168] W. A. Fernando, C. N. Canagarajah, and D. R. Bull, “A unified approach to scene change detection in uncompressed and compressed video,” *IEEE Trans.*

Bibliography

- on Consum. Electron.*, vol. 46, no. 3, p. 769–779, Aug. 2000. [Online]. Available: <https://doi.org/10.1109/30.883445>
- [169] N. Barman, N. Khan, and M. G. Martini, “Analysis of spatial and temporal information variation for 10-bit and 8-bit video sequences,” in *Proc. IEEE 24th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 2019, pp. 1–6.
- [170] S. Tomar, “Converting video formats with ffmpeg,” *Linux J.*, vol. 2006, no. 146, p. 10, Jun. 2006.
- [171] Q. Huynh-Thu and M. Ghanbari, “The accuracy of psnr in predicting video quality for different video scenes and frame rates,” *Telecommun. Syst.*, vol. 49, no. 1, p. 35–48, Jan. 2012. [Online]. Available: <https://doi.org/10.1007/s11235-010-9351-x>
- [172] Joint Collaborative Team on Video Coding (JCT-VC), “HEVC Test Sequences,” <https://hevc.hhi.fraunhofer.de/>, 2013, available from the JCT-VC or ITU-T upon request.
- [173] A. Mercat, M. Viitanen, and J. Vanne, “UVG dataset: 50/120fps 4K sequences for video codec analysis and development,” in *Proceedings of the 11th ACM Multimedia Systems Conference*. Istanbul, Turkey: ACM, May 2020, pp. 297–302.
- [174] H. Wang, W. Gan, S. Hu, J. Y. Lin, L. Jin, L. Song, P. Wang, I. Katsavounidis, A. Aaron, and C.-C. J. Kuo, “MCL-JCV: A JND-based H.264/AVC video quality assessment dataset,” in *2016 IEEE International Conference on Image Processing (ICIP)*. Phoenix, AZ, USA: IEEE, Sep. 2016, pp. 1509–1513.
- [175] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, “Video enhancement with task-oriented flow,” *Int. J. Comput. Vision*, vol. 127, no. 8, p. 1106–1125, Aug. 2019. [Online]. Available: <https://doi.org/10.1007/s11263-018-01144-2>
- [176] IEEE, “Ieee standard for floating-point arithmetic,” *IEEE Std 754-2019 (Revision of IEEE 754-2008)*, pp. 1–84, 2019.

Bibliography

- [177] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, “Overview of the versatile video coding (VVC) standard and its applications,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736–3764, 2021.
- [178] TensorFlow. (2025) EarlyStopping Callback — TensorFlow 2.x Documentation. Accessed: Oct. 7, 2025. [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/callbacks/EarlyStopping
- [179] [International Telecommunication Union], “Methodologies for the subjective assessment of the quality of television images,” ITU-R, Geneva, Switzerland, Recommendation BT.500-14, 2019.
- [180] Netflix, “Netflix Open Content,” <https://opencontent.netflix.com/>, 2025, accessed: 2025-05-08. [Online]. Available: <https://opencontent.netflix.com/>
- [181] International Telecommunication Union, “Subjective video quality assessment methods for multimedia applications,” International Telecommunication Union, ITU-T Recommendation P.910, October 2023, accessed: 2025-05-19. [Online]. Available: <https://www.itu.int/rec/T-REC-P.910-202310-I/en>
- [182] Video Quality Experts Group, “siti-tools: SI/TI calculation tools,” <https://github.com/VQEG/siti-tools>, 2024, accessed: 2025-05-29.
- [183] Google, “Google Colaboratory,” <https://colab.google/>, 2025, accessed: 2025-05-05. [Online]. Available: <https://colab.google/>
- [184] Joint Video Experts Team, “VVCSoftware_VTM,” 2025, accessed: 2025-05-05. [Online]. Available: https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM
- [185] J. Boyce, K. Suehring, X. Li, and V. Seregin, “Jvet-j1010: Jvet common test conditions and software reference configurations,” 07 2018.
- [186] Joint Video Experts Team, “HM,” 2025, accessed: 2025-05-05. [Online]. Available: <https://vcgit.hhi.fraunhofer.de/jvet/HM>

Bibliography

- [187] —, “ECM,” 2025, accessed: 2025-05-05. [Online]. Available: <https://vcgit.hhi.fraunhofer.de/ecm/ECM>
- [188] Alliance for Open Media, “AVM,” 2025, accessed: 2025-05-05. [Online]. Available: <https://aomedia.googleusercontent.com/aom>
- [189] C. Herglotz, H. Och, A. Meyer, G. Ramasubbu, L. Eichermüller, M. Kränzler, F. Brand, K. Fischer, D. T. Nguyen, A. Regensky, and A. Kaup, “The bjøntegaard bible why your way of comparing video codecs may be wrong,” *Trans. Img. Proc.*, vol. 33, p. 987–1001, Jan. 2024. [Online]. Available: <https://doi.org/10.1109/TIP.2023.3346695>
- [190] F. Bossen, K. Sühning, A. Wieckowski, and S. Liu, “VVC complexity and software implementation analysis,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3765–3778, 2021.
- [191] S. Ioffe and C. Szegedy, “Batch normalization: accelerating deep network training by reducing internal covariate shift,” in *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ser. ICML’15. JMLR.org, 2015, p. 448–456.
- [192] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [193] H. Kirchhoffer, P. Haase, W. Samek, K. Müller, H. Rezazadegan-Tavakoli, F. Cricri, E. B. Aksu, M. M. Hannuksela, W. Jiang, W. Wang, S. Liu, S. Jain, S. Hamidi-Rad, F. Racapé, and W. Bailer, “Overview of the neural network compression and representation (NNR) standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 3203–3216, 2022.
- [194] J. Hable, “Uncharted 2: HDR lighting,” Game Developers Conference, 2010. [Online]. Available: <https://www.gdcvault.com/play/1012459/Uncharted-2-HDR>

Bibliography

- [195] ———, “Filmic tonemapping with piecewise power curves,” 2017. [Online]. Available: <https://filmicworlds.com/blog/filmic-tonemapping-with-piecewise-power-curves/>
- [196] C. Cao, H. Yue, X. Liu, and J. Yang, “Unsupervised HDR image and video tone mapping via contrastive learning,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 2, pp. 786–798, 2024.
- [197] H. Yan, H. Zhang, M. Zhao, and G. Cao, “Video inverse tone mapping network with global colour mapping and multi-frame interaction,” *IEEE Transactions on Consumer Electronics*, vol. 71, no. 2, pp. 2930–2943, 2025.
- [198] F. Durand and J. Dorsey, “Fast bilateral filtering for the display of high-dynamic-range images,” *ACM Trans. Graph.*, vol. 21, no. 3, p. 257–266, Jul. 2002. [Online]. Available: <https://doi.org/10.1145/566654.566574>
- [199] ITU-R, “Parameter values for the HDTV standards for production and international programme exchange,” International Telecommunication Union, Recommendation BT.709, 2015. [Online]. Available: <https://www.itu.int/rec/r-rec-bt.709>
- [200] Eastman Kodak Company, “How to use a gray card,” 2026, accessed: 2026-05-23. [Online]. Available: <https://www.kodak.com/en/motion/page/gray-card/>
- [201] ITU-R, “Colour gamut conversion from recommendation ITU-R BT.2020 to recommendation ITU-R BT.709,” International Telecommunication Union, Report BT.2407, 2017. [Online]. Available: <https://www.itu.int/pub/R-REP-BT.2407>
- [202] L. Zhao, H. Bai, A. Wang, and Y. Zhao, “Multiple description convolutional neural networks for image compression,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2494–2508, 2019.
- [203] H. Li, L. Meng, J. Zhang, Y. Tan, Y. Ren, and H. Zhang, “Multiple description coding based on convolutional auto-encoder,” *IEEE Access*, vol. 7, pp. 26 013–26 021, 2019.

Bibliography

- [204] P. Samarathunga, U. Jayasinghe, Y. Ganearachchi, T. Fernando, T. Sivalingam, N. Rajatheva, and A. Fernando, “Semantic communication based multiple description coding for robust image transmission over wireless channels,” *IEEE Access*, pp. 1–1, 2026.
- [205] E. Arikan, “Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels,” *IEEE Transactions on Information Theory*, vol. 55, no. 7, pp. 3051–3073, 2009.
- [206] C. Tomasi and R. Manduchi, “Bilateral filtering for gray and color images,” in *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, 1998, pp. 839–846.
- [207] E. Agustsson and R. Timofte, “Ntire 2017 challenge on single image super-resolution: Dataset and study,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1122–1131.
- [208] MathWorks, *Communications Toolbox Channel System Objects: AWGNChannel, RayleighChannel, and RicianChannel*, The MathWorks, Inc., Natick, MA, 2024, [Online]. Available: <https://www.mathworks.com/help/comm/>. [Online]. Available: <https://www.mathworks.com/help/comm/>
- [209] C. E. Shannon, “Communication in the presence of noise,” *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949.
- [210] R. Keys, “Cubic convolution interpolation for digital image processing,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1153–1160, 1981.

Bibliography