From Contextual Understanding to Cognitive Adaptation: Multimodal AI for Intelligent and Adaptive Multimedia Systems.

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

Doctor of Philosophy

Orthalange Waruna Wangeesa De Silva

Supervisor: Professor Anil Fernando

Co-Supervisor: Professor Feng Dong

AI Research Group
Computer & Information Sciences
University of Strathclyde, Glasgow

February 2026

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.

Abstract

The rapid evolution of streaming and on-demand media has transformed video platforms into adaptive, data-driven ecosystems in which content presentation, interruption timing, and user experience must be continuously optimised. Despite advances in multimodal machine learning, contemporary media decision systems remain largely perception-driven, operating on surface-level cues without structured reasoning over semantic meaning, narrative organisation, and viewer cognitive state. Addressing this challenge requires an integrated modular multimodal framework capable of jointly modelling content understanding, narrative structure, and human-centred suitability.

This thesis develops a modular, integrated multimodal reasoning architecture for adaptive media systems, enabling structured semantic grounding, narrative coherence modelling, and viewer-centred cognitive assessment, and applies this framework to context-aware transition and advertisement placement decisions. The first contribution establishes a multimodal topic–taxonomy mapping framework combining visual and transcript-derived features within a BERTopic pipeline for contextual advertising. The second contribution extends this architecture with semantically richer embeddings, replacing Bag-of-Visual-Words features with CLIP-ViT visual representations and incorporating image captioning and LLM-generated topic explanations to improve interpretability, multilingual robustness, and taxonomy alignment.

Building on this foundation, the third contribution proposes a theme-aware segmentation framework that integrates deterministic perception modules with constrained, zero-temperature large language model reasoning. A controlled single-pass Tree-of-Thoughts mechanism induces global narrative themes and identifies boundaries corresponding to sustained thematic transitions. Unlike stochastic or heavily fine-tuned alter-

natives, the method remains training-free, reproducible, and computationally lightweight while demonstrating improved boundary stability and higher agreement with human annotations across diverse video genres.

The fourth contribution presents a multimodal cognitive-load estimation model that maps aligned audio–visual features onto NASA-TLX-inspired dimensions to generate a time-varying cognitive-suitability curve. Evaluation against human-annotated transition points shows that the model reliably identifies low-load narrative pauses and outperforms silence-based and rhythm-based heuristics, providing a complementary viewer-centred segmentation signal.

The fifth contribution synthesises these semantic, narrative, and cognitive components into an integrated modular segmentation and ad-break decision pipeline, operationalised through the AdBreakScore(t) framework. This integrated system jointly reasons about semantic coherence, thematic continuity, and cognitive readiness to produce context-aware boundary decisions under interpretability and deployment constraints.

Together, this thesis advances adaptive multimedia systems from heuristic content analysis toward structured multimodal reasoning, establishing a human-centred framework applicable to context-aware advertising, narrative-aware editing assistance, and intelligent media adaptation.

Contents

List of Figures

List of Figures

List of Figures

List of Figures

List of Tables

Acknowledgements

I would like to express my sincere gratitude to my principal supervisor, Prof. Anil Fernando, for his continuous guidance and support throughout my Ph.D. His valuable feedback and direction have played an important role in shaping this research and strengthening its overall quality.

I would like to thank my co-supervisor, Prof. Feng Dong, for his support, guidance, and for peer-reviewing my research articles before publication during the course of my PhD. I am also thankful to Dr. Andrew Abel for his support and for providing valuable insights into the research artifacts developed during this journey by conducting detailed peer reviews.

I would like to extend my thanks to all colleagues and friends who supported me through this research journey. Finally, I am thankful to my family, friends, and well-wishers for their support, motivation, and encouragement during my Ph.D. period.

Chapter 1

Introduction

1.1 Background and Motivation

Digital videos have become the dominant form in the online media space [1]. Billions of hours of digital videos are streamed every day [1]. These videos include content from platforms such as YouTube, Netflix, and others. Such videos have resulted in the urgent need to develop techniques to manage such content in terms of segmentation, indexing, as well as monetizing the videos [2]. Among the techniques required to manage videos is the placement of mid-roll ads. Such ads have to be strategically placed in such a manner that the flow of the videos is not compromised [3].

While this is an extremely valuable task, currently used automatic approaches to this task rely on superficial heuristics such as scene breakpoint detection, silence detection, or temporal windows, and do not exploit the underlying meaningful structure of this content [2].

1. **Semantic structure:** videos convey topics, subtopics, and taxonomically recognisable categories that reflect what the content is about;
2. **Narrative structure:** stories evolve through themes, transitions, and higher-level conceptual arcs that shape how the content progresses;
3. **Cognitive structure:** viewers experience fluctuating attention, engagement, and cognitive load, all of which influence when interruptions feel natural or jarring.

Human editors naturally operate in a way that takes these semantic, narrative, and cognitive aspects into account as they choose places to add transitions or places to cut for commercial breaks [4]. By contrast, if they are even considered in automated systems, they would be addressed individually in these models. The resulting boundaries often align with surface-level cues instead of deeper story logic or viewer readiness for change [5].

Recent advances in multimodal machine learning offer opportunities to address these limitations [6]. Vision–language models and transformer-based transcript analysis now assist in producing richer semantic representations than traditional metadata or unimodal embeddings. In parallel, recent evidence indicates that while multimodal large language models exhibit strong static visual grounding, they frequently fail to maintain consistent entity-level continuity across temporal transitions, exposing fundamental limitations in their narrative reasoning capabilities [7]. Meanwhile, research in cognitive load theory demonstrates that audio–visual signals can approximate a viewer's moment-to- moment mental effort, providing a principled basis for identifying segments where interruptions cause minimal disruption [8].

Together, these developments motivate a shift from purely perceptual segmentation toward *integrated multimodal reasoning*. This thesis argues that modern video segmentation and ad-break identification require the coordinated modelling of *semantic*, *narrative*, and *cognitive* layers within a modular, integrated framework. The aim is not simply to detect *when* content changes, but to understand *what* changes (semantic), *why* it changes (narrative), and *whether the viewer is cognitively ready* for a change (cognitive). These three layers operate at different timescales and require different modelling strategies; integrating them thus represents both a conceptual and a technical challenge.

However, to overcome this challenge, this thesis also considers the operational constraints that are normally not considered in research. There are constraints of determinism, re-producibility, latency, and safety that are imposed by the commercial media infrastructure, and these are difficult for the traditional black-box models or the learned models of segmentation to handle. Therefore, the modular framework used in this thesis has both theoretical justification as well as operational justification.

While advertisement placement serves as the primary applied use case, the resulting framework generalises to video editing assistance, narrative-aware content summarisation, recommendation timing, educational media design, and other human–AI multimedia interactions.

Scope and Assumptions

The proposed framework particularly focuses on pre-recorded medium-form as well as long-form video contents where transcripts and detected visual shots can be produced automatically. Scenarios involving live streaming, large-scale domain-based fine-tuning that require much deeper cultural knowledge will be beyond the scope of this thesis. These design choices ensure the realizability, repeatability, and generalisability of the proposed methods. While focusing on short to medium form content, evaluation includes diverse real-world datasets to test robustness.

1.2 Problem Statement

On the basis of the systematic review and comparative analysis of contemporary video segmentation and advertisement placement systems that have been presented in Chapter 2, it has been found that current automated systems have a number of basic structural limitations that make it difficult for them to find transition points that have a coherent narrative and that are appropriate for viewers. These include:

- **Dependence on shallow segmentation cues:** As has been evident in the existing breakpoint detection systems that have been discussed in section 2.4.1, boundaries have traditionally been detected in a manner that is related to perceptual boundaries like shot boundaries, silence detection, fades, or short range embedding similarity. While boundaries that are computationally efficient have been detected, they have not been strongly related to story boundaries.
- **Weak semantic grounding:** As described in Section 2.5.1, current models are unable to create a structured representation of meaning with well-defined semantics using existing content categorization frameworks. Without a structured rep-

resentation of meaning, models are unable to distinguish between real topic shifts and mere variations.

- **Limited narrative understanding:** The comparative analysis shown in section 2.4.4 reveals that most of the automated segmentation algorithms follow local similarity heuristics and have no provision for modelling global thematic continuity. Thus, the boundaries drawn by them fail to encode information about the global narrative progression or thematic closure, or story-level coherence.
- **Absence of cognitive modelling:** Based on the discussion in Section 2.5.2, the cognitive readiness of the viewers, the fluctuations in their engagement, and the cognitive load are not modelled in the usual segmentation pipelines. This has the effect of interrupting the viewers when the cognitive load is high.
- **Fragmented modelling paradigms:** A wider synthesis provided in Section 2.5 reveals that semantic, narrative, and cognitive models are normally developed separately and treated as non-overlapping elements. A non-integrative approach disregards inherent relationships between content meaning, narrative development, and audience experience, and therefore cannot perform reasoning about transition appropriateness.

Cumulatively, these issues cause the decisions on program segmentation and ad breaks to be often seen as arbitrary, intrusive, and not in sync with the semantic meaning and state of the viewer's mind. The problem here does not lie in the lack of individual solutions for semantic analysis, story modelling, and engagement assessment, but in the lack of an integrated strategy that can reason about these aspects together. This problem can be overcome by moving from fragmented, perception-centric heuristics to systematic multimodal approaches that can reason about semantic meaning, story, and the readiness of the viewer together. The next section will discuss the current state of the existing research work in addressing these issues, and the gaps that exist in the current state of the art.

1.3 Research Gap

Despite the existing body of research that tackles individual issues related to video segmentation and advertising insertion, some fundamental research gaps still remain. These research gaps are directly related to the inherent limitations pointed out in the Problem Statement (Section 1.2) but will be formulated below in the form of challenges that still exist in the research environment.

1. **Limited multimodal semantic integration:** Existing topic models and semantic understanding solutions fail to adequately incorporate visual, audio, and text data in a strong, interpretable semantic representation that is aligned to industry taxonomies [9]. This is because most of the state-of-the-art solutions to date are unimodal or weakly fused, as pointed out in various surveys of multimodal fusion techniques [10]. Additionally, the models are sensitive to noise in each of the individual modalities as well as in a multilingual setup [11]. This is supported by the findings of the literature review conducted in Section 2.4.2 of this thesis, pointing out that taxonomy-aligned, noise-robust multimodal topic models do not exist.

2. **Under explored theme-level narrative modelling:** Although existing segmentation approaches are increasingly leveraging semantic embeddings, their design is mostly focused on local similarities and short-term heuristics, with little capacity for modelling global story structure [12]. More recent chaptering and segmentation approaches based on large language models (LLM) have shown encouraging levels of reasoning, although these approaches are mostly dependent on stochastic decoding, heavy task-specific training, or dataset-specific fine-tuning [13, 14]. This renders these approaches nondeterministic, non-reproducible, and non-generalizable, and are also computationally intensive, rendering them impractical for implementation in a real-world media system environment. Chapter 5 proposes a framework that mitigates this limitation, where a training-free, light, and deterministic framework is employed to apply constrained LLM reasoning to develop global themes and identify fixed boundaries of the story.

3. **No direct implementation of cognitive load awareness in segmentation algorithms:** Although there is a theoretical foundation for modelling audio-visual information processing based on cognitive load theory [15], there is little evidence of direct implementation of cognitive load awareness into existing automatic video segmentation algorithms that are solely dependent on content-based cues for cognitive load estimation. This is because existing solutions are dependent on invasive physiological signals or engagement/attention-related side information with an implicit decision boundary [8]. Hence, there is little implementation of cognitive readiness awareness into transition detection algorithms. This issue is addressed in Chapter 6.

4. **Absence of integrated modular frameworks:** The current state-of-the-art is to address the issues of semantic understanding, story modelling, and cognitive analysis independently. This has led to a fragmented system which is unable to reason about the appropriateness of transitions. The need for a unifying paradigm is what has led to the development of the multimodal segmentation paradigm proposed throughout chapters 3 to 6 and synthesized in chapter 7.

Together, these gaps point to the need for a integrated research strategy, one which moves from perception-driven heuristics to structured multimodal reasoning. This thesis tackles the gaps by providing a series of models, each of which is semantic, narrative, or cognitive, meant to work together, as well as on their own, as part of a single, integrated segmentation process. While modular in architecture, the proposed layers are jointly reasoned within a integrated modular decision framework, thereby avoiding the fragmentation observed in prior work.

1.4 Research Aim

The overarching aim of this thesis is:

> **To develop a modular, multimodal segmentation framework that identifies optimal video boundaries by jointly modelling semantic**

content, narrative themes, and approximate cognitive load from content-derived audiovisual proxies.

The approach progresses from multimodal representation learning to narrative reasoning and finally to human-centred cognitive modelling.

1.5 Research Objectives

To achieve this aim, four primary objectives are defined:

1. **Semantic Objective**
   Build multimodal semantics with capabilities that fuse visual, textual, and audio features, such that topic-segment descriptors that align with the taxonomy can be derived.

2. **Narrative Objective**
   Model higher-order narrative structure via global theme induction using constrained LLM-based reasoning over aligned multimodal evidence.

3. **Cognitive Objective**
   Estimate cognitive load on the viewer from audio-visual clues, developing segment-level metrics that indicate low-load points with transition opportunities that are minimally intrusive.

4. **Integration Objective**
   Integrate the semantic, narrative, and cognitive layers into a integrated modular research-prototype segmentation and ad-break decision framework that operationalises their interactions while remaining interpretable and deployable at experimental scale.

These goals comprise a hierarchical stack with the foundations being provided by the semantic representation, the formation of global coherence being realised through narrative reasoning, viewer-centric suitability being achieved through cognitive modelling,

and its culmination being the integration of all three. Integration is ultimately operationalised through the AdBreakScore(t) formulation introduced in Chapter 2, which provides a structured scoring mechanism for jointly reasoning over semantic coherence, thematic continuity, and cognitive suitability.

Methodological Rationale

The modular, layered design is chosen for three reasons:

- *Interpretability and controllability:* industry deployment requires transparent and deterministic decision-making.
- *Architectural necessity:* LLMs cannot reliably infer cognitive states or execute segmentation without structured intermediate representations.
- *Technical complementarity:* semantic, narrative, and cognitive cues arise from distinct modalities and timescales, requiring specialised modelling.

1.6 Research Questions

The research is guided by the following questions:

1. How can multimodal topic models be designed to yield both robust embeddings and interpretable, taxonomy-aligned topics?

2. How can narrative themes be inferred from multimodal evidence using constrained LLM-based reasoning?

3. How can viewer cognitive load be estimated from audio–visual features in a way that generalises across content types?

4. How can semantic, thematic, and cognitive layers be integrated into a modular integrated, context-aware segmentation and ad-break decision pipeline?

1.7 Contributions

This thesis makes five key contributions:

1. **Multimodal semantic topic-modelling framework**

   A method that fuses visual, audio, and textual embeddings to produce interpretable, IAB-aligned taxonomies for video content.

2. **Explainable topic refinement using LLMs**

   A mechanism employing LLMs to improve topic clarity, coherence, and multilingual robustness.

3. **Theme-aware segmentation via constrained reasoning**

   A segmentation approach using cross-modal alignment and constrained Tree-of-Thoughts reasoning to infer themes and detect narrative transitions.

4. **Cognitive-load estimation from audio–visual features**

   A multimodal model mapping audio–visual cues to NASA–TLX dimensions and generating a cognitive suitability curve over time.

5. **A Modular, Integrated Multimodal Reasoning Pipeline**

   A complete framework integrating semantic, narrative, and cognitive layers into a coherent system for ad-break boundary selection.

1.8 Thesis Structure

The rest of this thesis is structured as follows, corresponding with the progressive development of the proposed multimodal segmentation technique:

- **Chapter 2** This chapter provides a comprehensive literature review of the topic, concentrating on the analysis of Semantic Videos, models of narratives, cognitive load estimation, as well as dynamic placement of advertising breaks. This chapter also introduces the AdBreakScore(t) scoring system.
- **Chapters 3 and 4** These two chapters address the *Semantic Objective*, whereby the concentration is on topic modelling pipelines design, taxonomy alignment algorithms, as well as topic explanations via LLMs, which form the semantic layer.

- **Chapter 5** meets the *Narrative Objective*, proposing a theme-aware segmentation architecture with a restricted multimodal reasoning system that achieves long-range coherence.
- **Chapter 6** addresses the *Cognitive Objective* through the presentation of a cognitive load estimation framework that is multimodal, as well as suitability measures for transition.
- **Chapter 7** realises the *Integration Objective* through the integration of the semantic, narrative, and cognitive parts via a integrated modular research prototype segmentation and ad-break decision pipeline that translates this interaction.
- **Chapter 8** Concludes the thesis with the summation of the main points, some limitations, as well as areas in which further studies can be undertaken.

The proposed framework provides a seamless link from the basic representation schemes for multimodal knowledge, through storytelling, reasoning, cognitive processing, all the way through to establishing a totally integrated human-centric segmentation system that would facilitate 'intelligent' ad break placement.

Chapter 2

Literature Review and Conceptual Foundations

Publication Notice.

This chapter is adapted from the author's peer-reviewed and published article:

De Silva, W., & Fernando, A. (2025). Toward Intelligent Ad Breaks: A Survey and Taxonomy of AI-Driven Ad Placement in Streaming Media. ***IEEE Access****, 13, 163844–163859. https://doi.org/10.1109/ACCESS.2025.3610662*

Although this chapter is based on a published survey, in this thesis, it is adapted as a conceptual preliminary to the a modular, integrated multimodal reasoning architecture developed in chapters 3-6. The taxonomy and AdBreakScore framework that follow, while intended as a review of existing work, also form a basis upon which the semantic, narrative, and cognitive layers are operationalized in the thesis.

2.1 Introduction to Intelligent Media Systems

The online streaming industry disrupts the conventional manner of interacting with the intended viewer through the platform by delivering experiences which vary from the linear storytelling model of episodic programming to social experiences within the domain of the short-form. Even as these models deliver highly personalized and engaging experiences to the viewer, these models depend on advertising as an integral mechanism. The conflict between the two continues: the presence of intrusive advertising pauses and negatively impacts viewer satisfaction, while the presence of ads which are contextual and appropriately placed improves viewer engagement and impact [16–21].

Former approaches based upon fixed intervals of scheduling, silence detection algorithms, or simple heuristics based upon scene changes simply could not scale or support the diversity that characterizes today's streaming environment [22, 23]. These methods overlook the cognitive and emotional dimensions of media consumption [21, 24] and cannot generalize across variable pacing, narrative rhythm, and production styles. Manual curation also fails to scale for real-time global content flows [25]. These limitations are also emphasized in the referenced a survey [26], which demonstrates that heuristic systems are insufficient for contemporary, multimodal streaming environments.

Recent developments in artificial intelligence (AI) and machine learning (ML) provide a more flexible foundation for adaptive advertising. Techniques from computer vision, natural language processing, affective computing, and reinforcement learning are increasingly employed to infer narrative context, estimate audience states, and recommend insertion points that are both semantically and cognitively appropriate [27–30].

Existing surveys have largely examined discrete components of this ecosystem such as targeting, personalization, or campaign optimisation without establishing an integrative framework that unifies technical, cognitive, structural, and ethical perspectives [30–33]. In particular, this research incorporates a holistic typology based upon the three interdependent stages of the survey: Data, Decision, and Delivery that define in totality the AI process pipeline relevant to ad placement and contextual adaptation. This framework is complemented by the $AdBreakScore(t)$ concept, a cognitively and

emotionally informed metric introduced in the survey for evaluating the appropriateness of advertisement insertion.

Within the context of this thesis, this work expands on these concepts and argues that such taxonomies are essential to the design of **intelligent and adaptive multimedia systems** in general. This generalization constitutes the thesis-level contribution, since, while the survey in this chapter has focused on the application domain of AI-assisted ad placement, here, the underlying concepts of multimodal sensing, reasoning, and cognitive modelling are abstracted and applied to general optimisation for media applications. The survey in this chapter describes the history from heuristic scheduling through to data-driven and AI-assisted adaptive media, which serves as the basis for the multimodal frameworks presented in the following chapters.

Together, these developments illustrate the shift from heuristic scheduling toward AI-driven contextual adaptation. In the context of this thesis, they motivate the need for the three-layer pipeline outlined in Chapter 1, semantic understanding (Chs. 3–4), narrative reasoning (Ch. 5), and cognitive suitability modelling (Ch. 6). This pipeline is a thesis-specific structuring of the broader concepts introduced in the survey and serves as the theoretical foundation for the multimodal frameworks developed later in this work.

2.1.1 Systematic Review Methodology

This chapter develops the taxonomy and detailed analysis after examining the state-of-the-art in AI-driven advertisement placement, multimodal video understanding, and adaptive media delivery.

This study selects more than 400 peer-reviewed research articles through extensive analysis present in reputable research archives like IEEE Xplore, ACM Digital Library, Springer Link, and Google Scholar. The combination of the research keywords was carried out through the use of the keywords: *ad placement*, *mid-roll optimisation*, *contextual video advert*, *multimodal analysis*, *engagement prediction*, and *reinforcement learning in the media*. The main review period was between 2015 and 2025. Earlier seminal works were also considered if relevant.

Screening Procedure. This study filters the initial corpus using the following steps:

- **Title and abstract screening** to remove papers unrelated to video, AI models, advertising, or media adaptation.
- **Full-text assessment** for technical relevance to ad placement, multimodal content understanding, cognitive modelling, or adaptive delivery.
- **Snowball sampling** of reference lists from key papers to expand coverage and include influential but hard-to-find works.
- **Expert inclusion** of foundational studies where necessary (e.g., cognitive load theory, affective responses, fairness and disclosure), even if not directly focused on advertising.

Final Corpus. After screening, over 125 papers were shortlisted for detailed analysis. Each paper is classified based on its contribution to breakpoint detection, engagement prediction, adaptive scheduling, multimodal understanding, and ethics, and has been linked to the three-step process: Data, Decision, and Delivery, discussed in this chapter.

This process ensures that the taxonomy and synthesis presented in this chapter are scientific, clear, and in line with the systematic review principles upon which the published survey is based.

2.2 Evolution of Advertisement and Content Personalization in Digital Media

This section examines how advertising within digital video environments has evolved in media delivery systems. Early broadcast models relied on fixed scheduling intervals and manual segmentation decisions, which were sufficient when content formats and audiences were relatively homogeneous. However, the rise of the "on-demand" and "streaming" services has disrupted this order and brought in a totally diverse form and pattern of viewing. Traditional heuristics such as detecting silences, fades, or shot transitions proved increasingly inadequate in such dynamic environments [22, 23].

These rule-based methods lacked the flexibility to account for variable pacing, narrative structure, and the cognitive states of audiences, leading to ad insertions that often felt arbitrary or intrusive.

The growing maturity in the cyber space environment has made personalization the prominent area in audience interaction. The first area of personalization involved the targeting of users based on metadata parameters, where the user profile and content type determined the advertisement but not the time of advertisement. Though the targeting approach has been effective in ensuring the product is relevant, the temporal element and the psychological process involved in viewing are two important aspects that the approach neglects [18, 21]. For example, well-suited commercials will hinder the immersion effect in viewing if they appear at the wrong point in the structure and emotional progression [18, 21].

The rise of large-scale data gathering and cloud computing made way for even more complex adaptive advertising methods to be accomplished. The process began to include multimodal cues - visual, auditory, and linguistic - to predict the content type, emotions, and audience engagement [27, 29, 30]. All the above methods laid the ground for the rise of *contextual advertising*, where machine learning-enabled methods began to predict suitable ad placements according to the similarity in content and emotions, unlike traditional methods that follow schedules [28].

Concurrently, computational progressions in the realm of cognitive and affective modelling started to make their presence in the domain of ad research. Cognitive models of viewer attention, emotional engagement, and narrative suspense studies showed that a well-placed advertisement break adheres to human cognitive rhythms instead of temporal markers [24, 34, 35]. This line of research has led to the development of cognitively informed models that attempt to evaluate advertisement break suitability based on narrative, emotional, and cognitive factors.

Recent studies have tried to combine the above advancements into a more comprehensive classification system regarding AI in advertising, but in most cases, they focus on a specific area, such as targeting, sentiment analysis, or advertisement optimisation, instead of integrating the ideas into a comprehensive media intelligence view of

the field [31–33]. The current fragmentation of the area, therefore, clearly shows the importance of a conceptual model that brings together ideas in content understanding and cognitive reasoning at a higher level.

Overall, the process of advertisement and content personalization has thus evolved from rule-based heuristics to data-driven contextual models and has now reached the cognition-aware and ethical AI systems phase. The historical process discussed in this section, however, is both a process of ad optimisation and a process where the need for semantic, narrative, and cognitive understanding has gained importance in a manner directly relevant to the pipeline discussed in Chapter 1. Based on the above historical process in this section, the need to structure this thesis in a layered form has been emphasized: semantic content representation in Chapters 3 & 4, narrative theme modelling in Chapter 5, and finally incorporating cognitive suitability assessment in Chapter 6. Section 2.3 continues based on the above lessons by studying their implementation through current AI paradigms in the context of a streaming platform.

2.3 Artificial Intelligence in Streaming and Advertisement Placement

Artificial intelligence has emerged as the driving force behind current streaming media platforms and has led to the optimisation of advertisement insertion instead of the traditional decision-making process. Modern pipelines no longer rely upon the simplicity of heuristics but instead model complex interactions involving the flow of narratives and the cognitive processes involved in viewer immersion, the ultimate aim of which is the insertion of advertisements in terms of time and manner to attain platform revenue goals.

Multimodal Learning for Context and Semantics

AI-based frameworks combine the signals of various modalities, namely visual, auditory, textual, and behavioural, to produce a semantic representation of the video content. Computer vision-based models detect the boundaries of the scenes, the presence

of certain objects, and visual tonality, and natural language models are able to generate the theme and emotions present in the caption/transcription of the video [22, 23, 29]. The combination of the aforementioned modalities provides the system the capability to identify the hidden structures of a story and pair the advertisement with the corresponding story context [27, 30], and this multimodal intelligence is crucial.

Affective-computing extensions enhance this layer by estimating viewers' emotional states through visual or acoustic cues. Emotion-aware ad selection has been shown to improve recall and persuasiveness by harmonizing tonal consistency between content and advertising [24,34,36]. Such adaptive matching forms the basis of emotionally congruent placement, wherein ads are chosen and timed to align with the current affective climate of the media stream.

Learning to Decide: Reinforcement and Cognitive-Guided Models

Aside from the understanding of static content, current work utilizes the application of reinforcement learning to improve placement decisions made in the form of a sequence of policies. The policy aims to utilize the signals received from the audience, known as viewer retention and click-through rates, in order to make placements with the highest long-term engagement [28]. Policy-gradient and deep-Q networks have been adapted for this purpose, balancing exploration of new placements with exploitation of known effective ones.

To integrate human-centric rationality, factors of cognition and narrative are being represented with increasing complexity as components of the decision-making process. Building upon this synthesis of the literature, this thesis formalises these dimensions into the AdBreakScore(t) framework, which integrates narrative fit, cognitive-load appropriateness, and emotional congruity into a integrated modular decision metric. Figure 2.1 provides a conceptual model representing the relationship of these factors with regard to their interaction and integration with softer penalties related to constraints of human-centric appropriateness. This helps artificial intelligence models adopt not only data-grounded regularities and appropriateness parameters but also parameters of cognitive appropriateness related to ad breaks [35, 37, 38].

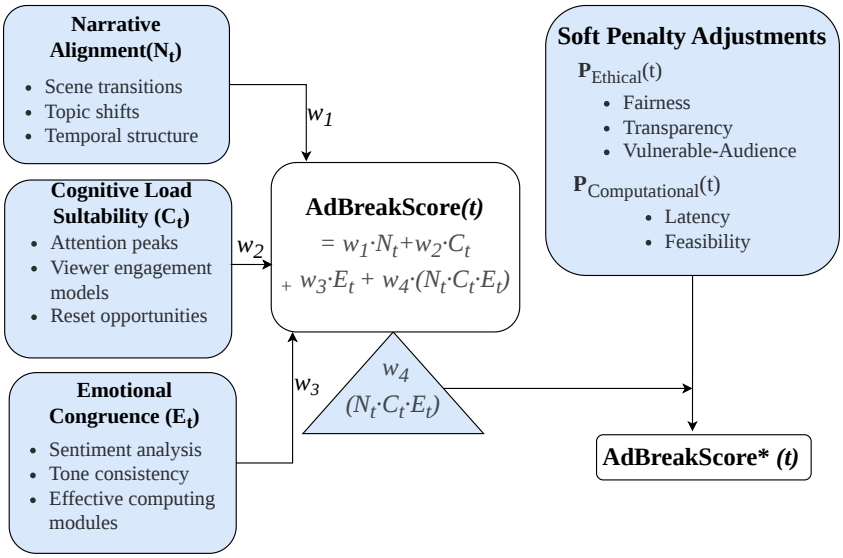


Figure 2.1: Conceptual framework for evaluating ad-break suitability using the *AdBreakScore*(t) model. The model combines narrative alignment (N_t), cognitive-load compatibility (C_t), and emotional congruence (E_t), moderated by ethical and computational penalty terms. Adapted from De Silva et al. (2025).

Operational Considerations in Real-Time Systems

For use in full-fledged streaming platforms, it's necessary for AI models to satisfy low latency requirements. Hybrid architectures were thus designed to integrate light front-end detectors based on silence or shot-level boundaries with deep contextual modules for refined processing [29,39]. This layered processing preserves contextual explanation capabilities while allowing for real-time processing. Edge-centric adaptations, for instance in LiveSense systems [29, 40], employ short video window processing with pre-trained models for real-time cloud-independent learning.

Ethical and Regulatory Dimensions

The integration of AI technology into advertising chains raises new ethical issues regarding transparency, fairness, and the protection of users. The use of personalization through algorithms could diminish the distinction between non-commercial and commercial content on the platform [28, 41]. Regulations like COPPA and other global

regulations attempt to safeguard sensitive groups, including minors [42–45]. The issues above highlight the importance of developing more compliant and responsible AI through the integration of constraints as first-class design parameters.

Summary

Overall, the application of advertisements by an AI system is a combination of multimodal perception, cognitive modelling, and ethics. With the shift from static heuristics to adaptive and cognition-sensitive frameworks, the current state of the art is closer to achieving intelligent media platforms that consider business needs and the well-being of the audience.

Although surveyed primarily within the context of advertisement placement, the techniques discussed in this chapter represent broader multimodal reasoning mechanisms applicable to adaptive media systems more generally. The same foundations of semantic grounding, narrative modelling, cognitive estimation, and operational constraints extend to video editing assistance, adaptive pacing, recommendation timing, and other human-centred multimedia applications.

The following section presents this progression in a formalized taxonomy structured as *Data, Decision, and Delivery.*

2.3.1 Unified Evaluation Model for Ad Placement

To enable consistent comparison of AI-based ad-placement strategies across narrative, cognitive, and affective dimensions, a unified evaluation model termed $AdBreakScore(t)$ is adopted. This formulation expresses the quality of an ad break occurring at time t as a weighted combination of three primary perceptual dimensions **narrative alignment** (N_t), **cognitive suitability** (C_t), and **emotional congruence** (E_t) along with their nonlinear interaction:

$$\text{AdBreakScore}(t) = w_1 N_t + w_2 C_t + w_3 E_t + w_4 (N_t \cdot C_t \cdot E_t) \tag{2.1}$$

In this scenario, w_i represents the weights learned or set through experimentation to

determine the importance of the components. The interaction term $w_4(N_t \cdot C_t \cdot E_t)$ captures synergy between narrative, cognitive, and emotional components, which rewards advertisements that are well-aligned, comfortable and affectively resonant.

To reflect deployment constraints, an adjusted version penalizes ethically or computationally suboptimal insertions:

$$\text{AdBreakScore}^*(t) = \text{AdBreakScore}(t)(1 - \lambda_{\text{eth}} P_{\text{Ethical}}(t) - \lambda_{\text{comp}} P_{\text{Computational}}(t)) \tag{2.2}$$

where $P_{\text{Ethical}}(t)$ and $P_{\text{Computational}}(t)$ denote normalized penalty factors (e.g., violation risk, latency, or resource cost), and λ_{eth}, λ_{comp} are their respective scaling coefficients. The adjusted score $AdBreakScore^*(t)$ therefore represents the holistic trade-off between perceptual quality and operational feasibility, serving as the reference criterion for the taxonomy discussed in Section 2.4.

Within the scope of the thesis, the AdBreakScore formula also serves as a structural bridge to the following chapters. Narrative alignment (N_t) is measured using the multimodal theme induction framework described in Chapter 5, while the measure of cognitive suitability (C_t) is calculated using the multimodal cognitive load model described in Chapter 6. The semantic vectors used as input to both N_t and E_t come from the multimodal topic and taxonomy models described in Chapters 3-4. In this way, *AdBreakScore* not only serves as a means of evaluation, but also as a structural bridge to the technical contributions of the thesis.

Building upon this evaluation model, the next section organises the diverse range of AI techniques for ad-placement optimisation into a structured taxonomy that spans data processing, decision modelling, and delivery orchestration.

It should be noted, however, that in the context of the current work, AdBreakScore(t) plays the role of a conceptual integration scaffold rather than an end-to-end optimisation objective, which has been fully and optimally defined and implemented. This is because the constituent parts of the AdBreakScore(t) corresponding to the components of narrative alignment (Nt), cognitive suitability (Ct), and semantic representations

related to emotional congruence (Et) are individually defined and addressed in Chapters 3-6 of the current work. However, end-to-end joint optimisation of all parts of AdBreakScore(t) under a integrated modular policy objective is not addressed in the current work and is instead recognized as an interesting direction for future work.

2.4 Taxonomy of AI Strategies for Ad Placement optimisation

AI-driven ad placement systems address challenges of narrative engagement, content understanding, computational tractability, and ethical compliance. They draw on a tiered set of techniques from low-level scene analysis to high-level decision making and delivery orchestration. Figure 2.2 shows the overall taxonomy grouped into three pipeline stages: *Data*, *Decision*, and *Delivery*. This taxonomy synthesizes insights from more than 125 studies published between 2015 and 2025 that were screened in the systematic review. Each class of techniques influences the $AdBreakScore(t)$ components - narrative alignment (N_t), cognitive suitability (C_t), and emotional congruence (E_t). Where relevant, the penalty-adjusted variant $AdBreakScore^*(t)$ is also used to indicate operational feasibility (latency, resource limits) and ethical compliance (age, fairness, disclosure).

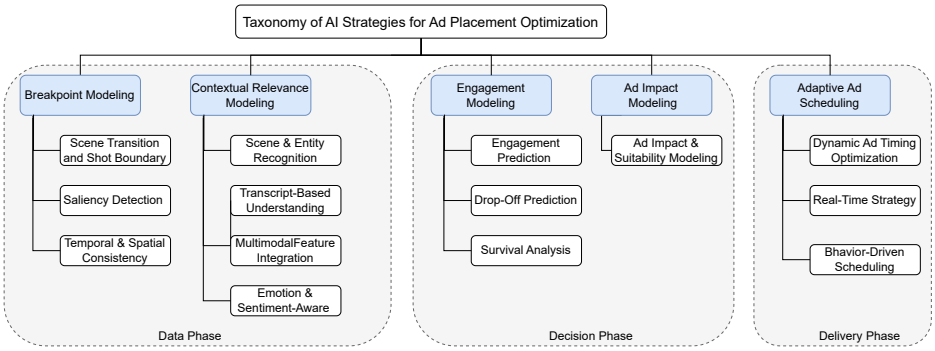


Figure 2.2: Taxonomy of AI strategies for ad-placement optimisation across *Data*, *Decision*, and *Delivery*, with representative subcomponents (modalities, learning paradigms, and deployment mechanisms).

This taxonomy is also conceptually connected to the methodological design approach in the thesis. The *Data* phase is relevant to the work on semantic modelling in

Chaps. 3–4, the *Decision* phase is relevant to the frameworks for narrative and cognitive reasoning in Chaps. 5–6, and the *Delivery* phase is relevant to the integrated end-to-end reasoning approach in Chap. 7. Therefore, the taxonomy is not only integrative in synthesizing previous work but is also structurally logically cohesive with the thesis.

2.4.1 Breakpoint Detection modelling (Data Phase I)

In this subsection, low-level content analysis methods used in identifying candidate breakpoints are discussed. It includes shot boundary detection, scene transition identification, saliency analysis, and maintenance of temporal/spatial coherence. There are breakpoints where the interruption is least intrusive, such as around a change of scene or topic. Computer vision, NLP, and multimodal approaches for this are included in the *Data* phase of the taxonomy.

Scene-Transition and Shot-Boundary Detection

Natural transitions detection is the basis for non-disruptive advertising. vADeo [46] and VideoSense [47] proposed transition detection using colour and motion features. Deep learning made it possible to apply semantic transition detection using CNNs, LSTMs, and attention mechanisms [48–51]. TransNet-V2 [52] is capable of real-time cut detection. Models using graphs and temporal convolutional networks explore long-term patterns. *AdBreakScore link:* Primarily elevates N_t by aligning breaks with narrative boundaries; can indirectly support C_t by reducing perceived disruption.

Saliency and Attention modelling

Saliency maps highlight low-attention frames suitable for ads. Classical models [53–55] evolved into deep saliency predictors [56, 57] and real-time attention-aware renderers such as LiveSense [29]. These minimize cognitive disruption (C_t) by avoiding focal areas. *AdBreakScore link:* Direct lift to C_t; may stabilize E_t by lowering perceived intrusiveness.

Temporal and Spatial Consistency

Stable, realistic ad insertions require temporal coherence. Optical-flow tracking [58, 59] and 3D camera tracking [60, 61] maintain alignment during motion and preserve occlusions. *AdBreakScore link:* Improves perceived quality and lowers penalties in *AdBreakScore*$^*(t)$ by mitigating visual instability and jitter.

2.4.2 Contextual Relevance modelling (Data Phase II)

Higher-level semantic modelling that interprets what the scene is about and how ads may align contextually. Covers object- and visual-context methods, transcript and language analysis, multimodal fusion, and affect/sentiment estimation. Contextual relevance ensures semantic and affective coherence with adjacent content, contributing mainly to N_t and E_t.

Object- and Visual-Context Systems

Techniques such as VideoAder [62], OLVA [63], identified objects in videos and mapped these to ad types. More recent approaches using CNNs [64, 65], improved upon recognition and personalization capabilities but remained object-centric. *AdBreakScore link:* Increases N_t through semantic linking, though it has little direct effect on C_t, except when used in the context of attention models.

Language- and Transcript-Based Scene Understanding

Scripts, captions, and ASR transcripts reveal scene topics. Early statistical approaches [66–68] have been complemented by transformer-based embeddings (e.g., BERT, RoBERTa) and lightweight caption-driven targeting [69]. *AdBreakScore link:* Strengthens N_t and, when coupled with sentiment analysis, E_t; robust to ASR noise with modern embeddings.

Multimodal Fusion for Ad Relevance

Integrating visual, audio, and text cues captures holistic context. Fusion may occur early, mid, or late in the pipeline. Representative mid-level systems include DEEP-

AD [70], SemanticAd [71], and Video eCommerce++ [72]. Graph-based and modular expert frameworks [73–75] extend scalability and interpretability. *AdBreakScore link:* Targets both N_t and E_t; some systems approximate the interaction term $w_4(N_t \cdot C_t \cdot E_t)$.

Emotion and Sentiment Analysis

Emotion and Sentiment Analysis Tone and mood are matched in affect-congruent advertising. Valence arousal models [24], behavioural signal inference [76], and multimodal transformers like TIMAN [77], MEDT [78], measure emotion dynamics for adaptive insertion. *AdBreakScore link:* Direct lift to E_t; may indirectly boost C_t by skirting around emotionally engaging portions.

2.4.3 Comparative Evaluation of Data-Phase Methods

To consolidate methodological diversity, representative systems are compared in terms of their stage within the taxonomy, dataset usage, performance metrics, and qualitative strengths and limitations (Table 2.1). The heterogeneity of approaches is consolidated into a maturity-oriented view spanning four design phases: *heuristics*, *hybrid systems*, *deep semantic models*, and *real-time adaptive systems*. Each phase contributes differently to optimizing N_t, C_t, and E_t in *AdBreakScore*(t); streaming-oriented designs more strongly affect the penalty terms in *AdBreakScore*$^*(t)$. These comparisons highlight trade-offs between semantic depth, interpretability, and computational feasibility dimensions that directly correspond to *AdBreakScore*$^*(t)$ subcomponents.

2.4.4 Evaluation of Existing Work Using the *AdBreakScore*(t) Criterion

To facilitate a structured comparison, representative approaches from the literature are mapped onto the key components of the *AdBreakScore*(t) framework: semantic coherence (N_t), contextual relevance (C_t), viewer engagement (E_t), their combined interaction ($N_t \cdot C_t \cdot E_t$), and the penalty terms (P_{Ethical} and $P_{\text{Computational}}$). This mapping highlights important trade-offs between contextual depth and practical deployment feasibility.

The $AdBreakScore(t)$ itself integrates these components through a weighted formulation involving parameters W_1–W_4 and λ. These weights were determined via controlled empirical tuning rather than a closed-form analytical solution, with the objective of achieving balanced and interpretable behaviour across diverse content types. In particular, the penalty weight λ regulates the influence of ethical and computational constraints on candidate breakpoint selection.

The chosen configuration reflects a deliberate balance between contextual fidelity and robustness. Sensitivity analysis further indicates that the framework remains stable under moderate variations in these weighting parameters.

Table 2.1: Taxonomy classification of AI approaches: scene/shot detection, saliency modelling, and object/context recognition.

Stage	System	Dataset	Evaluation	Strengths	Limitations
Sec. 2.4.1 : Scene/Shot Detection					
Heuristic	vADeo [46]	-	PoC[1]	Fast, interpretable	Brittle in complex narratives
	VideoSense [47]	-	PoC[1]		
	Bicda et al. [51]	Custom	F1 $\approx$ 92%[2]		
Hybrid	Wang et al. [79]	Custom	F1 $\approx$ 84%[2]	Interpretable with modest semantic depth	Limited generalization; mid-complexity only
	Liang et al. [50]	Custom	F1 $\approx$ 91%[2]		
	Yuan et al. [49]	TRECVid	F1 $\approx$ 95%[2]		
	SemanticAd [71]	TRECVid	$F_1 \approx 88\%$[3]		
	Chong et al. [80]	MovieScene	0.33 S-score		
	Ye et al. [81]	BBC	$\approx$ 0.65 F_co		
Deep Semantic	Rao et al. [82]	MovieScene	$\approx$ 47% AP[4]	Context-aware, narrative-sensitive	Heavy; slower for live streaming
	Chen et al. [48]	MovieNet	$\approx$ 53% AP[5]		
Real-Time Adaptive	TransNet V2 [52]	ClipShots	F1 $\approx$ 78%[6]	Real-time capable	Limited semantic/narrative alignment
Sec. 2.4.1 : Saliency / Attention modelling					
Heuristics	Rudoy et al. [55]	DIEM	$\chi^2 \approx 0.3$[7]	Understandable decisions	Limited generalization
Hybrid	AdOn [83]	Custom	Satisfaction = 4.1[8]	Saliency and face/text/motion for overlays	No deep learning; handcrafted features
	Dynamic Saliency (MT) [84]	SFU Eye-Tracking	NSS $\approx$ 1.4[9]	Scene-level attention	No narrative understanding
	Lu et al. [57]	Custom	SROCC $\approx$ 0.85[10]	3D CNN + attention	Some hand-tuned weights
Real-Time	LiveSense [29]	Custom (Twitch)	RMSE $\approx$ 0.15[11]	Real-time adaptive placement	Live-streaming focus
Sec. 2.4.1 : Temporal/Spatial Consistency					

Stage	System	Dataset	Evaluation	Strengths	Limitations
Heuristics	SfM [85]	Custom	N/A	Spatial alignment concept	Limited temporal handling
Analytical + Empirical	Lucas–Kanade [58]	Custom	Qualitative	Robust for smooth motion	Limited under abrupt motion
Hybrid	Wong et al. [59]	ADE20K	≈ 1.8 AP	Temporal/spatial consistency	Depth estimation sensitivity
	Efimova et al. [60]	Custom	IoU ≈ 0.83	Occlusion-aware 3D tracking	
	Bacher et al. [61]	Custom	Qualitative	3D geometry, occlusion handling	High compute cost
Sec. 2.4.2 : Object/Context Recognition					
Hybrid	VideoAder [62]	Custom	Precision $\sim$ 96%	Clear object–product links	Context blind; static
	OLVA [63]	Custom	Precision@10 ≈ 0.33	Context-aware, object-driven	Heuristic matching limitations
	DeepLink [65]	Custom	MRR $\approx 59\%$	Strong retrieval	Multiple CNNs; high compute
Real-Time Adaptive	Yes-Net [64]	VOC2012	mAP $\approx 79\%$	Real-time	Limited narrative understanding
Sec. 2.4.2 : Language/Transcript-Based Understanding					
Heuristics	TAS [69]	Custom	User-opinion $\approx 85\%$	Uses viewer attributes	Limited semantic depth
Hybrid	Ferreira et al. [66]	Custom	Precision@1 ≈ 0.26	TF/POS/PMI features	Noisy transcripts reduce accuracy
	Welch et al. [67]	Custom	Precision ≈ 0.44	Robust to noise	Limited semantics
	Yi et al. [68]	Custom	NDCG@5 ≈ 0.5	Situation-aware keywords	Shallow modelling
Sec. 2.4.2 : Multimodal Fusion					
Hybrid	Madhok et al. [86]	Custom	Acc $\approx 96\%$	Rules + learning	Scalability limits
	Kovaliuk et al. [87]	Custom	Qualitative	Lightweight	Lacks deep semantics
	ActVA [73]	Custom	$> 30\%$ "Not Intrusive"	Graph boosts relevance	Subtitle/vision quality dependency
	Chong et al. [80]	Custom	S-score ≈ 0.33	Uses pretrained models	Limited generalization

Stage	System	Dataset	Evaluation	Strengths	Limitations
	DEEP-AD [70]	Custom	F1 $\approx$ 83%	Scene consolidation	Compute cost
	ContextIQ [74]	Custom	MAP@5 $\approx$ 94%	Scalable expert retrieval	Independent embedding spaces
Deep Semantic	Song et al. [88]	Custom	Relevance $\approx$ 2.1	Content-agnostic	Fusion can be heavy
	Explainable Video Topics [75]	Custom	F1 $\approx$ 80%	Interpretable topics	Risk of semantic dilution
Real-Time Adaptive	LiveSense [29]	Custom	RMSE $\approx$ 0.15	Dynamic, viewer-adaptive	Shallow semantics possible
	Video eCommerce++ [72]	Custom	RMSE $\approx$ 0.43	Multi-factor personalization	Data dependency

2.4.5 Key Findings

1. **Narrative-first bias:** early methods optimise N_t via cuts/scene boundaries but under-model C_t and E_t.

2. **Affective growth, cognitive gap:** emotion-aligned insertion improves E_t, yet explicit C_t remains sparse.

3. **Limited interaction optimisation:** few systems model $(N_t \cdot C_t \cdot E_t)$ jointly.

4. **Depth–efficiency trade-off:** streaming systems raise $P_{\text{Computational}}$ compliance (low latency) at the cost of multi-dimensional depth; deep semantic systems show the reverse.

These gaps mirror the research objectives outlined in Chapter 1 and directly motivate the development of the semantic topic-taxonomy models (Chs. 3–4), theme-aware segmentation framework (Ch. 5), and cognitive suitability modelling (Ch. 6), all of which together form the integrated modular pipeline synthesized in Chapter 7.

2.4.6 Viewer Engagement modelling (Decision Phase)

Predictive modelling is used to anticipate when viewers are cognitively receptive to advertising and how they are likely to respond to interruptions. Using historical viewing

traces, interaction signals, and contextual metadata, these systems forecast receptivity and guide ad timing with minimal impact on experience [89–91]. Conceptually, such predictions primarily inform the C_t term of $AdBreakScore(t)$ (cognitive suitability), and when implemented under latency and resource constraints also affect the penalty term $P_{\text{Computational}}(t)$ in $AdBreakScore^*(t)$. These methods are therefore classified under the *Decision* phase of the taxonomy.

User Engagement Prediction

Optimizing ad timing depends on forecasting engagement to avoid fatigue and abrupt exits. Prior work positions engagement (e.g., fraction watched, replays, interactions) as central to recommendation, retention, and monetization [92]. McDuff and Berger [93] showed that automated analysis of facial expressions is predictive of ad sharing; high-arousal responses (e.g., disgust, joy) later in videos are strong indicators of intended engagement. In parallel, early industrial systems such as YouTube's Dynamic Ad Loads (DAL) [94] demonstrated that real-time activity signals can balance revenue and engagement through dynamic insertion policies.

These approaches motivate integrating affect and attention models [95, 96] (e.g., spatio-temporal features, LSTMs) into ad timing, while surfacing privacy considerations for biometric data. Complementary "cold-start" strategies predict relative engagement without user histories by relying on metadata (duration, topic, channel priors) [97]. Stappen et al. [98] further show that continuous affect traces (arousal, valence, trust) can estimate engagement where user histories are unavailable, enabling timing around affect troughs/peaks. Extending this line, DIGITWISE [92] introduces Digital Twin (DT) sensitivity profiles with XGBoost for user-conditional engagement modelling beyond generic QoE metrics.

Viewer Drop-Off Prediction

Relative engagement [97] can be used to position ads before likely exits. Dedieu et al. [99] forecast session length at login, enabling early interventions (pacing, skipability, teaser content) for high-risk sessions. From a network/QoE angle, Wassermann et al. [100]

show ensemble models can classify drop-off using only network indicators. Qiao et al. [101] find that *rebuffer frequency* and *bitrate switching*, rather than total stall time, dominate disengagement; their QoE-optimised ABR yields a 3.4% mean increase in session duration evidence that engagement-aware control is practical for ad timing. Lebreton et al. [102, 103] provide lab-calibrated models linking completion and quit dynamics to stalling position/duration and MOS, indicating that even brief stalling, especially late in a session, sharply elevates abandonment risk. Beyond network and behaviour, content salience affects retention: Huber et al. [104] show higher low-level salience (motion, colour, contrast) improves survival, suggesting salience-aware timing can complement drop-off models.

Survival Analysis

Survival methods predict time-to-event (e.g., time-to-quit) and expose retention windows suitable for ads. Aguiar et al. [105] report genre-dependent survival curves (e.g., sharper drop-offs in news vs. sustained entertainment), guiding time-window selection. Classical survival modelling [106] (Kaplan–Meier, Cox PH) suits disengagement trends. Qiu and Cui [107] model session end as an event with buffer features and popularity predictors; accuracy reaches 90%, supporting ad positioning before predicted exits. In live TV contexts, Lebreton and Yamagishi [108] combine temporal usage with bitrate/stall metrics to analyse quitting behaviour at scale. Survival analysis can also inform *post-click* engagement: Random Survival Forests predict landing-page dwell time [109], a principle extendable to *post-ad watch time* in video (dual optimisation of ad choice and timing).

2.4.7 Ad Suitability & Effectiveness modelling

Ad selection to match the most appropriate ad format, ad style, tone, and expected completion is a complement to timing policies. Attention-based deep CTR using GRUs [110] is a method that models temporal dependencies in ad engagement to make accurate click-through/rate predictions. In video scenarios where there is no engagement data available for the viewer, feature-based approaches (e.g., likes, dislikes, views, and com-

ments) are strong predictors of engagement classification in sentiment classification tasks on video platforms such as YouTube [111]. Viewability prediction using interpretable behavioural and contextual features further indicates how simple, auditable signals can yield strong performance.

For affect-aware effectiveness, Trans-LSTM [76] combines Transformers and LSTMs over sequential cues (likes, comments, shares, watch time) to preserve emotional dynamics for ad completion prediction, outperforming classical baselines. At session level, causal ML for *Bingeability* and *Ad Tolerance* [112] supports pacing and load decisions in streaming platforms. Together, these models primarily inform E_t and the interaction term in $AdBreakScore(t)$ by aligning ad characteristics with local context and predicted viewer state.

2.4.8 Comparative Evaluation of Decision-Phase Methods

To clarify how decision-phase methods influence $AdBreakScore(t)$ and its penalty-adjusted variant, representative approaches are mapped to to C_t, E_t, the interaction term ($N_t \cdot C_t \cdot E_t$), and ($P_{\text{Computational}}, P_{\text{Ethical}}$). A distinction is also made between *tactical* (short-term, session-level) versus *strategic* (long-horizon) predictions and *risk* (avoid churn) versus *opportunity* (maximize completion/receptivity) goals.

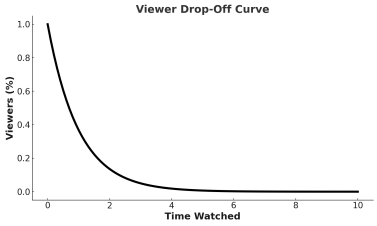
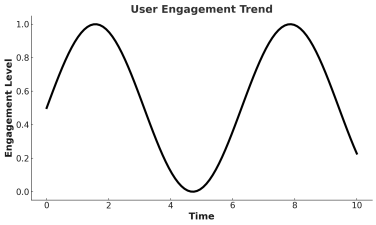
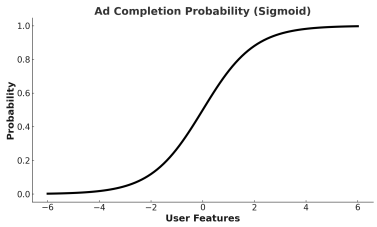
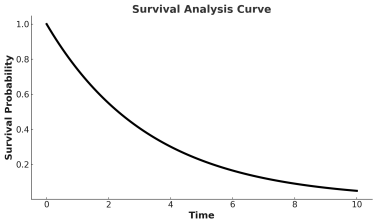
Key Findings from Decision-Phase Evaluation

1. **Cognitive/behavioural dominance.** Foundational systems [94,97,100] emphasize C_t (engagement/QoE/drop-off) and under-model N_t, yielding good immediate trade-offs but limited narrative sensitivity.

2. **Richer affect, limited integration.** Emotion-centric approaches [93, 98] lift E_t, and hybrid survival [109] adds post-ad engagement; yet C_t and E_t are often siloed.

3. **Underexplored interaction term.** Few models jointly optimise ($N_t \cdot C_t \cdot E_t$). DIGITWISE [92] approaches C_t+E_t via user sensitivities; survival models [105] blend N_t+C_t via genre/retention but full triad integration is rare.

Table 2.2: Classification of predictive models for user engagement and ad optimisation, by tactical/strategic scope and risk/opportunity intent.

Category	Tactical Prediction	Strategic Prediction
Risk Management	**Predicting Viewer Drop-Off** *Goal:* Detect imminent dropout *Method:* RF/ensembles, interpretable ML [99–103] *Action:* Immediate pacing/format intervention 	**Engagement Prediction** *Goal:* Understand engagement trends over time *Method:* LSTM, affect time-series, DT+XGBoost [92–94, 96–98] *Action:* Plan long-term engagement strategies 
Opportunity Maximization	**Ad Impact & Suitability** *Goal:* Predict ad completion/viewability *Method:* Logistic/CTR/Trans-LSTM [76, 110–113] *Action:* Dynamically select formats/types 	**Survival Analysis for Scheduling** *Goal:* Predict user survival (watch) time *Method:* Cox/Kaplan–Meier/RSF [105–109] *Action:* Optimise ad schedules and ad choice 

4. **Efficiency vs. comprehensiveness.** Real-time frameworks [94, 101] optimise P_{Comp} (low latency) at the cost of multimodal depth; affect-heavy or survival models raise accuracy but with higher compute.

2.4.9 Adaptive Ad Scheduling and Delivery optimisation (Delivery Phase)

The final stage of the taxonomy concerns *delivery optimisation* determining not only *where* but also *when*, *how frequently*, and *which* advertisements are to be shown. This stage is where the earlier predictions and insights come into play, turning them into policies for serving ads. Seen from the perspective of algorithms, these problems naturally fit into the category of sequential decision problems under uncertainty, where

Table 2.3: Evaluation of decision-phase approaches mapped to *AdBreakScore*(t) components.

Approach Cluster *(Representative Works)*	C_t	E_t	**Interaction** $(N_t \cdot C_t \cdot E_t)$	$P_{\textbf{Comp}}$	$P_{\textbf{Ethical}}$	**Key Notes**
Emotion-Driven Engagement (McDuff & Berger [93]; Stappen et al. [98])	High	High	Medium	Medium	Low	Biometric/affect timing; strong signals; privacy risk considerations.
behavioural Activity (YouTube DAL [94]; Wu et al. [97])	High	Medium	High	Low	Medium	Session-level activity + metadata; solid cold-start; limited fine temporal dynamics.
Drop-Off Forecasting (Dedieu [99]; Lebreton [102, 103])	High	Medium	Medium	Medium	Medium	Stalling-aware exit prediction; enables tactical pacing and skipability policy.
QoE-Optimised Retention (Wassermann [100]; Qiao [101])	High	Low	Medium	Low	High	Playback stability (rebuffer freq., bitrate switches) dominates engagement; practical for live contexts.
Content-Driven Survival (Aguiar [105]; Huber [104])	High	Medium	Medium	Medium	High	Genre/content windows for timing; salience improves survival; guides ad windows.
Hybrid Survival (Qiu & Cui [107]; Kapoor et al. [109])	High	Medium	High	Medium	Medium	Dual optimisation: timing + ad selection via watch-time/dwell survival models.

Legend: C_t/E_t/Interaction are qualitative coverage levels (Low/Medium/High). P_{Comp}: High means optimised for real-time; Low indicates heavy/latency-sensitive. P_{Ethical}: presence of explicit compliance/guardrails.

reinforcement learning (RL) is the appropriate framework. In this context, the learning agent continuously updates its policy to maximize a reward signal proportional to the adjusted *AdBreakScore*$^*(t)$, thereby balancing engagement gains against ethical and computational penalties.

Dynamic Ad Timing optimisation

Adaptive scheduling relies on optimizing ad timing dynamically according to user interaction states, engagement trajectories, and content rhythm. Two most popular paradigms in reinforcement learning that have been used to achieve optimisation in adaptive scheduling are Markov Decision Processes (MDPs) and Deep Q-Networks

(DQNs).

Markov Decision Processes (MDPs) When ad placement is expressed as a Markov Decision Process, each user–content interaction is modeled through a state–action–reward framework:

$$(S, A, P, R, \gamma), \tag{2.3}$$

where S denotes system states (e.g., user engagement, ad load, scene features), A defines admissible actions (e.g., insert, delay, skip), $P(s'|s, a)$ represents transition probabilities, $R(s, a)$ the reward linked to cognitive and emotional suitability (C_t, E_t), and γ the discount factor balancing short- and long-term objectives. Representative work includes Liu and Yang [114], who optimised ad policies through user behaviour logs and real-time feedback, and Zhao et al. [115], whose *DEAR* framework modeled ad decisions in short-video feeds (e.g., Douyin) as a two-level MDP combining recommendation and ad insertion. The later *RAM* framework [116] further integrated content and ad placement decisions, highlighting how hierarchical MDPs can jointly maximize revenue and engagement.

Deep Q-Networks (DQNs) Deep Q-Networks approximate the action–value function:

$$Q(s, a) = R(s, a) + \gamma \max_{a'} Q(s', a'), \tag{2.4}$$

MDP algorithms are extended by DQN to handle environments with high-dimensional state spaces, aiming to apply them to environments where there are millions of users and advertisement candidates. The application of DQN in ad scheduling, as proposed by systems like *DEAR* [115], involves the application of the Q-function, state value function, and advantage function to handle combinatorial actions like insertion, selection, and delay. Likewise, Ramagundam [117] made use of q-learn algorithms in a multi-format environment and illustrated the capability of deep reinforcement agents in controlling ad-loads and engaging the viewer while simultaneously.

Real-Time Ad Strategy Adaptation

For low-latency decisions, reinforcement learning can be simplified into Multi-Armed Bandit (MAB) models that capture the trade-off between *exploration* and *exploitation*. An ϵ-greedy policy, for instance, selects the best-known ad with probability $1 - \epsilon$ or explores a random alternative otherwise:

$$A_t = \arg\max_a Q(a) \quad \text{with probability } 1 - \epsilon, \text{ else random } a. \tag{2.5}$$

Advanced MAB variants such as Upper Confidence Bound (UCB) [118] and Thompson Sampling (TS) [119] introduce adaptive uncertainty handling, outperforming A/B testing in responsiveness and efficiency [120]. Recent developments integrate deep learning with MABs for cold-start and feedback-delay scenarios [121], making them viable for high-throughput environments like TikTok or YouTube Live. Their low computational footprint positions them as ideal cognitive controllers for "what-to-play" decisions under strict latency constraints.

Behaviour-Driven Placement and Sequencing

Hybrid models that combine reinforcement learning with deep supervised networks have gained attention for unifying content and behavioural modelling. For example, Liu and Yang [114] propose a CNN–RNN representation of user viewing patterns that feeds into an RL agent for ad recommendation. The DEAR [115] and RAM [116] systems embed GRU-based user sequence encodings into RL pipelines, allowing the model to jointly optimise content relevance and ad effectiveness. Other hybrid approaches incorporate NLP-based sentiment analysis to extract user preferences from textual data, which can be used to enhance personalization and advertising effectiveness [122].

Policy Gradient Methods for Continuous Control

Policy Gradient (PG) algorithms extend reinforcement learning to continuous control problems, such as adaptive ad frequency or intensity regulation. The expected return

$J(\theta)$ is optimised through:

$$\nabla J(\theta) = \mathbb{E}_{s\sim d^{\pi}, a\sim\pi_{\theta}} \left[\nabla_{\theta} \log \pi_{\theta}(a|s) Q^{\pi}(s,a)\right], \tag{2.6}$$

where $Q^{\pi}(s,a)$ represents the expected action value. Techniques such as DDPG [123], SAC [124], and PPO [125] can capture continuous behavioural tolerances e.g., when a viewer is likely to tolerate an ad without disruption. Although few implementations exist in video advertising due to sample inefficiency and simulation challenges, PG-based methods remain a promising direction for scalable, personalized ad systems aligned with long-term user satisfaction.

Evaluation of Delivery-Phase Approaches via the *AdBreakScore*(t) Criterion

Table 2.4 summarizes representative delivery-phase frameworks, mapped to the *AdBreakScore*(t) components: cognitive suitability (C_t), emotional congruence (E_t), interaction synergy ($C_t + E_t$), and penalty adjustments ($P_{\text{Comp}}, P_{\text{Ethical}}$).

Key Findings from Delivery-Phase Evaluation The synthesis of delivery-phase studies reveals four major insights:

1. **Latency–optimisation Trade-offs:** Real-time RL and MAB systems favor responsiveness and deployment scalability, while richer multimodal RL pipelines trade latency for interpretability and precision.

2. **Limited Ethical Formalization:** Ethical penalties (P_{Ethical}) are rarely explicit optimisation terms. Most systems rely on data anonymization or manual rule filters rather than fairness- or transparency-driven objectives.

3. **Sparse Adaptive Feedback Loops:** Despite theoretical support for continuous adaptation, most deployed RL frameworks (e.g., DEAR, RAM) operate via static, periodically updated policies.

4. **Emergence of Personalized Control:** Policy gradient and hybrid models signal a shift toward individual-level adaptation, blending behavioural, affective, and

Table 2.4: Evaluation of Adaptive Delivery Approaches (Delivery Phase)

Approach	C_t	E_t	$C_t + E_t$	$P_{\textbf{Comp}}$	$P_{\textbf{Ethical}}$	**Key Notes**
Markov Decision Process (MDP)	High	Medium	Medium	Medium	Medium	Sequential modelling with explicit reward functions; effective for engagement–revenue balance but moderately compute-heavy.
Deep Q-Network (DQN)	High	Medium	Medium	High	Medium	Scalable policy learning for large state spaces; mitigates ad fatigue through temporal Q-updates.
Multi-Armed Bandit (MAB)	Medium	Low	Low	Low	Medium	Lightweight real-time selection; ideal for latency-sensitive or resource-limited contexts.
Hybrid (RL + Deep Supervised Models)	High	Medium	High	High	Medium	Combines GRU-based behavioural encoding with RL for joint optimisation; increased complexity.
Policy Gradient (PG)	Medium	Medium	Medium	High	Medium	Continuous control of ad frequency and pacing; limited deployment but high personalization potential.

Note: Delivery-phase systems generally prioritize timing and frequency (*when/how*) rather than narrative coherence (N_t). Ethical components (P_{Ethical}) remain underrepresented across frameworks.

contextual insights for precision scheduling.

Summary Adaptive scheduling and delivery optimisation complete the AI-driven ad placement pipeline by translating perception and decision layers into real-time execution. Through reinforcement learning, bandit models, and hybrid behavioural architectures, modern systems are beginning to achieve self-optimizing ad ecosystems. Future work should embed ethical, transparent, and multi-objective optimisation directly into the learning loop to ensure sustainable viewer satisfaction alongside platform monetization goals.

2.5 Challenges and Future Research Directions in Adaptive Media Systems

This section synthesizes the three-phase taxonomy (*Data*, *Decision*, *Delivery*) with the *AdBreakScore*(t) and penalty-adjusted *AdBreakScore*$^*(t)$ criteria to surface open challenges and outline research directions. The focus is placed on how current pipelines negotiate narrative alignment (N_t), cognitive suitability (C_t), and emotional congruence (E_t) under computational and ethical constraints ($P_{\text{Computational}}(t)$, $P_{\text{Ethical}}(t)$), and identify where future work can tighten prediction–action coupling and system-level feedback.

2.5.1 Semantics at Scale: Open Problems in the *Data* Phase

Multimodal fusion, saliency modelling, and affect recognition have substantially improved contextual fidelity for ad placement; however, these gains often come with increased latency and hardware demands that complicate live operation. Deep semantic models provide strong N_t and E_t signals but remain difficult to deploy under tight $P_{\text{Computational}}(t)$ budgets in streaming settings. Hybrid pipelines that apply lightweight pre-filters before selective deep inference are a pragmatic compromise, yet they still require *portable* features that can be reused upstream and downstream across the pipeline.

A promising path is to move from shot-level descriptors toward *story-level* representations that are both semantically rich and compact. Emerging LM/VLM techniques for long-context video understanding [126,127] enable the extraction of coherent, time-extended structure that is more directly aligned with N_t while reducing redundant per-frame compute. Future work should standardize cross-phase feature interfaces (e.g., shared topic spans, affect arcs, engagement priors) so that the same representations can inform breakpoint detection (Data), engagement modelling (Decision), and pacing (Delivery) with minimal translation overhead.

Chapters 3 and 4 of this thesis address this need directly by developing reusable multimodal topic representations and taxonomy-aligned semantic structures that act as such story-level features and support downstream narrative and cognitive reasoning.

2.5.2 Closing the Prediction–Action Gap in the *Decision* Phase

Engagement, drop-off, and receptivity predictors are now quite capable, yet their outputs are not routinely bound to *explicit* scheduling rules or learning policies. While this thesis organised viewer models along tactical/strategic and risk/opportunity axes, few implementations instantiate clear mappings from predicted states to pacing/frequency decisions or to thresholded ethical constraints. This weak coupling limits the realised uplift of even accurate models.

Research should emphasize *policy-usable* predictions (with confidence, horizon, and sensitivity) and develop *executable decision templates* that translate forecasts into constrained actions (e.g., "delay mid-roll by Δt if predicted hazard $> \tau$" or "cap E_t mismatch at ϵ for protected cohorts"). Joint training of predictor + policy (via differentiable surrogates for *AdBreakScore**) is a promising route to shrink this gap.

The cognitive-load estimation framework introduced in Chapter 6 is designed with these concerns in mind, producing actionable cognitive-suitability boundaries that map directly onto decision rules within adaptive scheduling policies.

2.5.3 Explainability as Cross-Cutting Requirements

As optimisation becomes more automated, *explainability* must evolve from model-level add-ons to a *system property*. Attribution tools like SHAP and LIME [128, 129] can expose influential factors behind timing and pacing choices, but their performance and stability for multimodal, streaming inputs (gaze/audio/narrative/behaviour) remain underexplored. Explanations should reflect the *AdBreakScore* dimensions introduced in Section 2.3.1 (why C_t was high, why E_t mismatch triggered a penalty, etc.) rather than raw feature importances.

Equally, pipelines must shift from one-pass inference to *closed-loop learning*: downstream delivery signals (skip, rewatch, dwell, dissatisfaction events) should feed back to re-weight features, tune thresholds, and update priors. Reliable telemetry for these loops time-synced with narrative segments will be crucial to sustaining personalization through changing content and preferences while improving the actualized *AdBreakScore*$^*(t)$ over time.

The explainable topic-refinement methods in Chapter 4 and the constrained theme-reasoning framework in Chapter 5 and the cognitive-load estimation model in Chapter 6 exemplify this direction by integrating interpretability directly into multimodal inference.

2.5.4 Reinforcement Learning in the *Delivery* Layer: Promise and Pitfalls

Reinforcement learning (RL) is well-suited to adaptive scheduling, particularly for jointly managing ad timing, load, and selection under long-horizon objectives. Policy-gradient families [124,125,130,131] are attractive for continuous control (e.g., frequency curves), but deployment is constrained by sample inefficiency [132, 133], reward shaping [134], and the scarcity of realistic simulators [135].

A forward direction is to enrich rewards with *physiological proxies* (e.g., cognitive load, attention, affect) when available, to improve credit assignment and reduce data hunger [136]. This requires robust edge-side sensing/estimation and privacy-preserving aggregation. Practical questions include: how to fuse noisy biosignals with behavioural streams; how to compress and denoise at the edge; and how to guarantee latency budgets while keeping $P_{\text{Computational}}(t)$ low.

While reinforcement learning offers a promising mechanism for long-horizon optimisation in adaptive delivery systems, this thesis focuses primarily on the Data and Decision phases of the pipeline. Delivery-layer reinforcement learning is therefore discussed as a forward-looking extension rather than an implemented component of the present framework.

2.5.5 Coordination, Safety, and Governance

Learning-based delivery must be explicitly *safety-constrained*. Policies should adhere to guardrails avoid saturation, maintain content–ad separation, honor demographic protections, and respect narrative continuity so that $P_{\text{Ethical}}(t)$ is not an afterthought but a *first-class* objective. Safe RL mechanisms (constrained policy optimisation, shielded action sets, counterfactual evaluation) should be integrated into training and monitor-

ing.

Coordination across phases is equally important: Delivery should expose performance and constraint violations upstream; Decision should request specific *Data*-phase signals (e.g., updated affect spans) when uncertainty spikes; and *Data* should surface confidence and staleness indicators so downstream modules can degrade gracefully. This orchestration transforms independent blocks into a resilient, user-aligned control loop.

Summary of Priorities. Near-term priorities include: (1) story-level, reusable features that travel across phases; (2) executable decision templates aligning predictions with constrained actions; (3) closed-loop feedback to continually lift $AdBreakScore^*$; (4) safe RL with explicit ethical constraints; and (5) edge-compliant multimodal pipelines with privacy by default. Together, these directions move the field toward integrated, user-aligned ad placement systems that are contextually aware, computationally tractable, and ethically principled.

This chapter established the conceptual and technical landscape of AI-driven media adaptation and highlighted three core phases *Data*, *Decision*, and *Delivery* together with the *AdBreakScore* evaluation framework. The review identified a major gap in the **Data Phase**: current systems lack robust, scalable mechanisms for representing the semantic structure of video content in ways that are interpretable, multimodal, and aligned with industry taxonomies. Addressing this limitation is intended to support the first research objective defined in Chapter 1 and provides the foundation upon which the narrative (Ch. 5) and cognitive (Ch. 6) layers can operate.

The following chapter therefore introduces a multimodal content-taxonomy framework that operationalizes this semantic layer, enabling structured topic extraction and taxonomy alignment to support the integrated modular pipeline developed throughout the thesis.

Chapter 3

Multimodal Topic Modelling and Industry Taxonomy Mapping

Publication Notice.

This chapter is based on the following peer-reviewed open-access publication:

De Silva, W., Fernando, A., & Dong, F. (2026). Decoding content taxonomy in video for industry-specific contextual advertising. ***Multimedia Tools and Applications****. https://doi.org/10.1007/s11042-026-21243-4*

 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.

Chapter Abstract

The exponential growth of free video platforms is redefining content consumption and its advertising dynamics. Beyond this, the progressive irrelevance of cookies really underlines the necessity for finding a genuinely new and proper way of recording and analyzing user behaviour, with an implication for putting user privacy first by design, compliant with data protection and regulatory guidelines.

As a result, contextual advertising has become a suitable strategy for the delivery of relevant and personalized advertisements to users. In light of this evolving process, this chapter proposes a novel framework to facilitate the discovery of contextual information on video content within the industrial context. The proposed framework seamlessly integrates the visual and audio features that are extracted from video content to obtain a comprehensive comprehension of videos. This method optimises the presentation of ads within a context using a combination of multimodal analysis, industry taxonomy, and contextual advertising. The effectiveness of the framework is validated through experimental results using the YouTube-8M data set, demonstrating its potential to revolutionize contextual advertising by capturing the essence of video content and aligning it with the industry content taxonomy.

3.1 Introduction

The digital advertising ecosystem has experienced tremendous changes with the rise of free video streaming platforms that deliver hundreds of programs each day [137]. In addition, the transition to a cookie-free world presents a distinctive and urgent challenge, especially with respect to the access of behavioural data [138]. With cookies becoming obsolete, the ability to capture and analyse user behaviour is one of the major concerns for business players. These developments bring into focus the need for integrating data analysis requirements with user privacy needs and evolving data protection regulation needs. While much is laid on the standards required to make effective ad requests across varied systems, it is seen that the requirement of relevant and personalized ad targeting practices is more needed than ever in this environment [139].

In addition, there are also brand safety and user experience challenges. Such challenges have highlighted the importance of context analysis in identifying suitable advertisements [140–142]. Therefore, exploring contextual advertising is a proactive path to overcoming these issues and a technique to deliver successful targeted advertising of videos in industrial settings [139].

Traditional video advertising platforms often rely solely on video metadata [47, 143, 144]. However, assembling a relevant advertising selection using only metadata can

result in suboptimal results, since metadata are not always complete or comprehensive. The demand to implement contextual video advertising arises from the complexity of videos and TV categories, where understanding the broader picture goes beyond just metadata. Unlike other media, such as audio or text, video content benefits significantly from visual elements. These visual aspects of videos are often underused, which represents a critical oversight in contextual advertising platforms, as alignment with video content is critical.

In this chapter, the primary focus is the collection of diverse audio and visual multimodal data within video content, and then effective fusion [10] of these modalities to reflect a thorough contextual understanding of the video. This method enables a more complete understanding of contextual information. There has also been substantial advancement in different research domains regarding visual and auditory modalities in videos. In the visual domain, enormous gains have been made in contextual awareness in computer vision [145–148]. In audio, increasing emphasis is placed on Automated Audio Captioning (AAC), which involves the creation of coherent sentences in natural language [149, 150]. Audio signals and their constituents speech and ambient noise are similarly important, particularly in the context of robust Automatic Speech Recognition (ASR) [151] and transcript identification.

It is important to consider in the multimodal analysis of video media that the information capacity and the noise level can be different from one modality to another, as illustrated in Fig. 3.1. Each modality gives unique and complementary insight into the underlying content. With a realisation of the variation that videos bring in their auditive and visual content, the proposed method goes deeper into how these information-rich sources can be fully exploited to overcome the constraints imposed by missing or unreliable features. A deep understanding of information capacity helps to correctly interpret the message of a video, which is critical for accurate advertisement targeting.

This work addresses the problem of managing diverse multimodal data and handling varying information and noise. Video topic recovery is used to understand the underlying themes of the video and to go beyond simple video classification. The proposed approach uses topic modelling [152] for the extraction of latent semantic units and themes in a video. It classifies the underlying meaning by extracting subtopics,

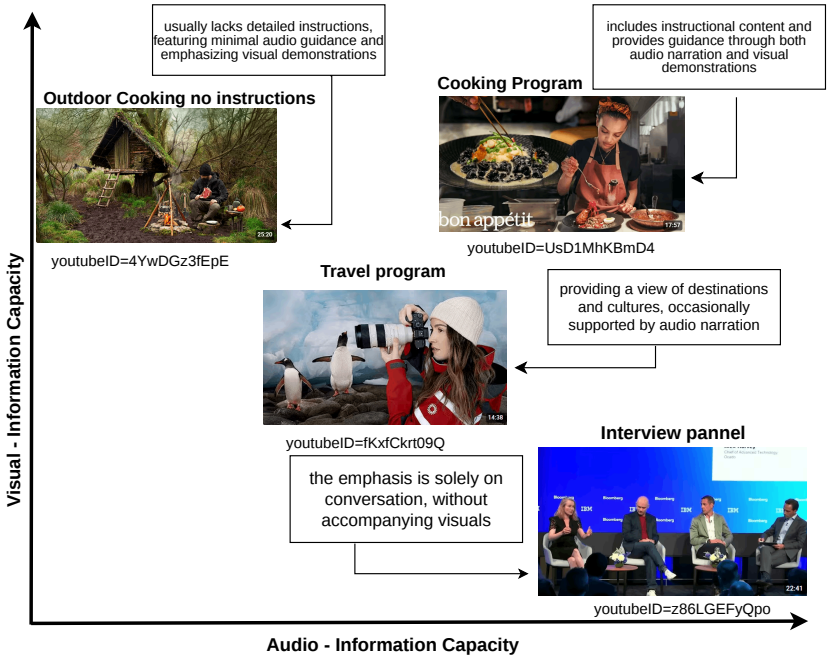


Figure 3.1: Understanding information capacity in diverse data modalities.

patterns, and relationships that may not be immediately recognisable. Topic recovery in a single modality is limited by the natural constraints of that modality. In contrast, multimodal fusion [10] gives access to richer information and implies more accurate and robust topic recovery. The joint use of audiovisual data also increases the richness of domain-specific features present in recovered topics.

Furthermore, the study is conducted in the context of recovering topic labelling using an industry-specific content taxonomy [153]. Since unsupervised topic modelling does not provide external labels, further refinement of the labelling process is necessary [154]. Although traditional methods such as pointwise mutual information [155] have been effectively used for cluster labelling, the proposed approach instead maps the identified video topics to an industry content taxonomy [153].

To validate the effectiveness of the proposed framework, experimental results are presented based on a YouTube-8M [156] sample dataset. These evaluations analyse each component of the framework, assess overall performance, and demonstrate its potential

to improve contextual advertising by aligning video content with standardized content taxonomies.

Chapter Contributions

The contributions of this chapter are consolidated into three main points:

1. **Integrated modular multimodal framework.** A practical pipeline is presented that fuses audio (ASR transcripts) and visual (Bag-of-Visual-Words) signals and adapts BERTopic for video, recovering semantically rich themes beyond metadata-only approaches.

2. **Standards-aligned taxonomy mapping.** A similarity-based mapping from discovered topics to industry taxonomies (e.g., IAB) using sentence embeddings with topic/term intensity weighting, yielding interpretable labels with confidence scores for ad targeting.

3. **Empirical validation at scale.** Using a YouTube-8M subset, high Top-5 success is demonstrated and late fusion is shown to outperform early fusion, evidencing scalability and deployment readiness.

3.2 Background and Theoretical Foundations

This section positions the proposed framework in the broader literature on contextual video advertising, multimodal learning, and topic modelling, and clarifies the research gaps that motivate the proposed approach.

3.2.1 Metadata-Based Approaches for Video Advertising

Earlier work on contextual video advertising proposed systems such as VideoSense [47]. The recommendation system of VideoSense comprised title, tags, queries, and local visual-aural features such as colour, movement, and audio. Another study [144] employed five classes of video metadata to extract informative advertisements, with the principal objective of reducing resource-hungry image and video processing. Metadata

attributes are weighted based on relevance and decisions taken according to a model of block importance. SalAd [143], in the same category, uses textual metadata to assemble a candidate set of informative advertisements; it further optimises the process using webpage saliency to decrease deliberate ignoring of advertisements.

A key drawback of these studies is that they do not conduct extensive video content analysis to extract contextual subtleties. The majority of existing methodologies depend critically on the source video's metadata or particular visual/audio features, and therefore fail to capture the complete semantic richness of video content. This drawback is especially problematic when metadata are incomplete, inconsistent, or poorly representative of the content.

3.2.2 Deep Learning Models for Video Content Analysis

Building on metadata-driven solutions, researchers later investigated deep learning methods for more complex video analysis. Some works examined advertisement insertion for videos associated with objects. For instance, object-level video advertising framework [63] associate detected objects with relevant advertisements, while deep CNN-based models such as YOLOv5 [157] have been used to detect objects in video frames to support context-aware advertisement insertion [80]. Another study [65] used deep neural networks for the recognition of the human body, poses, face, and clothes. While the study can prove useful for fashion-related advertisements, the authors state that the results may lack generality for unconstrained videos like movies and sport scenes.

Further progress was introduced in [79], where a CNN-based method performs content-targeted online video advertising by matching videos with advertisements using semantic descriptions. Even though deep CNNs have achieved strong results in many tasks, their application to video analysis remains constrained by the linear temporal progression inherent in videos. To capture longer-range temporal patterns, Recurrent Neural Networks (RNNs) have been considered [64]. Nonetheless, such techniques remain limited to relatively small datasets, and combining CNN and RNN models for end-to-end training on large-scale datasets, such as YouTube-8M, remains challenging.

3.2.3 Multimodal Fusion Techniques

As a solution to the deficiencies of unimodal techniques, multimodal fusion techniques for video commercials have been explored. Song et al. [88] introduced a framework to extract descriptive features from various modalities, including location, emotion, objects, audio and topics. These representations are fused into a single coherent representation to simplify scene matching. However, this approach lacks a thorough examination of the contribution of each modality to the most resilient semantic representation for contextual understanding.

Most research focuses on contextual understanding of local scenes in videos to derive ad-insertion points within specific contexts [70, 71]. While relevant in domain-specific or niche settings, such approaches become difficult to scale to a broader spectrum of video content. This limitation highlights the need for careful analysis of both hidden and global themes in videos enabling advertisers and downstream applications to make informed decisions and deliver appropriate ads with more holistic content understanding.

3.2.4 Topic Modelling in Video Context

Video thematic content identification has become a key element of video contextual advertising [31]. Classical topic modelling methods have been used for identifying hidden themes from video content, but they suffer from several critical issues in handling multimodal content [158]. While topic modelling of text is well established, applying these techniques to other modalities such as visuals and audio requires custom adaptations [159].

Merging topic modelling with industry taxonomies is a particularly underexplored area of video advertising research. Discovered topics must be framed such that they can be interpreted by downstream applications and advertisers and can coexist with popular industry taxonomies [153]. To the best of current knowledge, the robust mapping of video content to standardized taxonomies has not been fully explored in prior work, representing a significant gap.

3.2.5 Recent Advances in Video Understanding

Recent advances in self-supervised learning and Transformer architectures have opened promising avenues for video understanding [160]. Self-supervised learning methods have achieved state-of-the-art performance in learning rich video representations with minimal or no labelled data. Contrastive learning methods such as MoCo [161], SimCLR [162], and CLIP [163], originally developed for image-text alignment but influential in shaping multimodal contrastive frameworks, have been adapted to video and enabled the learning of temporal dynamics and semantic relationships that in some cases surpass supervised methods. Masked autoencoding approaches (e.g., MotionMAE [164], MGMAE [165], ViC-MAE [166]) further enhance this by jointly modelling spatial appearance and temporal motion in a self-supervised fashion.

Transformer architectures have likewise reshaped video understanding by modelling long-range spatio-temporal dependencies. Models such as VideoBERT [167] demonstrate the superiority of joint video-text representation learning, while TimeSformer [168] and ViViT [169] achieve high performance in efficient spatio-temporal modelling. These developments hold immense promise for contextual advertising systems by enabling more complete understanding of video information. Beyond these trends, transformer-based salient object detection (SOD) explicitly models global–local relations with compensative fusion; the Collaborative Compensative Transformer Network (CCTNet) alternates collaborative relation modelling with compensative fusion to improve saliency prediction [170]. Complementary to the privacy-by-design motivation, FedMDD calibrates global decision boundaries for federated long-tailed settings via multi-deliberation post-hoc calibration and local–global feature-contrast constraints; although orthogonal to the pipeline, these ideas are pertinent for future privacy-preserving, cross-partner deployments and for handling skewed taxonomy-label frequencies in decentralised data [171].

Despite these advances, only a few methods have been applied to contextual advertising; most are not specifically optimised for long-range video understanding or integration with ad-tech pipelines [160]. The integration of such models with advertising systems remains a major research avenue.

3.2.6 Summary of Research Gaps

Previous research on contextual video advertising shows several shortcomings. Methods that rely on metadata frequently fail to capture the complete semantic depth of video content because the available context is often incomplete or inaccurate. Deep learning-based methods are promising, yet they are constrained by issues of scalability and limited datasets. Multimodal fusion techniques have been proposed but lack robust tools for estimating the contribution of disparate modalities, leading to brittleness in challenging settings. Topic modelling has been widely used in text-based domains but not comprehensively adapted to multimodal video data. Finally, self-supervised learning and Transformers have not yet been widely applied to advertising applications due to prohibitive computational costs and limited optimisation for long-range video understanding.

These gaps suggest the need for a paradigm that goes beyond unimodal or metadata-based techniques, makes effective use of multimodal signals, and harmonizes extracted knowledge with the practices and standards of industrial communities.

3.2.7 Technical Challenges

Translating these research gaps into a practical system raises several technical challenges to which the proposed framework proposes solutions:

1. **Incomplete and unreliable metadata.** Existing systems rely heavily on titles, tags, and descriptions, which are often sparse, inconsistent, or misleading.

2. **Multimodal complexity.** Video combines a visual and an audio stream, each carrying semantic information and noise. Extracting, aligning, and synthesizing these sources into a single representation is non-trivial.

3. **Adapting topic modelling to video.** Classical topic models are text-based. Extending them to video involves projecting visual features into textual-like features and integrating them with audio transcripts while retaining semantic richness.

4. **Mapping to standardized taxonomies.** Latent topics from unsupervised models cannot be directly interpreted for advertising applications. Robust mapping onto traditional taxonomies (e.g., IAB) is necessary for deployment but technically challenging.

5. **Scalability and domain adaptation.** Efficient systems require the ability to handle large-scale data sets such as YouTube-8M without losing domain-specific subtleties. The challenge is to achieve a balance between computational efficiency and correctness with respect to the context.

These challenges form the foundation for the methodology detailed in the remainder of this chapter, which presents the multimodal topic modelling framework and its approach to taxonomy mapping.

3.3 Problem Definition

To address the technical challenges outlined in Section 3.2.7, The problem is formalised, and the objectives of the multimodal topic modelling and taxonomy-mapping framework are defined.

A video $V = F, A$ is considered, where F denotes visual frames and A denotes the accompanying audio track. The objective is to infer a set of industry-standard content taxonomy labels $\mathcal{Y}$ (e.g., IAB) and associated confidence scores $s \in [0,1]^{|\mathcal{Y}|}$ that reflect the primary and secondary themes present in V and are suitable for downstream ad targeting. Concretely, Visual features are extracted and transformed into a Bag-of-Visual-Words representation x_v, and the audio is transcribed into text x_a via ASR. A multimodal encoder produces topic representations T with weights w from (x_v, x_a); a taxonomy-mapping function then aligns T to $\mathcal{Y}$ using embedding-based similarity with topic/term intensity weighting, yielding ranked labels and scores:

$$(x_v, x_a) \xrightarrow{\text{encoder}} T, w \quad \text{and} \quad (T, w) \xrightarrow{\text{taxonomy map}} \{(y, s_y)\}_{y \in \mathcal{Y}}.$$

This two-stage decomposition introduces an explicit intermediate semantic repre-

sentation (T, w), rather than learning a direct mapping from (x_v, x_a) to (s, y). The motivation and implications of this design choice are discussed in Section 3.4.

This definition of video value explicitly recognizes both properties of video data (heterogeneity of information and noise in modalities, as well as varying resolution and audio qualities) and the necessity of model predictions to be meaningful according to standardized categorizations. Evaluation focuses on accuracy at top-k, coverage of relevant labels, and ranking quality, with robustness to modality imbalance and scalability to large corpora discussed in later sections.

3.4 Framework Overview

Figure 3.2 presents the overall architecture of the proposed multimodal topic modelling framework. The pipeline begins with feature extraction from both visual and audio modalities, followed by topic modelling, modality fusion, and mapping of the resulting topics onto industry-standard content taxonomies.

Video features are extracted using Amazon Rekognition [172], while audio transcripts are generated using OpenAI Whisper [173]. Two fusion strategies [174] early fusion and late fusion are implemented and evaluated. In the late fusion approach, topic modelling is performed independently on each modality to produce separate topic distributions, which are subsequently combined. In the early fusion approach, visual and audio features are concatenated at the input level before being passed through a shared BERTopic pipeline.

The choice between early and late fusion involves important trade-offs across four key dimensions: performance, robustness to modality-specific noise, computational efficiency, and stability. Early fusion concatenates features from both modalities prior to topic modelling, enabling the model to capture fine-grained cross-modal interactions. This can lead to stronger performance when the modalities are well-aligned and relatively noise-free. However, it is significantly more vulnerable to modality-specific noise and imbalance; errors or missing information in one stream (e.g., inaccurate visual labels or poor ASR quality) propagate directly into the joint representation, reducing overall

robustness and temporal stability.

In contrast, late fusion processes each modality independently before combining the resulting topic distributions (e.g., through weighted averaging or rank-based fusion). This strategy offers superior robustness to modality-specific noise, greater stability across diverse and heterogeneous video content, and improved interpretability, since the contribution of each modality can be examined separately. It is also generally more computationally efficient during inference, as the two modality-specific models can be executed in parallel and failures in one modality can be isolated. Its primary limitation lies in a potentially reduced capacity to model complex, low-level cross-modal dependencies.

Empirical evaluation in this work demonstrates that late fusion consistently outperforms early fusion across Top-N accuracy (Table 3.3) and taxonomy coverage metrics (Table 3.4), with statistical significance confirmed through paired Wilcoxon signed-rank tests (Table 3.5). These findings suggest that, for real-world user-generated videos characterised by varying noise levels and modality imbalance, the robustness and stability advantages of late fusion outweigh the potential benefits of early cross-modal modelling.

Following modality fusion, the derived topics are embedded in a shared vector space together with industry taxonomy categories (e.g., IAB Content Taxonomy). This framework establishes a standardised method for determining advertising categories by identifying the most semantically proximate topic terms to each taxonomy standard. This standards-based approach facilitates effective communication and seamless integration with integrated systems within the digital advertising ecosystem during the exchange of ad requests. This two-stage, topic-mediated architecture mapping multimodal inputs (x_v, x_a) first to topics (T, w) and subsequently to taxonomy labels (s, y) provides several advantages over direct end-to-end mapping. It enhances interpretability by maintaining an explicit semantic intermediate layer, increases modularity (allowing independent improvement of topic modelling and taxonomy mapping components), and improves flexibility in response to updates in taxonomy standards. Furthermore, the topic-based representation helps mitigate the impact of modality-specific noise, such as visual ambiguities or transcription errors.

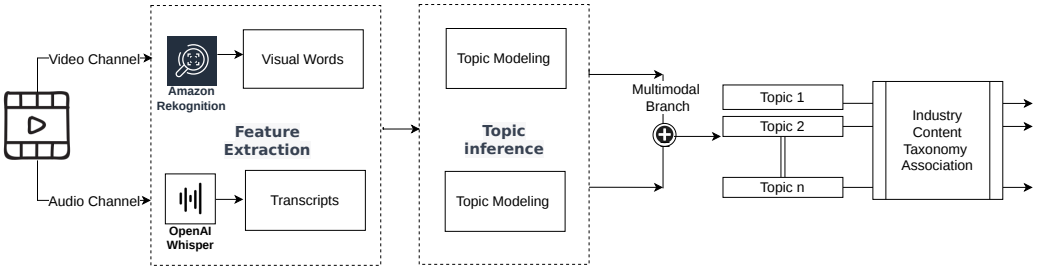


Figure 3.2: Flowchart of the proposed method (system overview). Visual and audio features are extracted, topics are inferred, modalities are fused, and topics are mapped to standardized taxonomies to produce ranked labels.

A notable characteristic of the visual processing pipeline is the transformation of detected concepts into a Bag-of-Visual-Words (BoVW) representation [175, 176]. Although topic modelling techniques were originally developed for textual data [177], this approach adapts the Bag-of-Words (BoW) model into a Bag-of-Features (BoF) framework [176] tailored for visual information. The BoF technique has been widely used in computer vision and machine learning competitions [178, 179]. Amazon Rekognition [172] is used to extract visual concepts and convert them into this standardised BoVW format, enabling seamless integration with the text-oriented BERTopic pipeline.

3.5 Multimodal Feature Extraction

This section details the extraction of visual and audio features that serve as inputs to the topic modelling component.

3.5.1 Visual Features: Bag-of-Visual-Words

For the visual modality, scene detection and keyframe extraction are first performed. A pre-trained object detection model (Amazon Rekognition [172]) is then applied to the extracted keyframes, with the detected object labels treated as visual "tokens". These labels are aggregated per video to construct a Bag-of-Visual-Words (BoVW) representation [175]. This approach effectively bridges the gap between low-level computer vision features and text-based topic modelling techniques.

This BoVW formulation was deliberately chosen to maintain compatibility with

the BERTopic pipeline. Since BERTopic operates on discrete tokenised representations treating each document as a distribution over words representing visual content as a "bag of visual words" enables seamless application of the same topic modelling framework across both textual and visual modalities.

3.5.2 Audio Features: ASR Transcripts

For the audio modality, the audio track A is transcribed using Whisper [173]. The resulting transcripts are preprocessed (lowercasing, removal of non-linguistic tokens, light punctuation filtering) to form a textual corpus that is naturally compatible with topic modelling. Whisper's robustness to noise and multilingual content is particularly advantageous for realistic video datasets.

Whisper has been chosen due to its excellent generalization capability in various acoustic scenarios such as differences in accent, noise, and casual speech, which are usually present in videos available online [173]. In contrast to traditional ASR models, Whisper has learned from vast and diverse data sources [173] and hence is capable of providing semantically meaningful transcripts in adverse situations [173]. This is especially relevant for multimodal video understanding tasks, where errors in transcripts could lead to flawed semantic modelling.

3.6 Topic Modelling with BERTopic

A topic model is used to capture the essence of video content. This requires a thorough understanding of contextual facets within the video and facilitates identification of numerous advertising opportunities embedded in the content. Topic models [180, 181] are effective in automatically identifying topics from features. They commonly use a BoW technique that quantifies the presence of words by studying occurrences and correlations [182, 183].

On the basis of the above, the primary topic modelling technique used in this study is the application of BERTopic [184] to the dataset. BERTopic incorporates neural components [185] and leverages pre-trained models such as BERT [186]. The decision to

use BERTopic is justified due to the following benefits that the tool will add in comparison to conventional topic modelling approaches such as LDA [187]: the application of semantic term representations, the reduction of vocabulary discrepancy [184], improved interpretability [188] for the topics, multilingual support, automatic determination of the number of topics, and greatly reduced hyperparameter tuning [188]. BERTopic also introduces the notion of a noise topic, which prevents unrelated documents from being forced into other topics, improving topic quality and reducing subjectivity [189].

The BERTopic pipeline aims to make document topics interpretable. First, the documents in this case, audio transcripts and visual features are passed through an encoder that generates document embeddings using pre-trained transformer-based language models. Next, dimensionality reduction is performed using Uniform Manifold Approximation and Projection (UMAP) [190], and clustering is performed using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [190] to generate clusters of semantically similar documents.

The choice of dimensionality reduction and clustering methods is critical for effective topic modelling. In comparison with Principal Component Analysis (PCA), UMAP is more suitable for topic modelling because it is able to preserve non-linear relationships among the high-dimensional semantic embeddings [184, 190]. Clustering algorithms, such as K-Means and Gaussian Mixture Models, need to be defined by a number of clusters in advance; therefore, they are not as flexible as HDBSCAN. This algorithm finds clusters of different densities without the specification of the number of clusters and is less sensitive to noise [191, 192].

Finally, topic representations are constructed using a class-based term frequency–inverse document frequency measure. In practice, this involves computing the class-based Term Frequency–Inverse Document Frequency (c-TF–IDF) [184]:

$$W_{t,c} = \mathrm{tf}_{t,c} \cdot \log\left(1 + \frac{A}{\mathrm{tf}_t}\right) \tag{3.1}$$

where:

- $\mathrm{tf}_{t,c}$ is the term frequency of term t in class/topic c (how often word t appears in documents belonging to topic c);

- tf_t is the total frequency of term t across all classes;
- A is the total number of all words in all classes (corpus length);
- $W_{t,c}$ is the importance (or weight) of term t in class/topic c.

After documents are embedded and clustered, c-TF–IDF (3.1) is used to create a class-based representation of topics by transforming the textual content of each topic cluster into a meaningful topic vector.

A key design choice at this stage is the adoption of c-TF–IDF [184] for topic representation, as opposed to alternatives such as Word2Vec [193] or KeyBERT [194]. While Word2Vec generates dense semantic embeddings, it does not naturally produce interpretable, cluster-level topic descriptors. Similarly, KeyBERT is effective for document-level keyword extraction but does not capture the overall topical structure of the corpus. In contrast, c-TF–IDF produces sparse, class-level representations that emphasise terms most distinctive to each cluster. This makes it particularly well-suited for generating interpretable topics that align effectively with the downstream taxonomy mapping task.

3.7 Topic Model Training

A non-traditional approach is employed to address difficulties caused by constrained vocabulary and limited context in BoVW representations [195]. Instead of training directly on video object labels, the proposed approach uses a large textual corpus in conjunction with pre-trained BERT-based models.

This corpus, rich in descriptive content, forms the foundation of the methodology. Through fine-tuning a BERTopic model with this textual dataset, the semantic understanding of language is improved. The fine-tuned model is then used to process object labels obtained from video frames in terms of semantic similarity to form clusters of video topics. This alternative method demonstrates greater contextual awareness and linguistic sophistication compared to traditional approaches [196].

The proposed approach successfully copes with the limitations of direct training on object labels, demonstrating feasibility and suggesting a path toward more nuanced and context-rich video topic analysis.

3.8 Multimodal Topic Representation and Taxonomy Mapping

While the topic modelling stage operates in an unsupervised manner to discover latent semantic structures, the subsequent taxonomy mapping stage bridges these emergent topics to industry-standard ontologies. This is achieved through embedding-based similarity rather than supervised classification, thereby preserving the flexibility of unsupervised discovery while enabling practical application.

Multimodal fusion is performed by integrating both topic probabilities (P_{tw}) and topic–term probabilities (P_{ttw}) within the taxonomy mapping process. As illustrated in Fig. 3.3, the Adjusted Similarity Score (ASS) is computed under both early- and late-fusion strategies. This intensity-weighted approach, inspired by the concept of topic intensity [197], assigns greater importance to high-probability topics and their representative terms.

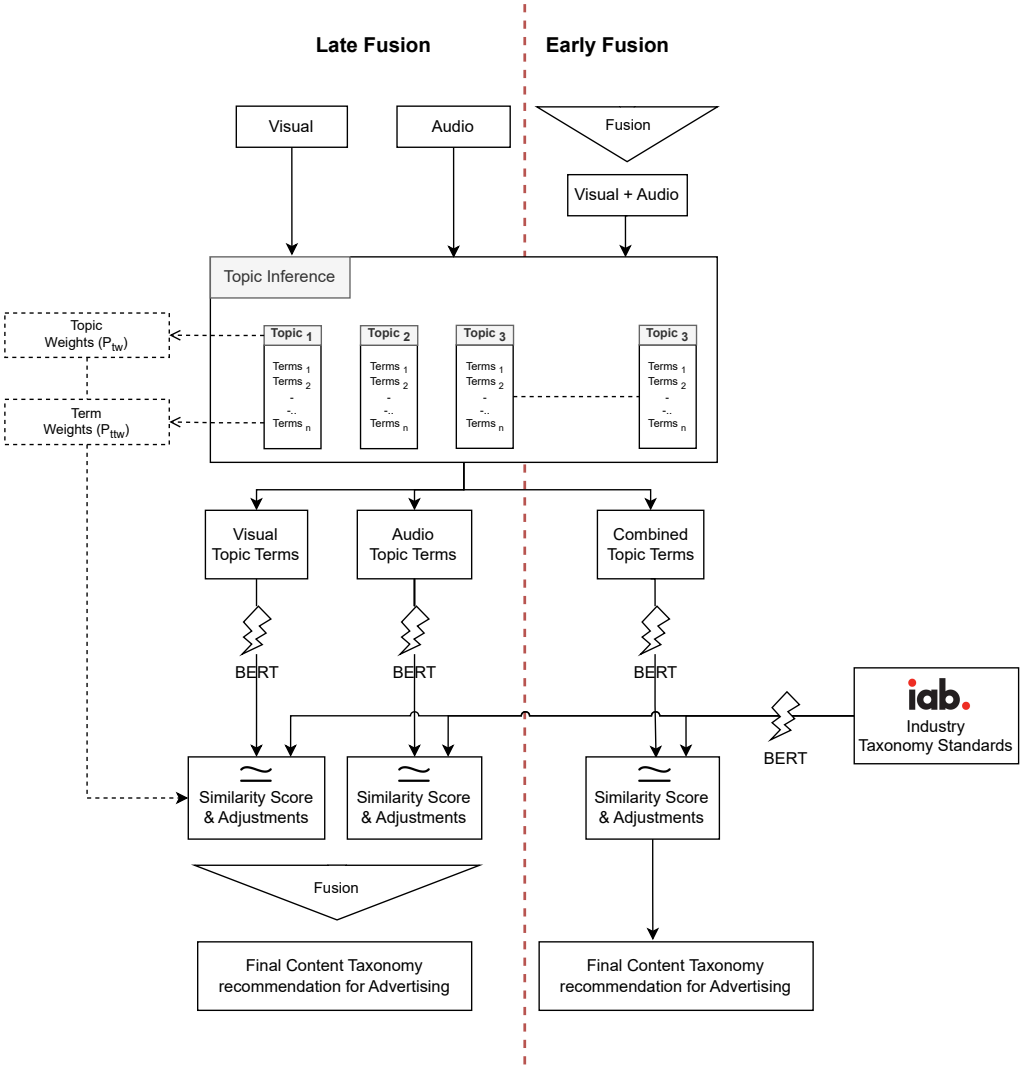


Figure 3.3: Detail of the multimodal fusion and intensity-weighted taxonomy mapping (corresponding to the central block in Fig. 3.2). It illustrates early/late fusion and the Adjusted Similarity Score (ASS) used for taxonomy mapping.

Pre-trained sentence embeddings from the BERT Sentence Transformer [186] are used to encode contextual taxonomy terms (EMB_C), visual topic terms (EMB_V), and audio topic terms (EMB_A) into a shared 384-dimensional semantic space. Fig. 3.4 provides a detailed view of the similarity computation sub-module.

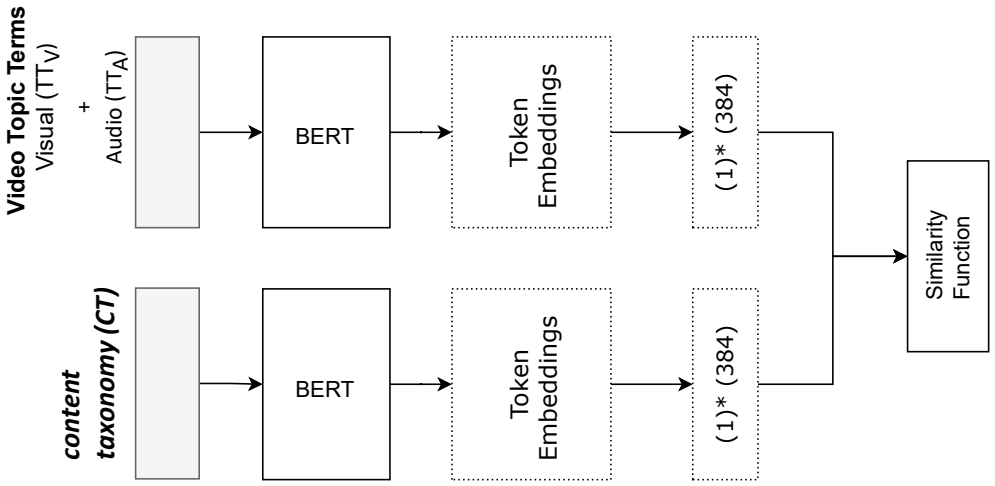


Figure 3.4: Zoom-in on the embedding-based similarity computation between topic terms and taxonomy terms (the scoring sub-module inside Fig. 3.3); it implements the cosine similarities in Eqs. (3.2)–(3.3).

Semantic similarity between the discovered topics and the industry taxonomy is quantified using cosine similarity. Specifically, two similarity scores are computed:

- $SS_{C,V}$: the semantic similarity between the **contextual taxonomy terms** (EMB_C) and the **visual topic terms** (EMB_V);
- $SS_{C,A}$: the semantic similarity between the **contextual taxonomy terms** (EMB_C) and the **audio topic terms** (EMB_A).

These scores measure how well the emergent visual and audio topics align semantically with the predefined industry taxonomy categories. Cosine similarity is chosen as the primary measure because it effectively captures directional semantic relationships in high-dimensional embedding spaces, focusing on the angle between vectors rather than their magnitude.

The similarity scores are formally defined as:

$$SS_{C,V} = \frac{\text{EMB}_C \cdot \text{EMB}_V}{\|\text{EMB}_C\| \cdot \|\text{EMB}_V\|} \tag{3.2}$$

$$SS_{C,A} = \frac{\text{EMB}_C \cdot \text{EMB}_A}{\|\text{EMB}_C\| \cdot \|\text{EMB}_A\|} \tag{3.3}$$

The final Adjusted Similarity Score incorporates topic and term probabilities to reflect relevance:

$$\text{ASS} = \text{SS} \times P_{tw} \times P_{ttw} \tag{3.4}$$

This design naturally accommodates the hierarchical and non-mutually exclusive nature of industry taxonomies (e.g., IAB), supporting many-to-many mappings. To handle ambiguity between similar taxonomy categories, the framework employs topic-term weighting combined with similarity thresholding. Overall, the proposed approach provides a flexible, interpretable mechanism for hierarchical multi-label taxonomy alignment in real-world multimedia content.

Statistical significance of the multimodal fusion results was assessed using paired Wilcoxon signed-rank tests ($\alpha = 0.05$), addressing potential dataset bias through rigorous hypothesis testing.

3.9 Algorithmic Realisation

To realise the multimodal fusion and taxonomy mapping framework described above, the overall process is summarised in Algorithm 1. The pseudocode encapsulates the end-to-end processing from feature extraction and topic modelling down to the final taxonomy ranking, applying both early and late fusion schemes with intensity-weighted similarity scoring. The algorithm acts as an implementation roadmap for the experimental verification reported later in this chapter.

Algorithm 1 Contextual Video Taxonomy Mapping via Multimodal Topic modelling

Require: Video $V = \{F, A\}$ ▷ F: Frames, A: Audio
Require: Industry Taxonomy Terms T_{taxonomy}
Require: FusionType $\in$ {'Late', 'Early'}
Ensure: Ranked Taxonomy Labels with Confidence Scores
$BoVW \leftarrow$ ExtractVisualFeatures(F) ▷ e.g., Amazon Rekognition
$ASR\_Text \leftarrow$ TranscribeAudio(A) ▷ e.g., OpenAI Whisper
if $FusionType ==$ 'Early' **then**
    $Fused\_Input \leftarrow$ Concatenate($BoVW, ASR\_Text$)
    $Topic\_List \leftarrow$ BERTopic($Fused\_Input$)
else
    $Visual\_Topics \leftarrow$ BERTopic($BoVW$)
    $Audio\_Topics \leftarrow$ BERTopic(ASR_Text)
    $Topic\_List \leftarrow Visual\_Topics \cup Audio\_Topics$
end if
$T\_terms \leftarrow$ ExtractTopTerms($Topic\_List$)
$EMB\_Topics \leftarrow$ BERT_Embed(T_terms)
$EMB\_Taxonomy \leftarrow$ BERT_Embed(T_{taxonomy})
for each taxonomy term t in T_{taxonomy} **do**
    for each topic term w in T_terms **do**
        $SSC[t][w] \leftarrow$ CosineSimilarity($EMB\_Taxonomy[t], EMB\_Topics[w]$)
        $ASS[t][w] \leftarrow SSC[t][w] \times P_{tw} \times P_{ttw}$
    end for
    $ASS\_Total[t] \leftarrow$ Aggregate($ASS[t][:]$) ▷ e.g., weighted sum
end for
$Ranked\_Labels \leftarrow$ SortDescending(ASS_Total)
return Top-N $Ranked\_Labels$ with Scores

3.10 Implementation Details

To facilitate reproducibility and clarify engineering choices, the end-to-end setup is summarised in Table 3.1. The table is organised by pipeline stage environment, preprocessing, feature extraction for visual/audio streams, BERTopic configuration, fusion procedure, taxonomy mapping, training data, and evaluation protocol so that default settings are visible at a glance. Unless otherwise stated in the ablation studies, all experiments use these defaults. A shared sentence-transformer encoder (384-d) is used for both topic modelling and taxonomy mapping to ensure a common embedding space; cosine similarity scores are adjusted by topic/term intensities as defined in Eq. (3.4).

Table 3.1: Implementation summary of the proposed framework.

Component	**Details**
Environment	Amazon SageMaker `ml.t3.xlarge` (CPU-only; 4 vCPU, 16 GiB RAM), Amazon Linux 2; Python 3.10. Key libs: `sentence-transformers`, `umap-learn`, `hdbscan`, `bertopic`, `numpy`/`scipy`, `whisper`, `boto3`.
Data preprocessing	• **Video:** decode and frame export via `ffmpeg`; scene segmentation with PySceneDetect (*content* detector); extract one keyframe per detected scene; resize keyframes to shorter side 256 px (preserve aspect ratio). • **Audio:** resample to 16 kHz mono for ASR; enable auto language detection when applicable.
Visual features (BoVW)	• Run AWS Rekognition only on keyframes; retain top-K=10 labels with confidence ≥ 0.55. • Normalise labels (lowercase, lemmatise, synonym map) and aggregate per-video to form a Bag-of-Visual-Words.
Audio features (ASR)	Whisper transcripts; normalise text (lowercase, remove non-linguistic tokens; light punctuation/number filtering).
Topic modelling (BERTopic)	Shared sentence-transformer encoder (384-d) for embeddings; UMAP for dimensionality reduction; HDBSCAN for clustering; c-TF–IDF topic representations; noise topic retained. Key hyperparameters: • **UMAP:** $n_{\text{neighbors}}$=15, min_dist=0.0, $n_{\text{components}}$=5. • **HDBSCAN:** $min\_cluster\_size$=15, $min\_samples$=10.
Training data for topic model	Domain adaptation of the pre-trained BERTopic model on the Food.com/Kaggle corpus (∼180k recipes & interactions) prior to evaluation on the YouTube-8M food subset.
Evaluation protocol	Metrics: Top-k success ($k \in \{1, 2, 3, 5, 10\}$) and Coverage. Statistics: paired Wilcoxon signed-rank test with α=0.05 comparing late vs. early fusion.
Data management & storage	All artefacts (keyframes, Rekognition JSON, ASR transcripts, manifests) stored in Amazon S3. A manifest (CSV) maps video ID → scene IDs, keyframe timestamps, S3 URIs, and derived outputs for reproducibility.

3.11 Experimental Setup and Datasets

Two main datasets are used to evaluate the model. The first is a rich textual corpus employed to train the topic model; the second comprises real videos for evaluating the topic analysis methodology. The model fine-tuned on the rich corpus is applied to real videos to identify and analyse topics.

3.11.1 Rich Corpus Training

The textual knowledge base on which the pre-trained BERTopic model has been trained is domain-specific regarding food and includes a vast array of articles that are capable of approximating various kinds of semantic relations and contexts. The primary task here is to endow the model with complete semantic, contextual, and linguistic knowledge. For that purpose, the Kaggle dataset [198], consisting of over 180K interactions related to cooking activities on Food.com, is utilised. The adaptability of the pre-trained BERTopic model on the aforementioned task has resulted in 4,349 topics.

3.11.2 YouTube-8M Videos for Evaluation

To assess the model, it is necessary to use a real-world video dataset with ground truth. Among various video datasets, the YouTube-8M video dataset is very appropriate, and that is because it provides video as well as the corresponding ground truth information [156]. The major issue with this video dataset is that it includes video candidates from various categories, and as such, it is not easy to determine the exact topics. To overcome such issues, an appropriate filter is applied that focuses on the food niche. The filtered YouTube-8M video dataset includes a wide variety of food videos that include both visual and audio details. This setting enables evaluation of cases where:

- audio transcripts are strong but visuals are noisy;
- visual information is strong while audio is noisy;
- both modalities contribute equally and coherently.

As indicated in the methodology, these videos are processed by first extracting visual features in textual format for the visual topic model and then audio transcripts for the audio topic model. The extracted visual words serve as input for the trained BERTopic model to identify topics within the visuals, while audio transcripts serve as input to identify topics within the audio, subsequently mapped to standard content taxonomies.

3.12 Results

3.12.1 Case Study: YouTube-8M Video Example

An example of the industry taxonomies representation for a YouTube-8M video under the proposed model is given in Table 3.2, which shows the topic probabilities used to generate the final contextual taxonomy predictions. The comparison between the ground-truth taxonomy and the predictions in Fig. 3.5 verifies that the model correctly maps the first two predicted taxonomies, covering four associated ground truth taxonomies. Even the missed category is close under human evaluation. The approach also discovers hidden themes not tracked in the ground truth, such as content related to children. The line graph further shows how each modality contributes to the final outcome.

3.12.2 Performance Evaluation Metrics

Evaluation is performed on a separate held-out dataset rather than an automatic process over the entire YouTube-8M dataset. Fifty random videos were manually selected from the YouTube-8M food category, aligning with the domain used for training.

Performance is evaluated based on two key measures:

- evaluation of the Top-N predicted video taxonomies for sampled videos;
- comparison of the proximity between predicted and ground-truth taxonomies.

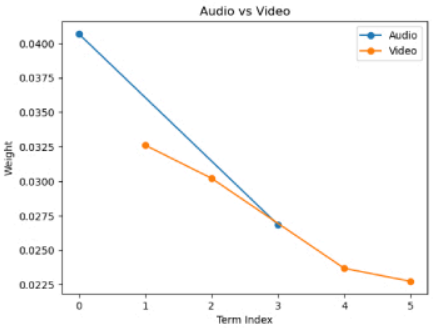
These measures emphasise the match between predicted and standard industry taxonomy mappings to provide a subtle and credible evaluation in the target domain.

Table 3.2: Example analysis data summary.

Example Analysis	
8M example video	
URL	`https://www.youtube.com/watch?v=AMe56hxTMHw`
Ground truth	'Food', 'Recipe', 'Cake', 'Cookie', 'Cupcake', 'Cake pop', 'Food', 'Recipe', 'Cake', 'Cookie', 'Cupcake', 'Cake pop'
Audio topics	[330, 1483, 899, 567, 79]
Audio topic weights	[0.68541145, 0.60544884, 0.59641486, 0.5956478, 0.59390974]
Video topics	[2701, 3929, 2752, 960, 2343]
Video topic weights	[0.4095165, 0.40553853, 0.40358847, 0.3900413, 0.35338795]
Video+audio topics	[330, 3239, 899, 567, 1367]
Video+audio topic weights	[0.60842794, 0.5249552, 0.5149646, 0.49994525, 0.49978724]

Prediction Taxonomies

	source	term	taxonomy_keyword	cosine_similarity	topic_weight	term_weight	final_weight
0	audio	treats	Desserts and Baking Food & Drink Desserts and ...	0.509367	0.596415	0.133958	0.040696
1	visual	foods	Food & Drink Food & Drink	0.620841	0.353388	0.148661	0.032616
2	visual	ideas	Model Toys Hobbies & Interests Model Toys	0.502574	0.403588	0.148953	0.030213
3	audio	candy	Gifts and Greetings Cards Shopping Gifts and G...	0.397023	0.685411	0.098664	0.026849
4	visual	everyday	Wellness Healthy Living Wellness	0.400660	0.390041	0.151518	0.023678
5	visual	rosengarten	Apprenticeships Careers Apprenticeships	0.295004	0.353388	0.218095	0.022737



Ground Truth Taxonomies

term	taxonomy_keyword	cosine_similarity
Food	Food & Drink Food & Drink	0.617123
Cake	Desserts and Baking Food & Drink Desserts and ...	0.534501
Cake pop	Desserts and Baking Food & Drink Desserts and ...	0.525742
Recipe	Cooking Food & Drink Cooking	0.502049
Cupcake	Desserts and Baking Food & Drink Desserts and ...	0.500328
Cookie	Desserts and Baking Food & Drink Desserts and ...	0.417892

Figure 3.5: Ground truth vs. predicted taxonomy comparison in a sample video, highlighting visual and audio modality contributions.

Evaluation of Top-N Predicted Taxonomies

This evaluation provides a comprehensive analysis of model performance at different confidence levels, including Top-1, Top-2, etc. The metric is derived from the first rank at which a correct prediction appears within the top n predicted labels and yields insight into reliability at various levels.

Let $P = \{\mathrm{L}_1, \mathrm{L}_2, \ldots, \mathrm{L}_m\}$ represent the set of ground-truth taxonomy labels for a video, and let $T_i = \{\mathrm{P}_1, \mathrm{P}_2, \ldots, \mathrm{P}_n\}$ denote the set of predicted labels in the top n positions. Define an indicator function δ_k for each top k as:

$$\delta_k = \begin{cases} 1 & \text{if } P_k \cap T_i \neq \emptyset \text{ (i.e., there is an intersection)} \\ 0 & \text{otherwise.} \end{cases}$$

The overall evaluation score is then:

$$\text{Overall Evaluation} = \sum_{k=1}^{n} \delta_k.$$

Given that YouTube-8M videos often include multiple topic labels (usually not exceeding 10), The analysis focuses on Top-N values of 1, 2, 3, 5, and 10. Experiments are conducted on 50 selected videos, and the results are presented in Table 3.3.

Table 3.3: Success rate for first correct predictions at different Top-N confidence levels.

Fusion	Top 1	Top 2	Top 3	Top 5	Top 10
Late	52.8%	73.6%	81.1%	90.7%	92.7%
Early	46.2%	69.8%	77.4%	81.1%	90.6%

The results suggest that the late-fusion model consistently outperforms early fusion, achieving 90.7% and 92.7% success for Top-5 and Top-10 predictions, respectively, across the 50 test videos.

Coverage Metric

Although Jaccard similarity is commonly employed to assess similarity between sets, it is less effective in the dataset due to the presence of many videos within the same niche domain. After taxonomy association, there are relatively few unique topic labels (as many converge to similar embeddings). Consequently, a few mismatches can disproportionately affect the evaluation, leading to an inaccurate representation of performance.

To address this, the *Coverage metric* is introduced, which calculates the ratio of predicted taxonomies to ground-truth-associated taxonomies. Let N_{V} be the total number of videos, N_{PT} the total number of predicted taxonomies, and N_{GT} the total number of taxonomies in the ground truth. The Average Coverage is:

$$\text{Average Coverage} = \frac{1}{N_{\text{V}}} \sum_{i=1}^{N_{\text{V}}} \left(\frac{N_{\text{PT}_i}}{N_{\text{GT}_i}} \right) \times 100.$$

This metric expresses the percentage of predicted taxonomies relative to the ground-truth-associated taxonomies and offers insight into the framework's ability to capture

relevant taxonomies. Results for early and late fusion are summarised in Table 3.4.

Table 3.4: Taxonomy prediction coverage (%).

Fusion	Coverage
Late	41.2%
Early	36.5%

The BERTopic model is fine-tuned on a domain-specific food corpus, hence topic inference leans towards food-oriented topics. While transfer learning and semantic generalisation may capture some other thematic elements, coverage remains constrained by the domain of the training corpus, contributing to relatively modest coverage scores (41.2% and 36.5%). Extending the training corpus to broader industry domains is a natural path to improving coverage.

Statistical Validation of Results

To rigorously assess the difference between late and early fusion, a paired Wilcoxon signed-rank test (the non-parametric analogue of the t-test, suitable for $n = 50$) is employed to evaluate two metrics:

- Top-5 success rate (primary metric);
- coverage metric.

The hypotheses are:

H_0 : Late and early fusion performance is not significantly different.

H_1 : Late fusion significantly outperforms early fusion.

The results are summarised in Table 3.5.

Table 3.5: Comparison of late and early fusion using paired Wilcoxon signed-rank tests.

Metric	**Late Fusion**	**Early Fusion**	**P-value**	**Effect Size (r)**
Top-5 Success	90.7%	81.1%	0.003	0.42
Coverage	41.2%	36.5%	0.018	0.31

With p-values < 0.05 for both metrics, H_0 is rejected and conclude that late fusion

significantly outperforms early fusion. The moderate effect sizes ($r > 0.3$) confirm practical significance beyond mere statistical significance.

3.13 Discussion

3.13.1 Multi-Dimensional Benchmarking with State-of-the-Art Methods

In order to further confirm the effectiveness of the performance of the proposed framework, a comparison with the current state-of-the-art works in the context of the online advertising domain is done. Due to the variety of the metrics used in video-based online advertising works in the past, a multi-dimensional method of comparing different works based on the methodology, functionality, and performance of each methodology is used in this work. Table 3.6 summarises a comparison with SemanticAd (2024) [71] and "What Modality Matters" (2023) [80].

Based on this comparison, the following observations have been made. First, while existing approaches may specialize in certain areas of contextual advertising, such as the importance of modalities and temporal segmentation, the current method integrates all the above areas to address the overall application of video to industry taxonomies in an industry-agnostic fashion. Second, the application of the embeddings from the BERT neural network enables the method to index and search video content efficiently with minimal need for retraining. Third, the method applies topic modelling using the BERTopic and enables the system to adapt to new video topics that may appear, allowing for the swift association of ads with topics that have popular appeal with minimal engineering upgrades to the current system.

3.13.2 Addressing Class Imbalance and Dataset Bias

Class imbalance and dataset bias are well-known challenges in deep learning systems trained on large-scale datasets such as YouTube-8M. As highlighted by Roccetti et al. [199], "bigger datasets are not always better" when the underlying data are biased or imbalanced. To mitigate these issues, this work incorporates several complementary

Table 3.6: Multi-dimensional comparison highlighting distinctive aspects of the framework.

Dimension	**SemanticAd (2024)**	**What Modality Matters (2023)**	**Our Framework**
Core Technical Approach	Temporal segmentation of video into story units	Modality importance analysis for relevance scoring	Topic modelling with industry taxonomy mapping
Semantic Understanding	Scene-level content analysis	Multi-modal feature extraction	Deep topic semantics using BERTopic with contextual fusion
Industry Readiness	Ad insertion point detection	Relevance scoring for ad matching	Standardized taxonomy output compatible with IAB standards
Technology Used	Shot-boundary detection, temporally constrained clustering, multimodal fusion	Pretrained CNNs (TSN, YOLOv5, PlacesCNN, EmotionNet, CNN14) with similarity matching	BERTopic topic modelling and inferencing, vector-space embedding similarity, IAB taxonomy mapping
Evaluation Approach	Custom dataset (50 TV news videos)	Movie content + 14 ads with expert study	YouTube-8M with ground-truth taxonomy labels
Performance Metrics	F1-score (92%+)	Relevance Score (S) (0.337)	Top-5 Success Rate (90.7%)

strategies spanning dataset design, model architecture, and evaluation methodology.

- **Targeted dataset selection:** Rather than using the full heterogeneous YouTube-8M collection, a food-domain subset is selected in order to reduce extreme label imbalance and domain variability. The evaluation set consists of 50 food-related videos drawn from the same semantic category, enabling more controlled experimentation and reducing cross-domain noise during assessment.
- **Model design:** BERTopic is used as the core topic modelling framework because it automatically determines the number of topics while also introducing noise top-

ics for semantically weak or irrelevant documents. This mechanism prevents noisy samples from distorting the latent topic structure, which is particularly advantageous in imbalanced datasets where minority semantic patterns may otherwise be overwhelmed by dominant classes.

- **Evaluation diversity:** The framework is assessed using multiple evaluation measures, including Top-N success and coverage metrics, alongside human evaluation. Combining automated and qualitative assessment helps reveal biases or limitations that may not be captured by a single quantitative metric alone.

Beyond these explicit mitigation strategies, the proposed framework also contains several implicit mechanisms that reduce the negative effects of class imbalance. Most importantly, the topic modelling stage constructs an abstract semantic representation (T, w) independently of the underlying class distribution. Instead of directly learning a mapping between input features and potentially imbalanced taxonomy labels, the system first identifies latent thematic structures within the multimodal data. Consequently, the learned representations are driven by semantic coherence rather than label frequency, reducing dependence on dominant classes.

In addition, the taxonomy mapping stage relies on embedding-based similarity between topic representations and taxonomy descriptors. This enables semantically meaningful alignment with less frequent or sparsely represented categories when strong semantic similarity exists. As a result, the framework is less likely to disproportionately favour majority classes compared to conventional supervised classifiers trained directly on imbalanced label distributions.

Despite these mitigation mechanisms, the evaluation is intentionally restricted to the food domain. This controlled scope allows the behaviour of the multimodal fusion and taxonomy mapping architecture to be analysed in isolation, without the confounding influence of large cross-domain variation. Importantly, however, the proposed semantic modelling framework is architectural rather than domain-specific. The approach is fundamentally based on embedding similarity and intensity-weighted topic representations, both of which are domain-agnostic by design. Broader generalisation considerations, including multilingual robustness and modality-specific noise sensitivity, are discussed

further in Chapter 4.

Nevertheless, the current evaluation may not fully capture the diversity and complexity of the complete YouTube-8M dataset, and dataset biases are not systematically analysed, such as the potential overrepresentation of particular video styles or content formats. These limitations are consistent with the concerns raised by Roccetti et al. [199] and represent important directions for future investigation.

3.13.3 Multimodal Fusion Performance and Framework Limitations

The statistical validation (Section 3.12.2) presents strong evidence that late fusion performs significantly better than early fusion ($p < 0.01$) and thus refutes potential doubts on the validity of the findings in light of the observed properties of the used datasets [199]. Both hypothesis tests validate the assumptions on the efficacy of multimodal fusion, despite natural constraints of the video data.

The proposed method shows that using both visual and audio information improves the precision and detail of content classification and taxonomy mapping. It is observed that the framework can identify hidden themes not captured in the ground truth due to missing labels, which negatively affects coverage scores. Audio transcripts frequently have a greater impact on topic probabilities than visual topics, leading to a higher contribution of audio topic terms to final predictions. However, visual themes become more prominent when audio is noisy. Additionally, the early fusion model performs well when there is coherent information sharing between visual and audio streams. These observations are consistent with current studies showing that multimodal approaches improve content recognition and classification outcomes.

One major limitation could be the assumption that the returned image objects from AWS Rekognition contribute as visual words to the system, which could potentially bring in noise, considering that all image objects may not add to the theme in the video. The reason for this could partially lie in the fact that the application has tried to use the full potential that can be tapped from the visual end using the BERTopic model, considering that the importance of image objects is still to be explored.

Furthermore, the use of Amazon's Rekognition system for feature extraction creates

a reproducibility issue, as the internal workings of the system and its updates are not entirely under the authors' control. Nonetheless, this is not an intrinsic aspect of the proposed framework since it only requires abstracted visual representations that can be obtained using other open-source models. This is discussed in Chapter 4, where CLIP-ViT embeddings are used instead of Amazon Rekognition, demonstrating that the framework generalises to fully reproducible feature extraction pipelines.

Another limitation is the way topics are represented, as current approaches represent topics by counting the most relevant words, which may not completely summarize the essence of the topics. Combining topic models with a way of representing that is more expressive may result in better comprehension of video information. Another promising area that may be explored is the use of large language models (LLMs), which can generate semantically meaningful representations of topics.

3.13.4 Positioning Against Synthetic Media

The increasing prevalence of synthetic media (e.g., deepfakes) presents new threats to contextual ad systems: adversarially crafted visual or audio streams could induce spurious topics and lead to unsuitable taxonomy labels with brand-safety implications. The proposed system counters some of these threats through:

- **Cross-modal coherence:** late fusion requires consistent evidence across modalities, making it harder for adversarial noise in one modality to dominate.
- **Topic-level semantics:** neural topic modelling prefers thematically coherent clusters over surface-level features, mitigating impact of isolated adversarial artefacts.
- **Taxonomy-level verification:** canonicalised label alignment can highlight inconsistencies typical of synthetic materials or unnatural distributions.

In practice, these features provide auxiliary support for auditing content and detecting anomalies, e.g., flagging videos where audio–visual semantics conflict or where taxonomy confidence profiles are atypical. While not a dedicated forgery detector, the system complements authentication checks; incorporating explicit deepfake detection modules is a natural extension for future work.

3.13.5 Limitations and Future Work

The framework presented in this chapter has several limitations that motivate future work:

- **Dependence on commercial APIs:**

  A notable implementation choice in the present chapter is the use of Amazon Rekognition for visual feature extraction (object labels subsequently converted into Bag-of-Visual-Words). While this service provides high-quality, production-grade recognitions that reflect real-world deployment conditions, it introduces significant reproducibility limitations. Commercial APIs are subject to internal model updates, versioning changes, and rate-limiting policies that lie outside the researcher's control. Consequently, exact replication of the reported experiments may vary over time even with identical input videos.

  To mitigate this concern and demonstrate the modularity of the proposed framework, Chapter 4 replaces Rekognition with fully open-source and reproducible alternatives (CLIP-ViT embeddings + ViT-GPT2 captioning). The strong performance of the revised pipeline confirms that the core methodological contributions BERTopic-based multimodal topic modelling and taxonomy mapping do not depend on proprietary services. Future implementations are therefore encouraged to adopt open models for complete methodological transparency and long-term reproducibility.
- **Noisy visual labels:** Generic detectors can yield noisy or irrelevant labels, biasing topics. Integrating scene understanding, salience scoring, or visual captioning could yield richer, more reliable descriptors.
- **Symbolic Visual Representation Constraints:** The use of object-label–derived Bag-of-Visual-Words representations provides interpretability and compatibility with text-based topic modelling, but it abstracts away fine-grained spatial and relational visual semantics. End-to-end multimodal encoders (e.g., joint vision-language models) may capture richer cross-modal dependencies. However, the present approach prioritises transparency, modularity, and taxonomy alignment over representational depth, consistent with the thesis' layered and interpretable

design philosophy.

- **ASR errors:** ASR errors under adverse acoustics can skew audio topics; robustness could be improved with confidence-aware fusion and speech enhancement.
- **Temporal granularity:** Current outputs are at video level; introducing temporal segmentation would enable time-stamped taxonomy labels for fine-grained ad insertion and brand safety.
- **Fusion strategy:** Fusion currently uses convex weighting rather than a learned, content-adaptive gating mechanism; lightweight multimodal transformers or gating modules could learn better fusion.
- **Confidence calibration:** Confidence scores are not explicitly calibrated; post-hoc calibration and abstention policies could improve reliability for downstream use.

These directions connect directly to the next stages of the thesis, where the progression moves from video-level taxonomy inference towards explainable topic models (Chapter 4), and subsequently to theme- and cognitive-load–aware segmentation.

3.14 Chapter Summary and Outlook

This chapter has presented a new approach for contextual video advertising that integrates multimodal topic modelling with standards-based taxonomy mapping. By effectively applying contextual video analysis with industry content taxonomies, the framework supports effective advertising in advanced advertising ecosystems. The combination of video and audio thematic analysis helps bridge the semantic gap between raw video content and human-interpretable, industry-compatible categories.

Neural components are employed for efficient association of video content with context-based taxonomies, and the framework is validated experimentally on a YouTube-8M subset. The model achieves a Top-5 prediction success of 90.7% and Top-10 success of 92.7% under late fusion, significantly outperforming early fusion ($p < 0.05$).

Finally, the chapter has identified limitations related to domain generalisation, visual and audio noise, temporal granularity, fusion strategies, and confidence calibration.

These limitations provide a direction for the advancement of this work in the future, in terms of incorporating large language models for explanations related to topics and finer segmentation in ad placement.

This is extended in the next chapter to provide an *explainable* multimodal topic modelling framework because the current chapter identifies limitations in the taxonomy mapping stage, in terms of transparency, robustness, and industry readiness.

Chapter 4

Explainable Multimodal Video Topics for Content Taxonomy

Publication Notice.

This chapter is based on the following peer-reviewed publication:

De Silva, W., & Fernando, A. (2025). Explainable Video Topics for Content Taxonomy: A Multimodal Retrieval Approach to Industry-Compliant Contextual Advertising. ***IEEE Access****, 13, 30597–30610. https://doi.org/10.1109/ACCESS.2025.3542562*

4.1 Introduction

As video consumption continues to grow, advertisement-based video-on-demand (AVoD) has become a dominant monetisation model in the streaming ecosystem [200]. AVoD platforms typically rely on real-time bidding (RTB) [201], where advertisers must make split-second decisions on which ads to serve based on the semantic context of the current video segment. Accurate and interpretable taxonomy mapping is therefore critical,

as it enables more effective content-to-ad matching, improves bidding precision, and supports brand safety. AVoD platforms have traditionally relied on behaviour-based targeting using cookies and user identifiers. However, privacy regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) [202], together with emerging frameworks like the Digital Services Act [203], are increasingly restricting the use of personal data and are driving the ecosystem towards privacy-preserving approaches.

Contextual advertising offers a compelling alternative: instead of targeting users, ads are matched to the context of the content being consumed [204]. For video, this requires methods that can effectively interpret rich multimodal signals (visual, audio, and textual) [205,206], operate robustly across languages [207], and remain interpretable for brand safety and regulatory auditing [208, 209].

Building directly upon the multimodal topic modelling and taxonomy mapping framework developed in Chapter 3, the present chapter addresses a critical remaining limitation: the lack of human-interpretable explanations for the discovered topics.

While Chapter 3 demonstrated that combining Bag-of-Visual-Words (BoVW) features with ASR transcripts within a BERTopic pipeline significantly outperforms unimodal and metadata-only baselines, the resulting topics remained opaque latent constructs. Although quantitative topic coherence measures can assess the internal consistency and interpretability of such topics [210], they still do not provide natural language explanations of the semantic rationale or the contribution of multimodal signals.

In the absence of clear explanation mechanisms, it becomes challenging to determine whether the inferred topics genuinely reflect significant content or emerge from spurious associations between modalities a well-known challenge in black-box models [211].

This chapter therefore advances the previous work by introducing explainability as a core design principle. Explainability becomes equally as important as predictive performance in this domain, as it enables advertisers, platforms, and auditors to understand and verify why particular topics and taxonomy labels are assigned.

Relationship to Chapter 3

Chapter 3 introduced a multimodal framework for contextual video advertising based on Bag-of-Visual-Words (BoVW) features derived from Amazon Rekognition labels and audio transcripts extracted via Whisper ASR. These features were jointly modelled using BERTopic [184] and mapped to an industry content taxonomy (e.g. IAB Tech Lab [212]) through embedding-based similarity with topic- and term-level intensity weighting. The focus of Chapter 3 was to demonstrate that multimodal topic modelling and taxonomy mapping outperform metadata-only and unimodal approaches in a cold-start, privacy-preserving setting.

The present chapter extends that foundation along three main dimensions:

- **Richer embeddings.** BoVW is replaced with CLIP-ViT visual embeddings [163] and image captions generated by a ViT-GPT2 model [213–215], while transcripts are embedded using a multilingual MiniLM sentence transformer [216, 217]. This yields semantically richer and better-aligned multimodal representations.
- **Explainability.** Large language models (LLMs), such as T5-Base [218] and PaLM 2 [219], are employed to convert BERTopic topics into short, human-readable descriptions. These explanations act as an interpretable interface between low-level embeddings and high-level taxonomy labels.
- **Multilingual and noisy robustness.** Robustness is explicitly evaluated under noisy visual conditions, noisy transcripts, and multilingual settings on a curated subset of YouTube-8M [156], and the temporal evolution of taxonomies within videos is analysed.

These enhancements directly respond to the limitations identified in Chapter 3 regarding interpretability, noise sensitivity, and multilingual generalisation, while maintaining the training-free, privacy-preserving philosophy of the overall thesis.

Chapter Contributions

The main contributions of this chapter are as follows:

1. **Explainable multimodal architecture.** An end-to-end framework is designed

that combines CLIP-ViT image embeddings, ViT-GPT2 image captioning, multilingual MiniLM sentence embeddings, and BERTopic to infer coherent topics from both visual and textual streams, and then map them to IAB taxonomies.

2. **Separate yet aligned topic models.** Two BERTopic models are trained, one for visual data and one for textual transcripts, allowing modality-specific hyperparameters and improved handling of heterogeneous noise characteristics.

3. **LLM-based topic explanations.** LLMs are integrated to convert topics (keywords plus representative documents) into concise natural-language labels, increasing interpretability for advertisers, platform operators, and auditors.

4. **Semantic taxonomy mapping.** A hierarchical taxonomy mapping procedure is proposed that concatenates IAB taxonomy tiers, embeds them in a shared semantic space, and uses cosine similarity and topic probabilities to produce multi-label taxonomy assignments.

5. **Robustness evaluation.** A detailed empirical study is conducted comparing unimodal and multimodal setups, analysing robustness to modality-specific noise and multilingual transcripts, and reporting performance at both top-level (T1) and granular (T4) taxonomies.

6. **Temporal taxonomy evolution.** An extension for tracking taxonomy frequencies over time within a video is sketched and empirically illustrated, enabling dynamic, content-aware ad placement.

The rest of the chapter is structured as follows. Section 4.2 presents the theoretical foundations, including explainability in topic modelling, vision–language pre-training, multilingual embeddings, LLM-based topic explanation, and content taxonomy standards. Section 4.3 formalises the problem and notation. Section 4.4 provides a high-level overview of the proposed framework. Sections 4.5 and 4.6 detail the visual and textual topic modelling pipelines, respectively. Section 4.7 describes LLM-based topic explanation. Sections 4.8 and 4.9 cover video topic inference and taxonomy mapping.

Section 4.10 explains the experimental setup. Section 4.11 presents the results, including robustness and multilingual analyses and visualisations. Section 4.12 discusses strengths, limitations, and practical considerations, and Section 4.13 summarises key takeaways and links to Chapter 5.

4.2 Background and Theoretical Foundations

4.2.1 Explainability in Topic Modelling

A classical topic modelling approach that utilizes Latent Dirichlet Allocation (LDA) [220] to express topics as distributions over words is not ideal when texts are short, when words in the dictionary are polysemous, or when many hyperparameters need to be tuned. Cross-topic modelling is proposed to apply embedding derived from BERT [186] to enhance semantic modelling [221].

BERTopic [184] builds on this line of work by combining contextual embeddings, UMAP dimensionality reduction [190], HDBSCAN clustering [192], and class-based TF–IDF (c-TF–IDF) topic representations. The resulting topics are typically more coherent and require less manual tuning than LDA [222]. However, even with c-TF–IDF, topics are still expressed primarily as ranked word lists, which can be opaque to non-expert users and difficult to align with industry taxonomies.

Explainable AI (XAI) research emphasises transparency and human-understandable explanations in machine learning systems, particularly in high-stakes domains [209, 223]. In the context of contextual advertising, explainability is needed both for brand safety and for regulatory compliance [208]. The proposed framework therefore enhances BERTopic with LLM-based topic descriptions and explicit mapping to standard taxonomies.

4.2.2 Vision–Language Foundations

Recent advances in vision–language models have transformed multimodal understanding. CLIP [163] learns a joint embedding space for images and text by contrastive training on large-scale image–caption pairs, allowing zero-shot classification and open-

vocabulary recognition. Vision Transformers (ViT) [213] treat images as sequences of patches and have demonstrated strong performance on image recognition tasks. CLIP-ViT variants combine ViT backbones with CLIP-style training, balancing semantic alignment and visual representational power [224, 225].

For generating textual descriptions of images, encoder–decoder architectures that couple visual encoders (e.g. ViT) with autoregressive language models (e.g. GPT-2 [214]) have become standard. ViT-GPT2-based captioning models, such as those available via the Hugging Face ecosystem [215], produce concise descriptions that can serve as textual surrogates for images.

By using CLIP-ViT embeddings for keyframes and generating captions with ViT-GPT2, the proposed framework leverages both visual and textual perspectives on the same content, enabling richer and more interpretable topics.

4.2.3 Multilingual Embeddings

Sentence-transformer architectures such as SBERT [216] extend BERT-like models to produce sentence-level embeddings optimised for similarity tasks. Multilingual variants, including `paraphrase-multilingual-MiniLM-L12-v2` [217], support dozens of languages and have been shown to provide robust cross-lingual representations [226]. Compared to generic multilingual BERT (mBERT) [227], MiniLM-based models are lighter and faster while maintaining strong performance in semantic similarity.

Because AVoD platforms serve global audiences, transcripts may contain code-switching, accented speech, and non-English languages. The use of multilingual MiniLM embeddings enables consistent topic modelling across languages, supporting the thesis objective of privacy-respecting, language-agnostic contextual advertising.

4.2.4 LLMs for Topic Explanation

Large language models such as T5 [218], PaLM 2 [219], and GPT-4 can summarise, paraphrase, and reframe textual input, making them natural candidates for generating human-readable topic labels. Recent work has used LLMs to refine or interpret topics, yielding more expressive and user-friendly representations [228].

In the proposed framework, LLMs are used in an *offline* fashion: during training and analysis, topic keywords and representative documents are fed to an LLM, which generates short descriptions. These descriptions are then embedded and used for taxonomy mapping and explanation. At inference time, the stored descriptions are reused, so there is no need to call the LLM in production, controlling cost and latency.

4.2.5 Taxonomy Standards for Contextual Advertising

To interoperate across platforms, supply-side platforms (SSPs), demand-side platforms (DSPs), and ad exchanges rely on shared content taxonomies. The IAB Tech Lab Content Taxonomy [212, 229] provides a standardised, hierarchical categorisation of content into tiers (T1–T4) covering topics such as news, sports, food, and finance. These taxonomies are embedded into programmatic protocols and are critical for ad eligibility, brand safety, and reporting.

Mapping video content to these taxonomies is therefore a core requirement for contextual advertising systems, and a central focus of both Chapter 3 and this chapter.

4.3 Problem Definition and Notation

The problem of explainable multimodal video taxonomy mapping is formalised as follows.

Let X_i denote the i-th video, with associated keyframes $V_i = \{k_1, \ldots, k_m\}$ and transcript segments $A_i = \{s_1, \ldots, s_n\}$. Each keyframe k_j is an RGB image; each transcript segment s_ℓ is a short text snippet produced by ASR.

A content taxonomy $\mathcal{C} = \{C_1, \ldots, C_L\}$ is given, where each C_ℓ represents a hierarchical label (e.g. an IAB Tier 4 path). The objective is to infer, for each video X_i, a set of taxonomies with associated weights:

$$\mathcal{Y}_i = \{(\mu_{i1}, C_{i1}), \ldots, (\mu_{ip}, C_{ip})\},$$

where $C_{ij} \in \mathcal{C}$ and $\mu_{ij} \in [0, 1]$ reflects the strength and frequency of evidence for that taxonomy in the video. This is a multi-label classification problem, where each label

corresponds to a taxonomy entry.

In addition, the following conditions must be satisfied:

- **Explainability:** Each intermediate topic should have a human-readable description that can be inspected and audited.
- **Multimodal consistency:** Visual and textual topics should be derived from modality-specific embeddings but mapped to a shared semantic space to support fusion.
- **Taxonomy compatibility:** Outputs must be expressed in the standard taxonomy used by the ecosystem (e.g. IAB).

The mapping pipeline can be written abstractly as

$$(X_i) \to (V_i, A_i) \to \{T_{ik}^{(V)}\}_k, \{T_{ik}^{(A)}\}_k \to \mathcal{Y}_i,$$

where $T_{ik}^{(V)}$ and $T_{ik}^{(A)}$ denote visual and textual topics inferred via BERTopic and explained via LLMs.

4.4 Proposed Framework Overview

Figure 4.1 presents a high-level overview of the proposed explainable multimodal topic-modelling framework. Visual and textual topic models are first trained offline on domain-specific image and text corpora. At inference time, keyframes and transcripts from a new video are processed by these models, producing topics which are then mapped to IAB content taxonomies.

Besides depicting the processing pipeline, Figure 4.1 emphasizes two important elements used to enrich the semantic meaning of the framework: CLIP-ViT-based visual embeddings and caption generation for images. CLIP-ViT visual embeddings represent keyframes in a joint vision-language embedding space [163], which allows the grouping of visually similar frames semantically rather than purely low-level features. This is particularly important for topic modelling, as it allows clustering algorithms (e.g., UMAP + HDBSCAN within BERTopic) to form semantically coherent visual topics.

Complementing the embeddings, an image captioning model (ViT-GPT2) gener-

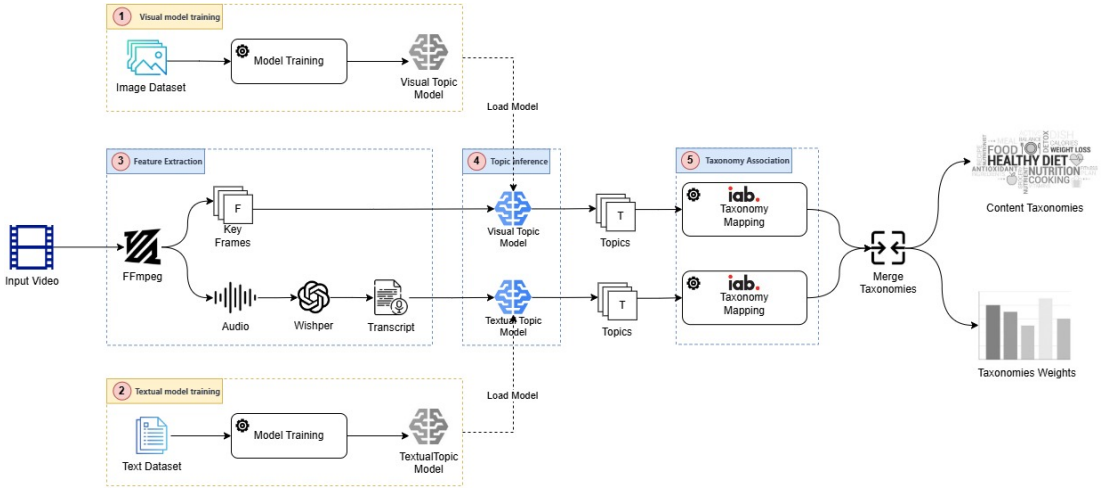


Figure 4.1: Explainable multimodal topic-modelling pipeline for contextual video analysis. (1) Visual topic model training on an image corpus. (2) Textual topic model training on a text corpus. (3) Feature extraction from video (keyframes and transcripts). (4) Topic inference for visual and textual streams. (5) Mapping topics to IAB content taxonomies and merging taxonomy weights.

ates concise natural-language descriptions for every keyframe. The captions are then considered as documents and used by BERTopic to build topic representations using contextualised c-TF-IDF over useful linguistic terms.

The combination of CLIP-based embeddings and captioning therefore provides both implicit semantic structure (via embedding-space clustering) and explicit semantic articulation (via textual descriptions). This dual representation improves topic coherence, enhances interpretability, and facilitates more accurate alignment with standardised IAB taxonomy categories.

The core steps are:

1. Train visual and textual BERTopic models using CLIP-ViT and multilingual MiniLM embeddings, respectively.

2. Extract keyframes and transcripts from each video using FFmpeg and Whisper.

3. Infer topics separately from visual and textual inputs.

4. Use LLMs to generate short descriptions for each topic.

5. Embed topic descriptions and taxonomy strings in a shared space and compute similarity.

6. Aggregate high-confidence matches into multi-label taxonomy assignments with weights.

Compared with the framework presented in Chapter 3, which relied on Bag-of-Visual-Words (BoVW) labels and word-based topics, this architecture operates directly on semantically rich, pre-trained embeddings and incorporates LLM-based explanations, resulting in significantly improved coherence, interpretability, and robustness.

4.5 Visual Topic Modelling Pipeline

The visual pipeline converts keyframes into topics using CLIP-ViT embeddings, image captioning, and BERTopic.

4.5.1 CLIP-ViT Embeddings

Keyframes are embedded using CLIP-ViT-B/32 [163]. This model uses a ViT-B/32 backbone [213] for images and a transformer text encoder, trained jointly to align images and text. CLIP-ViT embeddings capture both visual similarity and semantic relationships, enabling clustering of images by concept rather than only low-level features [224, 225].

4.5.2 Image Captioning with ViT-GPT2

To bridge visual and textual representations, a ViT-GPT2 image captioning model [213–215] is employed. For each keyframe, the model generates a short caption describing the scene (e.g. “chef slicing vegetables on a wooden board”). These captions serve as textual documents for BERTopic.

4.5.3 Visual BERTopic Model

Figure 4.2 illustrates the visual BERTopic pipeline.

The steps are:

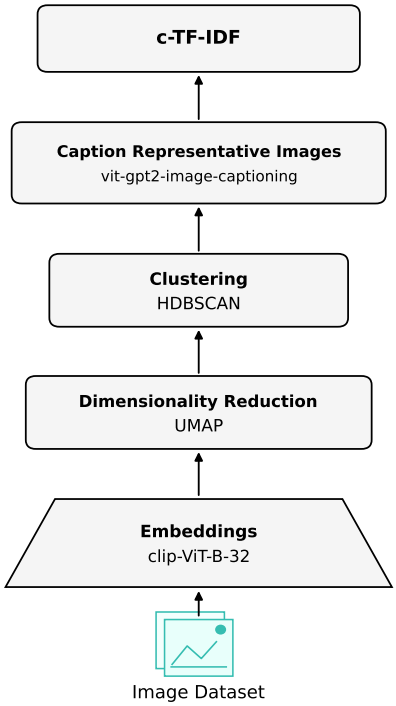


Figure 4.2: Visual topic modelling pipeline using CLIP-ViT embeddings, UMAP, HDBSCAN, and c-TF–IDF. Image captions support topic interpretability.

1. Compute CLIP-ViT embeddings for keyframes.

2. Optionally concatenate or align embeddings with their captions.

3. Apply UMAP to reduce dimensionality (parameters in Table 4.1).

4. Cluster reduced embeddings using HDBSCAN.

5. Build topic representations using c-TF–IDF over caption tokens.

The c-TF–IDF weight for word x in class c is

$$w_{x,c} = \mathrm{tf}_{x,c} \times \log\left(1 + \frac{A}{f_x}\right), \tag{4.1}$$

where $\mathrm{tf}_{x,c}$ is the frequency of x in class c, f_x is its frequency across all classes, and A is the average number of words per class [184].

Table 4.1: Optimal UMAP and HDBSCAN parameters for textual and visual topic models.

Model	UMAP parameters	HDBSCAN parameters
Textual embeddings	`n_neighbors`=100, `n_components`=5, `min_dist`=0.0, `metric`=`cosine`	`min_cluster_size`=200, `min_samples`=150, `metric`=`euclidean`
Visual embeddings	`n_neighbors`=3, `n_components`=4, `min_dist`=0.0, `metric`=`cosine`	`min_cluster_size`=30, `min_samples`=20, `metric`=`euclidean`

4.5.4 Hyperparameters and Challenges

The optimal UMAP and HDBSCAN parameters for the visual model are listed in Table 4.1. Compared with the textual model, the visual model uses smaller neighbourhoods and cluster sizes to capture finer-grained visual patterns.

Challenges include dealing with noisy or irrelevant keyframes and ensuring that clusters correspond to semantically meaningful visual concepts rather than purely low-level features. CLIP-ViT mitigates this by aligning images with text semantics, and c-TF–IDF over captions helps emphasise discriminative visual terms.

4.6 Text Topic Modelling Pipeline

The textual pipeline converts audio into transcripts and then into topics via multilingual embeddings and BERTopic.

4.6.1 Transcripts from Whisper

Audio tracks are extracted using FFmpeg [230] and transcribed with OpenAI Whisper [231]. Whisper supports multiple languages and is robust to noisy conditions typical of user-generated content. The output transcript is segmented into sentences or short phrases, each treated as a document.

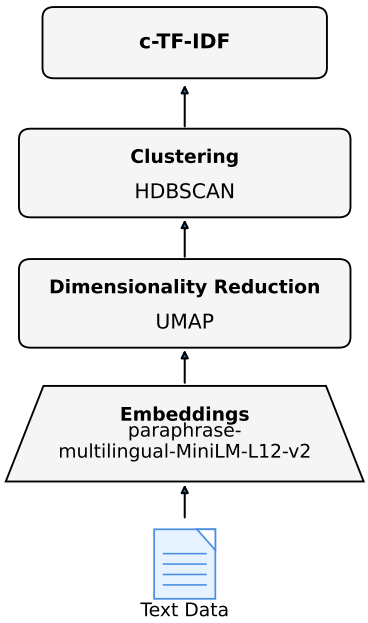


Figure 4.3: Text topic modelling pipeline using multilingual MiniLM embeddings, UMAP, HDBSCAN, and c-TF–IDF.

4.6.2 Multilingual MiniLM Embeddings

Transcript segments are embedded using the `paraphrase-multilingual-MiniLM-L12-v2` sentence-transformer model [216, 217]. This model supports over 50 languages, providing a unified semantic space for cross-lingual transcripts [226].

4.6.3 Text BERTopic Model

Figure 4.3 shows the textual BERTopic pipeline.

The steps mirror those of the visual pipeline:

1. Compute multilingual MiniLM embeddings for transcript segments.
2. Apply UMAP dimensionality reduction (parameters in Table 4.1).
3. Cluster reduced embeddings with HDBSCAN, capturing recurrent themes.
4. Compute c-TF–IDF topic representations using Eq. (4.1).

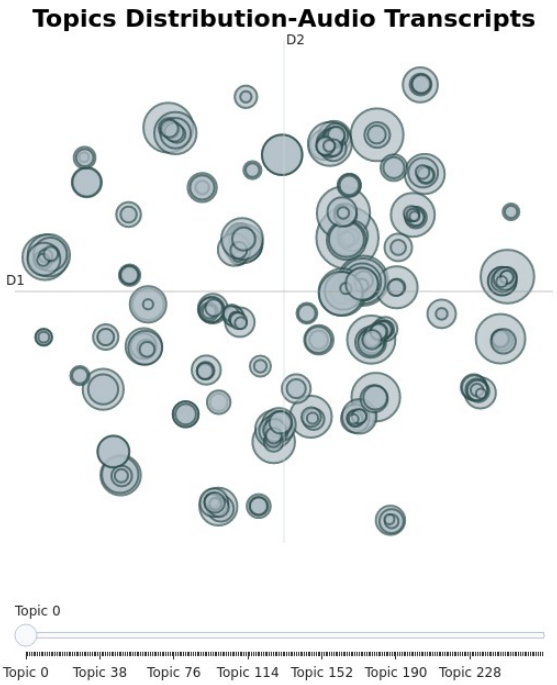


Figure 4.4: Topic distribution for the audio-based BERTopic model after UMAP dimensionality reduction. The plot projects all discovered topics onto a 2D UMAP embedding space (axes labelled D1 and D2). Each bubble represents one topic; bubble size is proportional to the number of document segments assigned to that topic. The interactive slider at the bottom (showing Topic 0, Topic 38, ...) allows selection of individual topics for closer inspection.

The textual topic distribution over the training corpus is visualised in Figure 4.4; the visual topic distribution is shown in Figure 4.5. Each bubble represents a topic, with size proportional to its frequency.

Figure 4.4 and Figure 4.5 display the topic distributions that have been generated through the use of audio-based and visual-based BERTopic model, respectively, after UMAP dimensionality reduction. Here, each data point represents a snippet of documents (that is, transcript or visual features), and the clusters indicate semantically related segments grouped using the topic modelling technique.

In Figure 4.4, the clustering of topics using audio-based data shows better separation among the clusters, due to the presence of semantic structure in texts. In spoken language, context is already encoded within the speech; this supports the effectiveness of transcript-driven topic modelling for capturing high-level semantic themes.

Figure 4.5: Topic distribution for the visual BERTopic model after UMAP dimensionality reduction. The plot projects all discovered topics onto a 2D UMAP embedding space (axes labelled D1 and D2). Each bubble represents one topic; bubble size is proportional to the number of document segments assigned to that topic. The interactive slider at the bottom (showing Topic 0, Topic 38, ...) allows selection of individual topics for closer inspection.

In contrast, as depicted in Figure 4.5, visual features are utilized to generate the distribution of topics. It can be seen from the figure that the clustering effect remains visible, although the separation between clusters are relatively less distinct compared to previous cases. This is expected because visual features may contain information that is more ambiguous or relative, resulting in a high degree of overlap in the embedded space. However, it can be seen that clusters do exist despite the increased level of ambiguity demonstrating the effectiveness of this method.

4.7 Topic Explanation via Large Language Models

To improve interpretability, each BERTopic topic is enriched with an LLM-generated natural-language description.

4.7.1 Prompting Strategy

For each topic, (i) top c-TF–IDF keywords and (ii) a small set of representative documents are collected (image captions for visual topics, transcript snippets for textual topics). These are then fed into an LLM using the following template:

```
I have a topic that is described by the following keywords: [
    KEYWORDS]
In this topic, the following documents are a small but
    representative subset
of all documents in the topic:
[DOCUMENTS]

Based on the information about the topic above, create a short
    description
of this topic with few words.
```

The LLM outputs a phrase such as "home baking desserts" or "healthy vegetarian cooking", which are stored as human-readable labels for the corresponding topics.

4.7.2 LLMs Used and Quality Control

Two LLMs are evaluated::

- **T5-Base** [218], an encoder–decoder model fine-tuned for text-to-text tasks.
- **PaLM 2** [219], a larger generative model capable of more nuanced summaries.

PaLM 2 generally produces more precise and domain-appropriate descriptions, and its use is observed to yield slightly higher downstream taxonomy performance (see Section 4.11). To control quality, degenerate outputs (e.g., repetitions, overly generic phrases) are identified, and filtering or regeneration is applied as needed.

4.7.3 Benefits and Risks of LLM-Based Topic Explanation

Though LLMs are an efficient approach for creating human-interpretable descriptions of topics, their application also introduces important risks that must be carefully managed.

A primary concern is the risk of hallucination, i.e. the generation of factually incorrect explanations that are not fully grounded in the underlying topic representations. In the context of multimodal topic modelling, this can lead to interpretations that extend beyond the actual semantic content captured by the topic model.

Moreover, the generated topic explanations could suffer from some level of bias due to the nature of model training on large-scale web data. Therefore, at times, explanations might highlight more dominant or stereotypical interpretations rather than being focused purely on the content described by the video topic itself.

In addition, generated topic explanations are highly dependent on the particular design of the prompt used, as small changes in the wording of the instruction can significantly affect reproducibility and consistency across runs.

Finally, it should be noted that LLM explanations are not tied to the topic modelling process itself and thus do not affect the core quality of the topic discovery task. They should therefore be treated strictly as an auxiliary interpretability layer rather than as the primary basis for decision-making.

To counteract the aforementioned limitations, controlled prompting techniques have been applied in this work, along with limiting the input to topic keywords and representative documents. Furthermore, all LLM outputs have been used only as an auxiliary aid for interpretability, while the actual taxonomy mapping continues to rely primarily on embedding-based similarity.

4.8 Video Topic Inference

During inference, the trained visual and textual topic models are applied to a new video X_i. Each keyframe and transcript segment is assigned to either a specific topic or to the noise class. Topic assignments are aggregated to obtain a probability distribution over topics.

The topic distribution for video X_i is given by

$$\text{TopicDist}(X_i) = \{(T_k, p_{i,k}) \mid k = 1, \ldots, n\}, \tag{4.2}$$

where T_k is a topic with an LLM-generated label and $p_{i,k}$ is the probability or normalised frequency of that topic in X_i. Separate distributions are obtained for visual and textual topics; these can be combined via late fusion (e.g. weighted sum or max-pooling) or treated independently in taxonomy mapping.

Late fusion is particularly helpful in noisy conditions, as each modality can compensate for deficiencies in the other, a finding consistent with Chapter 3 and with broader multimodal literature [232–234].

4.9 Taxonomy Mapping

4.9.1 Concatenating Taxonomy Tiers

Each row of the IAB content taxonomy [212] contains up to four tiers $T_{i,1}, \ldots, T_{i,4}$. To capture full hierarchical context, non-empty tiers are concatenated into a single string:

$$C_i = \text{trim}\left(\sum_{j=1}^{n}(T_{i,j} + \text{" "}) \cdot \mathbb{I}(T_{i,j} \neq \text{""})\right), \tag{4.3}$$

where $n = 4$ for T1–T4, and $\mathbb{I}$ is the indicator function. For example, a full path might read "Food Cooking Desserts and Baking Cakes".

4.9.2 Cosine Similarity and Thresholding

Both topic descriptions and concatenated taxonomy strings are embedded using the same sentence-transformer used in BERTopic. Let T_k denote a topic embedding and C_i a taxonomy embedding. Cosine similarity is computed:

$$\text{sim}(T_k, C_i) = \frac{T_k \cdot C_i}{\|T_k\| \, \|C_i\|}, \tag{4.4}$$

and retain only pairs with $\text{sim}(T_k, C_i) \geq \theta$, where θ is a similarity threshold (set to 0.7 in the experiments).

4.9.3 Final Taxonomy Weights

For each video X_i, evidence is aggregated from all topics whose probability exceeds a predefined threshold (set to 0.7 in the experiments). Let $\mu_{i,j}$ denote the resulting weight for taxonomy label C_j, obtained by summing or averaging similarity scores from supporting topics, optionally scaled by topic probabilities $p_{i,k}$:

$$\mathcal{Y}_i = \{(\mu_{i,1}, C_1), (\mu_{i,2}, C_2), \ldots, (\mu_{i,p}, C_p)\}. \tag{4.5}$$

These weighted labels constitute the explainable, taxonomy-aligned output of the framework and are used for evaluation in Section 4.11.

4.10 Experimental Setup

4.10.1 Datasets

Two sets of datasets are used: one for training the topic models and one for evaluating video taxonomy inference.

Training Data For textual topic model fine-tuning, a Kaggle Food.com dataset [198], comprising over 180k recipes and reviews, is used to provide rich food-domain semantics. For visual topic model training, the Food-101 dataset, consisting of 50k images across 101 food categories, is used.

Evaluation Videos Evaluation is conducted on a food-focused subset of YouTube-8M [156]. Specifically, a total of 135 videos labelled in the "Food & Drink" category are selected, and ground truth labels are refined by mapping each to one or more IAB taxonomy rows with the assistance of two domain experts. Dataset statistics are summarised in Table 4.2.

4.10.2 Training and Hardware

The topic models are trained on an AWS SageMaker `ml.t3.xlarge` instance (4 vCPUs, 16 GiB RAM) running Python 3.x and using libraries including `sentence-transformers`,

Table 4.2: Characteristics of the multimodal video evaluation dataset.

Attribute	Value
Number of videos	135
Number of languages	25
Total video duration	9h 30m 6s
Average duration per video	4m 10s
Number of taxonomies	18
Number of annotated labels	188

`umap-learn`, `hdbscan`, `bertopic`, and `transformers`. Once trained, the models are serialised and reused for inference.

4.10.3 Evaluation Metrics

Because each video may map to several taxonomy labels, Taxonomy predictions are evaluated as a multi-label classification task [235]. Let N denote the number of videos, L the number of labels, y_{ij} the ground truth (0/1) for video i and label j, and $\hat{y}_{ij}$ the prediction.

Hamming Loss

$$\text{HammingLoss} = \frac{1}{NL} \sum_{i=1}^{N} \sum_{j=1}^{L} \mathbf{1}[y_{ij} \neq \hat{y}_{ij}]. \tag{4.6}$$

Subset Accuracy

$$\text{SubsetAcc} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[y_i = \hat{y}_i]. \tag{4.7}$$

Micro-Averaged Precision, Recall, and F1 Let tp_j, fp_j, and fn_j denote true positives, false positives, and false negatives for label j; q is the number of classes.

$$\text{MicroPrec} = \frac{\sum_{j=1}^{q} \text{tp}_j}{\sum_{j=1}^{q} \text{tp}_j + \sum_{j=1}^{q} \text{fp}_j}, \tag{4.8}$$

$$\text{MicroRec} = \frac{\sum_{j=1}^{q} \text{tp}_j}{\sum_{j=1}^{q} \text{tp}_j + \sum_{j=1}^{q} \text{fn}_j}, \tag{4.9}$$

$$\text{MicroF1} = \frac{2 \times \text{MicroPrec} \times \text{MicroRec}}{\text{MicroPrec} + \text{MicroRec}}. \tag{4.10}$$

These metrics are reported for both top-level (T1) and granular (T4) taxonomies.

4.11 Results and Evaluation

4.11.1 Multimodal versus Unimodal Performance

Tables 4.3 and 4.4 present results for unimodal and multimodal configurations at T1 and T4 levels respectively under PaLM 2 and T5-Base topic-explanation settings.

Multimodal fusion improves clearly over unimodal setups, particularly at T1, where PaLM 2 with transcript+visual achieves an F1-score of 0.80 and subset accuracy of 0.66. At T4, the task is more challenging, but fusion still yields the best F1-scores.

4.11.2 Robustness to Modality-Specific Noise

Tables 4.5 and 4.6 report performance under noisy conditions (noisy transcripts, noisy visuals, and both noisy) on a subset of 30 videos.

The results show that the framework degrades gracefully under noise. When at least one modality remains informative, F1-scores are still strong. The most challenging case is when both modalities are noisy, particularly at T4, where fine-grained distinctions require cleaner signal.

4.11.3 Multilingual Performance

Tables 4.7 and 4.8 break down performance for videos with English transcripts versus non-English transcripts.

Performance on non-English videos is slightly lower but remains strong, confirming that multilingual MiniLM embeddings effectively support cross-lingual topic modelling.

4.11.4 Visualisations and Temporal Taxonomy Evolution

To illustrate the dynamic nature of content taxonomies within a video, a sample cooking video is segmented into time windows, and taxonomies are inferred for each segment.

Table 4.3: Top-level (T1) taxonomy inference using PaLM 2 and T5-Base for topic explanations.

Condition	PaLM 2					T5-Base				
	SubsetAcc	HammingLoss	Precision	Recall	F1	SubsetAcc	HammingLoss	Precision	Recall	F1
Transcript only	0.54	0.13	0.57	0.55	0.56	0.53	0.13	0.55	0.53	0.54
Visuals only	0.43	0.15	0.59	0.80	0.68	0.43	0.15	0.59	0.80	0.68
Transcript + Visual	**0.66**	**0.05**	**0.71**	**0.90**	**0.80**	**0.54**	**0.08**	**0.66**	**0.91**	**0.76**

Table 4.4: Granular (T4) taxonomy inference using PaLM 2 and T5-Base for topic explanations.

Condition	PaLM 2					T5-Base				
	SubsetAcc	HammingLoss	Precision	Recall	F1	SubsetAcc	HammingLoss	Precision	Recall	F1
Transcript only	0.21	0.17	0.55	0.31	0.38	0.21	0.15	0.52	0.32	0.36
Visuals only	0.22	0.17	0.62	0.53	0.57	0.22	0.17	0.62	0.53	0.57
Transcript + Visual	0.22	**0.15**	**0.63**	**0.59**	**0.61**	**0.23**	**0.12**	**0.62**	**0.56**	**0.59**

Let the video be partitioned into segments $G_1, \ldots, G_x$, each with its own taxonomy set:

$$\begin{aligned} &\{G_1 : [\mu_{11}T_{11}, \ldots, \mu_{1p}T_{1p}], \\ &G_2 : [\mu_{21}T_{21}, \ldots, \mu_{2p}T_{2p}], \\ &\ldots, \\ &G_x : [\mu_{x1}T_{x1}, \ldots, \mu_{xp}T_{xp}]\}. \end{aligned} \tag{4.11}$$

Figure 4.6 shows how taxonomy frequencies (e.g. "Cooking", "Desserts and Baking", "Healthy Cooking and Eating") change over time, alongside a strip of representative keyframes.

This temporal perspective suggests opportunities for fine-grained ad insertion, where ads for particular categories (e.g. healthy food products) are aligned with corresponding segments in the video.

Table 4.5: T1 taxonomy inference under noisy conditions.

Condition	PaLM 2					T5-Base				
	SubsetAcc	HammingLoss	Precision	Recall	F1	SubsetAcc	HammingLoss	Precision	Recall	F1
Noisy transcript + structured visuals	0.51	0.11	0.70	0.81	0.77	0.47	0.13	0.65	0.78	0.71
Noisy visuals + clean transcript	0.55	0.08	0.68	0.67	0.75	0.50	0.10	0.68	0.62	0.68
Both noisy	0.35	0.19	0.55	0.49	0.53	0.29	0.15	0.60	0.59	0.51

Table 4.6: T4 taxonomy inference under noisy conditions.

Condition	PaLM 2					T5-Base				
	SubsetAcc	HammingLoss	Precision	Recall	F1	SubsetAcc	HammingLoss	Precision	Recall	F1
Noisy transcript + structured visuals	0.23	0.19	0.60	0.52	0.57	0.21	0.26	0.45	0.38	0.46
Noisy visuals + transcript	0.19	0.19	0.56	0.48	0.46	0.13	0.24	0.47	0.42	0.41
Both noisy	0.01	0.29	0.25	0.22	0.28	0.01	0.29	0.22	0.21	0.23

Table 4.7: T1 taxonomy inference for English versus non-English transcripts.

Condition	PaLM 2					T5-Base				
	SubsetAcc	HammingLoss	Precision	Recall	F1	SubsetAcc	HammingLoss	Precision	Recall	F1
English transcripts + visuals	0.69	0.05	0.77	0.91	0.83	0.68	0.07	0.70	0.88	0.81
Non-English transcripts + visuals	0.60	0.08	0.75	0.79	0.77	0.55	0.13	0.69	0.76	0.73

Table 4.8: T4 taxonomy inference for English versus non-English transcripts.

Condition	PaLM 2					T5-Base				
	SubsetAcc	HammingLoss	Precision	Recall	F1	SubsetAcc	HammingLoss	Precision	Recall	F1
English transcripts + visuals	0.27	0.10	0.64	0.66	0.63	0.24	0.12	0.64	0.66	0.63
Non-English transcripts + visuals	0.25	0.12	0.59	0.63	0.56	0.22	0.15	0.55	0.52	0.50

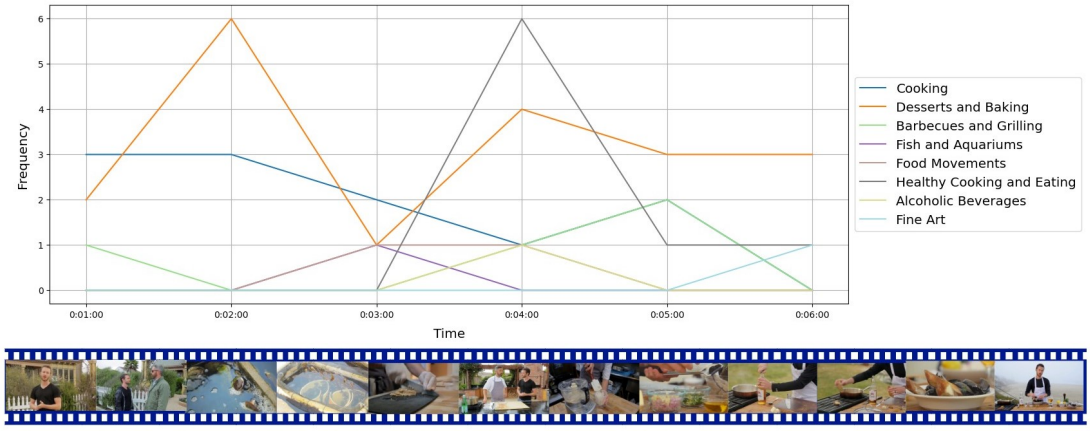


Figure 4.6: Example of taxonomy evolution over time in a cooking video. Top: taxonomy frequency curves over time. Bottom: representative keyframes from the same video.

4.12 Discussion

4.12.1 Strengths and Relationship to Chapter 3

Compared with the framework in Chapter 3, which relies on BoVW labels and word-based topics, the present chapter advances the thesis in several ways:

- **Semantic richness.** CLIP-ViT and multilingual MiniLM embeddings provide deeper semantic representations that capture high-level concepts and cross-lingual equivalences more effectively than BoVW and raw ASR words.
- **Explainability.** LLM-generated topic descriptions create an interpretable layer between embeddings and taxonomies, supporting human auditing, brand safety, and regulatory transparency [209].
- **Robustness.** Multimodal fusion and noise-aware evaluation show that the framework can handle noisy user-generated content and still produce meaningful taxonomy assignments.
- **Standards alignment.** The explicit mapping to hierarchical IAB taxonomies ensures compatibility with existing programmatic advertising workflows [212].

In short, Chapter 3 establishes the feasibility and benefits of multimodal topic–taxonomy mapping; Chapter 4 turns this into an explainable, embedding-rich, and multilingual solution suitable for real-world deployment.

4.12.2 Limitations

Despite its strengths the framework has several limitations:

- **Domain specificity.** Training data and experiments are focused on the food domain; extension to additional domains can be achieved through lightweight embedding adaptation and domain-specific re-clustering.
- **Granular taxonomy performance.** T4 performance is notably lower than T1, especially under noise, reflecting the difficulty of fine-grained topic discrimination and limitations of training data coverage.

- **LLM dependency at training time.** Although the framework relies on the power of LLMs in the offline process of topic description generation, the inference of topics as well as taxonomy mapping is entirely performed by the embeddings, without any need to call the LLM.
- **Temporal granularity and segmentation scope.** The temporal taxonomy experiment in this chapter operates over fixed-duration windows to illustrate semantic evolution across time. It does not attempt content-adaptive or theme-aware boundary detection. Structural segmentation based on global thematic continuity is introduced separately in Chapter 5. Thus, Chapter 4 establishes time-indexed semantic tracking rather than narrative-level segmentation.

4.12.3 Practical Considerations

From an engineering perspective, the framework is computationally feasible on modest cloud hardware, thanks to compact sentence embeddings and keyframe sampling. Deployment requires integration with existing video processing pipelines (for FFmpeg and Whisper) and with ad-tech components that consume IAB taxonomies. Because the system does not depend on user identifiers, it is naturally aligned with privacy-by-design principles.

Deployment and Domain Adaptability. The proposed framework is architecturally modular and supports efficient domain adaptation without requiring full end-to-end retraining. Because semantic representation is embedding-driven and clustering-based, new domains can be incorporated through lightweight embedding replacement or fine-tuning, guided topic injection, and re-clustering of domain-specific corpora. This enables incremental vertical expansion (e.g., finance, health, automotive) while preserving the overall system architecture and taxonomy alignment mechanisms. As a result, the framework is operationally deployable in real-world ad-tech environments where content categories evolve over time and rapid adaptation is required without heavy computational overhead.

4.13 Summary

The explainable multimodal topic modelling approach outlined in this chapter enables the mapping of video content to content taxonomies that conform to the industry standard. The approach uses CLIP-ViT embeddings, image captioning tasks, multilingual MiniLM embeddings, BERTopic, and topic descriptions provided by an LLM.

Empirical results on a food-domain subset of YouTube-8M show that multimodal fusion clearly outperforms unimodal baselines, that the system is robust to noise and multilingual transcripts, and that it can capture temporal evolution of taxonomies within videos. These results further develop the core multimodal taxonomy mapping framework of Chapter 3 and move the thesis closer to deployable, industry-grade contextual advertising solutions.

Chapter 4 directly addresses Research Question 1 by demonstrating how multimodal topic models can be designed to produce embedding-rich, interpretable, and taxonomy-aligned representations. Robustness is operationalised here in three specific dimensions: (1) multimodal robustness through fusion-based improvements over unimodal baselines, (2) resilience to modality-specific noise, particularly ASR degradation, and (3) multilingual stability through MiniLM embeddings.

Although experiments are conducted within a specific domain, the embedding-based and taxonomy-aligned architecture is inherently domain-agnostic, enabling lightweight knowledge extension without structural modification, the present framework establishes the architectural and methodological foundations for robust, interpretable, taxonomy-aligned multimodal topic modelling within a controlled evaluation setting. These foundations are subsequently extended in Chapters 5 and 6 through narrative and cognitive integration.

In Chapter 5, the existing foundations are strengthened to investigate theme-aware segmentation, which explores the application of the proposed techniques in an end-to-end advertising decision pipeline.

Chapter 5

Proposed Multimodal Theme-Aware Segmentation Framework

5.1 Introduction

The methodological contribution of this thesis is a multimodal, theme-aware segmentation framework designed to support adaptive and minimally disruptive mid-roll advertisement insertion in long-form video content. While commercial platforms increasingly rely on automated content analysis, existing segmentation technologies remain predominantly focused on low-level visual change detection or local linguistic cues (see § 2.4.1). Such systems are effective at identifying camera cuts or lexical topic shifts, but they do not model how *themes* evolve across a video narrative, nor do they reason about whether a detected transition represents a momentary fluctuation or a meaningful, sustained change in story direction. These shortcomings make contemporary approaches unsuitable for applications such as adaptive advertising in which segment boundaries must be both semantically justified and narratively non-disruptive (see § 1.2).

The methodology presented in this chapter addresses this gap by framing segmentation as a *reasoning* problem rather than a purely perceptual one. The central insight is that the insertion of advertisements in long-form videos (e.g., podcasts, documentaries, instructional media, or vlogs) requires identifying *semantic story units* (addressing the Narrative Objective in § 1.5): segments that maintain internal coherence and correspond

to stable thematic regions within the narrative structure. Unlike domain-specific advertising heuristics that rely on silence duration, black frames, or hand-designed thresholds (as discussed in § 2.4.1), the proposed framework constructs a multimodal representation of each part of the video integrating visual content, linguistic structure, speaker information, and temporal continuity and then applies a constrained Tree-of-Thoughts (ToT) reasoning [236] procedure to determine which transitions are sufficiently strong and sufficiently sustained to justify placement of an advertisement.

The chapter proceeds by decomposing the system into three conceptual layers: (1) a *multimodal perception layer*, which extracts visual scene boundaries, produces rich keyframe descriptions, and generates word-level ASR transcripts with diarization; (2) a *structural alignment layer*, which reconciles inconsistent temporal units by merging shots only when supported by bridging audio, forming aligned audio–visual windows that preserve the editorial integrity of both perceptual and linguistic cues; and (3) a *reasoning layer*, which examines window-level semantic summaries and applies a constrained ToT mechanism to differentiate momentary topic oscillations from stable thematic transitions.

This decomposition reflects the dual nature of the task. On one hand, video segmentation requires grounding in perceptual modalities, since themes are often signalled through changes in scene, environment, or speaker. On the other hand, determining whether a segmentation point is *appropriate for advertisement insertion* requires contextual reasoning: the system must not only detect change but also justify that the change marks a meaningful narrative boundary. The constrained ToT framework introduced in this thesis is therefore essential, as it allows the model to evaluate multiple reasoning paths, enforce stability criteria, and prioritise "fewer, stronger" breaks that preserve the viewer's cognitive experience (an aspect further explored in Chapter 6).

A high-level overview of the entire pipeline is provided in Figure 6.1. This diagram emphasises how each architectural stage contributes to the overall goal: perceptual analysis constructs the raw multimodal substrate; structural alignment ensures consistency across modalities; and the reasoning layer imposes thematic coherence and interpretable decision-making. In the sections that follow, each of these stages is examined in detail,

including the theoretical motivation for the design choices, implementation considerations, and their relationship to the research questions posed in Chapter 1.

Overall, this methodology chapter in particular presents not only an explanation of how the system is set up to function but also a rationale with regard to just which architectural elements are in fact required to be included in it in order to meet a particular need and provide a particular benefit. Such integration of multimodal perception and LLM-based reasoning leads to a more stable and interpretable segmentation aligned with narratives through the proposed framework.

Chapters 3–4 provide the semantic substrate, and Chapter 5 introduces the reasoning and structural modelling required for narrative segmentation suitable for adaptive media applications.

5.2 Motivation and Design Requirements

The design of the proposed segmentation framework is driven by the core research challenge identified in this thesis: the need for *reliable*, *interpretable*, and *narratively coherent* segmentation of long-form video for adaptive advertisement insertion (formalised in the Problem Statement in § 1.2). Existing approaches in video analysis, including scene-cut detection, transcript-based segmentation, and embedding-driven clustering, are highly effective for their respective tasks (as reviewed in § 2.4.1) but remain insufficient for contexts in which segment boundaries must correspond to meaningful story units that preserve viewer engagement. This section formalises the motivation behind the framework and identifies the design requirements that shaped the methodological choices developed in later sections (which directly address the Research Questions posed in § 1.6).

5.2.1 Limitations of Contemporary Segmentation Approaches

Traditional video segmentation pipelines predominantly rely on low-level perceptual cues such as shot boundaries, colour histogram differences, or abrupt differences in motion vectors.

Tools such as PySceneDetect [237] provide accurate shot detection, a shot boundary does not necessarily correspond to a meaningful thematic transition.

In many forms of content (e.g., interviews, podcasts, educational videos), shots may change while the underlying topic remains constant. Conversely, significant thematic transitions may occur without any obvious visual change, such as when a speaker introduces a new subject or shifts from exposition to demonstration. Visual-only segmentation therefore fails to capture the semantic structure required for effective advertisement placement.

Linguistic segmentation methods, such as TextTiling [238], TopicTiling [239], or more recent embedding-based approaches [240], focus primarily on lexical cohesion but often ignore visual cues and speaker structure. Moreover, automatic speech recognition (ASR) transcripts frequently contain disfluencies, filler words, and transient topic fluctuations, which further complicate coherence-based segmentation.

In long-form content, reliance on lexical cohesion alone tends to produce excessive fragmentation, many of which reflect momentary digressions rather than genuine narrative breaks.

Recent LLM-based chapter-generation systems such as Vid2Seq [13] or Chapter-Llama [14] have shown promise in modelling higher-level narrative structure. However, these approaches typically require extensive fine-tuning, impose substantial computational costs, and remain inherently stochastic producing varying outputs across runs on identical inputs. Such instability is incompatible with large-scale advertising pipelines, where decisions must be reproducible, explainable, and consistent over time.

A core limitation shared by most existing techniques is their heavy dependence on local similarity within small temporal windows. While effective at detecting abrupt perceptual changes (e.g., shot cuts or silence), these methods rarely account for the high-level semantic and thematic development of the content. In practice, natural story boundaries often occur at points of thematic completion rather than at locations of maximum perceptual difference.

As a result, such approaches commonly suffer from over-segmentation of coherent narrative units or under-segmentation of semantically significant transitions.

Collectively, these limitations highlight the need for a segmentation approach that moves beyond local perceptual and lexical cues toward theme-level understanding grounded in stable, multimodal reasoning.

5.2.2 Adaptive Advertising as a Context with Distinct Constraints

The problem of advertisement insertion introduces specific constraints that are not present in general video segmentation tasks. Most critically:

1. **Viewer Experience Constraint:** The ads must be placed at locations of least disruption to the cognitive processes. Discontinuities and unjustified breakpoints have been found to be damaging to satisfaction and retention levels, as well as to perceptions of program quality.

2. **Narrative Continuity Constraint:** Segments need to conform to the narrative rhythm of the video. The point at which a segment ends should relate to a point of narrative resolution or transition, not just some change in imagery or linguistic pattern.

3. **Semantic Coherence Constraint:** Each segment must have an intelligible topic or sub-topic. Incomprehensible portions, such as incomplete explanations or sentences that have been interrupted, lead to substandard advertising and make the segmenting unintelligible.

4. **Reproducibility and Stability Constraint:** Segmentation must be consistent across runs, devices, and content ingestion pipelines. Stochastic LLM outputs are therefore insufficient; the reasoning mechanism must be tightly constrained.

5. **Low-Latency and Low-Cost Processing Constraint:** Commercial pipelines may handle thousands of videos per hour. For the model to be economical, it definitely cannot involve extensive fine-tuning or multi-pass generative decoding.

These constraints collectively define a methodological space in which neither purely perceptual nor purely linguistic nor purely LLM-driven approaches are adequate. A

hybrid architecture is necessary one that fuses multimodal signals while constraining LLM reasoning to produce stable, interpretable outputs.

5.2.3 Requirements for a Theme-Aware Segmentation Framework

To satisfy these constraints, the methodology must fulfil the following requirements:

R1: Multimodal Integration The system should also be able to handle both visual and linguistic information, as changes in the themes of the content of the video are often indicated by the interplay of both visual and linguistic information. For instance, the speaker could be in the same location while discussing different topics, or the location could change while the speaker is still discussing the same topic.

R2: Alignment Across Modalities Visual and linguistic units must be merged into consistent windows that do not fragment meaningful semantic units. For example, shot boundaries should not split spoken sentences, and long monologues should not be arbitrarily divided by minor visual changes. An alignment mechanism is required to reconcile these modal inconsistencies.

R3: Stable and Interpretable Reasoning The system must reason about narrative flow in a manner that is stable, consistent, and explainable. While LLMs are powerful, unconstrained generative reasoning introduces variability. A constrained Tree-of-Thoughts (ToT) procedure not a free-form generative model is required to ensure consistency and interpretability.

R4: Sensitivity to *Sustained* Theme Changes Segmentation must be triggered only at transitions that represent a genuine, ongoing shift in theme. Momentary digressions, rhetorical questions, or short side comments must not cause segmentation. This necessitates explicit lookahead and persistence criteria.

R5: Low Inference Cost The pipeline must rely on lightweight reasoning, reusing perceptual information where possible, and avoiding high-dimensional embeddings or large transformer models that must be fine-tuned or sampled repeatedly.

R6: Operational Suitability for Advertising Pipelines Outputs must be stable across runs, use constrained schemas, and require no post-hoc filtering. The system design must enable auditing (why a boundary was chosen) and deterministic behaviour under fixed prompts and models.

5.2.4 Justification for a Hybrid Perception–Reasoning Architecture

These needs lead to the natural design of the three-layer architecture, described in this chapter:

1. **A Multimodal Perception Layer** provides grounded, factual descriptions of the video's visual and linguistic structure without interpretation.
2. **A Structural Alignment Layer** builds the window structures to maintain the integrity of the visual cuts and the linguistic units, thereby producing meaningful atomic units for reasoning.
3. **A Constrained Reasoning Layer** employs thematic ToT reasoning to establish if the transition between windows signals corresponds to a meaningful thematic shift that is appropriate for advertisement insertion.

This decomposition helps a system leverage the best parts of Data-driven perception (robustness, determinism, modality-specific fidelity) combined with the advantages of LLM-based reasoning (semantic abstraction, theme identification, and interpretability).

5.2.5 Connection to the Thesis Research Questions

The methodology explicitly tackles the primary research questions identified in Chapter 1:

- **RQ1:** How can thematic transitions in long-form video be identified in a manner that is stable, interpretable, and computationally lightweight?
- **RQ2:** How can multimodal cues, visual setting, linguistic structure, and speaker information be integrated to produce coherent story units?

- **RQ3:** Can constrained LLM reasoning provide a practical and reliable alternative to heavy embedding-based or fine-tuned models for segmentation?
- **RQ4:** How can a segmentation framework be designed to support real-world, large-scale adaptive advertising applications?

The rest of this chapter will elaborate these requirements in concrete methodological terms, demonstrating how each can be made to rest upon empirical constraints, theory, and the affordances of modern multimodal AI systems.

5.3 Architectural Overview

The proposed framework adopts a modular, multimodal architecture designed to produce stable, reproducible, and contextually justified segmentation boundaries suitable for adaptive advertising workflows. As illustrated in Figure 5.1, the overall system integrates four major subsystems *perception and alignment*, *structural windowing*, *controlled multimodal reasoning*, and *schema-guided decision outputs* to model narrative flow using a blend of algorithmic determinism and constrained LLM-based inference. For clarity, the perception and alignment components together constitute the structural preprocessing layer described earlier in Section 5.2.4. Whereas existing LLM-driven chaptering systems typically rely on stochastic sampling or multi-pass generation, the architecture developed in this thesis emphasises *controlled inference*: every reasoning step is performed under fixed prompts, fixed temperatures, and a single-pass reasoning structure. This ensures that the system exhibits high operational consistency and can be deployed reliably across large-scale content libraries where reproducibility and interpretability are essential.

The framework's architectural design stems directly from the requirements articulated in earlier chapters: (i) segmentation must preserve thematic continuity to prevent conflicting ad placements; (ii) segmentation must remain aligned to linguistic units to avoid disrupting sentence-level semantic flow; and (iii) segmentation must be resource-efficient and model-agnostic to support in consumer-media environments. In traditional video segmentation, the pipelines that use low-level features often fail to produce bound-

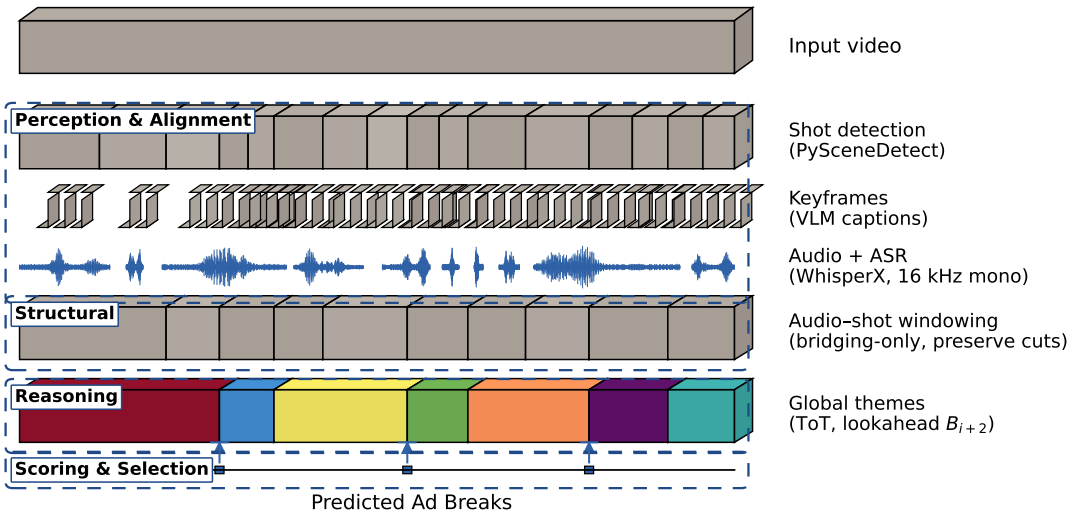


Figure 5.1: Overall architecture of the proposed theme-aware multimodal segmentation framework. The pipeline integrates deterministic perception, structural alignment, and constrained LLM reasoning to produce stable, interpretable, and semantically coherent story units.

aries that correspond to the narrative, although they correspond to the transitions in the video [237, 241, 242].. This is because these pipelines do not capture semantic flow, which is important in the insertion of advertisements.

Conversely, purely transcript-driven segmentation methods, including classical lexical cohesion models such as TextTiling and its descendants, may overlook visual continuity and often fragment scenes that remain coherent at the visual or narrative level [238, 239, 243]. Similarly, multimodal embedding-based temporal clustering approaches such as those that apply CLIP to align visual and textual features can produce unstable boundaries due to embedding drift, context-blind similarity thresholds, or insufficient modelling of narrative progression [240]. More recently, chapter-generation systems using large language models (LLMs), including VidChapters-7M and Vid2Seq [13, 244], together with model-specific variants such as Chapter-Llama [14], have demonstrated impressive semantic capabilities but remain dependent on unconstrained generative decoding. As a result, they tend to over-segment, produce hallucinated boundaries, or exhibit inter-run variability limitations directly at odds with the operational consistency required in commercial ad-placement scenarios.

The proposed architecture addresses these issues by means of a sound decomposition of the segmentation process based on high-level reasoning. The initial component of this decomposition is *perception and alignment*, which provides complementary cues from both visual and auditory media. Shot boundaries can then be detected based on deterministic visual reasoning and then aligned with the helps of WhisperX. Audio can then be allied with WhisperX for transcription at a word level and on a global temporal basis to help firmly ground linguistic alignment with visual alignment on a temporal basis.

The second subsystem carries out *audio-shot structural windowing*, which is a crucial task in maintaining continuity of narrative flow in cinematographic cuts [245]. It is often the case that cuts at shot boundaries emerge due to standard cinematographic editing patterns that do not correlate with the theme of progression. Thus, directly applying cuts at shot boundaries as segment boundaries can lead to breaking down linguistic elements. To counter this issue, structural windowing combines two sequential shots only if uninterrupted speech passes through the boundary.

The third subsystem constructs a compact multimodal textual representation for each window through *window-level re-summarisation*. Each window's content is summarised by combining keyframe-derived visual descriptions and ASR-derived linguistic content into a single, text-only synopsis. Importantly, this summarisation operation is executed using low-variance, zero-temperature LLM inference under rigid prompt constraints. The LLM is used strictly as a compression mechanism never as an open-ended generator thereby ensuring stable interpretations that are reproducible across runs. This step abstracts local content while retaining thematic signals necessary for downstream reasoning.

Finally, the system uses *controlled multimodal reasoning* to detect the boundaries of the theme shift. The reasoning module uses a controlled, one-pass version of the Tree-of-Thoughts (ToT) reasoning paradigm [236]. Unlike the full-fledged ToT reasoning frameworks, where the LLM performs multi-stage expansion, sampling, and ranking on a massive search space, in this framework, the LLM's search space is restricted to a very small number of candidate thoughts, with the selection of only the best among

them done in the same function call. The reasoning process has a fixed lookahead window to ensure that a putative transition between a pair of windows does not constitute a brief excursion into a secondary topic, but rather a period of sustained theme change, a criterion that fundamentally underlies the suppression of spurious segmentations and the marking of significant changes in the narrative, suitable for the insertion of advertisements.

Together, these four subsystems form a coherent, pipeline-level architecture that captures narrative hierarchy from low-level visual transitions to high-level thematic structure while maintaining strict operational reproducibility. By treating segmentation as a temporally anchored, multimodal reasoning process rather than a purely visual, textual, or generative modelling task, the proposed architecture offers a practical, interpretable, and commercially viable solution for adaptive ad break placement in long-form consumer video content.

5.4 Perception and Alignment

The perception and alignment subsystem constitutes the foundation of the proposed segmentation framework. Its primary role is to transform the raw video stream into a set of temporally grounded, multimodal representations that can later support controlled reasoning over thematic structure. As emphasised in Section 5.3, this subsystem is intentionally designed to be algorithmic, stable, and free from stochastic generative behaviour. By establishing a common temporal reference across the visual and auditory modalities, all subsequent operations - structural windowing, window-level abstraction, and theme-based reasoning inherit a consistent representational structure. This ensures that segmentation decisions are traceable to observable cues rather than generative variance.

From a conceptual standpoint, the perception and alignment stage consolidates two independent but complementary processes: (i) *visual scene detection and summarisation*, which encodes cinematographic structure and salient visual semantics, and (ii) *audio transcription and linguistic alignment*, which anchors dialogue and narration to precise timestamps. Together, these processes address fundamental limitations of prior

segmentation approaches. Classical visual methods, for instance, detect abrupt transitions but do not capture narrative intent [241, 242]. Conversely, transcript-only segmentation techniques such as TextTiling [238] and TopicTiling [239, 243] track lexical cohesion but ignore cinematic boundaries and visual context. The purpose of the multimodal perception layer is thus not only to extract these cues but also to maintain their distinct contributions in a temporally aligned form, creating the conditions under which controlled LLM reasoning can reliably infer thematic structure.

5.4.1 Scene Detection and Visual Summarisation

The visual analysis process starts with the deterministic detection of shot boundaries through the use of **PySceneDetect**, a technique which ensures a consistent and repeatable segmentation of the video into meaningful visual chunks. The theory behind PySceneDetect is the shot boundary detection theory where scene changes are detected by the visual difference between successive frames, thereby providing a reliable low-level structural signal for subsequent narrative segmentation. The decision to use a deterministic, non-learning-based detector is deliberate: shot boundaries must act as anchors for multimodal alignment, and their stability is essential for repeatable downstream reasoning. In contrast to learning-based alternatives, which may introduce variability due to training data dependence, PySceneDetect offers consistent performance without requiring labelled data, making it suitable for large-scale and domain-agnostic deployment within the proposed modular framework. Each detected shot C_i forms the basis of a hierarchical visual representation composed of two levels of abstraction: *keyframe descriptions* and *scene summaries*.

For each shot, a predefined set of "unique keyframes" is extracted using fixed selection heuristics to avoid redundant queries. Each keyframe $f_m \in C_i$ is then processed by a vision–language model, queried under a rigid prompt template to produce a structured textual description d_m. The prompt design emphasises objective elements (setting, characters, visible objects, on-screen text, and mood), ensuring that the LLM acts as a controlled visual captioning engine rather than an imaginative generator. The resulting

ordered set of per-frame descriptions

$$D_i = \{d_m\}_{m=1}^{M_i}$$

preserves visual chronology without using embedding similarity or clustering, thereby avoiding the variability often observed in embedding-based multimodal segmentation methods [240].

A second abstraction level condenses the above-described information into a compact *scene summary* S_i, extracted using a text-only function call with the same LLM with a separate prompt template, in a zero-temperature and low-variance regime. This summarization is in line with the controlled inference approach. It omits the redundant information present in the images while retaining the relevant information required in the reasoning of the subsequent levels. The dual-level summarization enables the encoding of a semantically rich compact representation.

5.4.2 Audio Processing, Transcription, and Linguistic Alignment

The auditory channel offers critical semantic information - dialogue, narration, speaking turns that is often orthogonal to visual variation. Audio-driven segmentation methods such as silence detection [47] have historically been used in advertising systems, but silence alone is not a reliable proxy for thematic transitions. The goal here is therefore to extract not mere acoustic patterns but precise linguistic units anchored to the video timeline.

The audio is processed using **WhisperX**, which was selected on the basis of its ability to generate word-level timestamps with low drift even on very long videos. The transcription service processes the audio to create a unified timing reference by converting it to a mono, 16 kHz PCM stream. The words are represented as a tuple

$$w_\ell = (\text{text}_\ell,\, s_\ell,\, e_\ell,\, q_\ell)$$

and all timing offsets are resolved. This word-level representation allows the video shots to be aligned with the text perfectly and avoids boundary misalignment that would

happen with approximate sentence segmentation based on chunk-level timestamps.

Sentences (utterances) are constructed by grouping consecutive words into meaningful linguistic units with start and end times defined. Unlike embedding-based textual similarity approaches, which may be sensitive to lexical variation or paraphrasing, word-level alignment provides a deterministic structure that is both human-interpretable and robust to minor transcription errors.

To further enrich the multimodal representation, speaker diarisation identifies speaker turns $D_k = (\tau_k^s, \tau_k^e, \text{spk}_k)$ and assigns speakers to words through temporal overlap. This enhances the representation of conversational structure useful, for instance, when thematic boundaries coincide with speaker changes or discourse shifts.

The result of the audio processing chain is a structured sequence of sentences with annotated text, temporal cues, speaking identity, as well as word-level metadata. Such representation is vital in the subsequent process of structural windowing. The system guarantees thematic segmentation that splits sentences or merges independent discourse into self-contained chunks by correlating sentences with video shot boundaries.

5.4.3 Cross-Modal Alignment as a Foundation for Controlled Reasoning

The final stage of perception and alignment is the temporal integration of the visual and auditory streams. By placing keyframe-derived summaries and ASR-derived utterances on the same timeline, the system produces a unified multimodal substrate that enables reasoning about narrative continuity. Unlike prior chaptering systems that rely on unconstrained LLM inference [13, 14], the present framework isolates all generative operations to a later, controlled reasoning module, ensuring that no boundary decisions are influenced by stochastic generation or hallucinated content.

This multimodal alignment is more than a preprocessing step: it is a design principle. By ensuring that all subsequent reasoning operates on temporally grounded, non-stochastic representations, the system achieves operational consistency and supports transparent, analytically traceable segmentation. The representational structure produced here becomes the backbone for the structural windowing procedure in Sec-

tion 5.5, which merges shots based on linguistic continuity and thereby preserves the integrity of narrative flow.

5.5 Audio–Shot Structural Windowing

A key challenge in multimodal video segmentation lies in the mismatch between cinematographic structure and narrative structure. Shot boundaries often reflect editorial conventions, camera angle changes, cutaways, or B-roll insertion rather than genuine thematic transitions. Many classical segmentation methods that rely primarily on visual change detection [241, 242] therefore produce overly fragmented units that fail to align with linguistic or conceptual coherence. Conversely, linguistically driven approaches [238, 239] may entirely ignore visual edit structure, resulting in semantically coherent but visually implausible segments. For systems aiming to support adaptive advertising, neither extreme is acceptable: a segmentation must respect editorial boundaries while preserving semantic integrity, ensuring that ad breaks occur at points consistent with both narrative flow and visual rhythm.

The objective of this section is to establish a *structural* segmentation layer that precedes all high-level reasoning. Specifically, a set of temporally contiguous windows $\{W_k\}$ is constructed by merging consecutive shots only when there is explicit evidence of linguistic continuity across their boundary. This process, denoted *audio–shot structural windowing*, acts as a corrective mechanism that compensates for the visual fragmentation induced by editing while avoiding the pitfall of merging visually unrelated scenes. The resulting windows become the basic units for window-level summarisation and sustained-theme reasoning in later stages of the pipeline.

The deterministic structural consolidation procedure that constructs these windows prior to reasoning is detailed in Appendix A.1. The segmentation outputs produced by the proposed model are subsequently validated and normalised using a deterministic post-processing procedure, described in Appendix A.3.

5.5.1 Motivation and Design Rationale

The motivation for windowing based on structure relies on the fact that the relevant continuity in video content is most evident in language. Speech tends to be in a slower tempo compared to the video cuts, while the audio transitions typically signal changes in topic, speaker intention, or semantic highlighting. The proposed windowing method relies on *Active Speech Continuity* as a mean to derive meaningful flows of narratives around video cuts, as opposed to methods based on silence, proposed by Mei et al. in [47].

Formally, let the video be segmented into N visually defined shots $\{C_i\}_{i=1}^{N}$ and let the ASR pipeline output R atomic audio units $\{A_r\}_{r=1}^{R}$ (typically corresponding to sentence-level utterances). Each shot C_i and audio unit A_r is associated with a temporal span on a common timeline. A window W_k is defined as an ordered union of consecutive shots that together form a linguistically unified region. This definition ensures that the segmentation never violates visual edit boundaries while avoiding misalignment with ongoing linguistic content.

From a systems perspective, this step creates the *intermediate temporal granularity* that the ToT-based reasoning module requires. Windows are long enough to encode thematic evidence and short enough to support meaningful forward lookahead. Without this structural pass, the reasoning layer would be forced to operate either at too fine a granularity (leading to over-sensitivity to local visual noise) or too coarse a granularity (leading to lost detail and poor interpretability).

5.5.2 Bridging Audio Across a Cut

The fundamental operation in structural windowing is the detection of "bridging" speech-speech segments whose temporal span overlaps a shot boundary. Such bridging segments indicate that the narrative content persists across edits, implying that the shots belong to a single semantic unit.

Let the temporal spans of shots and audio segments be

$$\text{span}(C_i) = [c_i^-, c_i^+], \qquad \text{span}(A_r) = [a_r^-, a_r^+],$$

with all times referenced to the common 0-based ASR timebase. A small temporal tolerance $\varepsilon > 0$ accounts for decoding jitter.

For the boundary between consecutive shots C_j and C_{j+1}, the set of *bridging* audio units is defined,

$$\mathcal{B}(j) = \Big\{ A_r \;\Big|\; a_r^- \le c_j^+ + \varepsilon \;\wedge\; a_r^+ \ge c_{j+1}^- - \varepsilon \Big\}. \tag{5.1}$$

If $\mathcal{B}(j)$ is non-empty, it is inferred that continuous speech crosses the boundary, and therefore that the two shots likely belong to the same narrative segment. If $\mathcal{B}(j)$ is empty, the boundary is treated as a candidate structural break.

This bridging-based rationale is more linguistically grounded than similarity- based methods: it does not require embeddings, lexical cohesion metrics, or topic models, and therefore avoids the instability often reported in embedding- driven segmentation literature [240]. It relies purely on observed continuity of human speech a stable, interpretable, and cross-modal cue.

5.5.3 Greedy Window Construction

Windows are constructed by scanning the shot sequence from left to right and greedily merging shots whenever bridging speech is detected. For window index k, the following is defined:

$$W_k = \bigcup_{i=i_k}^{j_k} C_i, \qquad \text{span}(W_k) = [c_{i_k}^-,\, c_{j_k}^+].$$

The greedy growth rule is:

$$\text{merge } C_j \to C_{j+1} \quad \Longleftrightarrow \quad |\mathcal{B}(j)| > 0. \tag{5.2}$$

The scan terminates when a cut without bridging speech is encountered, ensuring that each window is a maximal region of audio-supported continuity. The resulting

sequence of windows is:

- **disjoint** by construction, - **complete**, covering all shots, - **editorially consistent**, never subdividing a shot, - **linguistically grounded**, preserving ongoing speech units.

This process creates natural demarcations of narrative flow that are far better aligned with human perception of story transitions than raw shot boundaries.

5.5.4 Window–Audio Association

Each finalised window W_k is associated with all audio segments that overlap its temporal span:

$$A(W_k) = \Big\{ A_r \;\Big|\; a_r^+ \geq c_{i_k}^- - \varepsilon \;\wedge\; a_r^- \leq c_{j_k}^+ + \varepsilon \Big\}. \tag{5.3}$$

This mapping ensures that the subsequent window-level summarisation step in Section 5.5.2 receives all spoken content relevant to the window, even if some utterances begin slightly before or end slightly after the visual window's boundaries.

5.5.5 Properties, Variants, and Theoretical Considerations

The structural windowing procedure satisfies several desirable properties:

1. **Multimodal compatibility**. Windows are aligned with visual edits but grounded in linguistic continuity, enabling coherent multimodal summarisation.

2. **Stability and reproducibility**. Because merging decisions depend purely on deterministic timing comparisons, the windowing behaviour is fully reproducible across runs.

3. **Interpretability**. Each merge decision can be explained in terms of overlapping utterances, providing an immediate human-interpretable justification.

4. **Narrative correspondence**. Windows approximate *micro-themes*: they are longer than shots but shorter than full story arcs, providing the ideal granularity for ToT-based reasoning.

A variant of this method replaces bridging with a silence-based threshold, merging shots whenever the inter-shot gap is below a fixed silence duration δ. While heuristically useful in some domains, this approach is not adopted, as it lacks the semantic grounding provided by active speech continuity.

Overall, the structural windowing module establishes the intermediate temporal units on which the controlled multimodal reasoning procedure in Section 5.6 operates. Without this layer, the reasoning system would have to interpret raw shots too fine-grained for sustained-theme inference or rely on long transcript blocks that obscure visual edit structure. The windowing module thus forms an essential bridge between perceptual stability and thematic reasoning.

5.6 Constrained Multimodal Reasoning for Theme Segmentation

Having established a temporally aligned and structurally coherent representation of the video through perception, alignment, and audio–shot windowing (Sections 5.4–5.5), the next objective is to identify *thematic* boundaries points in the narrative at which the underlying topic, intent, or semantic structure undergoes a meaningful shift. Traditional segmentation methods that rely on visual transitions, lexical cohesion, or topic modelling often treat boundaries as a purely statistical phenomenon, detecting changes in appearance or local word distributions [238, 239]. However, thematic change in long-form video content - documentaries, vlogs, lectures, how-to videos rarely coincides with abrupt lexical or visual changes. Instead, themes tend to evolve gradually, with smooth transitions that require reasoning over a span of multiple windows.

To address this challenge, the proposed framework introduces a *controlled multimodal reasoning module* that uses a constrained variant of Tree-of-Thoughts (ToT) reasoning [236] to infer sustained thematic transitions. Unlike unconstrained LLM-based approaches, which can suffer from stochastic variation, hallucination, or over-segmentation [13,14], this module enforces operational consistency through fixed prompts, zero-temperature decoding, and bounded reasoning paths. The result is a segmentation

strategy that provides coherent narrative breaks appropriate for adaptive advertising while maintaining interpretability and reproducibility.

All constrained Tree-of-Thought reasoning in this section assumes fixed reasoning units produced by the deterministic structural consolidation procedure documented in Appendix A.2.

5.6.1 Motivation: Why Reasoning (Not Embeddings) Is Required

Prior chaptering and segmentation approaches can be broadly categorised into two main families.

- **Embedding-driven methods**, which compute similarity between adjacent chunks using models such as CLIP or language embeddings [240]. While computationally efficient, these techniques are highly sensitive to noise, stylistic variation, and minor lexical or visual fluctuations. As a result, they frequently suffer from over-segmentation and struggle to maintain sustained thematic continuity over longer sequences [246].
- **Generative LLM-based methods**, such as Vid2Seq, VidChapters-7M, or Chapter-Llama [13,14,244], which produce chapter boundaries through open-ended generation. However, these models typically rely on stochastic or maximisation-based decoding strategies (e.g., sampling or beam search), which have been shown to produce unstable and degenerate outputs in open-ended settings [247]. Consequently, the boundaries generated can be semantically reasonable but non-deterministic, reducing suitability for commercial ad placement pipelines requiring stable outputs.

A fundamental limitation shared by both families is the absence of an explicit mechanism for assessing whether a theme *persists* across multiple temporal windows. Embedding-based similarity can be disrupted by transient changes, while generative approaches risk reflecting model biases rather than observable multimodal evidence.

To address these shortcomings, the proposed framework introduces a *sustained theme requirement*. This criterion stipulates that a thematic transition is only accepted when validated against both the immediate successor window and a short look-ahead

horizon. Such bounded temporal validation aligns naturally with constrained Tree-of-Thoughts (ToT) reasoning, which enables deterministic, explainable, and reproducible inference while remaining grounded in multimodal observations.

This controlled reasoning approach not only ensures output stability but also improves segmentation quality by balancing robustness against local noise with sensitivity to genuine thematic shifts, leading to boundaries that align more closely with human annotations.

5.6.2 Objective of the Reasoning Module

Let the sequence of multimodal windows produced earlier be

$$\{B_1, B_2, \ldots, B_N\},$$

where each block B_i contains:

$$B_i = \big(W_i,\ U_i^{\text{aud}},\ \widehat{S}_i\big),$$

including (i) the window span W_i, (ii) the concatenated transcript text U_i^{aud}, and (iii) the window-level multimodal summary$\hat{S}_i$.

The goal of the reasoning module is to determine whether the boundary at (B_i, B_{i+1}) corresponds to a *sustained thematic transition*. Formally:

> *A boundary is accepted only if* (1) *there exists a thematic change between* B_i *and* B_{i+1}*, and* (2) *the new theme persists into* B_{i+2} *(when available).*

This rejects short-lived topical fluctuations common in expository or conversational video content and yields fewer but more meaningful boundaries.

5.6.3 Global Theme Conditioning and Implicit Assignment

To support robust breakpoint reasoning, each video is associated with a fixed set of Global Themes $G = \{g_1, \ldots, g_K\}$, representing coarse-grained semantic categories corresponding to the major narrative phases of the video. These themes are derived in a

prior preprocessing stage from full-video context and remain fixed during batch-level evaluation. Typical cardinality satisfies $3 \leq K \leq 7$.

During breakpoint evaluation, the LLM is prompted to reason over each window block B_i as belonging to one of the provided Global Themes. This induces an implicit theme assignment function

$$\pi : \{B_1, \ldots, B_N\} \to G,$$

where $\pi(B_i)$ denotes the Global Theme contextually associated with window B_i during reasoning. The mapping is not externally stored as a segmentation output; rather, it serves as a structured semantic anchor within the prompt to guide local breakpoint decisions.

By limiting the reasoning to movements between these major themes, the model is less sensitive to subtopic fluctuations that are characteristic of conversational and expository texts. The Global Themes function as stable narrative containers that permit the subsequent decision rule to capture significant thematic changes as opposed to within-theme variations.

5.6.4 Continuity Scoring with Lookahead

To determine whether a candidate boundary between blocks B_i and B_{i+1} constitutes a sustained thematic transition, the reasoning module evaluates a local temporal neighbourhood centred on the boundary.

Specifically, the LLM is provided with up to six adjacent blocks:

$$\{B_{i-3}, B_{i-2}, B_{i-1}, B_i, B_{i+1}, B_{i+2}\},$$

when available within the valid range $[1, N]$.

This sliding-window context enables the model not only to detect a potential global theme change between B_i and B_{i+1}, but also to assess whether the new theme persists into subsequent blocks.

The inclusion of the lookahead block B_{i+2} is particularly important for distinguishing

genuine thematic transitions from temporary digressions or subtopic fluctuations.

Within the prompt, the model generates a small number (typically two to three) of constrained breakpoint hypotheses and evaluates them before producing a single structured decision. This single-pass, in-prompt branching provides a lightweight approximation of Tree-of-Thoughts reasoning without requiring multiple external inference iterations.

By incorporating both preceding and lookahead context, the continuity scoring mechanism promotes temporal coherence across neighbouring blocks, ensuring that visual and linguistic signals are jointly considered. This approach reduces the likelihood of selecting inconsistent or abrupt boundary positions.

At the same time, the evolving representation of the global theme across windows introduces a limited but effective form of long-range narrative awareness.

Importantly, this method does not attempt to model arbitrarily long-range dependencies as in full attention or recurrent architectures. Instead, it offers a computationally efficient and reproducible compromise that balances narrative sensitivity with the practical constraints of real-world deployment.

5.6.5 Constrained (Single-Pass) ToT Reasoning

The constrained ToT module operates according to the schematic shown in Figure 5.2. In contrast with full ToT implementations that expand deep trees of reasoning steps and compare multiple paths through stochastic sampling [236], the proposed variant is designed as a controlled reasoning mechanism that balances deliberative capability with determinism and efficiency.

The rationale behind the design stems from the constraints of narrative-aware segmentation, where consistency and interpretability are valued above exploration. Traditional ToT models are based on multi-branch searches and probabilistic sampling, thereby inducing noise and complexity. By contrast, the novel framework confines reasoning to one branch, enforces a single-pass inference structure for reasoning, and requiring schema-consistent outputs. Furthermore, reasoning is implicitly informed by global themes, thus ensuring that decisions are made on the basis of narrative transi-

tions.

In practice, this variant imposes the following constraints:

1. **Single-pass inference**. All reasoning occurs in a single LLM call per batch of windows.

2. **Bounded branching**. The LLM is instructed to enumerate exactly 2–3 candidate thoughts, each representing a hypothetical thematic interpretation.

3. **In-prompt selection**. The model must select a single final decision using evidence from the candidates, removing the need for external scoring or tree traversal.

4. **Zero-temperature decoding**. This enforces operational consistency across runs, eliminating stochastic variability that would violate the commercial deployment requirement.

5. **Sustained-theme requirement**. Boundary acceptance requires:

$$\pi(B_i) \neq \pi(B_{i+1}) \quad \wedge \quad \pi(B_{i+2}) = \pi(B_{i+1}).$$

This yields a structured reasoning process that preserves the advantages of deliberative LLM reasoning while avoiding the instability of unconstrained generation.

5.6.6 Decision Rule

Let $\pi(B_i)$ denote the implicit Global Theme associated with window block B_i during reasoning, as described in Section 5.6.3. A candidate boundary between B_i and B_{i+1} is accepted only if it corresponds to a sustained Global Theme transition:

$$\pi(B_i) \neq \pi(B_{i+1}) \quad \wedge \quad \pi(B_{i+2}) = \pi(B_{i+1}), \tag{5.4}$$

whenever lookahead B_{i+2} is available. This constraint ensures that a breakpoint is a real change from one major phase to another and that the new theme extends beyond a single window. Changes within a subtopic of a Global Theme are not allowed.

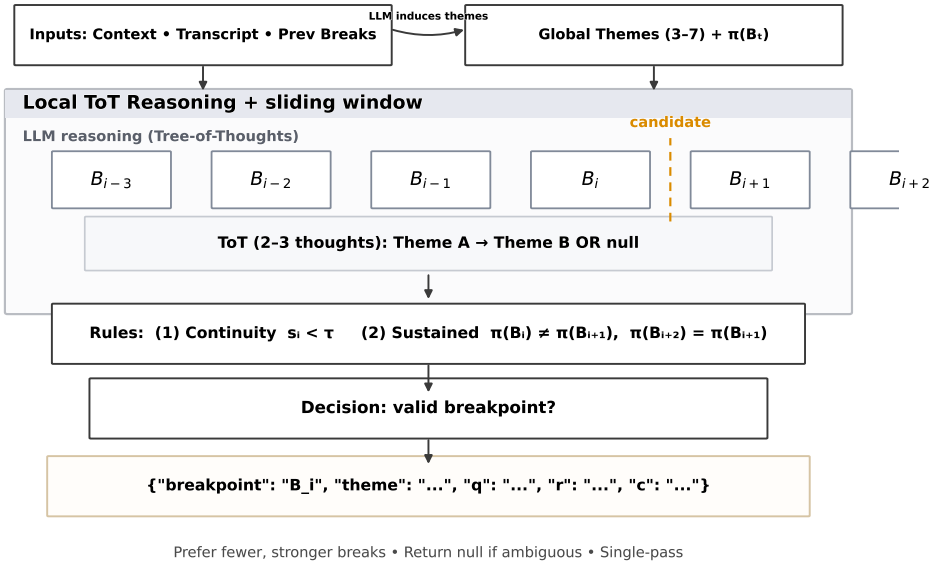


Figure 5.2: Detailed schematic of the proposed constrained (single-pass) Tree-of-Thoughts (ToT) reasoning mechanism for sustained-theme segmentation. The figure illustrates the complete flow from block batching through candidate-thought generation to final breakpoint selection, complementing Algorithm 2 and forming the core of the controlled-inference reasoning layer used in the proposed framework.

All windows within the current batch that satisfy this sustained-transition condition, and are not redundant with previously selected breakpoints, are added to the candidate set $\mathcal{C}_t$. The in-prompt reasoning module then generates multiple breakpoint hypotheses and selects a single final breakpoint by jointly considering Global Theme transitions, lookahead validation, full-video context, and proximity to prior breakpoints.

For the selected candidate the model outputs calibrated evaluative scores: segment quality q_i, flow integrity r_i, and confidence c_i.

$$q_i \in \{1, \ldots, 100\}, \quad r_i \in \{1, \ldots, 100\}, \quad c_i \in \{0, \ldots, 100\},$$

The confidence score is a measure of the clarity of the theme boundary, strength of sustained continuation in lookahead, agreement with transcript evidence, and separation from previously selected breakpoints. The scores are used for structured confidence calibration and are not part of the decision rule as constraints.

5.6.7 Algorithmic Specification

The complete reasoning procedure is formalised in Algorithm 2. The method operates over sliding batches of windows and uses up to three preceding blocks and two subsequent blocks (when available) to validate sustained Global Theme transitions. Each batch returns either a single accepted boundary or a null marker indicating that no sustained thematic change was detected.

5.6.8 True Non-ToT Baseline: A Controlled Single-Path Reasoner

To distinguish genuine ToT behaviour from simple chain-of-thought reasoning, a strict non-ToT baseline is defined. In this mode, the LLM:

- produces *no* multiple candidate thoughts,
- performs *no* internal branching or comparison,
- uses only local context (B_i, B_{i+1}),
- applies *no* forward verification via B_{i+2},
- outputs a single reasoning trace with no selection step.

Algorithm 2 Constrained (Single-Pass) In-Prompt ToT Segmentation with Lookahead

Require: Blocks $B_{1:N}$, provided Global Themes $\mathcal{G}$, batch size W, previously selected breakpoints $\mathcal{P}$

for each batch index t with start s and end $e = s + W - 1$ **do**
  $\mathcal{C}_t \leftarrow \varnothing$
  Construct prompt input using: batch blocks $\{B_s, \ldots, B_e\}$, lookahead context (if available), $\mathcal{G}$, and $\mathcal{P}$
  Query the LLM once to induce an implicit theme mapping $\pi(\cdot)$ over the batch (and lookahead) and to generate 2–3 in-prompt breakpoint hypotheses
  for $i \leftarrow s$ **to** $e - 2$ **do** ▷ requires B_{i+2} for persistence
    if $\pi(B_i) \neq \pi(B_{i+1}) \wedge \pi(B_{i+2}) = \pi(B_{i+1})$ **then**
      Add i to $\mathcal{C}_t$
    end if
  end for
  if $\mathcal{C}_t = \varnothing$ **then**
    Output `breakpoint = null`
  else
    Select a single $i^\star \in \mathcal{C}_t$ using in-prompt comparison of candidate hypotheses
    Output `breakpoint = window_`$i^\star$, `theme =` $\pi(B_{i^\star+1})$, and justification
    Record calibrated scores $q_{i^\star} \in \{1, \ldots, 100\}$, $r_{i^\star} \in \{1, \ldots, 100\}$, $c_{i^\star} \in \{0, \ldots, 100\}$
    Update $\mathcal{P}$ with `window_`$i^\star$
  end if
end for

This baseline is crucial for demonstrating that ToT reasoning with its sustained-theme constraint and bounded branching provides measurable improvements over chain-of-thought alternatives.

5.6.9 Schema-Constrained Output and Reproducibility

To ensure consistent interfacing with downstream evaluation and commercial systems, all outputs from both ToT and non-ToT variants adhere to a fixed JSON schema. For a valid boundary, the system outputs:

- `breakpoint = window_`i,
- `theme =` $\pi(B_{i+1})$,
- textual justification (for interpretability).

If no sustained thematic change is detected, the module returns `breakpoint = null`. Schema validation ensures complete reproducibility of the reasoning process.

5.7 Summary of Methodology

This chapter introduced a complete, end-to-end framework for theme-aware multimodal segmentation designed specifically for adaptive advertising and consumer-media applications. The methodology integrates perception, alignment, structural windowing, and controlled reasoning into a single coherent architecture that directly addresses the limitations identified in the literature (Chapter 2). Rather than treating video segmentation as a low-level change-detection problem or as an open-ended generative task, the proposed system models segmentation as a *thematic* decision process, requiring evidence of sustained semantic change across visual and linguistic modalities.

Multimodal Perception and Alignment

The pipeline begins by jointly processing the video's visual and auditory streams. Shot boundaries are detected with PySceneDetect and converted into temporally consistent units enriched with structured visual descriptions obtained from a vision–language model. Parallel audio processing via WhisperX yields word-level timings and diarized speaker attribution, enabling sentence-level alignment. This ensures that all subsequent reasoning operates on temporally harmonised data where scene boundaries, linguistic units, and speaker transitions share a common timeline. This step directly addresses fragmentation issues typical of naïve shot-based systems.

Audio–Shot Structural Windowing

Recognising that visual cuts often occur mid-sentence, Section 5.5 introduced a bridging-based windowing mechanism that groups consecutive shots whenever continuous audio crosses a cut. This ensures that linguistic units remain intact and that windows reflect the editorial structure rather than artefacts of scene detection. The resulting windows provide the minimal structural units used for higher-level reasoning, balancing the granularity of shot segmentation with the continuity required for theme inference.

Window-Level Representation

Each audio–shot window is transformed into a compact but semantically rich representation integrating (i) visual scene summaries, (ii) utterance text from the transcript, and (iii) a combined window-level summary produced via text-only LLM compression. This representation reduces multimodal signals into a stable semantic form that is resilient to noise, redundancy, and momentary topic drift. Crucially, these window summaries capture narrative intent rather than isolated events or lexical patterns, making them suitable for theme-level inference.

Controlled Reasoning for Thematic Segmentation

The core methodological contribution, presented in Section 5.6, is a constrained, single-pass Tree-of-Thoughts (ToT) mechanism that performs *controlled multimodal reasoning* to detect sustained theme transitions. The reasoning module induces global themes, evaluates each proposed boundary through a bounded lookahead window, and accepts a boundary only if the new theme persists beyond a single block. This sustained-theme requirement is fundamental: it filters out short-lived subtopic shifts and conversational fluctuations, producing robust and semantically meaningful segmentation boundaries.

By enforcing bounded branching, zero-temperature decoding, and in-prompt selection, the ToT variant ensures *operational consistency* a property absent in unconstrained LLM chaptering systems, which suffer from variability, over-segmentation, or hallucinated structure. A strictly non-ToT baseline was also defined to clarify the contribution of constrained reasoning relative to simple embedding- or local-contrast-based methods.

Alignment With Thesis Objectives

Taken together, the components of this methodology directly address the research questions posed in Chapter 1:

- The system **achieves segmentation at the thematic level**, moving closer to human-perceived thematic units compared to purely perceptual baselines of story units than traditional visual or textual methods.

- The architecture **ensures reproducibility and stability** through deterministic processing stages essential for commercial ad placement pipelines.
- The integration of multimodal cues and constrained reasoning **reduces over-segmentation** and prevents spurious boundaries due to lexical noise, visual clutter, or brief topic diversions.
- The design produces **interpretable decision traces**, supporting transparency and human auditing.

This chapter will introduce the narrative layer, which will be combined with cognitive modelling in Chapter 6. The approach is characterized by the combination of multimodal perception, structural windowing, semantic compression, and LLM-based reasoning. Each of the layers has been specifically designed to address a particular set of shortcomings of existing segmentation methodologies. The following section will discuss the empirical evaluation of the approach, which will show the effectiveness of the approach compared to state-of-the-art baselines as well as its applicability to adaptive mid-roll advertising.

5.8 Experimental Evaluation

5.8.1 Overview and Positioning

This chapter evaluates the proposed theme-aware segmentation framework on long-form, real-world video content. In contrast to end-to-end learned approaches such as Vid2Seq [244] and Chapter-Llama [14], the proposed system is a *training-free, controlled inference pipeline* that uses only off-the-shelf modules WhisperX for ASR, PySceneDetect for visual cuts, lightweight vision–language captioning, and a constrained LLM-based reasoning layer with fixed prompts and deterministic decoding. No fine-tuning or gradient updates occur at any stage.

This design choice has important methodological implications. First, the system must be evaluated not only for segmentation accuracy but for its operational properties consistency, and runtime efficiency. Second, because the model operates deterministically (temperature $t = 0.0$, top-$p = 1$), the output for a given video is suitable

for commercial and regulatory settings where non-deterministic LLM outputs are untenable. Finally, the experiments position this work as a complementary alternative to data-hungry generative systems: rather than approximating a dataset's distribution through supervised learning, the method uses fixed reasoning constraints to enforce bounded LLM variance and thematic consistency.

Lightweight here refers to the absence of task-specific training, iterative search, or stochastic decoding, rather than the absence of pretrained foundation models.

Therefore, in light of this, an evaluation methodology is adopted that considers:

- segmentation accuracy in comparison with established baselines,
- theme coherence quality and appropriateness in ad placement,
- runtime efficiency,
- robustness with respect to different types of long-form content,

These analyses jointly assess whether a deterministic, training-free design can achieve accuracy competitive with learned segmentation models at a fraction of the computational expense.

It is, however, important to note that, in this chapter, the Ad-Suitability (%) metric is based on narrative appropriateness. In this sense, it considers if the identified narrative boundaries are aligned with thematic transitions, which are appropriate for content interruption at a story level. It does not incorporate cognitive load or viewer-state modelling, which is introduced separately in Chapter 6. Thus, this metric should be interpreted as a structural narrative proxy for ad placement rather than a full measure of viewer disruption.

5.8.2 Datasets and Evaluation Setup

Evaluation is conducted on **VidChapters-7M** [244], the largest and most established benchmark for open-domain chapter boundary detection. The dataset contains 817k videos and 7M human-annotated chapter boundaries spanning tutorials, vlogs, documentaries, interviews, how-to videos, news commentary, and talk shows. The dataset was not used for training any component of the system; instead, it provides an industry-standard test bed for evaluating chapter detection performance.

A 20-video evaluation subset is created from VidChapters-7M, ensuring:

- diversity across genres (tutorials, vlogs, documentaries, talk formats),
- variation in visual complexity and speaking style,
- preservation of the ground-truth chapter annotations.

Similar to the prior study [14,244], the assessment is performed on a pre-selected set that mainly consists of "Medium" videos, i.e., those with a duration of 15-30 minutes . To further ensure the consistency of the approach and the fairness of the benchmarking, the results are also verified on the test set, allowing for direct and reproducible comparisons with the existing results. Although VidChapters-7M contains over 800,000 videos, full-scale benchmarking was not the objective of this study. The aim was detailed structural evaluation under controlled conditions, including qualitative inspection, boundary-level analysis, and cross-method comparison. The curated subset therefore prioritises interpretability and methodological control over dataset-scale statistical generalisation.

While evaluation is conducted on 20 videos, each contains multiple annotated segments (over 120 total decision points), enabling fine-grained assessment. The dataset spans diverse genres and narrative styles, providing indicative robustness of the proposed framework.

5.8.3 Baselines

A broad set of representative baselines is used for comparison covering classical, embedding-driven, and deep-learning paradigms:

- **TopicTiling [216, 238, 248]**: transcript-only - adopt a neural variant of TextTiling [238], replacing lexical similarity with dense sentence embeddings from Sentence-BERT [216] computed on medium-sized blocks, and detect topic boundaries at local minima in the inter-block cosine similarity curve [248].
- **CLIP+KTS [163, 249, 250]**: visual-only, a training-free temporal segmentation pipeline that combines semantically rich frame embeddings from CLIP [163] with Kernel Temporal Segmentation (KTS) [249] for unsupervised change-point detection on the pairwise similarity matrix, yielding variable-length, semantically coherent segments [250].

- **Vid2Seq** [244]: a supervised transformer model trained on 7M examples.
- **Chapter-Llama** [14]: a LoRA-finetuned Llama-3 model designed specifically for chapter boundary generation.

The first three represent *training-free segmentation*, while the latter two represent *learned, dataset-specific* methods.

This mixture enables evaluation of whether a controlled inference pipeline can match or exceed the performance of systems that rely on millions of labelled examples.

5.8.4 Metrics

To measure the performance of the segmentations, the standard metrics for temporal boundary detection are used, along with a human-centric ad-suitability metric. Inspired by [14], the tIoU and F1 scores are used, as they allow for continuous evaluation at multiple IoU thresholds, are more interpretable, and offer a more.

(a) **F1 score**: compute precision and recall at multiple IoU thresholds ranging from 0.5 to 0.95 (in steps of 0.05). At each threshold, a predicted segment is considered correct (true positive) if its temporal IoU with any ground-truth segment exceeds the threshold (with greedy one-to-one matching to avoid double-counting). Precision is the ratio of correct predictions to the total number of predictions; recall is the ratio of matched ground-truth segments to the total number of ground-truth segments. The F1 score at each threshold is the harmonic mean of precision and recall. The final F1 metric is the average F1 across all thresholds, multiplied by 100 to obtain a percentage.

(b) **tIoU (temporal Intersection over Union)**: The optimal one-to-one matching is first determined between predicted and ground-truth chapter segments by greedily pairing those with the highest individual IoU scores. The tIoU is then the mean IoU value computed across all successfully matched pairs (intersection over union of their temporal durations), multiplied by 100 to obtain a percentage.

(c) **AvgSegLen**: Average predicted segment length.

(d) **Ad-Suitability@majority**: Percentage of boundaries marked as suitable for ad placement by three independent human raters, with reliability measured by Fleiss' κ.

According to [14], several advantages are provided by tIoU and F1 scores, including continuous evaluation of the model at multiple thresholds, interpretability, and a more complete evaluation of the model.

5.8.5 Implementation Details

WhisperX [251] is employed for ASR and diarization, PySceneDetect for visual cut detection, and `Gemma-3-27B-IT (QAT)` [252] for all reasoning and summarisation. All modules operate with fixed prompts and zero-temperature decoding, enabling deterministic behaviour.

Key parameters include:

- lookahead depth: up to two windows ahead are included in the context, with a minimum of one required for continued theme verification,
- decoding temperature: $t = 0.0$,
- JSON schema validation for both ToT and non-ToT modes.

All experiments run on `AWS g4dn.xlarge` via SageMaker, ensuring consistent compute conditions. No fine-tuning or gradient updates are done in the proposed approach.

5.8.6 Main Results

This section presents quantitative boundary accuracy and computational efficiency results. Training-free methods evaluated under a consistent protocol are first discussed, followed by contextual comparison with supervised LLM-based systems reported in prior work.

Training-Free Methods. Table 5.1 summarises boundary accuracy and computational efficiency for training-free segmentation approaches. Boundary accuracy is reported using F1 and mean-tIoU. Inference time is measured as the wall-clock time required to process a one-hour input video, expressed relative to input duration.

Table 5.1: Boundary accuracy and efficiency for training-free methods. Bold = best within block.

Method	F1	mean-tIoU	#Breaks	AvgSegLen (min)	Inference Time (× video duration)
Ours (controlled inference)	**15.8**	**49.7**	8.1	2.4	≈ (1.20–1.80)×
TopicTiling	9.9	35.2	10.2	1.8	≈ (0.10–0.25)×
CLIP+KTS	11.9	38.9	9.8	1.9	≈ (0.50–0.70)×

In the training-free scenario, the proposed controlled inference framework shows the highest boundary accuracy with respect to F1 and mean-tIoU. Compared with traditional segmentation methods such as TopicTiling and CLIP+KTS, the proposed framework shows superior boundary alignment, validating the effectiveness of sustained theme multimodal reasoning.

The primary computational cost arises from the vision–language model used for structured keyframe description generation. In the current implementation, a locally deployed `Gemma-3-27B-IT (QAT)` [252] model running on a g6.12xlarge EC2 [253] instance is used to produce semantically grounded frame-level captions that support downstream reasoning. While this introduces additional inference time relative to lightweight baselines, it enables the global thematic modelling responsible for the observed accuracy gains.

The VLM stage represents a clear optimisation direction. Potential improvements include lighter multimodal captioning architectures, or the use of highly optimised frontier API-based vision–language models [254]. Depending on deployment scale and workload patterns, such alternatives may reduce latency and offer favourable cost trade-offs, though these factors would require empirical benchmarking.

Reported Supervised Systems. For contextual positioning, Table 5.2 presents boundary accuracy metrics reported by recent supervised LLM-based chaptering systems. These results are drawn from larger-scale evaluations in prior work and are presented separately, as they are not directly comparable to the controlled subset evaluation in Table 5.1.

Note on baselines: Vid2Seq and Chapter-Llama metrics are reported from the original papers on the medium-length split of VidChapters-7M [14, 244]. These values

Table 5.2: Boundary accuracy reported by supervised LLM-based systems in prior work.

Method	F1 (reported)	mean-tIoU (reported)
Vid2Seq [14, 244]	19.0	53.3
Chapter-Llama [14]	46.7	72.3

represent averages over the full medium test subset (∼1.7k videos), from which the curated 20-video evaluation set is sampled and designed to be representative (emphasizing ad-relevant genres such as tutorials, vlogs, and talks with clear thematic breaks). No baseline re-runs were conducted on the exact subset due to infrastructure and computational constraints; however, the reported F1 scores serve as a reliable and conservative benchmark for the performance gap between training-free and fully supervised approaches.

In comparison with these supervised methods, the framework demonstrates promising performance in boundary alignment in strictly deterministic and training-free conditions. Even though the peak segmentation accuracy of supervised methods is higher, these methods are trained with large-scale data and much more computational power.

The results demonstrate that multimodal reasoning in controlled conditions with sustained theme verification improves segmentation quality in comparison with classical heuristic methods, yet maintains inference efficiency and behaviour.

5.8.7 Ablation Analysis

Several ablations of the pipeline are evaluated to isolate the contribution of:

- multimodal fusion,
- constrained ToT reasoning,
- bridging-based windowing.

Because ToT behaviour requires enumerating and comparing multiple candidate thoughts [236], a strict *non-ToT baseline* is defined, which removes branching, lookahead, scoring, and selection.

Table 5.3 isolates the contribution of each major architectural component. The results reveal a consistent and interpretable performance hierarchy aligned with the

design rationale of the framework.

Firstly, the performance of the full multimodal pipeline outperforms all other models in every measure, reinforcing the argument that segmentation performance derives from the interplay of multimodal evidence, constrained reasoning, and persistence validation rather than from individual components in isolation. Significant performance degradation in F1, mIoU, and ad-suitability for the audio-only and visual-only models serves to emphasize that neither modality alone is sufficient to identify sustained narrative structure. Moreover, the relative performance of the audio-only model over the visual-only model indicates that linguistic continuity may be a stronger indicator of thematic persistence than changes in visual style.

Non-ToT baseline is seen to perform competitively but is still below the full Constrained (Single-Pass) ToT configuration. This is an indication that the use of bounded branching and the comparison of in-prompt alternative breakpoint hypotheses add value in terms of stability and precision. Most importantly, the improvement from the use of Constrained (Single-Pass) ToT is not disproportionate; it is incremental.

Lastly, the bridging-only windowing baseline performs the worst, validating that structural alignment by itself is not adequate without semantic reasoning. Although windowing maintains linguistic integrity across cut scenes, there is no mechanism in windowing to identify sustained thematic changes.

Overall, the ablation results support the central thesis claim: robust segmentation requires not only multimodal evidence but also constrained reasoning and explicit persistence validation. The sustained-theme constraint and bounded ToT reasoning together improve both quantitative boundary alignment and perceived ad suitability, demonstrating that narrative continuity must be enforced rather than inferred implicitly from local similarity cues.

5.8.8 Human Evaluation

Another important aspect is that human evaluation is a qualitative measurement, and there is no way to quantify this using boundary-based metrics. This is particularly important for ad insertion, where the smoothness of transition is as important as the

Table 5.3: Ablation study on major design components.

Configuration	F1	Seg-tIoU	Ad-Suit. (%)
Full pipeline (Multimodal)	**15.8**	**49.7**	**64**
– Audio only	11.3	42.9	49
– Visual only	8.7	41.5	47
Non-ToT	11.5	42.5	58
Windowing (bridging-only)	9.7	38.4	37

accuracy of segmentation. In this evaluation, three raters were used for each video, and each of them evaluated five boundaries for transition suitability.

Inter-rater agreement, measured using Fleiss' κ, was **0.71**, indicating substantial agreement and confirming the reliability of the human judgments.

With respect to ad suitability, the full multimodal sustained-theme model achieved an overall suitability score of **64%**, outperforming unimodal variants (audio-only: 49%, visual-only: 47%) and the controlled Non-ToT single-path baseline (58%). The improvement over the single-path reasoner is incremental rather than disproportionate, reflecting the stabilising effect of bounded branching and structured in-prompt hypothesis comparison.

Ad suitability, in the context of this chapter, means the narrative appropriateness and thematic coherence of transitions, and the modelling of cognitive load is not included at this point and is introduced separately in Chapter 6

Although the achieved suitability remains below the 75–80% approval levels reported in controlled broadcast-quality mid-roll placement studies [19], the proposed framework operates without supervised fine-tuning, dataset-specific optimisation, or manual editorial intervention. The results nevertheless indicate that the sustained-theme multimodal model produces transitions that are perceptually meaningful and minimally intrusive, while preserving methodological determinism, reproducibility, and deployment feasibility.

5.9 Discussion and Limitations

The proposed framework for video segmentation shows the potential of a controlled, training-free multimodal reasoning model for achieving a level of performance comparable to a large supervised model, along with several other advantages. The experimental results highlight the viability of using fixed, deterministic language-model reasoning as an alternative to large-scale supervised chaptering systems, especially in advertising and media-delivery contexts where reproducibility and explainability are critical. At the same time, several conceptual and practical limitations of the framework must be acknowledged.

A further consideration concerns the comparison with supervised state-of-the-art methods. In this work, such comparisons rely on performance figures reported in the literature rather than direct re-implementation under a unified experimental protocol. This introduces constraints on strict comparability, as differences in datasets, preprocessing pipelines, and evaluation metrics may significantly influence results. However, this approach reflects the intended positioning of the proposed framework as a training-free and reproducible alternative, rather than a fully supervised system optimised for benchmark performance.

5.9.1 Interpretation of the Controlled Reasoning Paradigm

A central objective of this chapter has been to investigate whether structured and constrained reasoning can effectively replace or supplement purely data-driven approaches to video segmentation.

The experimental results presented in Section 5.8 indicate that this is indeed possible, albeit with important qualifications.

- The sustained-theme constraint promotes *semantic persistence*, substantially reducing false positives caused by momentary lexical or visual fluctuations.
- The single-pass constrained Tree-of-Thoughts mechanism introduces *bounded variance*, which enables the system to yield consistent results in multiple evaluations.
- Finally, the multimodal integration of audio structure and visual summaries at

the shot level captures aspects of narrative continuity that cannot be detected unimodally.

These features collectively support the broader thesis argument that carefully controlled reasoning can guide large language models toward robust, interpretable segmentation without the need for task-specific training.

However, the determinism that enables this stability also constrains flexibility, and this trade-off naturally yields several limitations.

A notable practical limitation, evident in Table 5.1, is the increased inference time of the proposed method, which is approximately twice that of several baseline approaches. This overhead stems primarily from the multimodal reasoning components, particularly the generation of visual captions using vision–language models.

While this cost reflects a deliberate trade-off for improved semantic coherence and boundary stability, it can be mitigated through several optimisation strategies.

These include employing smaller or distilled captioning models, caching generated captions and intermediate representations, parallelising preprocessing steps, and using lightweight heuristics to pre-filter candidate segments before applying the more expensive reasoning module.

Thus, although the current design prioritises interpretability and structured reasoning, future work can focus on improving computational efficiency from both model and system perspectives, thereby enhancing scalability for real-world deployment.

5.9.2 Structural Limitations of Shallow Lookahead

The algorithm uses a shallow lookahead window of depth two (B_{i+2}) to verify whether a thematic transition persists. This design contributes to robustness and computational efficiency, yet it restricts sensitivity to long-range narrative dependencies.

In practice:

- Slow-building thematic transitions (for example, multi-step tutorial phases or narrative arcs) may only become semantically clear several windows later.
- Videos with extended expository or conversational segments may exhibit latent themes that unfold gradually and therefore escape detection within two windows.

- The system may delay legitimate boundaries or cluster multiple small transitions into a single larger segment when long-range coherence is ambiguous.

A deeper lookahead, or a hierarchical memory structure, could improve consistency for complex, multi-topic videos, but would also increase computational cost and complicate determinism.

5.9.3 Dependence on Explicit Linguistic and Visual Cues

Although multimodal, the system fundamentally relies on explicit, surface-level features extracted from transcripts and visual captions:

- Shot boundaries and keyframe captions reflect visible changes in setting or subject matter, but cannot capture more abstract cinematic signals such as emotional buildup, narrative tension, or audio–visual rhythm that often guide human judgments about where breaks feel natural.
- WhisperX transcripts provide reliable lexical content, but do not model prosodic cues (intonation, pacing, emphasis) that humans use to perceive narrative boundaries.
- The visual summariser captures objects, scenes, and coarse actions, but not tonal elements such as mood, colour grading, or affective composition.

As a result, the system may misinterpret moments where the narrative structure is driven by implicit, stylistic or emotional cues rather than explicit topic changes. This limitation is especially apparent in dramatic storytelling, music-driven sequences, travel vlogs, and instructional videos where the visual setting remains static but conceptual phases shift.

5.9.4 Cognitive Considerations and Viewer Experience

An important aspect of mid-roll advertisement placement often studied in media psychology is the role of cognitive load and viewer attention. The present system does not incorporate cognitive metrics, yet such factors influence perceived smoothness:

- High-load segments (e.g., dense explanations or emotionally intense scenes) create "attention binding", making interruptions feel more disruptive.

- Low-load or transitional phases often provide natural resting points that align well with ad placement.

Absent the explicit modelling of such viewer-centered variables, it's possible the system will choose points to segment the video that are semantically correct but not optimal from the viewer's point of view. Future work could include the use of simple models for cognitive load (e.g., based on linguistic complexity or acoustic energy) to increase perceived appropriateness from the viewer's viewpoint.

5.9.5 Error Propagation Across Modular Components

The pipeline's modular design offers clarity and interpretability but also introduces cascading dependencies:

- Errors in ASR timing propagate into window construction.
- Incoherent visual captions can distort window summaries.
- Shot-detection failures (too many or too few cuts) distort the downstream reasoning context.

Although each component performs strongly in isolation, inaccuracies in any stage can subtly influence the downstream segmentation. Because the system is deterministic, such errors repeat identically across runs: favourable for auditing but limiting for robustness.

5.9.6 Limited Adaptability Without Fine-Tuning

The paradigm of controlled inferences deliberately excludes training. While this method promotes replicability, it also limits adaptability in the following way:

- The system is not capable of extracting the patterns associated with video information, like transitions and video styles.
- Domain-specific cues, such as sports broadcasts versus cooking videos versus lectures, remain unexploited.
- Emerging content formats (e.g., hybrid vlog–tutorial mixes) may degrade segmentation accuracy unless explicitly accommodated.

A potential direction for future work is lightweight, non-destructive domain adaptation for example, learning a small number of prompt-selection rules or decision thresholds while preserving overall determinism.

5.9.7 Summary of Limitations

In summary, while the system achieves strong performance for a training-free approach, its limitations can be grouped into four categories:

1. **Structural constraints:** fixed shallow lookahead prevents modelling long-range theme drift.
2. **Feature-level constraints:** reliance on explicit cues limits sensitivity to emotional, affective, and stylistic signals.
3. **Human-experience constraints:** absence of cognitive modelling restricts viewer-aligned ad-suitability predictions.
4. **Architectural constraints:** modular design introduces potential error propagation and limited adaptability.

Rather, these weaknesses enumerate the directions for further research, such as hierarchical theme modelling, more informative affective vectors, cognitive load computation, or mixed deterministic-adaptive reasoning schemes. The aforementioned weaknesses do not undermine the value of the controlled reasoning paradigm, but they outline the directions for future studies.

5.10 Chapter Summary

In this chapter a holistic and training-free multimodal solution for theme-aware segmentation of long-form video content has been proposed. The proposed solution combines scene-level visual features, audio-text alignment, and a constrained reasoning approach for effective and human-aligned breakpoints for mid-roll ad placement.

The chapter began by presenting the architectural motivations and design principles underpinning the system, emphasising the need for reproducibility, interpretability,

and low operational cost. A detailed breakdown of the pipeline components followed, including shot detection and visual captioning, WhisperX-based transcription, bridging-based window construction, window-level summarisation, global theme induction, and the constrained Tree-of-Thought (ToT) reasoning model with a sustained-theme constraint.

The major methodological achievement of this chapter is to propose a unified one-pass ToT process which reaches a balances between semantic sensitivity and deterministic behaviour. The verification procedure for sustained themes also ensures that boundary descriptions capture real transitions in narratives. The chapter also detailed on how a consistent and verified output needs to be ensured through certain implementation details such as using a fixed prompt and non-sampling decoding.

Moreover, the experimental evaluation demonstrated that the proposed model achieves competitive segmentation performance with other training-free models, and its performance is superior to the performance of classical heuristic models, with two orders of magnitude less computational cost than supervised LLM-based chaptering models. Ablation experiments also verified the significance of multimodal fusion, sustained theme constraints, and constrained reasoning. Human evaluation also validated the ability of the framework to generate ad-suitable transitions with high inter-rater agreement.

Finally, the discussion highlighted the system's methodological implications as well as its limitations, including shallow lookahead, reliance on explicit cues, and the absence of cognitive or affective modelling. These insights motivate future extensions involving deeper contextual reasoning, enriched non-linguistic features, and viewer-centric models of cognitive load and narrative rhythm.

In conclusion, chapters 5.7 and 5.8 together show how it is possible to have a deterministic model or system which performs segmentation without any training, as an effective substitute where the requirements of real life converge into accuracy, interpretability, reliability, as well as cost-effectiveness.

Chapter 6

Cognitive Load-Aware Video Segmentation for Minimally Intrusive Ad Placement

Publication Notice.

This chapter is based on the following peer-reviewed conference publication, presented by the author:

De Silva, W., Dong, F., Abel, A., & Fernando, A. (2026). Cognitive-Load-Aware Multimodal Video Segmentation for Adaptive Consumer Media Systems. In ***Proceedings of the 2026 IEEE International Conference on Consumer Electronics (ICCE)****, Dubai, UAE.*

The material has been revised and expanded for integration into this doctoral thesis, including extended methodological detail, additional analysis, and alignment with the overarching research framework.

6.1 Motivation and Problem Statement

The rapid expansion of online streaming, social video platforms, and personalised media delivery has foregrounded the need for systems that adapt not only to the *content* of a video but also to the *cognitive experience* of its viewers. Contemporary user experience

research consistently shows that a viewer's perceived comfort, narrative immersion, and tolerance for interruptions are strongly shaped by their moment-to-moment cognitive load [255, 256]. In long-form content, where narrative arcs vary in complexity, pace, and information density, cognitive load naturally fluctuates over time. These fluctuations create short, low-load intervals that function as "natural breathing points" in the story moments at which viewers are most amenable to smooth transitions, pauses, or mid-roll interventions [257].

To align this chapter with the Integrated modular adaptive media framework introduced in Chapter 2, the *cognitive suitability* component C_t is explicitly addressed of the AdBreakScore(t) model. While Chapters 3–5 developed methods for understanding semantic content and identifying theme-aligned structural boundaries, none of these approaches quantify the moment-to-moment mental effort experienced by viewers. The present chapter fills this gap by operationalising C_t using a multimodal, sensor-free cognitive load estimation framework, thereby enabling breakpoint selection that is not only semantically coherent but also cognitively appropriate.

Despite this, existing video segmentation and ad-insertion pipelines continue to rely on surface-level criteria such as fixed time intervals, shot boundaries, or audio silence detection. While computationally simple, such heuristics are misaligned with the psychological factors that shape a viewer's sense of continuity. Interrupting a dense explanation, a visually complex scene, or a rapid sequence of narrative events can introduce mental fatigue [258]. As a result, these methods often produce segmentation boundaries that feel abrupt or intrusive to end users, especially in educational, documentary, and conversational formats [259].

Recent progress in video understanding has shifted from low-level change detection toward narrative-aligned segmentation [260]. Approaches such as chapter boundary prediction model structural transitions in the story rather than merely visual discontinuities. However, these methods remain fundamentally *content-centric*: they predict where the story changes, but not whether that moment is cognitively suitable for an interruption. In other words, they describe the narrative, but they do not consider the viewer's cognitive experience *as part of the segmentation signal*. This omission is

significant because a narrative transition may occur during a period of high cognitive demand for example, during dense exposition or visually rich action making it a poor choice for viewer-oriented pacing or ad placement.

Cognitive load theory, originating from Sweller [38], provides a principled foundation for modelling mental effort during multimedia consumption. Traditional implementations rely on post-hoc questionnaires such as the NASA Task Load Index (NASA–TLX) [261], which measure perceived Mental Demand, Effort, and Temporal Demand. While suitable for controlled experiments, these methods are impractical for real-time or large-scale media applications. Physiological measurement techniques - EEG, EDA, heart-rate variability, or eye-tracking [262, 263] offer richer temporal resolution but require wearable sensors, raise privacy concerns, and are incompatible with passive, consumer-facing video platforms.

This chapter addresses these limitations by adopting a non-invasive, multimodal approach to cognitive load estimation that operates directly on the *content of the video itself*. Building on research showing that speech prosody, visual motion, scene complexity, and pacing correlate with perceived workload [264, 265], a lightweight framework is developed to map synchronized audiovisual cues to NASA–TLX–inspired dimensions. Speech rate, loudness variability, visual motion intensity, and shot pacing are aggregated to reflect fluctuations in Mental Demand, Effort, and Temporal Demand at the segment level. Unlike binary high/low-load classifiers commonly used in educational settings [266], the proposed approach models cognitive load as a *continuous* time series, enabling fine-grained, frame-aligned segmentation of viewer demand.

The main novelty of the proposed approach is the inclusion of a reasoning-assisted fusion mechanism based on a prediction of cognitive load from multimodal descriptors with the help of a large language model (LLM). Rather than predicting load through a learned model requiring annotated training data, the LLM interprets audio–visual cues using structured prompts and deterministic (temperature = 0) decoding. This makes the system training-free, reproducible, and operationally lightweight properties essential for large-scale streaming platforms and commercial deployments. Determinism in this context refers to fixed prompt configuration and temperature-zero decoding under a

specific pretrained model checkpoint. Reproducibility is therefore guaranteed provided the same model version and inference parameters are used. As with any foundation model, behaviour may vary across model revisions.

Chapters 3 and 4 established a multimodal framework for understanding *what* a video is about its topics, semantic content, and taxonomy-aligned descriptors while Chapter 5 introduced a reasoning-driven method for segmenting videos according to *how* narrative themes evolve over time. However, semantic and thematic cues alone are insufficient for determining *when* a viewer can be interrupted without degrading the viewing experience. This chapter extends the earlier contributions by modelling the viewer's cognitive state directly from audio–visual signals. The proposed cognitive-load–aware segmentation framework therefore complements the semantic analyses of Chapters 3–4 and the theme-aware segmentation of Chapter 5 by adding a human-centred dimension grounded in cognitive ergonomics. In effect, Chapters 3–5 describe the structure and meaning of the video itself, while Chapter 6 operationalises the cognitive suitability component of the AdBreakScore formulation introduced in Chapter 2, enabling segmentation that is simultaneously theme-aligned and cognitively appropriate.

Research Problem. The central question addressed in this chapter is:

> *How can viewer cognitive load be inferred from audio–visual cues alone, without wearable sensors or supervised training, and use this estimation to identify cognitively suitable segmentation points in long-form video?*

This problem directly advances the broader thesis objective of enabling user-aware and semantically coherent video segmentation. While Chapter 5 introduced a theme-driven segmentation framework based on multimodal perception and constrained reasoning, the current chapter extends the segmentation paradigm to incorporate *viewer-centred factors*. Together, these chapters address two complementary aspects of adaptive multimedia understanding:

- **Content-centric coherence** (Theme-aware segmentation; Chapter 5)
  Ensuring boundaries align with narrative transitions.

- **Viewer-centric suitability** (Cognitive-load–aware segmentation; Chapter 6) Ensuring boundaries occur at moments that minimise cognitive disruption.

Chapter Contribution. This chapter contributes a novel, training-free, multimodal framework that:

1. **Estimates time-varying cognitive load** based on synchronized audio-visual cues with dimensions inspired by NASA-TLX, without requiring any physiological sensors.
2. **Employs an LLM-based reasoning module** to interpret multimodal descriptions and achieve reproducible trajectories of cognitive load under a fixed model configuration.
3. **Identifies low-load intervals** that corresponding to narrative gentle pauses and human-rated breakpoints in comfort.

Although advertisement placement is a motivating application, the framework has broader implications for adaptive streaming, user-oriented pacing, accessibility-aware playback, and cognitively sensitive media interfaces. The remainder of the chapter develops the methodology, evaluation strategy, and findings that demonstrate the potential of cognitive-load–aware segmentation as an integral component of future consumer video systems.

6.2 Cognitive Load Theory and Multimodal Estimation

Understanding how viewers process, interpret, and mentally manage video content is central to this thesis, particularly because the goal of Chapter 6 is to model *time-varying cognitive load* using non-invasive audiovisual cues. While Chapter 2 provided a broad survey of multimedia analysis, multimodal fusion, and semantic segmentation, it did not address the psychological or cognitive foundations that underpin user–video interaction. This section therefore introduces the theoretical background on cognitive load, reviews sensor-based and sensor-free estimation methods, and situates the proposed framework within contemporary research on user-centric media systems.

6.2.1 Foundations of Cognitive Load Theory

Cognitive Load Theory (CLT), introduced by Sweller [38], posits that human working memory has a limited processing capacity and that both learning effectiveness and perceptual comfort depend on keeping mental demand within an optimal range. CLT divides cognitive load into three types: *intrinsic* load (task complexity), *extraneous* load (presentation inefficiencies), and *germane* load (effort dedicated to understanding). Although originally developed for instructional design, CLT has since been widely adopted in multimedia learning, video-based education, and interactive systems to explain why certain visual or auditory patterns impose higher mental effort on the viewer [267].

A key instrument derived from CLT is the NASA Task Load Index (NASA–TLX) [261], which quantifies perceived workload across several dimensions, including Mental Demand, Effort, and Temporal Demand. These aspects are highly relevant to video content. For instance, moments of dense dialogue, fast visual motion, changes in pacing, or high task urgency (e.g., countdowns or fast narration) are highly correlated with reported mental strain. Previous work has shown that breaking videos into smaller segments reduces cognitive load and improves comprehension [268]. While these results are based on post-hoc subjective responses, they are not feasible for online segmentation.

6.2.2 Physiological and behavioural Approaches to Cognitive Load Estimation

For supporting continuous estimation, there has been considerable research on the physiological indicators of cognitive loads. Tasks such as electroencephalography (EEG), electrodermal activity (EDA), heart rate variability, pupil dilation, and eye-tracking are commonly used in laboratory settings to identify variations in cognitive loads [262]. Data sources such as ADABase [263] have shown that multimodal physiological parameters can accurately identify workload if time-aligned with task stimuli.

Within the context of instructional videos, models based on EEG have been employed in detecting overload situations and cognitively demanding areas of videos [266]. Similarly, models based on blink rates, eye movements, or micro-expressions have been identified as having great possibilities in e-learning systems [269]. However, because

of their accuracy, such systems require wearable technology or a camera close to the viewer, which makes it difficult to align them with streaming technology.

Consequently, physiological approaches offer important conceptual insights but are not feasible for the non-invasive, scalable segmentation framework targeted in this thesis.

6.2.3 Sensor-Free Cognitive Load Estimation from Audiovisual Signals

Recent research is attempting to directly infer cognitive load from *multimedia content* without relying on any user sensors. This is similar to the approach taken in this chapter. Research suggests that audiovisual patterns that reliably correlate with mental workload include fast speech, increased prosodic variability, fast-motion video, and highly edited video, as well as visually complex environments [270].

More recently, Nguyen et al. [265] introduced a multimodal, multitask framework that integrates deep audiovisual speech and visual scene representations for real-time, non-intrusive cognitive load estimation without relying on physiological sensors. These studies prove the possibility of content-based estimation; they always tend to be related to learning and provide a very broad classification: low versus high cognitive load.

For consumer video applications particularly adaptive advertising such binary predictions are insufficient, as well-timed ad-breaks depend on detecting localized *drops* in cognitive load that correspond to natural narrative pauses.

6.2.4 Cognitive Load in Video Segmentation and Media Interruption

Video segmentation research has historically focused on low-level cues such as shot boundaries, audio silence, or optical flow [271]. While computationally simple, these heuristics often misalign with the viewer's cognitive state; for example, a hard cut during a tense or information-heavy moment may be visually appropriate yet cognitively disruptive.

Cognitively grounded segmentation research exists but is normally done via physiological feedback. Related research on applying EEG to the detection of cognitive states to guide interactive system adaptation was conducted by Zander et al. [272].

In advertising studies, models seeking to place advertising during the low-load periods typically make use of manually designed signals of user engagement or post-view survey self-reports [273]. However, these approaches lack a generalizable means of estimating load continuously from raw audiovisual content.

This gap creates a need for frameworks that: (a) infer viewer cognitive load from content alone, (b) operate continuously at the segment level, and (c) align with narrative semantics rather than isolated low-level cues.

6.2.5 Positioning of the Proposed Framework

The method introduced in this chapter extends prior research by combining cognitive theory with modern multimodal signal processing and LLM-guided reasoning. Unlike physiological approaches, it requires no sensors; unlike binary educational load models, it produces fine-grained temporal trajectories of cognitive demand; and unlike classical segmentation, it identifies low-load intervals that explicitly reflect NASA–TLX-inspired mental processes.

This positions the framework as a scalable, user-focused alternative for cognitively adaptive media systems. The framework not only addresses the motivating application of ad insertion, but it also addresses other applications in user-aware personalized streaming, the pacing of narratives, assistive interfaces, as well as the optimisation of the viewer experience. This advancing the thesis objective of developing intelligent, user-aware segmentation tools grounded in multimodal understanding.

6.3 Overview of the Proposed Multimodal Framework

Building on the theoretical foundations and literature reviewed in Section 6.2, this section introduces the proposed multimodal framework for estimating segment-level cognitive load in narrative video. The central objective of the framework is to infer how cognitively demanding different regions of a video are *for a typical viewer*, using only the audio–visual properties of the content itself. Unlike physiological approaches that rely on wearable sensors, the present system is designed to operate passively and scalably,

making it suitable for deployment in consumer media platforms, adaptive streaming, and narrative-aware content editing.

The framework is based on the premise that cognitive load during video consumption is reflected not only in low-level sensory patterns such as prosody, motion, pacing, and scene variability but also in higher-level semantic and narrative cues. For example, dialogue-heavy sequences may impose greater *Mental Demand*, while visually intense or rapidly edited scenes increase *Effort* and *Temporal Demand*. Each of these load dimensions corresponds directly to NASA–TLX factors reviewed in Section 6.2. The framework leverages multimodal feature extraction and reasoning-assisted fusion to translate these cues into a smooth, time-varying cognitive load curve.

System-Level Architecture

Figure 6.1 provides a high-level overview of the complete inference pipeline. The processing flow proceeds through five conceptually linked stages:

1. **Multimodal Feature Extraction.** The audio and visual streams are decomposed into perceptual descriptors capturing sound intensity, speech rhythm, motion dynamics, shot pacing, and visual scene characteristics (details in Sections 6.6 and 6.7). WhisperX provides time-aligned transcripts, while a vision–language model (VLM) summarises keyframes to capture semantic context.

2. **Contextual Summarisation via LLM.** To jointly interpret perceptual cues and semantic content, the extracted features, transcripts, and captions are fused into structured prompts. A lightweight instruction-tuned LLM condenses local narrative context, emotional tone, and scene intent into a unified representation.

3. **NASA–TLX Mapping and Reasoning.** The LLM produces estimates of Mental Demand, Effort, and Temporal Demand within a normalised $[0, 100]$ scale. Importantly, the LLM also proposes adaptive weighting coefficients that reflect the relative importance of each dimension in the current scene. This yields a cognitively grounded multimodal estimate for each temporal window.

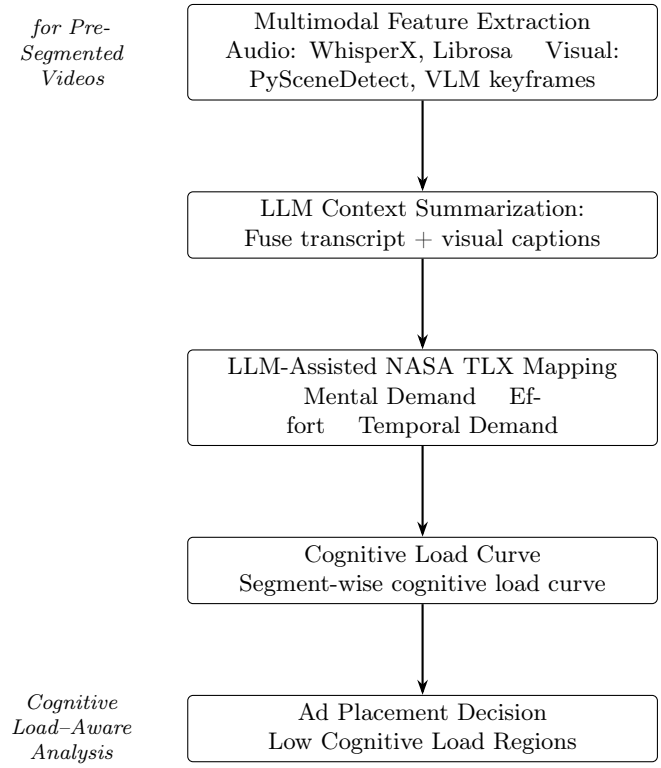


Figure 6.1: Overview of the proposed cognitive load–aware segmentation framework. Audio and visual features are extracted and mapped to NASA–TLX dimensions to produce a time-varying cognitive load curve. Low-load points aligned with narrative pauses are selected as candidate transition positions.

4. **Cognitive Load Curve Construction.** The dimension-weighted scores are combined into a continuous cognitive load curve. Temporal smoothing filters out the short-term oscillations, pointing to the stable cognitive areas which tend to be the natural narrative "breathing points".

5. **Segment-Level Interpretation and Applications.** The load curve is aligned with narrative segmentation units (e.g., chapter or scene boundaries). Low-load intervals immediately following a boundary are interpreted as cognitively comfortable transitions, making them suitable for ad placement or pacing adjustments.

Conceptual Relation to Thesis Objectives

The multimodal framework serves three major goals of this thesis:

- **User-Aware Segmentation:** By modelling cognitive load rather than relying solely on shot boundaries or pause detection, the system identifies narrative pauses that align with human comfort.
- **Scalability and Non-Invasiveness:** The framework extracts cognitive signals purely from content no sensors, user data, or personalisation are required, addressing the deployment limitations of prior work.
- **Interpretability Through Cognitive Dimensions:** The use of subscales from NASA-TLX provides interpretable analysis of the reasons why a region is cognitively heavy or light, providing diagnostic value for research and industrial use.

The above design principles enable the system to produce temporally aligned cognitive load curves that take account of both the intensity and semantic richness of the content. This forms the methodological foundation for the feature extraction (Sec. 6.6–6.7), LLM-assisted estimation (Sec. 6.4.5), and segment-level decision rules (Sec. 6.9) presented in the remainder of the chapter.

6.4 Multimodal Feature Extraction: High-Level Overview

A central premise of this chapter is that *viewer cognitive load is inherently multimodal*. No single signal neither audio, visual, nor textual adequately captures the mental effort, attention, or temporal pressure experienced by a viewer as a narrative unfolds. Cognitive load emerges from interactions between spoken content, visual pacing, scene complexity, and the semantic clarity of the underlying story. Consequently, any model that aims to infer time-varying cognitive load from video must integrate heterogeneous perceptual and linguistic cues in a manner that approximates human interpretation.

The proposed framework adopts a **content-derived, non-invasive** approach that fuses *three synchronized modalities*:

1. **Audio features** capturing prosody, speech rhythm, loudness variation, and silence structure.
2. **Visual features** capturing motion intensity, shot pacing, cut density, and scene composition.

3. **Semantic features** derived from transcripts and vision–language captions, capturing conceptual difficulty, narrative tension, and contextual grounding.

These modalities are aligned on a shared temporal axis and summarized at the level of *analysis windows*, which form the basis for downstream NASA–TLX mapping. The following subsections motivate why each modality is essential, how they jointly encode cognitive phenomena, and how multimodal fusion provides a more reliable foundation for segment-level load estimation than signal-only heuristics.

6.4.1 Why Cognitive Load Requires Multimodal Signals

Cognitive load theory (CLT) offers conclusive evidence on mental effort as a function of:(i) the intrinsic difficulty of the content, (ii) the perceptual requirements associated with its processing, and (iii) Temporal pressure due to pacing. The variables that are not viewable in one modality [15, 38, 274, 275]:

- **Audio alone** encodes intensity and speech rate, but not visual complexity or semantic depth.
- **Visual signals** capture pacing and motion but fail to characterize conceptual or linguistic difficulty.
- **Textual semantics** represent the complexity of text, but not the sensory involvement of the viewer.

Therefore, a cognitively meaningful representation requires *integrated perceptual and semantic cues*. This multimodal foundation mimics human viewing conditions: the cognitive effort of following a scene is a joint function of what is *heard*, *seen*, and *understood*.

Furthermore, in story-based videos (movies, documentaries, vlogs, etc.), sound and visuals don't behave in a single consistent way. Instead, they keep switching between different "modes". In each mode, audio and visual information contribute differently, and often in complementary ways. Because of this, you can't rely on just one kind of feature (only audio, only visuals, only motion, only text, etc.) to understand what's going on. A multimodal approach ensures that cognitive-load estimates remain stable and interpretable across diverse video genres.

6.4.2 Temporal Alignment Across Modalities

To facilitate common reasoning based on the various cues, all streams are operated on within the same timeline. Audio timestamps come from the timebase of the ASR system in WhisperX [251], visual features are based on shot boundaries identified by PySceneDetect [237], and textual semantics inherit the same boundaries via WhisperX alignment and VLM-based captioning. Each analysis window W_k therefore aggregates:

- standardized audio features (Section 6.6),
- shot-aligned visual metrics and scene-level VLM captions (Section 6.7),
- transcript snippets describing dialogue and narrative context(Section 6.6).

This alignment ensures that cognitive-load reasoning reflects the true perceptual and semantic state at each moment of the video, enabling consistent NASA–TLX estimation [261].

6.4.3 From Perceptual Signals to Cognitive Indicators

Each mode provides distinctive information about cognitive demand:

- **Prosodic cues** (speech rate, articulation, pause ratio) correlate with mental effort and linguistic density [270, 276].
- **Loudness and dynamic range** reflect intensity and emotional arousal, contributing to perceived effort [270].
- **Motion, optical flow, and shot pacing** correlate with temporal demand, especially in visually complex or rapidly changing scenes [277, 278].
- **Semantic embeddings and caption-based descriptions** features can be used to quantify conceptual density, emotional tone, and narrative structure in multimedia content [279]. These features are strongly associated with learners' cognitive load in multimedia learning environments [280].

By combining these perspectives, the model generates a *cognitive load signature* for each window, capturing both low-level perceptual difficulty and higher-order semantic demands.

The multimodal mapping from perceptual audio–visual features to the NASA–TLX-

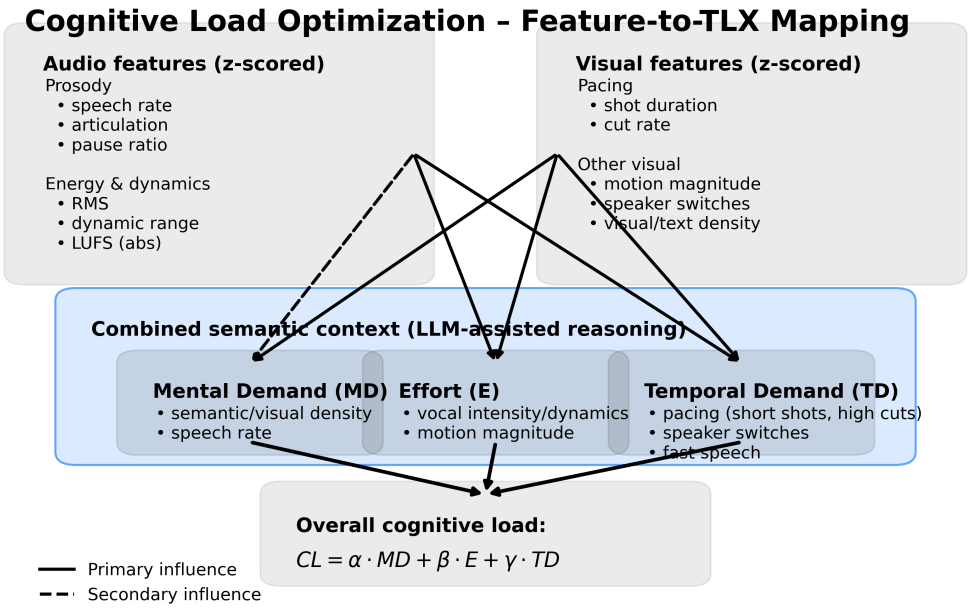


Figure 6.2: Feature-to-TLX mapping for cognitive load estimation. Z-scored audio and visual cues are mapped to NASA–TLX-inspired dimensions of Mental Demand (MD), Effort (E), and Temporal Demand (TD) through LLM-assisted multimodal reasoning. Primary and secondary influences are indicated, and the overall cognitive load is computed as a weighted combination of these dimensions.

inspired cognitive dimensions is summarised in Fig. 6.2, which illustrates the primary and secondary influences of pacing, motion, and prosodic cues on Mental Demand, Effort, and Temporal Demand prior to weighted fusion. Low-level audio and visual features are extracted and normalised deterministically prior to LLM inference; implementation details are provided in Appendix A.4. The cognitive-load scores corresponding to these dimensions are obtained using a fixed structured LLM prompt, documented in Appendix A.5, which is instantiated programmatically for each segment.

6.4.4 Why Multimodal Fusion Outperforms Signal-Only Approaches

Signal-only heuristics such as longest pause detection, energy dips, or shot-change rules rely on isolated acoustic or visual events and therefore do not generalise reliably across narrative contexts. A moment can be acoustically quiet yet cognitively demanding due to dense exposition; conversely, visually dynamic montage sequences may impose

relatively low cognitive load when their semantic content remains predictable. Rapid shot transitions similarly do not necessarily correspond to elevated cognitive demand if narrative progression is simple and expected.

The proposed framework mitigates these ambiguities through multimodal fusion guided by structured LLM-assisted reasoning, which jointly considers the full perceptual–semantic state of each narrative segment. As illustrated in Fig. 6.2, complementary audio–visual cues are mapped onto NASA–TLX-inspired dimensions of Mental Demand, Effort, and Temporal Demand. Prosodic and intensity-based audio features primarily influence Effort and Mental Demand, while visual pacing and motion cues dominate Temporal Demand, with secondary cross-modal influences refining the joint estimate through semantic interpretation of scene and dialogue context.

This multimodal formulation enables the model to distinguish between visually calm but cognitively demanding didactic segments, visually intense but semantically lightweight montage sequences, and dialogue-heavy narrative transitions cases in which unimodal heuristics systematically fail. The fusion architecture therefore provides both improved empirical accuracy and a principled cognitive justification for the segmentation performance gains reported in Section 6.9.

6.4.5 Role of LLM-Assisted Reasoning

While traditional audiovisual models rely solely on numeric features, a large language model provides *interpretive structure* by:

- summarizing multimodal cues,
- assigning relative importance to features based on context,
- mapping cues to NASA–TLX dimensions (Mental Demand, Effort, Temporal Demand),
- producing numerically stable and interpretable cognitive-load scores.

This reasoning layer allows the framework to capture subtle semantic states such as resolution, narrative closure, conceptual difficulty, or emotional de-escalation, which would otherwise be invisible to signal-level models.

The extraction of multimodal features sets the base for meaningful load estimation in

terms of cognition and enables the system to make the connection between the low-level perceptual stimulus and the higher-level narrative reasoning. The following subsections elaborate on the feature extraction for each of the modalities and describe how these cues contribute to the overall reasoning framework.

The structured prompt template that defines the controlled multimodal reasoning operator used for cognitive-load estimation is provided in Appendix A.4.2.

6.5 Semantic Feature Extraction (Semantic Cognitive Sensing)

Existing methods for estimation of cognitive load generally focus on the perception of motion, pacing, and acoustic intensity. However, in video narratives, the cognitive load not only depends on the perception of information, but also on the semantic complexity of information being processed. To include this, the proposed framework also considers a semantic sensing modality, which acts as a first-class cognitive sensor, in addition to auditory and visual perception.

For each segment-aligned window W_k, a fused textual representation, denoted as *window_context*, is constructed by temporally aligning shot-level scene descriptions and word-level transcript excerpts. This representation provides a compact narrative abstraction that encodes discourse structure, topic density, and conceptual continuity within each narrative unit.

Rather than being treated as auxiliary context, the semantic modality directly informs the NASA–TLX Mental Demand (MD) and Effort (E) dimensions. Linguistic complexity, topic compression, and discourse cohesion captured in *window_context* are interpreted by the LLM reasoning module to infer semantic workload characteristics. These semantic cues further modulate the adaptive weighting of TLX dimensions, thereby influencing both the magnitude and composition of the final cognitive load estimate.

Let $\mathrm{Sem}(W_k)$ denote the semantic sensing function operating on *window_context*. The resulting semantic workload representation contributes to the TLX fusion as:

$$MD_k, E_k \sim \mathrm{Sem}(W_k), \tag{6.1}$$

where MD_k and E_k represent the Mental Demand and Effort components of the NASA–TLX formulation for window W_k.

By explicitly modelling semantic perception as a cognitive sensing channel, the proposed framework extends NASA–TLX beyond its traditional questionnaire-based usage and operationalises it as a multimodal cognitive fusion operator. This formulation enables segment-level cognitive load estimation to reflect perceptual stimulation.

Window-Level Semantic Construction

Let $W_k = [t_0^k, t_1^k)$ denote a segment-aligned analysis window. Let $\mathcal{S}_k = \{S_i\}$ be the set of visual scene summaries whose timestamps overlap W_k, where each S_i is produced by the vision–language captioning and summarisation pipeline described in Chapter 5. Similarly, let $\mathcal{T}_k = \{u_j\}$ be the set of transcript utterances overlapping W_k, obtained from word-level ASR alignment.

A raw multimodal semantic trace for the window is formed by concatenating these ordered elements:

$$X_k = \mathrm{concat}\big(\mathcal{S}_k, \mathcal{T}_k\big), \tag{6.2}$$

where ordering follows the global video timeline.

To obtain a stable, noise-resilient representation suitable for controlled reasoning, the trace X_k is compressed using a constrained LLM summarisation operator $g(\cdot)$ (temperature = 0, fixed prompt), yielding the fused semantic window representation:

$$\mathit{window\_context}_k = g(X_k). \tag{6.3}$$

This construction matches the window-level re-summarisation mechanism introduced in Chapter 5 and provides the semantic sensing input used in the NASA–TLX reasoning procedure in this chapter.

6.6 Audio Feature Extraction

The audio stream is a primary carrier of narrative information and therefore a critical modality for estimating time-varying cognitive load. Spoken language conveys temporal pacing cues, discourse rhythm, emotional tone, and structural markers of narrative organisation, emotional tone, and prosodic markers of pacing all of which influence the viewer's perceived mental effort. In many genres, including documentaries, tutorials, commentary videos, and interviews, the audio track serves as the dominant channel for information transfer. Consequently, any cognitively meaningful segmentation framework must treat audio as a first-class signal rather than as an auxiliary input [281].

This section details the audio-processing component of the framework, explaining not only *how* the features are extracted but, crucially, *why* they represent valid indicators of cognitive load in narrative video. Particular emphasis is placed on the factors that connect raw prosodic patterns to the NASA–TLX dimensions employed in the reasoning stage (Mental Demand, Effort, and Temporal Demand).

6.6.1 Motivation: Why Audio Reflects Cognitive Load

Theories on cognitive psychology propose that listeners' mental workload is heavily affected by the temporal and structural characteristics of speech. Quick speech rate, high information density, and low instances of silence require the viewer to dedicate more resources to the mental processing of the information. Slow speech rate, consistent loudness levels, and intervals of silence provide the viewer 'breathing space' to process the information [282].

6.6.2 Processing Pipeline

Figure 6.1 situates audio processing at the first stage of the multimodal pipeline. For each video, WhisperX [251] provides a high-precision alignment of speech segments, timestamps, and speaker identities. This alignment serves as the temporal backbone onto which additional audio features are computed. Low-level audio descriptors including RMS energy, dynamic range, loudness (LUFS), speech rate, articulation rate,

and pause ratio are computed using the Librosa audio analysis library [283] prior to transcript alignment and multimodal fusion.

The audio signal is resampled to a common format, and frame-wise descriptors are computed using short-time analysis windows (typically 25–50 ms). Features are then aggregated into the larger analysis windows W_k used for downstream TLX reasoning. To ensure that values are comparable across videos with different recording conditions, loudness levels, or microphone profiles, all features are z-scored within each video [284].

6.6.3 Feature Categories and Cognitive Interpretation

The extracted audio descriptors fall into three conceptual categories, each associated with a particular aspect of cognitive load:

(1) Prosodic and Temporal Speech Features

These are characteristics of the rhythm, density, and structure of spoken language:

- **Speech rate**: rate in syllables or words per second. High rates will increase *Mental Demand* because the viewer must process information faster [285].
- **Articulation rate**: Speech rate without pauses. This measure of rate has the advantage of being able to distinguish fast speech from pause-rich regular speech [286].
- **Pause ratio**: silence duration between verbal elements. A high pause ratio indicates a reduction in *Temporal Demand*. This relief can precede transitions in the narrative [287].

The role of the prosodic markers lies mainly in the *Mental Demand* and *Temporal Demand* dimensions within the TLX procedure.

(2) Loudness and Intensity Features

These descriptors encode changes in perceptive energy:

- **RMS energy**: short-time signal power.
- **LUFS loudness**: perceptually weighted loudness measure.

- **Dynamic range**: difference between high energy and low energy regions.

Studies in auditory attention have determined that as the volume and variability of the auditory cue increase, listener effort and vigilance also increase . Both RMS and LUFS contribute to the TLX *Effort* subscale, and dynamic range reduction is typical of calm narrative sections [288].

(3) Voice Activity Detection and Turn-Taking Structure

Voice activity detection (VAD) helps in distinguishing speech and silence [289]:

Patterns of speech affect *Temporal Demand* (when rapid alternation forces processing pressure) [290] as well as *Effort* (when high speech overlap forces a high perceptual load) [291].

6.6.4 Segment-Level Aggregation

Audio features are first computed at fine temporal resolution and then summarised within each analysis window W_k. For windows spanning multiple WhisperX segments, feature values are aggregated using statistics such as mean, variance, and silence proportion. These aggregated descriptors form the audio portion of the structured reasoning prompt (A.4.2) used by the LLM.

This aggregation is essential for segment-level load estimation. The goal is not merely to detect instantaneous spikes in energy or speech rate, but to model sustained perceptual patterns that correspond to narrative rhythm. Windows that exhibit stable speech rates, consistent loudness, and ample pauses tend to be cognitively easier, while those with compressed prosody or high dynamic intensity correspond to cognitively demanding intervals.

6.6.5 Connection to the Research Question

Accurately estimating audiovisual cognitive load requires identifying the perceptual conditions under which a viewer is more or less likely to tolerate an interruption. Audio features offer a direct lens into the *temporal structure* of information flow, making them indispensable for:

- detecting implicit narrative boundaries,
- differentiating high- and low-pressure segments,
- supporting the LLM's TLX reasoning step with grounded perceptual cues,
- increasing precision in identifying low-load segments suitable for ad placement.

Thus, audio processing contributes not only raw values but an interpretive scaffold for cognitive load estimation, enabling the broader multimodal framework to answer the central thesis question: *Can we identify cognitively light segments using only content-derived cues?* The results in Chapter 6 demonstrate that audio features, when fused with visual and semantic information, greatly improve segmentation reliability and user-experience alignment.

Formal Definition of Audio Sensing

Let $W_k = [t_0^k, t_1^k)$ denote a segment-aligned analysis window and let $\mathcal{A}(t)$ be the continuous audio signal. For each window, a vector of low-level prosodic descriptors is extracted:

$$\mathbf{a}_k = \Phi(\mathcal{A}(t) \mid t \in W_k),$$

where $\Phi(\cdot)$ denotes the feature extraction operator computing speech rate, articulation rate, pause ratio, RMS energy, LUFS loudness, dynamic range, and VAD-based turn-taking statistics.

To ensure cross-video comparability, each feature dimension is standardised via per-video z-scoring:

$$\hat{\mathbf{a}}_k = \frac{\mathbf{a}_k - \mu(\mathbf{a})}{\sigma(\mathbf{a})}.$$

The resulting standardised audio sensing vector $\hat{\mathbf{a}}_k$ constitutes the auditory perceptual input to the NASA–TLX reasoning module and primarily informs the *Temporal Demand* and *Effort* dimensions during TLX fusion.

6.7 Visual Feature Extraction

While audio governs much of the information density in narrative video, the visual channel plays an equally important role in shaping the viewer's cognitive experience. Editing style, motion patterns, shot duration, and scene composition all influence how much perceptual effort a viewer must allocate at any given moment [292]. In film and media psychology, rapid visual changes, kinetic motion, and dense scene layouts are strongly associated with elevated perceptual load and increased attentional demand [292]. Conversely, long static shots and visually simple scenes allow viewers to relax, consolidate information, and transition between narrative units with minimal cognitive friction [278, 292].

For these reasons, visual analysis is a key component of the proposed multimodal framework. This chapter details how the visual stream is processed, why the extracted features relate to cognitive load, and how they support the overall goal of identifying naturally low-load segments suitable for unobtrusive transitions or ad insertion.

6.7.1 Motivation: Why Visual Cues Matter for Cognitive Load

Vision can be considered in relation to the 'effort' and 'temporal demand' components of cognitive load. The notion that editing tempo, movement on screen, and the complexity of the 'mise-en-scène' impact moment-to-moment processing needs on the viewer can be described by film editing techniques/practices [293]. The important concepts in film studies from a perception perspective include:

- **Shot duration regulates cognitive pacing:** shorter shots require rapid attentional switching, increasing *Temporal Demand* [294, 295].
- **Motion intensity increases visual effort:** high optical flow necessitates continuous perceptual tracking, elevating *Effort* [295].
- **Scene complexity influences interpretive load:** more objects, people, or concurrent activities require additional mental organisation [296].

These elements will help determine which moments a viewer will find cognitively "busy" versus "calm" about a scene. A calm moment will likely correspond with a scene

break, making these ideal candidates for commercials.

6.7.2 Shot Boundary Detection and Its Importance

The initial step in the process is to identify shot boundaries through the use of `PySceneDetect` with the aim of determining cuts through histogram difference [237]. Shot units are basic units of vision, comprising continuous arrays of images shot using the same configuration [297].

Identifying shots provides three benefits for cognitive-load modelling:

1. It segments the video into perceptually meaningful units aligned with editorial decisions rather than arbitrary frame windows.
2. It allows measurement of shot duration, a well-known correlate of perceptual load and narrative pacing.
3. It provides anchor points for keyframe sampling and captioning.

By organising the visual analysis around shot structure, the framework mirrors the viewer's natural perceptual segmentation of the video.

6.7.3 Low-Level Visual Features

For each shot, several visual descriptors are extracted to capture perceptual dynamics [295]:

(1) Motion Intensity (Optical Flow)

Optical flow quantifies pixel-level movement between frames. High motion typically increases both perceptual load and attentional vigilance [295]. In narrative content, intense motion often corresponds to:

- action sequences,
- rapid demonstrations or fast-paced tutorials,
- chaotic or emotionally heightened scenes.

Normalised motion measurements thus directly inform the *Effort* component of the NASA–TLX mapping.

(2) Shot Duration and Cut Frequency

The number of shots in a sequence can be seen as a marker for *Temporal Demand*. Short shot sequences literally require the viewer's constant refocusing on visual information [293], whereas shots with considerable lengths, such as medium-length interviews or instructional explanations delivered in a classroom-like setting, facilitate cognitive processing, which in turn functions as an indicator of transitions within the narrative. The number of cuts per second indicates this information [295].

(3) Scene Complexity Indicators

To approximate the visual richness of each shot, a set of lightweight heuristics is computed:

- **Face count:** Because people and their social relations are primary anchors of attention and event model construction in visual narratives [298], the number of visible faces is used as a lightweight proxy for social and conversational complexity.
- **Keyframe entropy or texture variability:** indicates whether the scene is visually simple or dense [299].
- **Foreground–background contrast:** It involves how much perceptually separate are the major components of visuals (the foreground) and the overall scene context (background). It implies that there is an obvious difference between important objects/actions and the background, such as a distinction based on colours, movements, or space allocation. Strong contrast may indicate highly perceptually noticeable/attention demanding areas while weak contrast demands more perceptual processing efforts [293].

As face count acts as an effective surrogate measure for cognitive load arising from social and interaction-related factors, there will be situations when it is not possible to apply the measure, due to the absence of human characters in the frames under consideration. Indeed, for non-human-centric video clips such as landscapes or object-

based sequences, it might not be possible to derive any valuable conclusions based on the presence and number of faces. This issue can be overcome by applying a range of additional measures for generalization of the framework. For instance, the density of objects, degree of clutter, and motion can be utilized as alternative proxies for cognitive demand.

These features serve as proxies for how much visual information a user must integrate in a short time, influencing both *Mental* and *Effort* demands.

6.7.4 Semantic Visual Features via VLM Captioning

Low-level motion and pacing features capture perceptual load but miss higher-level semantic cues. To complement them, the system extracts representative keyframes from each shot and generates descriptions using a vision–language model (VLM) [300]. Each caption captures elements such as:

- the scene setting (e.g., "a wide static shot of a kitchen"),
- the activity (e.g., "a presenter demonstrating a tool"),
- emotional tone (e.g., "calm," "intense"),
- presence of on-screen text or graphics,
- indicators of transitions (e.g., "fade to black," "establishing shot").

These captions enable the LLM to interpret the visual context and structural role of each shot, linking perceptual information to narrative segmentation cues rather than conceptual workload estimation. This step is essential for detecting cognitive shifts that arise not only from movement but from the narrative content itself.

6.7.5 Segment Aggregation

Although features are extracted at the shot level, decisions are made at the level of larger narrative segments S_b, formed by the annotated human boundaries. Within each segment, shot-level descriptors are aggregated. This aggregation reduces sensitivity to short-lived spikes such as a single rapid cut and better represents the cognitive experience over coherent narrative units.

Formal Definition of Visual Sensing

Let the video frames be $F(t)$ and let $\{s_i\}_{i=1}^{N}$ denote the set of detected shots, where each shot s_i spans an interval $[a_i, b_i)$. For each segment-aligned analysis window $W_k = [t_0^k, t_1^k)$, a visual sensing vector is constructed by aggregating shot-level descriptors that overlap the window:

$$\mathbf{v}_k = \Psi\big(\{s_i : [a_i, b_i) \cap W_k \neq \emptyset\}\big), \tag{6.4}$$

where $\Psi(\cdot)$ includes pacing and structural statistics such as cut count, cut rate, and mean shot duration computed over the overlapping shots.

To model semantic visual change, each shot is represented by a keyframe caption c_i produced by a vision–language model (Section 6.7.4). Let $e_i = \mathrm{Emb}(c_i)$ denote a caption embedding. A caption-change proxy is then computed via cosine distance between consecutive shot embeddings:

$$d_i = 1 - \cos(e_i, e_{i+1}). \tag{6.5}$$

Given a time increment Δt_i between consecutive captions, the window-level semantic-motion proxy is obtained by a time-weighted aggregation:

$$\mathrm{motion}_k = \frac{\sum_{i \in \mathcal{I}_k} d_i \, \Delta t_i}{\sum_{i \in \mathcal{I}_k} \Delta t_i}, \tag{6.6}$$

where $\mathcal{I}_k$ indexes consecutive caption pairs whose timestamps fall within (or overlap) W_k. This quantity acts as a lightweight proxy for visually meaningful change, capturing shifts in depicted content and scene role in addition to raw motion.

As in the audio modality, cross-video comparability is achieved by per-video standardisation:

$$\hat{\mathbf{v}}_k = \frac{\mathbf{v}_k - \mu(\mathbf{v})}{\sigma(\mathbf{v})}. \tag{6.7}$$

The standardised visual sensing vector $\hat{\mathbf{v}}_k$ constitutes the visual input to the NASA–TLX reasoning module and primarily informs *Temporal Demand* (via pacing) and *Effort* (via motion and structural complexity), while shot captions provide additional evidence

for structural segmentation cues.

6.8 Refined Semantic Integration or Semantic–TLX Mapping

Traditionally, cognitive load is estimated based on perceptual information such as motion, pacing, and acoustic intensity. However, for narrative videos, cognitive load is not just driven by information that stimulates the senses but is also affected by the semantic complexity of the information being processed. This is addressed by the proposed framework with the introduction of a semantic sensing modality that is a first-class cognitive sensor, along with auditory and vision perception.

For each segment-aligned window W_k, a fused textual representation, denoted as *window_context*, is constructed by temporally aligning shot-level scene descriptions and word-level transcript portion. This representation provides a compact narrative abstraction that encodes discourse structure, topic density, and conceptual continuity within each narrative unit.

Rather than being treated as auxiliary context, the semantic modality directly informs the NASA–TLX Mental Demand (MD) and Effort (E) dimensions. Linguistic complexity, topic compression, and discourse cohesion captured in *window_context* are interpreted by the LLM reasoning module to infer semantic workload characteristics. These semantic cues further modulate the adaptive weighting of TLX dimensions, thereby influencing both the magnitude and composition of the final cognitive load estimate.

Let $\mathrm{Sem}(W_k)$ denote the semantic sensing function operating on *window_context*. The resulting semantic workload representation contributes to the TLX fusion as:

$$MD_k, E_k \sim \mathrm{Sem}(W_k), \tag{6.8}$$

where MD_k and E_k represent the Mental Demand and Effort components of the NASA–TLX formulation for window W_k.

By explicitly modelling semantic perception as a cognitive sensing channel, the pro-

posed framework extends NASA–TLX beyond its traditional questionnaire-based usage and operationalises it as a multimodal cognitive fusion operator. This formulation enables segment-level cognitive load estimation to reflect perceptual stimulation.

Window-Level Semantic Construction

Let $W_k = [t_0^k, t_1^k)$ denote a segment-aligned analysis window. Let $\mathcal{S}_k = \{S_i\}$ be the set of visual scene summaries whose timestamps overlap W_k, where each S_i is produced by the vision–language captioning and summarisation pipeline described in Chapter 5. Similarly, let $\mathcal{T}_k = \{u_j\}$ be the set of transcript utterances overlapping W_k, obtained from word-level ASR alignment.

A raw multimodal semantic trace for the window is formed by concatenating these ordered elements:

$$X_k = \mathrm{concat}\big(\mathcal{S}_k, \mathcal{T}_k\big), \tag{6.9}$$

where ordering follows the global video timeline.

To obtain a stable, noise-resilient representation suitable for controlled reasoning, the trace X_k is compressed using a constrained LLM summarisation operator $g(\cdot)$ (temperature = 0, fixed prompt), yielding the fused semantic window representation:

$$window\_context_k = g(X_k). \tag{6.10}$$

This construction matches the window-level re-summarisation mechanism introduced in Chapter 5 and provides the semantic sensing input used in the NASA–TLX reasoning procedure in this chapter.

6.9 Segment-Level Load Statistic and Acceptance Rule

A central component of the proposed framework is the transformation of the continuous cognitive load estimates into a segment-level representation that can be used to make actionable decisions about where transitions such as advertisement insertions or playback pauses should occur. While the cognitive-load curve $\tilde{CL}(t)$ provides temporally

localized estimates, reliable boundary decisions require reasoning at a larger temporal granularity that reflects how humans perceive narrative flow. This section introduces the segment-level statistic used for this purpose and the acceptance rule that determines whether a given narrative boundary is considered a cognitively suitable transition point.

6.9.1 Narrative Segments as the Decision Unit

In narrative video, cognitive load does not fluctuate arbitrarily; rather, its overall intensity can be meaningfully characterized at the level of semantically coherent storytelling units [257]. Human editors naturally align transitions with these units, placing breaks just after periods of low perceptual and semantic intensity. To reflect this, the narrative segment (i.e., the interval between two adjacent annotated chapters) is adopted as the fundamental decision region. Let b denote a boundary and $S_b = [b, b_{\text{next}})$ the segment immediately following that boundary.

Cognitive load is estimated over the duration of each segment S_b, and the resulting segment-level value is assigned to the terminal boundary b_{next}. The question addressed here is: given the cognitive demands experienced throughout the preceding segment, does the upcoming boundary b_{next} represent a cognitively suitable point for interruption?

This decision rule is consistent with established practices in editing and media psychology: interruptions are best tolerated immediately after segments that, considered holistically, allow the viewer to cognitively settle, rather than after periods of heightened narrative or perceptual demand.

6.9.2 Defining the Segment-Level Statistic

To quantify the overall cognitive characteristics of each narrative segment S_b, the window-level cognitive load scores within the segment are aggregated into a single representative value:

$$L(S_b) = \text{stat}\Big(\{\tilde{C}L(t) : t \in S_b\}\Big). \tag{6.11}$$

where $\tilde{C}L(t)$ denotes the cognitive load score of the analysis window containing time t.

The window-level cognitive load is computed as a weighted combination of three NASA-TLX-inspired dimensions: Mental Demand (MD), Effort (E), and Temporal Demand (TD), using adaptive coefficients (α, β, γ). These coefficients are not fixed a priori but are dynamically inferred for each window through LLM-assisted reasoning. This adaptive weighting is guided by qualitative analysis of cognitive load trajectories and their alignment with human-annotated transition points, allowing the model to capture meaningful variations in viewer cognitive effort across different content styles.

This formulation reflects the observation that different modalities contribute unevenly to cognitive load (for example, dialogue-heavy segments tend to increase mental demand, while rapid visual edits increase temporal demand).

In the default configuration, stat is the arithmetic *mean*, which provides a balanced representation of the viewer's overall cognitive experience throughout the segment.

However, because some segments may contain brief high-load spikes that can be diluted by averaging, a robust alternative based on the *25th percentile* (representing the cognitively easiest portions of the segment) is also evaluated in the ablation study. This allows assessment of the decision rule's sensitivity to noise and momentary fluctuations under the current windowed estimation approach.

6.9.3 Acceptance Rule for Transition Suitability

Once the segment statistic $L(S_b)$ is computed, the framework determines whether boundary b constitutes a *suitable* transition point. A boundary is accepted if the cognitive load of the upcoming segment is below a threshold:

$$\hat{y}(b) = \begin{cases} 1, & L(S_b) \leq \tau_{\text{seg}}, \\ 0, & \text{otherwise.} \end{cases} \tag{6.12}$$

Here, τ_{seg} represents an operating threshold that reflects the desired balance between *comfort* and *coverage*. A high threshold (e.g., 80) accepts more boundaries and thus increases the number of potential ad or pause positions, whereas a low threshold (e.g., 60) accepts only the "safest" boundaries those followed by segments that are very easy

for the viewer to process. Choosing τ_{seg} therefore provides a direct control knob for user experience: platforms with aggressive monetisation goals may adopt a lenient setting, while premium or subscription-based services may prefer stricter settings that prioritise comfort.

6.9.4 Why the Segment-Level Approach Is Necessary

Evaluating load at the segment level is essential for three reasons:

1. **Narrative continuity.** Cognitive demand is shaped by story flow, not merely instantaneous audiovisual cues. Segment aggregation reflects how humans process information over meaningful units.

2. **Avoiding shallow dips.** Instantaneous minima brief decreases in cognitive load may occur even when the overall segment remains intense. Evaluating $L(S_b)$ ensures transitions are not placed at misleading dips that precede cognitively heavy content.

3. **Robustness and interpretability.** Segment-level statistics produce stable, explainable decisions. A threshold on $L(S_b)$ is easier to analyse, tune, and justify than frame-level heuristics.

This reasoning aligns with findings in multimedia cognition, where viewers' ability to tolerate interruptions depends less on momentary fluctuations and more on whether the *upcoming* narrative unit is cognitively demanding.

6.9.5 Algorithmic Specification of Cognitive Load–Aware Segmentation

The preceding sections introduced the multimodal audio and visual feature extraction procedures (Sections 6.5–6.6), the segment-level cognitive load statistic (Section 6.7.2), and the acceptance rule for cognitively suitable segmentation boundaries (Section 6.7.3). This subsection presents an explicit algorithmic specification of the proposed cognitive load–aware segmentation framework.

As illustrated in Fig. 6.2, synchronized audio–visual pacing cues are mapped to NASA–TLX–inspired workload dimensions and fused through an LLM-assisted reasoning mechanism. For each audio–shot window $W_k = [t_0^k, t_1^k)$, standardized multimodal descriptors are interpreted to obtain Mental Demand (MD_k), Effort (E_k), and Temporal Demand (TD_k), together with adaptive importance weights $(\alpha_k, \beta_k, \gamma_k)$ such that $\alpha_k + \beta_k + \gamma_k = 1$. These quantities are combined to produce a window-level cognitive load value

$$CL_k = \alpha_k MD_k + \beta_k E_k + \gamma_k TD_k, \tag{6.13}$$

which defines a piecewise-constant cognitive load trajectory over the video timeline. Optional smoothing yields the continuous load curve $\tilde{CL}(t)$. All boundary acceptance decisions and quantitative evaluations are derived directly from the unsmoothed window-level cognitive load values CL_k. No smoothing is applied prior to computing segment-level statistics or threshold-based acceptance.

Candidate segmentation boundaries b, derived from the theme-aware reasoning layer introduced in Chapter 5, are evaluated using the segment-level statistic $L(S_b)$ defined in Eq. (6.1). A boundary is accepted only if the corresponding segment load satisfies the acceptance rule in Eq. (6.2), ensuring that final segmentation points occur at narrative intervals of low cognitive demand. This algorithmic formulation provides a deterministic and training-free operational definition of the cognitive suitability component C_t of the AdBreakScore framework (Table 6.3).

Algorithm 3 summarises the complete cognitive load–aware segmentation pipeline, consolidating multimodal perception, LLM-assisted NASA–TLX reasoning, and segment-level acceptance into a single executable procedure.

6.9.6 Summary

The segment-level load statistic and acceptance rule provide the final step in the cognitive-load–aware segmentation pipeline. By integrating perceptual features, semantic reasoning, and narrative structure into a unified criterion, this mechanism identifies the transition points that align most closely with human expectations of comfortable, non-intrusive editing. The next section evaluates how effectively this decision process

matches human judgements across a diverse video dataset.

Algorithm 3 LLM-Assisted Multimodal NASA–TLX Cognitive Load Estimation and Segmentation

Require: Video $V = (A, F)$; windows $\{W_k = [t_0^k, t_1^k)\}_{k=1}^K$; candidate boundaries $\mathcal{B}$; threshold τ_{seg}; statistic stat$(\cdot)$
Ensure: Window loads $\{CL_k\}_{k=1}^K$; curve $\tilde{CL}(t)$; accepted boundaries $\mathcal{B}_{\text{acc}}$

1) Tri-modal feature extraction (Semantic, Audio, Visual)
for $k \leftarrow 1$ to K **do**
    Construct semantic window representation $window\_context_k$ by aligning scene captions and transcript excerpts within W_k.
    Extract visual cues in W_k (cut rate, mean shot duration, motion / caption-change proxy, speaker switches).
    Extract audio cues in W_k (speech rate, pause ratio, loudness/RMS, LUFS, slope).
end for

2) Per-video standardisation
for $k \leftarrow 1$ to K **do**
    Z-score numeric window features to obtain standardized descriptor $\mathbf{z}_k$.
end for

3) LLM-assisted NASA–TLX mapping
for $k \leftarrow 1$ to K **do**
    Query LLM with $(\mathbf{z}_k, window\_context_k)$ to obtain $(MD_k, E_k, TD_k, \alpha_k, \beta_k, \gamma_k)$.
    $CL_k \leftarrow \alpha_k MD_k + \beta_k E_k + \gamma_k TD_k$ with $\alpha_k + \beta_k + \gamma_k = 1$.
end for

4) Construct load curve
Define $CL(t) = CL_k$ for all $t \in [t_0^k, t_1^k)$ and optionally smooth to obtain $\tilde{CL}(t)$.

5) Segment-level scoring and acceptance
$\mathcal{B}_{\text{acc}} \leftarrow \emptyset$
for all $b \in \mathcal{B}$ with corresponding segment S_b **do**
    $L(S_b) \leftarrow \text{stat}\big(\{\tilde{CL}(t) : t \in S_b\}\big)$.
    if $L(S_b) \leq \tau_{\text{seg}}$ **then**
        $\mathcal{B}_{\text{acc}} \leftarrow \mathcal{B}_{\text{acc}} \cup \{b\}$.
    end if
end for

6.10 Evaluation Methodology

This section describes the methodological framework used to assess the accuracy, stability, and practical usefulness of the proposed multimodal cognitive-load–aware segmentation approach. The evaluation is designed to answer three key questions:

- **How well does the model identify segment boundaries that human viewers consider cognitively comfortable?**
- **Does multimodal reasoning offer improvements over signal-only heuristics commonly used in video segmentation?**

To address these questions, the system is evaluated on narrative-aligned content using human annotations of acceptable transition points together with both quantitative and qualitative metrics.

6.10.1 Dataset Selection and Rationale

To analyse the proposed method, a carefully selected subset of 20 videos from the *VidChapters-7M* dataset [244] is used, which is currently one of the largest public datasets for chapter segmentation in videos. A subset of 20 videos was selected to enable detailed dual-annotator cognitive suitability labelling, which is infeasible at full dataset scale. While the evaluation is conducted on 20 videos, each video is segmented into multiple annotated units (over 120 total decision points), enabling fine-grained assessment of transition suitability. These videos include annotated chapter or scene markers as a set of *narrative reference points*. Human-annotated chapter boundaries from VidChapters-7M are treated as ground-truth narrative segmentation units. These boundaries are not ground truth for cognitive load, but serve as validated narrative structure over which cognitive load is estimated. VidChapters-7M is different from existing datasets in the following manner:

- multi-minute to hour-long videos,
- heterogeneous genres (documentaries, interviews, explainers, vlogs),

- real-world editing styles, including hard cuts, slow fades, and smooth transitions.

These parameters closely reflect the environments in which consumer-facing segmentation systems need to work, thereby making this dataset a test bed for the proposed cognitive load approach.

6.10.2 Human Annotation Procedure

To establish a human-grounded reference for suitable transition points, two independent annotators (senior undergraduate students with training in video analysis and media psychology) reviewed each video. For every candidate narrative boundary, annotators assigned a binary label: "OK" (interruption unlikely to disrupt narrative flow or viewer experience) or "Not-OK" (interruption would feel abrupt or intrusive). Annotations were informed by clear criteria including semantic closure, pacing, perceived cognitive load, and overall viewer comfort.

Inter-rater reliability was substantial (Fleiss' $\kappa = 0.76$), indicating strong consistency beyond chance. Disagreements were resolved through discussion and majority vote, yielding a final consensus ground truth. While human annotations of perceptual and cognitive qualities are inherently subjective and may vary with individual differences in media consumption habits, the high inter-rater agreement supports their reliability as a proxy for viewer experience. This approach follows established practices in video segmentation and affective computing research, where direct physiological measurement at scale is impractical. The observed alignment between model outputs and these human judgments provides meaningful evidence of ecological validity.

6.10.3 Experimental Pipeline

For each video the full processing pipeline described in Chapter 6 is executed end-to-end:

1. **Audio analysis:** computation of prosodic, loudness, temporal, and rhythm features.

2. **Visual processing:** shot detection, motion analysis, and VLM keyframe captioning.

3. **Text extraction:** WhisperX-based transcription with word-level timestamps.

4. **LLM-assisted integration:** reasoning-based estimation of NASA-TLX dimensions (MD/E/TD) per temporal window. All LLM reasoning in this stage is executed with temperature = 0 to ensure deterministic decoding under a fixed model version, fixed prompts, and identical input features. Outputs are constrained via a predefined schema enforcing structured JSON-formatted responses. Given fixed prompts and identical input features, the resulting TLX weights (α, β, γ) and dimension scores are therefore reproducible and deterministic under a fixed model version, fixed prompts, and identical input features

5. **Cognitive-load curve generation:** smoothing and normalization.

6. **Segment-level reasoning:** computation of the segment statistic $L(S_b)$ for each annotated boundary.

7. **Binary decision:** determining whether a boundary is acceptable based on threshold τ_{seg}.

The restriction of boundaries of candidate shots to annotated narrative points is reminiscent of real-world video editing where transitions between shots tend to take place at structural boundaries rather than arbitrary timestamps.

6.10.4 Operating Points and Threshold Selection

To determine how strictly the system filters candidate transitions, the following threshold settings are evaluated:

$$\tau_{\text{seg}} \in \{60, 70, 80\}.$$

These correspond to:

- **Lenient (80):** Recall is maximized, and more potential ad opportunities are accepted.
- **Medium (70):** Precision and recall are balanced.

- **Strict (60):** User comfort and narrative smoothness are prioritised.

To ensure fair evaluation and prevent test-set leakage, threshold selection was performed on a held-out development subset via grid search over the above candidate values. The acceptance threshold τ_{seg} was then fixed to 70 for all subsequent experiments. No per-video or per-genre tuning was performed.

This global operating point demonstrated stable performance across diverse narrative styles, supporting its generalisability and deployment suitability.

6.10.5 Evaluation Metrics

To quantify performance, the following metrics are used:

- **Precision** - proportion of boundaries predicted as OK that humans also judged as OK.
- **Recall** - proportion of human-acceptable boundaries correctly found by the model.
- **F1-score** - harmonic mean of precision and recall, balancing both.
- **Average Segment Cognitive Load (Avg. Seg. CL)** - average value of $L(S_b)$ over all boundaries accepted by the model.

While the first three metrics measure agreement with human judgement, Avg. Seg. CL provides an important sanity check: a lower value indicates the model selects boundaries that truly correspond to cognitively light intervals, even if no ground truth were available.

6.10.6 Rationale for This Evaluation Strategy

The evaluation methodology is intentionally aligned with human cognitive processing and real-world media workflows. It assumes that:

- viewers perceive content at the level of segments, not frames;
- interruptions should occur before cognitively easy stretches;
- cognitive load can be estimated from content-derived audiovisual cues;
- human annotation provides a trustworthy reference for acceptable transitions.

By combining human-grounded labels, a narrative-aligned dataset, multimodal pro-

cessing, and both quantitative and qualitative evaluation metrics, the methodology provides a comprehensive assessment of how effectively the proposed framework models cognitive suitability in real-world video content.

6.11 Results and Performance Analysis

This section presents a detailed analysis of the model's behaviour on the 20-video evaluation subset, examining both quantitative performance metrics and qualitative patterns that emerge from the cognitive-load–aware segmentation process. Building on the methodological framework defined in Section 6.10, the analysis reports how the model compares to content-based heuristics, how its predictions align with human judgments, and how stable the system is across heterogeneous narrative styles.

The results demonstrate three overarching findings:

- the proposed multimodal cognitive-load model consistently outperforms classical signal-only heuristics;
- threshold choice (τ_{seg}) provides an interpretable means of balancing the number and comfort of predicted boundaries;
- the system exhibits low variance across videos, indicating robustness to genre diversity and editing conventions.

6.11.1 Binary Prediction Performance

Table 6.1 reports precision, recall, and F1 for three operating thresholds. The strongest performance occurs at $\tau_{\text{seg}} = 70$, achieving an F1-score of **0.66**, which represents a meaningful improvement over signal-level baselines such as pause-length or energy dips.

High-level observations include:

- **Low thresholds (60):** favour precision. The model accepts only boundaries followed by exceptionally low-load segments, resulting in fewer predictions with high confidence. However, this strictness reduces recall, as some human-acceptable transitions are excluded.
- **High thresholds (80):** favour recall. A broader set of boundaries is accepted,

Table 6.1: Binary evaluation of transition suitability using the next-segment criterion ($L(S_b) \leq \tau_{\text{seg}}$).

Method	Precision ↑	Recall ↑	F1 ↑	Avg. Seg. CL ↓
Random-Boundary	0.33	0.30	0.31	64.2
Longest-Pause Boundary	0.45	0.39	0.42	59.8
Energy-Dip Boundary	0.49	0.42	0.45	58.7
CL-Minima (τ_{seg}=80)	0.50	0.72	0.59	62.1
CL-Minima (τ_{seg}=70)	**0.69**	0.63	**0.66**	58.8
CL-Minima (τ_{seg}=60)	0.78	0.51	0.62	**54.5**

capturing more potential insertion opportunities; however, this comes at the cost of increased false positives, resulting in reduced precision and a lower overall F1-score relative to human judgement.

- **Medium thresholds (70):** achieve the best trade-off between precision and recall. This balanced operating point maximises the F1-score, reflecting strong alignment with human judgement while maintaining practical insertion coverage.

Across all thresholds, the **Average Segment Cognitive Load (Avg. Seg. CL)** decreases as the threshold tightens, confirming that the acceptance rule systematically selects cognitively easier intervals when stricter settings are applied.

6.11.2 Stability Across Narrative Styles

In addition to absolute performance, the stability of the model across the 20 evaluation videos is examined. Table 6.2 reports mean and standard deviation for precision, recall, F1-score, and Avg. Seg. CL using the best threshold configuration.

The standard deviations observed (**0.05–0.06**) are small relative to the metric ranges, indicating:

- strong consistency across genres such as documentaries, talk shows, instructional videos, and narrative vlogs;
- resilience to variations in pacing, shot density, and speech patterns;
- reliable behaviour even when the narrative structure is unconventional (e.g., expository monologues or conversation-driven content).

This is particularly important for consumer media systems, where the variety of

Table 6.2: Performance stability for CL-Minima (τ_{seg}=70). Mean ± SD across 20 videos.

Metric	Prec.↑	Rec.↑	F1↑	Avg. Seg. CL↓
Mean ± SD	0.69 ± 0.06	0.63 ± 0.06	0.66 ± 0.05	58.8 ± 0.4

content is extremely broad and models must generalise beyond curated or domain-specific datasets.

6.11.3 Qualitative Behaviour of Accepted and Rejected Boundaries

A qualitative inspection of the accepted boundaries reveals that the model typically selects transitions followed by segments exhibiting:

- slower or more uniform prosody,
- reduced visual activity (longer shots, low optical flow),
- semantic indicators of closure or topic completion.

By contrast, boundaries rejected by the decision rule usually precede segments characterised by:

- rapid or overlapping speech,
- high motion intensity or dense shot changes,
- captions or transcripts signalling unresolved tension, open questions, or escalation.

These observations indicate that the model captures not only low-level audiovisual patterns but also higher-order narrative signals derived from LLM reasoning.

6.11.4 Qualitative Visualization of Cognitive Load Curves

In addition to the quantitative evaluation, qualitative visualizations of the segment-wise cognitive load trajectories computed by the framework are provided. These visualizations provide an intuitive view of the way multimodal features and the LLM-assisted reasoning are mapped to a continuous cognitive load signal over time.

Figure 6.3 illustrates a representative segment-wise NASA–TLX cognitive load curve for a narrative-aligned video. The horizontal axis corresponds to playback time, while the vertical axis denotes the estimated cognitive load score. Vertical red markers indicate human-annotated narrative boundaries, and dashed horizontal lines show the

operating thresholds ($\tau_{\text{seg}} = 60, 70, 80$) used for transition suitability decisions. The blue curve represents the smoothed segment-level cognitive load produced by the proposed model.

Across the trajectory, variations in the cognitive load signal reflect changes in the audiovisual and semantic properties of successive segments. Regions with relatively stable or slowly varying load values correspond to segments characterised by lower temporal pacing or reduced multimodal activity, while sharper increases in load are associated with segments exhibiting denser speech patterns, increased visual motion, or heightened narrative progression.

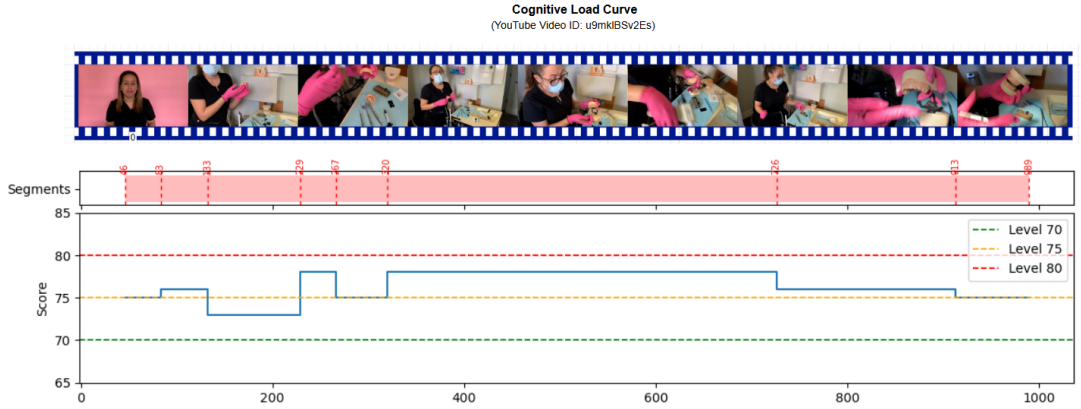


Figure 6.3: Segment-wise cognitive load curve for a representative narrative-aligned video. Vertical red lines denote human-annotated boundaries and dashed horizontal lines indicate operating thresholds.

Figure 6.4 shows an additional illustration of the process in which the same cognitive load process will have varying sets of potential candidate transitions based upon evaluation against varying cognitive loads. The figure demonstrates the predicted boundary positions based upon varying threshold settings of low, medium, and high threshold settings, providing insight into the process in which threshold settings will have varying impacts upon the overall process without having an impact upon the cognitive loads that are estimated.

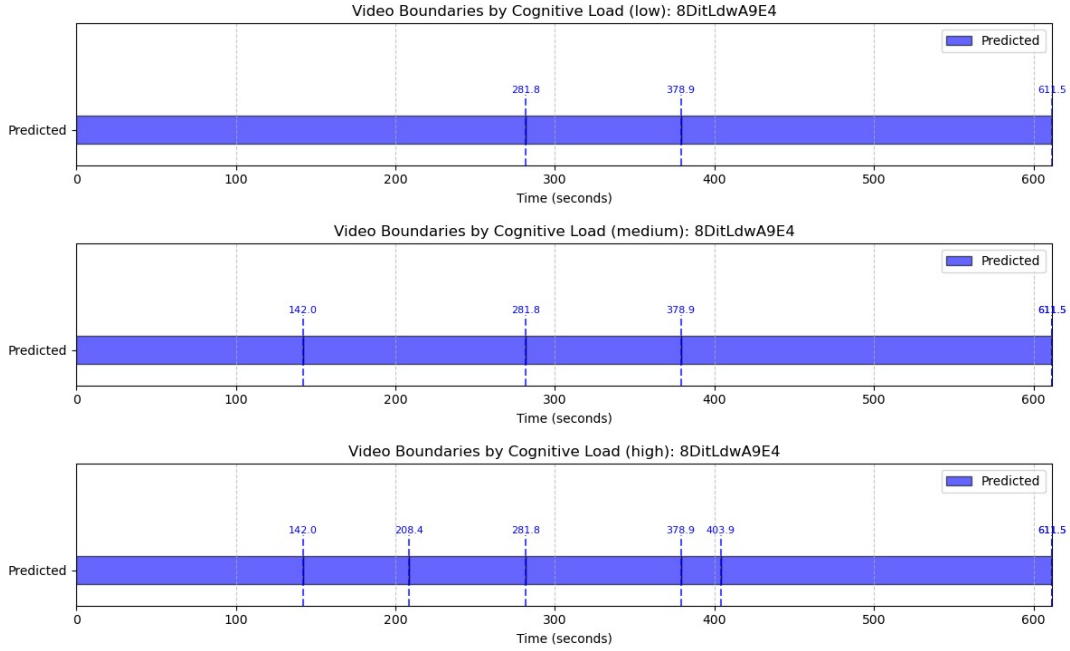


Figure 6.4: Illustration of candidate transition windows obtained under low, medium, and high cognitive load thresholds for the same underlying cognitive load trajectory.

These visualizations show that significant patterns exist and that these patterns can be explained and accounted for, providing interpretable evidence for its effectiveness.

6.11.5 Impact of LLM-Assisted TLX Reasoning

The proposed framework estimates cognitive load for each video segment k as

$$CL_k = \alpha_k MD_k + \beta_k E_k + \gamma_k TD_k, \tag{6.14}$$

where the LLM dynamically assigns the weighting coefficients $(\alpha_k, \beta_k, \gamma_k)$ according to the semantic and perceptual characteristics of the segment. Unlike conventional fixed-weight NASA–TLX formulations, this approach allows the relative influence of Mental Demand (MD), Effort (E), and Temporal Demand (TD) to vary contextually.

Illustrative Adaptation Behaviour. Table 6.3 provides representative examples of how the LLM reallocates weights depending on scene characteristics. Conceptually demanding dialogue segments increase α_k, while visually intense or fast-paced sequences

Table 6.3: Illustrative examples of adaptive NASA–TLX weighting. Weights sum to 1.0.

Scene type	**Feature cues (summary)**	**Weights** (α, β, γ)	**Primary emphasis**
Calm dialogue	Low motion; dense technical speech; static visuals	(0.60, 0.20, 0.20)	Mental Demand (MD)
Fast chase	High motion; frequent cuts; loud dynamics; sparse dialogue	(0.20, 0.40, 0.40)	Effort (E) and Temporal Demand (TD)
Montage transition	Moderate motion; short shots; limited speech	(0.30, 0.30, 0.40)	Temporal Demand (TD)
Didactic slide	Low motion; jargon-dense narration; long shots	(0.55, 0.25, 0.20)	Mental Demand (MD)

elevate β_k and/or γ_k. These examples clarify the qualitative reasoning mechanism underlying adaptive weighting.

Empirical Window-Level Adaptation. To demonstrate that adaptive reweighting occurs in practice rather than only in illustrative examples, Table 6.4 reports real window-wise coefficients extracted from representative videos. The variation of $(\alpha_k, \beta_k, \gamma_k)$ across windows shows that rather than applying static weights, the LLM actually dynamically balances the importance of MD, E, and TD based on segment-level features.

Across windows, the non-zero variation in α_k, β_k, and γ_k confirms that the model performs context-sensitive reweighting rather than applying fixed coefficients. This dynamic behaviour helps prevent reliance on a single modality, supports generalization across genres, and is responsible for the improvement in F1-score compared to non-adaptive baselines. Ablation studies also show that disabling adaptive weighting or any demand dimension reduces classification performance.

6.11.6 Comparison with Baseline Heuristics

The model outperforms all classical baselines evaluated, including:

- longest-pause detection,
- energy-dip selection,
- random selection (as a sanity check).

The signal-only heuristics have a problem because they do not incorporate any semantic understanding. For instance:

- There may be a long pause during a tense situation despite a heavy narrative load.
- An interval of low energy may also indicate background silence rather than the end of a narrative.

The proposed system combines the meaning of transcripts, VLM scene descriptions, and multimodal information to overcome the aforementioned failures and make a more human-like decision.

6.11.7 Summary of Findings

The performance shows these factors:

- Features obtained from multimodal content may estimate the cognitive load of the viewer without the use of physiological sensors;
- adaptive LLM reasoning strongly enhances boundary choice;
- the model generalises well to varied real-world data;
- Cognitive load minima offer a meaningful and interpretable control signal for video segmentation as well as the placement of ads.

These results demonstrate the validity of the proposed approach and the feasibility of the ideas presented, and offer a starting point in the pursuit of creating a cognitively aware media system.

Table 6.4: Window-wise weighting coefficients and qualitative descriptors aligned with $CL_k = \alpha_k MD_k + \beta_k E_k + \gamma_k TD_k$.

Video ID	**Window**	α_k	$\mathbf{MD}_k$	β_k	$\mathbf{E}_k$	γ_k	$\mathbf{TD}_k$
cV64pv4tdfU	1	0.3	Moderate complexity with contrasting scenes	0.5	High intensity from visuals and audio dynamics	0.2	Moderate pacing with frequent scene changes
	2	0.2	Casual conversation with low complexity	0.5	High vocal intensity and visual activity	0.3	Moderate pacing with frequent cuts
	3	0.3	Casual dialogue with moderate complexity	0.5	High visual activity and vocal intensity	0.2	Moderate pacing with frequent cuts
	4	0.3	Moderate complexity with personal narrative and visuals	0.5	High vocal intensity and engaging visuals	0.2	Moderate pacing with some quick cuts
ixqYwKuNAq4	1	0.4	Moderate complexity in planning methods presented	0.5	High vocal intensity and engaging visuals	0.1	Steady pacing with minimal rapid changes
	2	0.4	High info density with structured planning	0.4	Moderate vocal intensity and visual focus	0.2	Steady pacing with few rapid changes
	3	0.4	Moderate complexity with structured planning visuals	0.4	High vocal intensity and visual activity	0.2	Moderate pacing with steady shot durations
	4	0.4	Moderate complexity in planning visuals and spoken content	0.4	High vocal intensity and visual activity	0.2	Moderate pacing with steady visual transitions
	5	0.3	Moderate complexity with planning details	0.5	High vocal intensity and visual engagement	0.2	Fast pacing with frequent visual changes
	6	0.4	Complex planning with multiple elements increases mental demand	0.5	High vocal intensity and visual activity elevate effort	0.1	Moderate pacing and manageable temporal demand

6.12 Discussion

The results presented in Section 6.11 demonstrate that cognitive-load–aware segmentation offers a viable alternative to conventional, signal-driven approaches for identifying natural and comfortable transition points in narrative video. Importantly, this framework is not intended to replace structural segmentation systems such as those evaluated in Chapter 5, which focus on boundary alignment with narrative themes. Rather, it operates as a complementary suitability layer, refining candidate boundaries by assessing whether interruptions are cognitively appropriate from a viewer-centred perspective. By integrating audio–visual descriptors with transcript semantics and LLM-assisted NASA–TLX reasoning, the proposed framework approximates aspects of viewer mental workload in a way that is both lightweight and interpretable. This section discusses the implications of these findings, the strengths of the model, and their relevance to wider applications in consumer media systems.

6.12.1 Interpreting Cognitive-Load Curves as a Segmentation Signal

A central finding of the evaluation is that the time-varying cognitive-load curve encodes meaningful perceptual and narrative structure. Low-load regions tend to co-occur with moments of narrative resolution, topic closure, or pacing reduction. This produces a segmentation signal that is inherently aligned with human comfort. Unlike shot-change heuristics or pause detection, which may trigger during high-tension sequences or rapidly shifting scenes, cognitive-load minima more reliably correspond to points where interruptions feel least disruptive.

This feature becomes even more important in long-form streaming, where pause points that disrupt narrative pacing need to be avoided. This, in addition to the segment-aware policy, which focuses on the predicted load of the *following* segment, prevents a brief, transient drop from being confused with a low-load state.

Table 6.5: Why multimodal fusion is essential. 4-segment ablation. Removing audio inflates CL by avg. +11.5 (up to +26 in dialogue); removing visual by +6.5.Both modalities are required explains 0.09 F1 gain.

Scene	Audio	Visual	CL	No Aud.	No Vis.
Dialogue	Slow, jargon	Static	42	68 (+26)	48 (+6)
Chase	Rapid, loud	High motion, cuts	78	70 (-8)	88 (+10)
Montage	Music, sparse	Short shots	65	62 (-3)	72 (+7)
Didactic	Dense narration	Long static	58	75 (+17)	61 (+3)

6.12.2 Role of Multimodal Fusion and Adaptive TLX Weighting

The ablation experiment outcomes emphasize the need for a fusion of audio, vision, and semantic features. Removing any stream leads to a bias, with audio-only streams underestimating the mental load when the video is intense, and the vision-only streams face linguistic overload when dialog is linguistically intricate. NASA–TLX, a weighting scheme based on the LLM, eliminates the biases by locally adjusting the significance of *Mental Demand*, *Effort*, and *Temporal Demand*.

This adaptive behaviour is a key differentiator of the framework. Rather than relying on fixed fusion coefficients that must be tuned manually, the reasoning model varies its weighting in response to scene structure (e.g., elevating Mental Demand during dense technical exposition, or emphasising Effort during high-motion sequences). This enables a richer and more human-aligned interpretation of cognitive load than threshold-based or linear fusion strategies.

Table 6.5 provides an ablation-based explanation for the observed performance improvements of the proposed cognitive-load segmentation model. Removing audio cues significantly inflates estimated cognitive load in dialogue-heavy and didactic segments (up to +26 points), while removing visual cues inflates load during high-motion sequences such as chase scenes (+10 points). These results confirm that audio and visual modalities capture complementary cognitive determinants and that neither modality alone is sufficient for reliable cognitive-load estimation. The combined multimodal formulation therefore provides a principled explanation for the F1 gains reported in Section 6.11.

6.12.3 Generalisation Across Narrative Styles

Analysis of variance data also highlights that there is not much variance in performance over a wide array of videos that are shot in different styles, varying in pace, and belonging to different genres. The consistency in performance in varying storytelling types ranging from monologues to action scenarios in the outdoors implies that the model is not dependent on patterns in speech or motion but on a combination of multiple cues and inferences based on contexts.

This generalisation capability is critical for consumer media ecosystems, where content varies widely and manual tuning for each genre is infeasible. The system's robustness also supports wider integration into recommendation and delivery pipelines, including adaptive playback and context-aware interface design.

6.12.4 Comparison with Prior Cognitive-Load Modelling Paradigms

Traditional cognitive-load estimation techniques typically rely on physiological measurements such as EEG, EDA, or eye-tracking, which, while accurate under controlled conditions, are impractical for general consumer-facing applications. The proposed framework avoids these hardware and privacy constraints by leveraging content-derived signals alone. Although not a direct replacement for physiological sensing in fine-grained research settings, the model provides a functional, scalable, and low-cost alternative for real-world video systems.

Importantly, because the model is trained zero-shot and operates without domain adaptation, it can be deployed rapidly without collecting user-specific calibration data. Its semantic reasoning layer enables broader applicability than binary classification schemes focused on educational or experimental tasks.

6.12.5 Implications for Ad Placement and Consumer Media Systems

In addition to segmentation accuracy, the results highlight the benefit of cognitive load-aware decision-making in streaming services. The threshold τ_{seg} serves as a parameter to control the following functions:

- **inventory-focused operation** (higher thresholds): more potential advert slots with user comfort.
- **premium experience operation** (lower thresholds): fewer but exceptionally smooth breakpoints,
- **balanced operation** (mid-range thresholds): optimal point between ad views and storyline unity.

This adaptability is what is required by operational commercial streaming platforms. This is especially the case when it comes to contexts where retention rates are a critical concern with regard to intrusiveness.

Beyond advertising, the proposed model has further applications in:

- *adaptive streaming*, where playback pacing can be modified based on estimated cognitive demand;
- *viewer-aware summarisation*, where cognitively heavy sections can be compressed or annotated;
- *content navigation interfaces*, which may highlight cognitively light entry points for skimming or re-entry.

These directions reflect a growing interest in human-centred media engineering where systems adapt dynamically to cognitive and attentional patterns.

6.12.6 Broader Significance

The broader contribution of this work lies in demonstrating that cognitive modelling does not need to rely on physiological measurement; meaningful approximations of cognitive load can be inferred directly from multimodal content signals when enhanced with structured reasoning. By unifying signal-level descriptors with transcript semantics, the framework bridges traditional signal processing and modern LLM-based inference. As a result, it lays the foundations for future systems capable of reasoning about viewer experience without requiring active user input or invasive sensing.

Overall, this chapter illustrates that cognitive-load–aware segmentation is both practical and effective, offering a scalable pathway to more human-aligned multimedia systems in consumer electronics.

6.13 Limitations

While the proposed multimodal cognitive-load framework demonstrates promising performance and offers several practical advantages over heuristic or physiological-sensor–based approaches, it also exhibits a number of limitations that should be acknowledged. These limitations arise from assumptions built into the feature extraction pipeline, constraints of the reasoning model, and the evaluation protocol used in this work.

6.13.1 Dependence on Content-Derived Signals

One of the appeals of the framework is that it draws inferences regarding the cognitive load of the viewers solely on the basis of the audio-visual content and transcript semantics without the use of physiological data. Thus, the framework is scalable, and the model is deployable, though the model has the capability of approximating the viewers' *perceived* cognitive load, rather than the *actual* cognitive state of the viewers. Factors such as viewer familiarity with the topic, prior knowledge, emotional investment, fatigue, or personal preferences are not captured.

Although narrative structure, pacing, and prosodic cues are predictors of cognitive demand at the population level, variation across users limits the degree to which the framework can claim to model cognitive load at an individual level.

6.13.2 Sensitivity to Transcription and Caption Quality

The reasoning module relies heavily on WhisperX transcripts and VLM-generated scene descriptions. Errors in either component propagate into the TLX mapping step. For example:

- inaccurate transcripts (e.g., in noisy environments) can distort estimates of Mental Demand;
- imprecise VLM captions may misrepresent the scene's semantic content;
- low-quality keyframes may fail to capture key visual transitions.

While the LLM reasoning step mitigates some errors through contextual inference, the overall accuracy of cognitive load estimation remains bounded by the fidelity of

these upstream processes.

6.13.3 Shot-Level Visual Features and Limited Higher-Order Visual Understanding

The visual processing pipeline employs classical shot-boundary detection, optical flow, shot duration, and face-count estimates. Although effective for capturing pacing and activity, these descriptors do not account for deeper visual semantics such as emotional tone, narrative symbolism, or subtle visual tension. LLMs ingest captions derived from keyframes, but the quality of this semantic layer still depends on the VLM's ability to capture nuanced or abstract visual cues. As a result, the framework may underestimate cognitive load in visually symbolic sequences or editorial montages where surface motion does not fully reflect narrative tension.

Advances in the field of self-supervised learning in computer vision with models such as DINO [301] and other transformer-based models allow us to represent rich visual information including object relationships without any supervision. It is quite possible that by using such embeddings in the visual modality, improved semantic fidelity and more accurate cognitive load estimation from visual information can be achieved. But these models also have certain disadvantages because of the extra processing power required for such computations which might not align with the objectives of scalability and efficiency in multimedia applications. Exploring hybrid approaches that combine lightweight features with selectively applied high-capacity representations represents a promising direction for future work.

6.13.4 Zero-Shot Reasoning Constraints

The NASA–TLX mapping is performed in a zero-shot reasoning mode without task-specific fine-tuning. While this approach provides generalisation across domains, it also means that:

- the LLM does not have explicit cognitive-load training data;
- its interpretation of TLX dimensions is based on prior knowledge rather than supervised grounding;

- the adaptive fusion weights (α, β, γ) may not align with empirically validated cognitive models in all scenarios.

This limits the interpretability of absolute cognitive-load values, although relative trends remain meaningful.

6.13.5 Short-Term Memory and Local Context Assumptions

The framework processes windows of audio–visual and semantic features sequentially, with limited lookback. This design choice reduces computational cost but restricts the model's ability to incorporate long-range narrative dependencies. For example:

- viewer load may be influenced by tension accumulated across several minutes;
- emotional arcs or multi-scene sequences may be flattened into local estimates;
- abrupt tonal shifts may be under-recognised without global context.

This limitation could be addressed through hierarchical reasoning or narrative-level summaries in future work.

6.13.6 Behaviour Under Highly Stylised Content

Finally, the framework is optimised for conventional narrative content and may behave differently for highly stylised or experimental videos such as:

- rapid symbolic montages,
- music videos with no dialogue,
- heavily edited vlogs with artificial pacing,
- abstract animations or Computer-Generated Imagery(CGI)-heavy sequences.

In such cases, surface audio–visual cues may not reflect cognitive load, and the TLX mapping may need specialised adjustments.

6.13.7 Summary

Overall, while the proposed system provides a practical and interpretable solution for cognitive-load–aware segmentation, its performance is constrained by the fidelity of upstream feature extraction, the assumptions of zero-shot reasoning, and the use of

content-derived proxies for cognitive demand. These limitations highlight several opportunities for future work, including sensor-free personalisation, richer semantic reasoning, hierarchical narrative modelling, and expanded benchmark datasets.

6.14 Conclusion

This chapter has presented a multimodal, sensor-free framework for estimating segment-level cognitive load in narrative video content and using this signal to identify viewer-friendly transition points. Building on the premise that cognitive load fluctuates dynamically with changes in prosodic intensity, visual pacing, and semantic complexity, the framework integrates audio–visual feature extraction with LLM-driven reasoning to produce a continuous cognitive-load curve that aligns with human judgements of interruption suitability and narrative tension.

A key contribution of this work is the mapping of content-derived features onto selected NASA–TLX dimensions (Mental Demand, Effort, Temporal Demand) using an instruction-tuned large language model operating in zero-shot mode. This approach maintains scalability and avoids the hardware burdens associated with physiological measurement, while still capturing perceptually meaningful variation in cognitive demand across diverse scenes. The adaptive weighting mechanism introduced in the reasoning step further enhances robustness by dynamically interpreting which modalities dominate cognitive load in each context dialogue-heavy, motion-heavy, or conceptually dense scenes.

Evaluation on a curated subset of the VidChapters-7M dataset demonstrates that the proposed cognitive-load–aware method consistently outperforms signal-level heuristics such as longest-pause and energy-dip baselines. The segment-aware acceptance rule, which evaluates load in the *next* narrative unit, was shown to be particularly effective in avoiding disruptive boundaries that follow tension peaks. Both quantitative metrics and qualitative inspection indicate that points flagged as low-load correspond to natural periods of de-escalation, conversational closure, or structural pauses mirroring human judgements of suitable mid-roll insertion points.

While it excels in many ways, it is not without limitations. First, it is based on

content-derived cues, not directly on cognitively measured load. That is, it computes content-implied, not empirically supported, cognitive demand. Further, it is dependent on transcript or caption quality, and zero-shot reasoning is not supported with any grounding in empirical cognitive data. These limitations are precisely why research directions in Chapter 7 focus on personalised modelling, hierarchical narrative context, and semantic understanding.

Overall, the results of this chapter show that cognitive load traditionally viewed as a psychological construct requiring invasive instrumentation can be approximated effectively through content analysis alone when supported by multimodal signals and structured reasoning. More broadly, the findings reinforce the thesis vision that viewer-centred metrics can be integrated into media segmentation workflows, supporting applications in advertisement insertion, adaptive pacing, and cognitively aware user experience design. Chapter 7 extends these insights by situating both segmentation frameworks - theme-aware and cognitive-load–aware within a unified discussion and exploring broader implications for next-generation consumer media systems.

Chapter 7

General Discussion

7.1 Introduction

This chapter synthesises the core contributions presented throughout the thesis and situates them within a Integrated modular conceptual and operational framework. While Chapters 3, 4, 5 & 6 introduced distinct multimodal learning and reasoning mechanisms - topic modelling, explainable taxonomy mapping, theme-aware segmentation, and cognitive-load-aware segmentation these contributions collectively form an integrated pipeline for intelligent video understanding. The aim of this chapter is therefore two fold: (1) to articulate how the individual components developed in earlier chapters interact and reinforce one another, and (2) to interpret their combined implications for adaptive, human-centred multimedia systems.

For this purpose, the chapter proposes an integrated approach to multimodal reasoning to connect the work on semantic representation learning, narrative constrained reasoning, and cognitive state modelling. This integrated approach will address the conceptual divide that may exist when the use of multimodal AI methods is considered separately and will show that the work described by the thesis contributes to the overall research agenda presented in Chapter 1.

7.2 A Integrated modular Multimodal Reasoning Pipeline

The thesis proposes that a full solution for intelligent media segmentation involves these three complementary elements:

1. **Semantic understanding** – extracting the meaning, topics, and taxonomic structure from content.
2. **Narrative understanding** – recognising how meaning evolves over time into coherent themes.
3. **Cognitive understanding** – assessing when interruptions are least disruptive for viewers.

Chapters 3, 4, 5 & 6 each contribute one of these capabilities. Figure 7.1 summarises the integrated pipeline.

7.2.1 From Topics to Taxonomies (Chapters 3–4)

Chapters 3 & 4 are the foundation chapters that introduce multimodal pipelines to extract *semantic topics* from video data using visual, transcript, and multimodal embeddings. These are then mapped to taxonomies, which are compliant with the Interactive Advertising Bureau (IAB), and the inclusion of large language models (LLM) or multilingual embeddings ensures robustness and eliminates ambiguity regarding the explanation of the topics.

The results from these chapters, namely topic embeddings, descriptive text, and distributions over taxonomies, will become the *semantic state space*.

7.2.2 From Topics to Themes (Chapter 5)

Semantic topics alone do not capture the higher-order structure of narrative flow. Chapter 5 therefore introduces a constrained, multimodal theme-segmentation framework that infers *latent global themes* governing story progression. Using cross-modal alignment (via audio–shot windowing), structured natural-language summarisation, and a

single-pass Tree-of-Thoughts (ToT) reasoning mechanism, the method assigns each window to one of a small number of global themes.

This allows transitions to be detected not merely at points of visual change but at moments of *narrative discontinuity*, satisfying a requirement long identified in the literature: ad breaks should align with natural story shifts rather than arbitrary content boundaries.

7.2.3 From Themes to Cognitive Suitability (Chapter 6)

Even well-aligned thematic transitions cannot ensure the audience feels comfortable. Based on the NASA–TLX dimensions of multimodal characteristics, Chapter 6 proposes a model for *cognitive load*. As a result, the time-variant curve of *cognitive load* is obtained, and the audience's states of attention peak, immersion, and restoration are determined. The lowest points of the curve are the most appropriate locations for the thematic transition, while the highest parts of the curve should not be chosen, regardless of whether thematic boundaries are available.

Chapter 6 thus translates the C_t feature of the AdBreakScore(t) model proposed in Chapter 2 into the context of cognitive suitability based on measurable multimodal cues.

7.2.4 Integrating Semantic, Narrative, and Cognitive Layers

The Integrated modular pipeline combines, the results of Chapters 3 to 6 into a single segmentation approach. Within this framework,

- **Semantic topics and taxonomy mappings** constrain theme induction and ensure that narrative boundary formation is conditioned on structured semantic grounding.
- **Theme transitions** Theme transitions based on reasoning with constraints form the basis for candidate structural boundaries.
- **Cognitive-load minima** identify temporally appropriate moments for viewer-friendly interruptions.

Topics (Ch. 3/4) → Themes (Ch. 5) → Cognitive Load (Ch. 6) → Final Boundaries (Ch. 7)

Figure 7.1: Integrated modular multimodal segmentation pipeline across the thesis. Multimodal topic modelling and taxonomy mapping (Chapters 3–4) provide semantic structure; theme-aware segmentation (Chapter 5) induces narrative units; cognitive-load–aware segmentation (Chapter 6) identifies viewer-friendly interruption points; and this chapter integrates these components into a coherent strategy for final boundary selection.

- It employs a structured decision mechanism that integrates semantic grounding, narrative continuity modelling, and cognitive suitability constraints to determine final segmentation points.

Chapters 5 and 6 were evaluated independently in order to isolate and quantify the narrative and cognitive contributions of the proposed framework. This separation enables clearer methodological attribution of performance gains to specific architectural components. The Integrated modular pipeline presented in this chapter synthesises these layers conceptually and operationally through the AdBreakScore(t) formulation; however, it is not introduced as a separate supervised benchmark task. Rather, the integration demonstrates architectural coherence and deployment feasibility across semantic, narrative, and cognitive dimensions.

Accordingly, the thesis not only offers several independent models but also a comprehensive system for intelligent segmentation of media.

7.3 Broader Implications for Consumer Media Systems

The integrated pipeline developed in this thesis has several implications for emerging AI-driven media technologies.

7.3.1 Implications for Streaming and On-Demand Services

Through the united model, context-aware, cognitive-aligned advertising can be performed. This is because the online platforms will not need to depend on rigid programming or simplistic scene change detectors. Rather, they can use a multilayer reasoning

pipeline that will take into account the story structures as well as cognitive loads to decide on suitable positions for advertising.

7.3.2 Implications for Monetisation and Advertisement Ecosystems

The semantic structure provided by the taxonomy mapping makes the alignment of the advertisements with the content categories a clear, interpretable one. The theme-aware segmentation minimizes the cost of disruption by breaking the story at units of significance, while the cognitive segmentation places the advertisement insertions at the right places of attention of the human attentional cycles.

7.3.3 Human-Centred Quality of Experience

The health and well-being of the viewer is supported by the addition of cognitive load as an objective of modelling. This is consistent with the trend toward more regulatory and ethics-based concerns related to the transparency and acceptability of algorithms. Such an approach allows the system to avoid obtrusive insertions that lower the overall QoE in types such as educational, news programmes or emotionally sensitive content.

7.3.4 Implications for Content Production and Editing Tools

Video editing systems could use the Integrated modular pipeline to automatically identify narrative beats, semantic transitions, and cognitive pauses, providing intelligent editing suggestions or dynamic summarisation. For example, editors might be presented with suggested cut points that optimise both narrative continuity and cognitive comfort, or with alternative edit structures informed by theme and load profiles.

7.3.5 Ethical and Regulatory Considerations

Through the combination of cognitive modelling for explainability, the pipeline avoids potential dangers of non-transparent advertising targeting/manipulative content dissemination. The proposed pipeline also aids in the fulfillment of the demand for media automation system transparency by allowing the construction of understandable expla-

nations for the segmentation and placement of advertisements. At the same time, the same techniques could be utilized for crafting more subtle advertising approaches.

7.4 Limitations Across the Integrated Pipeline

While the individual components developed in this thesis offer significant advantages, several common limitations emerge when the full multimodal pipeline is considered as an integrated system.

- **Dependence on content-derived signals**: The framework relies exclusively on video and audio information. It does not incorporate personal or physiological data, and therefore cannot account for individual differences in cognitive or affective responses.
- **Sensitivity to noisy inputs**: Performance can degrade under challenging conditions due to errors in Whisper transcripts, vision–language model captions, or visual embeddings (e.g. heavy background noise, cluttered scenes, or extreme lighting).
- **Constraints of LLM reasoning**: Although the use of constraints in ToT increases reliability, the LLMs still exhibit some degree of variability, lack long-range memory, and can reflect inherent biases.
- **Dataset and benchmarking limitations**: The scarcity of publicly available datasets annotated for both thematic structure and cognitive-load-based transitions limits direct comparison with prior work and comprehensive generalisation studies.
- **Generalisation challenges**: Highly stylised genres, experimental cinematography, or non-narrative formats may disrupt theme induction and cognitive-load estimation, leading to suboptimal boundaries.
- **Computational considerations**: The full multimodal pipelines remain heavy for real-time deployment without model compression, quantisation, or architectural simplification.
- **Interpretability trade-offs**: Explainability adds architectural complexity be-

cause more layers of reasoning and validation need to be added.

A notable methodological limitation is that the integrated pipeline has not been evaluated empirically as a fully unified end-to-end system. While each component is rigorously validated in its respective chapter, a complete joint evaluation would require new unified benchmarks that simultaneously capture semantic, narrative, and cognitive suitability benchmarks that are currently unavailable. Consequently, the integration presented in this thesis should be viewed primarily as an architectural and conceptual contribution, supported by strong component-level evidence, with full system-level validation reserved for future work.

These limitations also provide an opportunity to reflect on the extent to which the research gaps identified in Section 1.3 have been addressed.

Substantial progress has been made in tackling weak semantic grounding through the multimodal topic modelling and taxonomy alignment framework (Chapters 3 and 4). Similarly, the theme-aware segmentation approach in Chapter 5 advances narrative understanding by moving beyond local similarity heuristics toward global thematic continuity. However, challenges related to long-range narrative modelling remain.

The cognitive modelling gap (Chapter 6) is addressed through a practical, content-derived proxy approach suitable for large-scale deployment, though it cannot fully capture individual viewer differences. Overall, while the thesis makes meaningful strides toward unifying semantic, narrative, and cognitive layers, full integration and comprehensive empirical validation remain important areas for future research.

7.5 Opportunities for Deep Integration

Despite the limitations, the components developed in this thesis exhibit several complementary strengths, enabling tighter integration in future work.

7.5.1 Strengths of Combined Semantics, Themes, and Cognitive Load

Topic modelling provides fine-grained semantic descriptors; theme-aware segmentation structures these into coherent units; cognitive modelling filters them through viewer

comfort. Together, they form an adaptive media pipeline that is more robust than any individual component considered in isolation. Semantic and thematic layers reduce the risk of content mismatch, while cognitive modelling reduces the risk of temporal disruptive interruptions.

7.5.2 Toward a Modular, Integrated Multimodal Reasoning Layer

A shared reasoning layer could jointly process semantic topics, theme transitions, and cognitive signals, reducing fragmentation and ensuring consistent decision criteria. Rather than chaining separate modules, a unified multimodal reasoning component could jointly optimise segmentation with respect to semantic, narrative, and cognitive constraints.

7.5.3 Hybrid Boundary Scoring Functions

The pipeline could employ a quantitative scoring mechanism combining:

- semantic discontinuity,
- thematic shift strength, and
- cognitive suitability,

yielding a single unified metric for break selection. Such a metric would be a practical instantiation of AdBreakScore(t), embedding both perceptual and operational considerations into an actionable decision rule.

7.5.4 Mutual Reinforcement Across Layers

Theme stability may sharpen cognitive-load estimates, and cognitive fluctuations may highlight meaningful theme boundaries. This mutual reinforcement presents an opportunity for co-training or joint optimisation, in which errors in one layer are corrected by signals from another. For instance, persistent high load around a putative boundary may indicate an incorrectly inferred theme transition.

7.5.5 Toward Fully Adaptive Media Pipelines

Ultimately, the current thesis seeks to lay the architectural groundwork for a more adaptive media system in which segmentation, advertising, and transition are guided by a

combination of semantic, narrative, and cognitive reasoning, as opposed to heuristic scheduling. While the current framework is designed as a relatively static and offline-based system, future work could consider various strategies for scaling inference and deployment to further extend the current framework towards a more adaptive media system. In these extensions, the current boundary controls could be dynamically regulated based on the changing characteristics of the media content or the operational requirements of the system, while still maintaining the multimodal grounding established in the current work. This would be a further extension of the current architecture, not a departure from it, and would be a more scalable evolution of the principles established in the current work.

7.6 Future Directions

Several future research opportunities emerge from the unified perspective:

1. **Personalised cognitive-load models**, supporting segmentation based on user-specific baselines and preferences, rather than making generalization at a population level.

2. **Light-weight adaptation layers** suitable for real time execution on devices with constrained resources, using distillation, pruning, or quantisation.

3. **Long-range temporal modelling** using hierarchical transformers or memory-augmented networks to capture dependencies over entire episodes or films.

4. **Real-time feedback loops** (for instance, skips, repeat requests, and abandonments) with themes and cognitive judgments.

5. **Synthetic data generation** to train models of cognitive load and theme on a large scale, which could be achieved through the use of controllable simulators/-generative models to address the issue of annotated data scarcity.

6. **Ethical auditing frameworks**, to guarantee secure and transparent usage in a business setting for bias, over-personalization, and manipulation of users.

7.7 Summary

This chapter has described a Integrated modular view of the thesis, which combines multimodal semantic representation learning, theme induction in a narrative, and cognitive model-driven computational process for intelligent and human-centric media segmentation. The described architecture not only provides an advancement in the state-of-the-art solution but also forms a conceptual basis for future-adaptive systems that provide a balance among various semantically pertinent, human-centric themes and viewer well-being.

Chapter 8

Conclusion

This thesis explored the segmentation of long-form video content in a manner that integrates its semantic meaning, story structure, and cognitive experience. By using a modular and hierarchical framework that covers multimodal semantic understanding, thematic story inference, and cognitive modelling, the thesis demonstrated that segmentation systems based solely on content and using LLM reasoning are comparable to those that are trained in terms of quality but are more interpretable and efficient. The results of the thesis are part of a new trend in video intelligence that focuses on human-centered video intelligence in which video editing, segmentation, and playback are based not only on content but also on cognitive experience.

8.1 Summary of Contributions

The thesis provides five interrelated pieces of work to address human-centered video segmentation. First, it proposes a multimodal semantic topic modelling framework that incorporates both visual, audio, and text modalities into interpretable representations of video data that are aligned with IAB Ad taxonomy. Second, it demonstrates how large language models can be used in a controlled and explainable manner to refine and interpret multimodal topics, improving coherence, multilingual robustness, and semantic transparency. Third, the thesis proposes a deterministic, training-free theme-aware segmentation approach using constrained LLM-based reasoning over audio-visual

alignments to pick up narrative transitions without resorting to stochastic decoding or fine-tuning. Fourth, it provides a multimodal model for estimating cognitive load that maps audiovisual cues derived from content to the dimensions of the NASA-TLX and is able to identify low-demand points that correlate with the closure of narratives and viewer comfort. Finally, these components are integrated into a Integrated modular multimodal segmentation pipeline that demonstrates how semantic understanding, narrative reasoning, and cognitive modelling can jointly inform segmentation and ad-break decisions. This integration reframes video segmentation as a human-centred process that balances content structure, narrative coherence, and viewer experience while remaining transparent and operationally viable.

8.2 Key Findings

Several key insights emerged across the experimental chapters:

- **Constrained deterministic inference is more deployable and reproducible.** Using fixed prompts, reasoning paths that are bounded in scope, and zero temperature during the decoding process made the reasoning system behave in a stable and predictable way when segmenting, and that LLM-assisted reasoning can satisfy the requirements of being stable, auditable, and reproducible in the context of industrial video services.
- **Semantic modelling improves interpretability and structured content understanding.** Multimodal topic modelling and taxonomy mapping produced more coherent, interpretable semantic representations than unimodal or weak-fusion baselines, improving robustness to modality-specific noise and supporting explainable content characterisation at scale.
- **LLM-assisted fusion compensates for the absence of training.** Both frameworks rely on LLM reasoning to contextualise and reconcile multimodal cues. This allowed competitive performance relative to learned models despite no fine-tuning or dataset exposure.
- **Cognitive-load variation aligns with narrative rhythm.**

The consistent correspondence between low-load troughs and well-positioned chapter boundaries rated as appropriate for ad placement also lends support to cognitive load curves as an important control signal for narrative editing decisions.

Taken together, these results suggest that training-free, modular multimodal reasoning can be used to structure video data in a meaningful and cognitively consistent manner, which supports a move towards transparent and human-centric video intelligence.

8.3 Implications for Research and Industry

The outcomes of this thesis have several broader implications for both research and practice.

For **multimedia and media-systems research**, the work demonstrates that high-level semantic, narrative, and cognitive structure can be modelled using training-free, content-derived pipelines, challenging the prevailing reliance on end-to-end learned segmentation and highlighting the value of modular, interpretable reasoning architectures.

For **media engineering and streaming platforms**, deterministic multimodal segmentation enables scalable, low-latency applications such as ad scheduling, chapter generation, content pacing, and recommendation timing, while supporting reproducibility and operational auditability. For semantic content analysis and contextual advertising, taxonomy-aligned multimodal topic modelling provides a robust and explainable foundation for content categorisation, brand safety, and contextual relevance, even in multilingual and noisy environments.

For **semantic content analysis and contextual advertising**, taxonomy-compliant multimodal topic modelling, a strong and interpretable basis for content classification, brand safety, and context relevance even for multiple languages and noisy data.

For **cognitive-aware interfaces and adaptive playback**, the proposed cognitive-load modelling framework suggests a path toward viewer-centric adaptation driven solely by audiovisual content, removing the need for invasive biometric sensing or explicit user instrumentation.

For **creative production and editorial workflows**, theme-driven segmentation can benefit editors by helping to identify places of narrative closure, coherence breaks, or "breathing spaces" in long-form content.

Finally, for **regulatory and ethical contexts**, limited and deterministically based LLM-assisted reasoning provides an auditable and transparent approach that is more accountable than an unconstrained probabilistic model.

These implications show the relevance of combining the understanding of semantics, story-based reason, and cognitive models in a multimedia system with a focus on humans.

8.4 Limitations

However, there are still some limitations that remain despite having such promising results in the semantic, narrative, and cognitive aspects that have been shown in this thesis.

At the **Semantic Level**, the proposed framework for topic modelling in a multimodal scenario is purely dependent on content-driven visual, audio, and text information and does not use world knowledge, cultural context, or fine-grained domain ontologies beyond industry taxonomies. This might result in a *semantically low-quality representation when dealing with very abstract or symbolic content.* Additionally, although industry taxonomies make the task more robust, they might make the system less expressive when dealing with a scenario that doesn't map well to a taxonomy.

At the **narrative level**, the theme-aware segmentation model uses shallow lookahead and local thematic continuity constraints. Though this architecture enables determinism, replicability, and efficiency in computation, it constrains the model's capacity to model longer-term narrative trajectories, multi-episode narratives, and themes that evolve over longer periods of time. The model further presupposes that transitions in narratives are observable in terms of changes in the audiovisual/linguistic data , and does not maintain an explicit persistent narrative state across distant segments. As a result, transitions that rely primarily on implicit meaning, symbolism, or sustained visual repetition may not be reliably detected.

At the **cognitive modelling level**, cognitive-load estimation is based on heuristically selected audiovisual features mapped to a subset of NASA–TLX-inspired dimensions, primarily Mental Demand, Temporal Demand, and Effort, inferred through LLM-assisted reasoning. Dimensions such as Physical Demand, self-assessed Performance, and Frustration are intentionally excluded, as they require subjective reporting or affective sensing beyond what can be reliably inferred from content-derived signals alone. Consistent with the empirical evaluation, this formulation is effective at identifying relative low- and high-load intervals suitable for segmentation, but does not support absolute or individualised estimates of mental effort. As a result, while the model does not capture richer affective states, musical tension, or discourse-level complexity, nor does it provide a direct physiological measurement of mental effort. In addition, viewers are treated as a generic population, without accounting for individual differences in attention, expertise, cultural background, or viewing preferences.

From the **evaluation point of view**, the lack of large-scale datasets annotated for ground truth on cognitive load makes it difficult to validate on objective measures of mental effort. While human annotation is helpful for proxy tasks, it also introduces subjectivity. On the other hand, the quality of segmentation is mainly measured at the boundary level and not on long-term engagement measures.

Lastly, in the **system-integration level**, the modular approach on purpose foregoes end-to-end optimisation. Though this is helpful in achieving transparency, interpretability and reproducibility, it might result in suboptimal performance in comparison to extensively trained models in the same domain.

These limitations delineate the current boundaries of the proposed approach and motivate the future research directions discussed in the following section.

8.5 Future Work

There are also several avenues for extending the proposed framework in future work to address the methodological limitations of the current work, as outlined in this thesis. At the level of cognitive modelling, for instance, the investigation of personalised cognitive load estimation through the incorporation of user baselines and attentional characteris-

tics could provide more flexibility in segmentation strategies to accommodate individual differences in expertise and viewing or interaction behaviour.

Such extensions would allow more tailored content structuring while maintaining the privacy-aware and sensor-free principles of the current approach.

At the semantic and narrative levels, hierarchical and memory-augmented reasoning mechanisms offer a promising avenue for modelling longer-term structure. The addition of a semantic/narrative state could potentially allow modelling of long-term narrative structure, themes, and thematic development that may be hidden from view on a locally bounded scale of reasoning. Such a direction might be of particular interest when modelling long-form content, episodic, or multi-part content.

Additional degrees of freedom could also be achieved through the application of lightweight adaptation of the cognitive models to improve the quality of the underlying reasoning behaviour for particular domains or genres of content.

These techniques would allow controlled adaptation without compromising determinism, interpretability, or reproducibility, which are central to the framework's design philosophy.

Another important direction concerns the development of weakly supervised or synthetic datasets for cognitive-load modelling. Such datasets could support more direct learning of workload-related patterns without reliance on physiological sensors, for example through structured annotations, controlled content manipulations, or proxy behavioural signals. This would improve scalability while preserving ethical and practical feasibility.

Lastly, one very important aspect of future research will focus on the possibility of real-time adaptive segmentation using a lightweight and computationally efficient model. This is especially true considering the use cases in streaming and on-demand media where placing advertisements and transitions between segments requires a certain level of speed and time efficiency. While the existing solution has proven itself to be highly effective in structured multimodal reasoning, several steps need to be taken to implement it into practice due to several computational inefficiencies associated with feature extraction, topic modelling, and taxonomy alignment.

Moreover, the usage of the proposed technique in practical settings will likely involve edge computing, mobile computing, and large scale infrastructure for online streaming which imposes limitations on memory usage, CPU, and power efficiency. Therefore, it may be necessary to explore approximations and more efficient methods in order to achieve higher computational efficiency and speed.

Collectively, these avenues point to the potential for future media systems to increasingly leverage knowledge of semantic structures, narratives, emotionally salient features, and cognitive processes of the human brain, while also being tractable and understandable in real-world settings.

8.6 Closing Remarks

This thesis has shown the potential of training-free deterministic multimodal reasoning to produce segmentation results that are well-aligned to both narrative structures and cognitive human experience. This work challenges the conventional wisdom that large-scale training is required to achieve high-quality video understanding and shows that feature design, multimodal synchrony, and constrained LLM reasoning can provide a viable and principled alternative.

As media consumption grows more dynamic and personalised, the methods developed here offer a foundation for future systems that adapt to narrative rhythm and viewer cognitive states supporting more comfortable, coherent, and human-centred multimedia experiences.

Appendix A

Supplementary Material

A.1 Overview

A.2 Datasets

This section provides a consolidated description of the datasets used across the experimental chapters of this thesis. While each chapter introduces its datasets within the context of a specific modelling objective, this appendix serves as a central reference summarising dataset provenance, modality composition, scale, and their role within the semantic (Chapters 3–4), narrative (Chapter 5), and cognitive (Chapter 6) layers of the proposed multimodal segmentation framework.

A.2.1 Overview of Dataset Usage

Table A.1 summarises the datasets employed throughout the thesis, indicating their primary modalities and the chapters in which they are used.

A.2.2 Food.com Textual Training Corpus

Source and Availability. The dataset publicly available and was obtained from Kaggle.com with cooking-related content originating from Food.com.

Content and Modalities. The data set is purely text-based and contains information such as recipes, reviews, and user activities. It also contains a wealth of natural

Table A.1: Summary of datasets used across the thesis.

Dataset	Modalities	Chapters	Primary Purpose
Food.com (Kaggle)	Text	3, 4	Textual topic model training
Food-101	Image	4	Visual topic model training
YouTube-8M (food subset)	Video, Audio, Text	3, 4	Semantic evaluation
VidChapters-7M	Video, Audio, Text	5, 6	Narrative structure and cognitive load priors
Cognitive load evaluation set	Video, Audio	6	Human validation of ad suitability

language information regarding various activities that take place while cooking, such as the use of various ingredients and culinary contexts.

Scale and Composition. The corpus provides over 180,000 recipes and corresponding interactions, giving us significant linguistic diversity for large-scale topic induction.

Usage in This Thesis. This dataset is only utilized in Chapters 3 and 4 to fine-tune the textual BERTopic model for domain adaptation purposes.

A.2.3 Food-101 Visual Training Dataset

Source and Availability. Food-101 is an image dataset that is public available and consists of food images that are categorized into 101 classes.

Content and Modalities. The data contains still images only, without any corresponding text or audio.

Scale and Composition. It includes approximately 50,000 images of widely varying kinds of foods and styles of visual presentation.

Usage in This Thesis. This Thesis The Food-101 is used in Chapter 4 to train the visual BERTopic model, allowing visual topic induction based on the semantics of food. It is not used for the evaluation or segmentation tasks involving videos.

A.2.4 YouTube-8M Evaluation Subset

Source and Availability. A representative subset of the publicly accessible YouTube-8M dataset is chosen for evaluation purposes.

Content and Modalities. Each video includes visual frames, automatically extracted audio features, and transcript-derived textual representations generated using automatic speech recognition.

Scale and Composition. This selected subset focuses only on the category "Food & Drink," with varying narrative structures and characteristics in the videos.

Usage in This Thesis. This data is used in Chapters 3 and 4 to assess multimodal topic coherence, taxonomy mapping quality, multilingual robustness, and the presence of modality-specific noises. This data is not used in the training process.

A.2.5 VidChapters-7M Narrative Segmentation Dataset

Source and Availability. VidChapters-7M is a publicly available large-scale benchmark for chapter boundary detection in long-form videos.

Content and Modalities. The dataset contains videos with audiovisual stream alignment, as well as human-annotated chapter boundaries.

Scale and Composition. The entire set consists of around 817,000 videos and over 7 million annotated boundaries in the following genres: tutorials, vlogs, documentaries, interviews

Usage in This Thesis. VidChapters-7M is used in Chapter 5 to evaluate the proposed theme-aware segmentation framework against established baselines. In Chapter 6, its chapter boundaries are repurposed as narrative-aligned segments over which cognitive load is estimated. These segments provide a stable structural prior but are not treated as ground truth for cognitive load. No training or fine-tuning is performed on this dataset.

A.3 Implementation Details for Theme-Aware Segmentation (Chapter 5)

A.3.1 Pre-Reasoning Structural Consolidation for Theme Segmentation

Listing A.1: Deterministic structural consolidation performed prior to constrained Tree-of-Thought reasoning.

```
This step carries out deterministic structural consolidation prior to any LLM-based
    reasoning. Its goal is to assemble stable reasoning windows from scenes that
    are adjacent and not semantically or editorially separable, relying on
    objective audiovisual continuity rules.



### Inputs
- Ordered list of detected scenes/shots {C1, C2, ..., CN}, each of which includes:
    - Start timestamp, End timestamp
     Per-scene visual summaries
- Aligned transcript segments with word-level timestamps
- Indicators for audio continuity across the scene boundary

### Objective
The Create a sequence of contiguous windows {W1, W2, ..., WK} that maintains
    linguistic integrity while preserving the audiovisual continuity. This sequence
    of windows becomes the basis on which the subsequent downstream reasoning steps
    are carried out.

### Algorithmic procedure
1. Initialise an empty list of windows.
2. Start a new window Wk at the first scene Ci.
3. For each consecutive scene boundary (Ci, Ci+1) ;
   a. Check for bridging audio across the boundary, defined as the
      presence of speech or continuous audio overlapping both scenes.
   b. If bridging audio exists ;
      - Merge Ci+1 into the current window Wk.
   c. Otherwise:
```

```
      - Finalise Wk.
      - Start a new window Wk+1 at Ci+1.
4. Continue until all scenes have been processed ;
5. Output the final sequence of windows ;

### Structural constraints
- Scene boundaries occurring mid-sentence are not permitted as window boundaries.
- No trimming or splitting is done in the original scene.
- All merges operate strictly at scene granularity.
- Windows are disjoint and their union covers the full video timeline.

### Determinism and scope
- This procedure is fully deterministic and contains no stochastic
  elements;
- No semantic inference, theme induction, or hypothesis generation is
  performed.
- No alternative windowings are explored or evaluated.
- Given identical inputs, the output window sequence is identical.

### Output
Each window Wk is defined by:
- Start and end timestamps
- Concatenated visual summaries of its constituent scenes
- Concatenated transcript content aligned to the window duration

These windows form the fixed reasoning units for the constrained
Tree-of-Thought segmentation stage. By enforcing structural coherence
before reasoning, this step prevents the reasoning model from reacting
to transient stylistic edits or editorial artefacts.
```

A.3.2 Constrained ToT prompt (segmentation)

Listing A.2 provides the fixed structured prompt template used to operationalise the constrained single-pass Tree-of-Thought reasoning process for theme-aware video segmentation.

Listing A.2: Public prompt template used for constrained ToT theme segmentation.

```
You are performing **theme-aware video segmentation** for adaptive ad insertion.
Use **constrained single-pass Tree-of-Thought / Chain-of-Thought style reasoning**
    to decide
whether to place a breakpoint between consecutive multimodal windows.

### Inputs
- Video context: {full_video_context}
- Global themes (if provided): {global_themes}
- Window names (in order): {window_names}
- Current batch windows (in order): {windows_list}
- Lookahead windows (for verification): {lookahead_context}
- Previously accepted breakpoints (JSON): {previous_breakpoints_json}
- Batch / step identifier: {batch_number}

### Task
    Internally consider a small number (2-3) of alternative, reasoned candidate
    decisions
    about whether (and where) a breakpoint should be placed within this batch only,
    then commit to a single final decision.
For each candidate boundary between consecutive windows in the *current batch*:
1) Infer/confirm the theme of each window.
2) Determine if a boundary is justified *only* when there is a **sustained theme
    transition**.

### Sustained-theme constraint (lookahead)
Accept a boundary between B_i and B_{i+1} only if:
- theme(B_i) != theme(B_{i+1}), AND
- theme(B_{i+1}) == theme(B_{i+2})  (verified via lookahead)

### Additional constraints
- Prefer **fewer, stronger breaks** over many weak breaks.
- Reject transient subtopic shifts, stylistic changes, quick cutaways, or short
    asides.
- Do not place breaks too close to previously accepted breakpoints unless strongly
    justified.
- If uncertain, return **no breakpoint** (null decision) rather than guessing.
```

```
- Base decisions strictly on the provided window summaries; do not invent facts.

### Output format (STRICT)
Return **JSON only** (no prose, no markdown, no code fences, no extra keys).
Use this schema:
{
  "breakpoint": {
    "decision": "accept" | "reject",
    "between_windows": ["<window_name_i>", "<window_name_i+1>"],
    "pre_theme": "<theme_label>",
    "post_theme": "<theme_label>",
    "confidence": 0-1,
    "rationale": "<= 25 words grounded in evidence>"
  }
}

If no breakpoint is acceptable in this step, return:
{ "breakpoint": null }

### Reasoning constraints
- Perform all reasoning in a **single pass** (no iterative refinement).
- Do NOT reveal intermediate chains of thought; output only the final decision
    schema.
```

A.3.3 Output schema and validation

Listing A.3: Deterministic post-reasoning output validation and preparation prior to evaluation.

```
This step performs deterministic post-reasoning output validation and
normalisation prior to quantitative or qualitative evaluation. It operates
exclusively on the outputs of the segmentation stage and does not modify
the underlying segmentation decisions.

### Inputs
- Predicted segmentation outputs (JSON) generated from the Constrained
    Tree-of-Thought segmentation layer
```

```
- Video-level metadata (e.g., duration,
- Reference annotation files used for evaluation
- Temporal Matching Tolerance Delta
- Minimum segment duration tau



### Objective
This involves converting the raw segmentation output into a standard and verified
    format to facilitate fair, consistent, and comparable evaluation with the
    reference segmentation.

### Algorithmic overview
The validation procedure enforces schema correctness, removes redundant or invalid
    boundaries, and aligns predictions with reference annotations using fixed
    temporal tolerances. All operations are deterministic and order-preserving.

### Algorithm A.X: Output Validation of the Post-Reasoning Process

input,
- Predicted Breakpoints P = {p1, p2,., pn}
- Video metadata M
- Reference annotations R
- Temporal tolerance delta
- Minimum segment duration tau

Output
 - Validated breakpoint set P'
- Canonical Prediction-Reference Pairs for Evaluation

Procedure:
1. Initialise an empty list of valid predictions P_valid.
2. For each predicted breakpoint p in P:
   - If p conforms to the expected output schema, add p to P_valid.
3. Sort P_valid by timestamp.
4. Initialise an empty list P_dedup.
5. For each breakpoint p in P_valid:
   - If no breakpoint in P_dedup lies within delta of p, add p to P_dedup.
```

```
6. Clamp all breakpoints in P_dedup to the valid video duration [0, M].
7. Initialise an empty list P_filtered.
8. For each consecutive pair of breakpoints (pi, pi_plus_1):
   - If (pi_plus_1 - pi) >= tau, retain pi in P_filtered.
9. Align P_filtered with reference annotations R using temporal tolerance delta.
10. Convert aligned predictions and references into a unified, canonical
    representation suitable for downstream metric computation.

### Validation constraints
- All operations are fully deterministic.
- No semantic reasoning, optimisation, or re-scoring is performed.
- Predicted boundaries are never shifted to improve evaluation scores.
- Reference annotations remain unchanged.

### Output
- A validated and normalised set of predicted breakpoints per video
- Canonical prediction-reference pairs ready for evaluation
- Structured logs of discarded or invalid predictions

This process is important so that the evaluation is actually a measure of the
    quality of the segmentation model rather than an artifact of other
    characteristics like formatted data, redundancy, or temporal misalignment.
```

A.3.4 Results

Figure A.1 and A.2 : Example segmentation output from the proposed deterministic theme-aware pipeline, illustrating boundary placement under sustained-theme constraints.

This appendix provides representative qualitative examples of segmentation outputs produced by the proposed theme-aware framework. The examples are intended to complement the quantitative results reported in Section 5.8 by illustrating typical boundary placement behaviour under sustained-theme verification. All outputs are generated using fixed prompts and deterministic decoding.

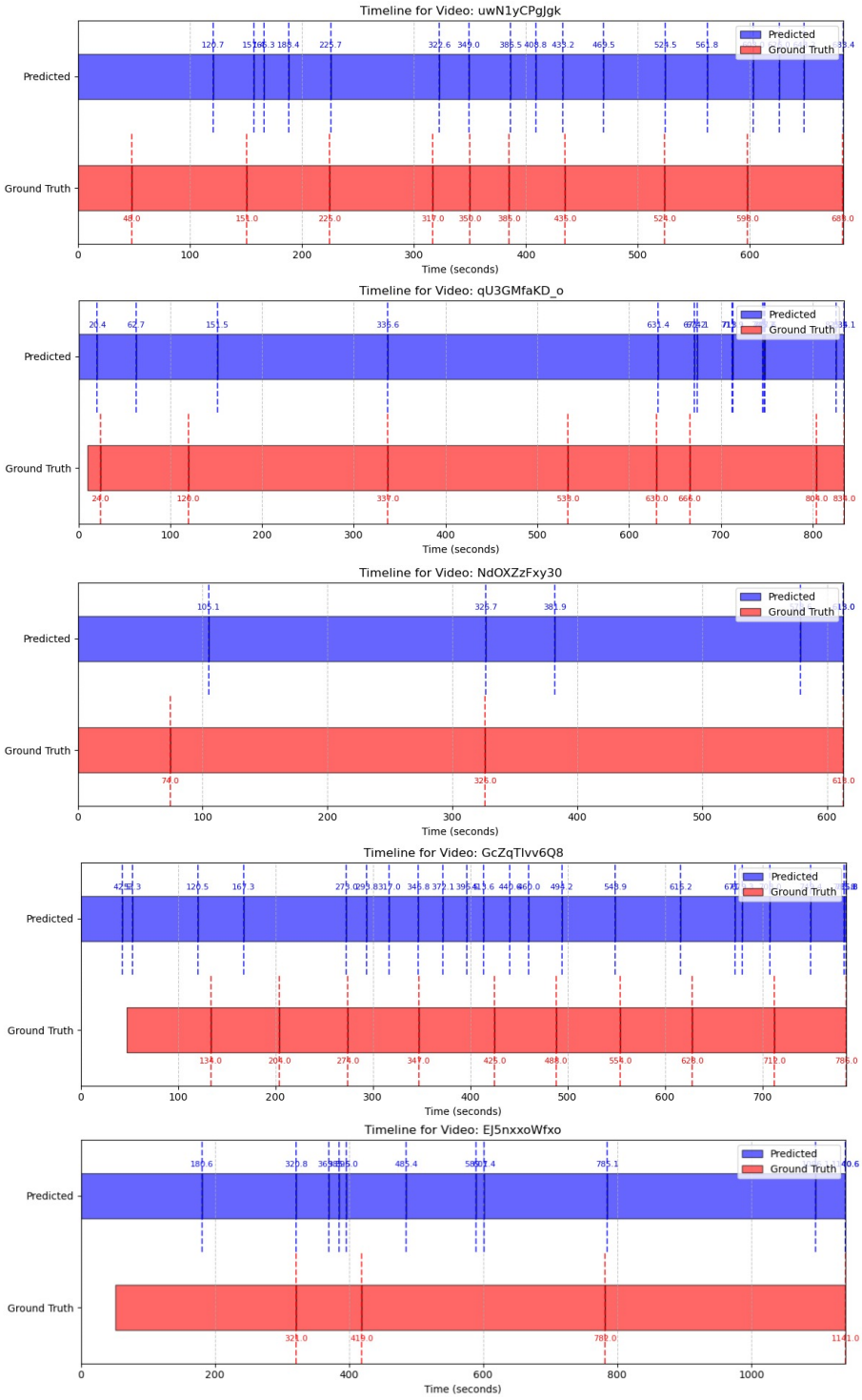


Figure A.1: Sample qualitative output of the proposed theme-aware segmentation framework, illustrating deterministic boundary placement under sustained-theme constraints.

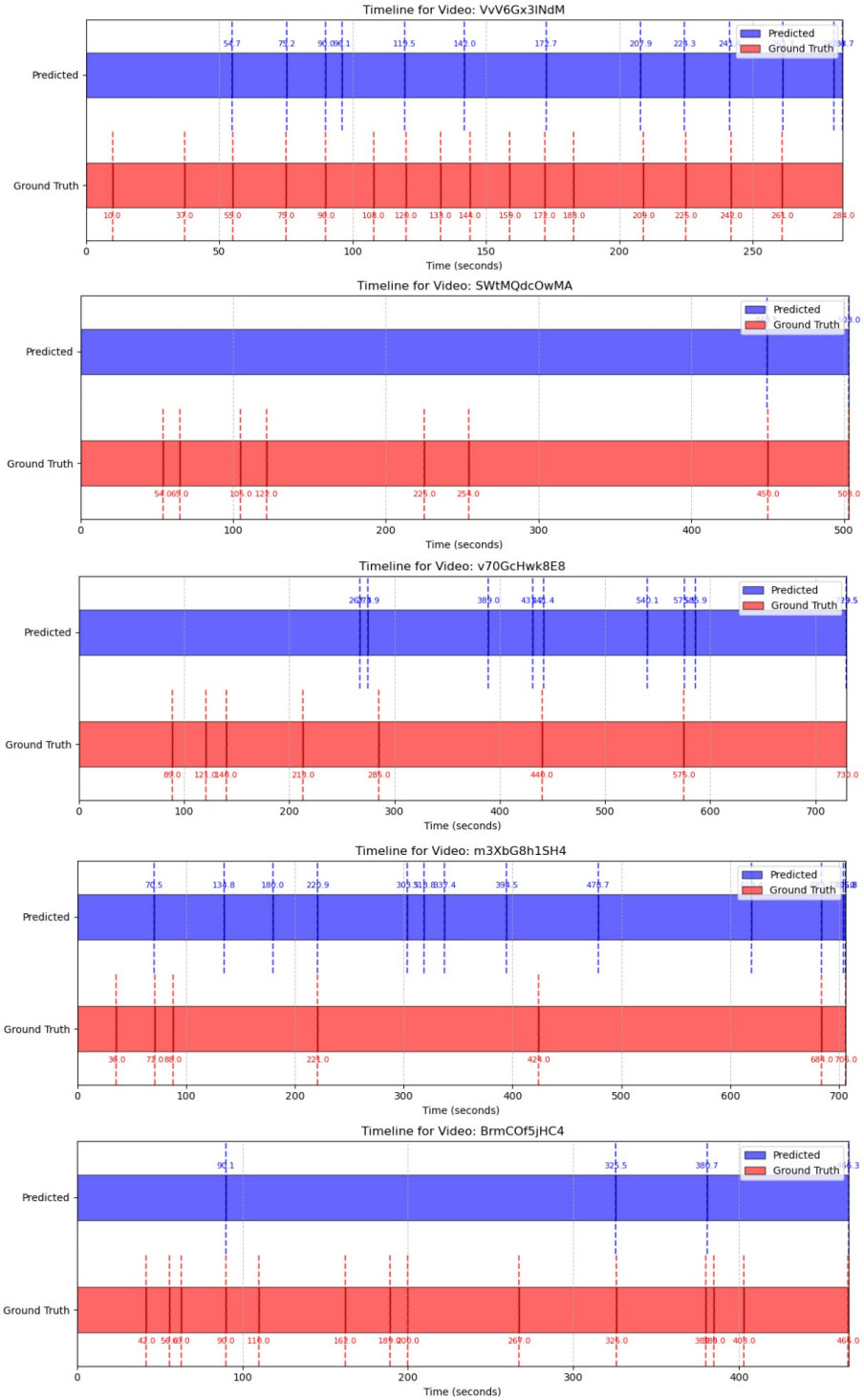


Figure A.2: Sample qualitative output of the proposed theme-aware segmentation framework, illustrating deterministic boundary placement under sustained-theme constraints.

A.4 Implementation Details for Cognitive Load Estimation (Chapter 6)

A.4.1 Feature Extraction normalisation

Listing A.4: Deterministic feature extraction and normalisation prior to cognitive-load inference.

```
This section summarises the low-level audio and visual feature extraction
and normalisation procedures applied prior to LLM-based cognitive-load
inference. These operations are deterministic and do not involve any
learned or adaptive components.

### Audio feature extraction
For each segment, the following prosodic and temporal audio features are
computed from the aligned speech signal and transcript:

- Words per second
- Articulation rate
- Pause ratio
- Mean RMS loudness
- Dynamic range (in dB)
- Integrated loudness (LUFS, absolute)

These features characterise speaking rate, vocal intensity, and temporal
regularity, which are known correlates of perceived effort and temporal
demand.

### Visual feature extraction
For each segment, the following visual pacing and activity features are
computed from shot-level metadata and motion statistics:

- Mean shot duration
- Shot cut rate
- Time-weighted motion magnitude
- Speaker switch rate
```

```
These features capture editing pace, visual motion, and conversational
dynamics relevant to temporal pressure and attentional demand.

### Feature normalisation
To ensure scale consistency across segments within the same video, most
audio and visual features are normalised using within-video z-scoring:

  z = (x - mu) / sigma

where mu and sigma denote the mean and standard deviation computed over all
segments of the corresponding video.Absolute loudness (LUFS) is retained
in its original scale.

Normalisation is performed deterministically and independently for each
video, preventing cross video leakage and preserving relative intra-video
variation.

### constraints
- Feature extraction and normalisation are fully deterministic.
- No semantic inference or learning is performed at this stage.
- The LLM operates only on normalised feature summaries provided via the
  structured prompt.
- These steps serve solely to prepare scale-consistent inputs for the
  cognitive-load inference process described in Chapter 6.
```

A.4.2 Structured prompt for cognitive-load reasoning

Listing A.5 provides the fixed structured prompt template used to compute the NASA–TLX-inspired cognitive dimensions (MD, E, TD) and their adaptive weights for each segment. All variables in braces are programmatically populated from the aligned multimodal feature extractor.

Listing A.5: Structured prompt template used for cognitive-load reasoning.

```
You are evaluating cognitive load in a video clip based on NASA-TLX.
The task for you is to rate three dimensions:
```

```
 - Mental Demand (MD): complexity, and density of the information.
- Effort (E): Intensity of the audio and visual channels; attentional energy.
- Temporal Demand (TD): pacing, time pressure, rapidity of changes.

### Segment metadata
Time: {t0} to {t1} (duration={duration})

### Combined semantic context
{context_summary}

### Audio (prosody) features (z-scored within this video unless noted)
- words_per_sec_z: {words_per_sec}
- articulation_rate_z: {articulation_rate}
- pause_ratio_z: {pause_ratio}
- rms_db_mean_z: {rms_db_mean}
- dyn_range_db_z: {dyn_range_db}
- lufs (absolute): {lufs}

### Visual (low-level) features (z-scored within this video unless noted)
- shot_duration_mean_z: {shot_duration_mean}
- cut_rate_z: {cut_rate}
- motion_magnitude_z: {motion_timeweighted}
- speaker_switch_rate_z: {speaker_switch_rate}

### Instructions
1. Assign numeric ratings (0-100) for MD, E, and TD.
2. Suggest adaptive weights (alpha, beta, gamma) (all >= 0, sum = 1).
3. Compute CL = alpha*MD + beta*E + gamma*TD.
4. Provide brief rationales; (< 15 words each).
5. Prefer measured features for E/TD; let context refine MD.
```

A.4.3 Output schema and validation

Listing A.6: Output schema and deterministic validation rules for LLM-assisted cognitive-load inference.

```
The section defines the output schema and validation for the inference of cognitive
    loads with the assistance of the LLM and the NASA TLX mapping. These are the
    output validation checks for the inference.



### Expected output fields (per analysis window / segment k)
The LLM response must provide:
- MD_k: Mental Demand rating in [0, 100]
- E_k: Effort rating in [0, 100]
- TD_k: Temporal Demand rating in [0, 100]
- alpha_k, beta_k, gamma_k: non-negative weights with alpha_k + beta_k + gamma_k = 1
- CL_k: composite cognitive-load score computed as:
        CL_k = alpha_k * MD_k + beta_k * E_k + gamma_k * TD_k
- brief rationales for MD_k, E_k, and TD_k (short, evidence-grounded)

### Validation procedure (deterministic)
1. Convert the LLM's output into the specified structured format.
2. Check for presence;
   - Ensure every required field exists and that none of them are null.
3. Type check:
   - Verify that MD_k, E_k, TD_k, alpha_k, beta_k, gamma_k, and CL_k are numeric.
4. Range check;
   - Enforce 0 <= MD_k, E_k, TD_k <= 100.
   - Enforce 0 <= alpha_k, beta_k, gamma_k <= 1.
5. Weight-sum constraint;
   - Verify |(alpha_k + beta_k + gamma_k) - 1| <= eps_w, where eps_w is a small
    tolerance.
6. Consistency check;
   - Recompute CL_k_hat = alpha_k * MD_k + beta_k * E_k + gamma_k * TD_k.
   - Verify |CL_k_hat - CL_k| <= eps_cl.
7. Rationale verification (lightweight);
   - Ensure that rationales are present and concise (e.g., each no longer than 15
    words).
8. Handling failures;
   - If any requirement is not met, mark the segment output as invalid and either:
     (a) rerun the LLM using the same fixed prompt, or
     (b) produce a null/invalid record so it can be omitted from analysis.
```

```
   - The chosen failure-handling method must be kept consistent within a single
     experiment to preserve comparability.

### Notes- determinism and scope
- These validation checks are deterministic and do not change the underlying
     features.
- No optimisation or re-scoring is done during validation.
- The LLM is limited to outputting only the required fields, and all extra text is
     ignored.
```

A.4.4 Qualitative Results

Figure A.3: Example deterministic cognitive load outputs from the proposed framework, illustrating segment-wise NASA–TLX estimation and low-load transition selection.

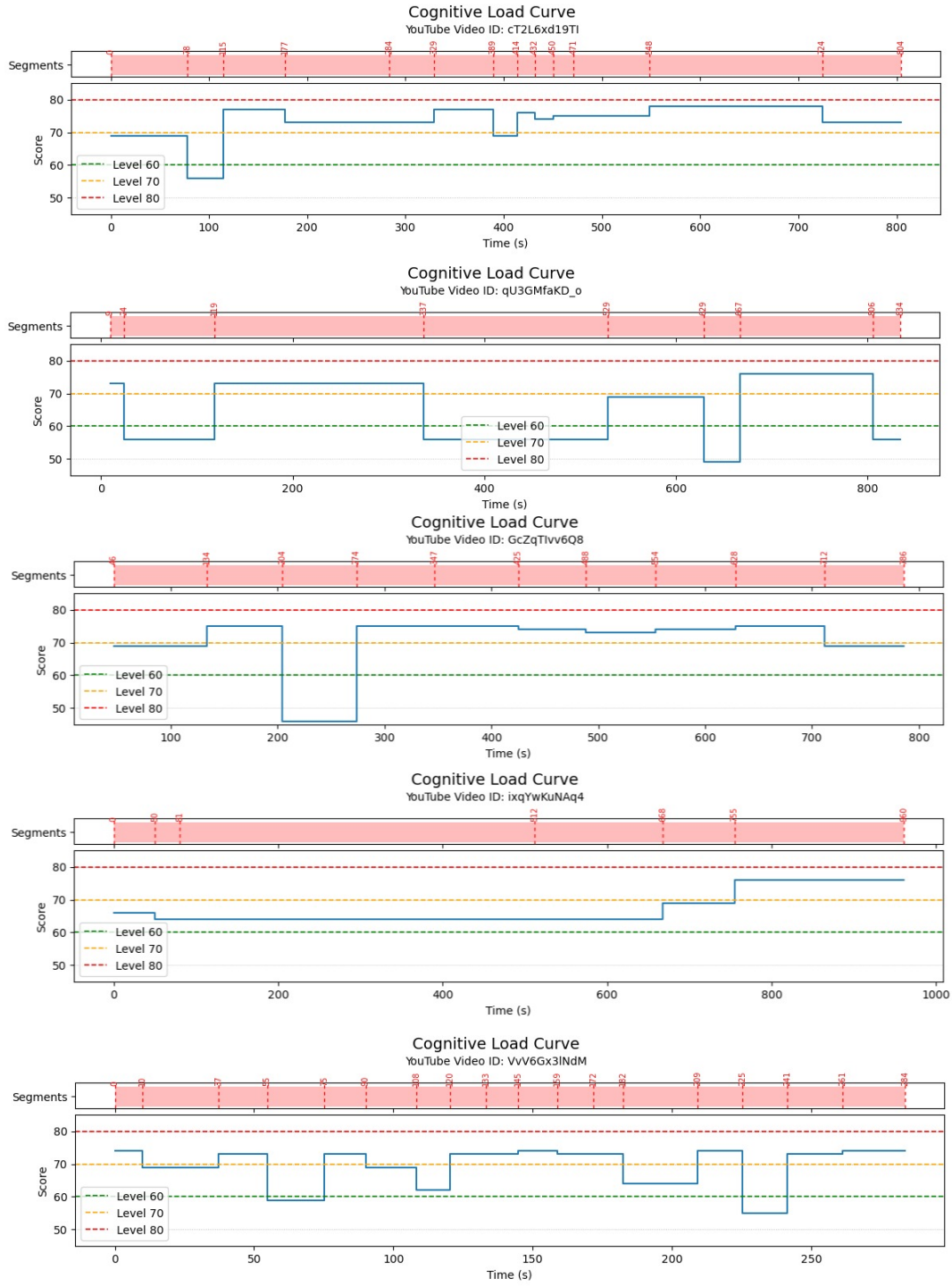


Figure A.3: Representative qualitative examples of segment-wise cognitive load trajectories produced by the proposed cognitive-load–aware segmentation framework.

A.5 Published Papers and Conference Submissions

Parts of the research presented in this thesis have been peer-reviewed and published in international journals and conference proceedings, or accepted for presentation. These publications represent both foundational and extended contributions to the thesis and are summarised below.

1. **Decoding Content Taxonomy in Video for Industry-Specific Contextual Advertising.** Waruna De Silva and A. Fernando. Accepted for publication in *Multimedia Tools and Applications*, Springer. At the time of thesis submission, this manuscript has been formally accepted following peer review and is awaiting production. The work contributes to the semantic understanding layer of the thesis, particularly the multimodal taxonomy mapping approaches developed in Chapters 3 .

2. **Toward Intelligent Ad Breaks: A Survey and Taxonomy of AI-Driven Ad Placement in Streaming Media.** Waruna De Silva and A. Fernando. *IEEE Access*, vol. 13, 2025. This survey offers a comprehensive taxonomy of AI-based advertisement placement in the context of the streaming media. The Data-Decision-Delivery approach and the AdBreakScore are introduced as the conceptual and methodological basis of the literature review and the problem statement discussed in Chapter 2.

3. **Explainable Video Topics for Content Taxonomy: A Multimodal Retrieval Approach to Industry-Compliant Contextual Advertising.** Waruna De Silva and A. Fernando.

   IEEE Access, vol. 13, 2025. This paper extends multimodal topic modelling by introducing explainable topic representations and taxonomy-aligned retrieval for contextual advertising. The methods and findings directly correspond to the explainable multimodal topic modelling framework developed in Chapter 4.

4. **Cognitive-Load-Aware Multimodal Video Segmentation for Adaptive Consumer Media Systems.** Waruna De Silva, Feng Dong, Andrew Abel, and

Anil Fernando.

Accepted for presentation at the *IEEE International Conference on Consumer Electronics (ICCE) 2026*. This conference paper introduces a cognitive-load-aware segmentation framework that estimates viewer cognitive readiness from audiovisual cues. The work forms the basis of the cognitive modelling and segmentation approach presented in Chapter 6.

Bibliography

[1] E. Lukan, "50 video statistics you can't ignore in 2025," https://www.synthesia.io/post/video-statistics, 2024, accessed: 2026-01.

[2] J. Schwenzow, J. Hartmann, A. Schikowsky, and M. Heitmann, "Understanding videos at scale: How to extract insights for business research," *Journal of Business Research*, vol. 123, pp. 367–379, 2021.

[3] S. S. Krishnan and R. K. Sitaraman, "Understanding the effectiveness of video ads: A measurement study," in *Proceedings of the 2013 conference on Internet measurement conference*, 2013, pp. 149–162.

[4] J. Caballero and C. Sora-Domenjó, "Automation and creativity in ai-driven film editing: the view from the professional documentary sector," *Communication & Society*, pp. 201–218, 2024.

[5] T. Swenberg and P. E. Eriksson, "Effects of continuity or discontinuity in actual film editing," *Empirical Studies of the Arts*, vol. 36, no. 2, pp. 222–246, 2018.

[6] P. P. Liang, A. Zadeh, and L.-P. Morency, "Foundations & trends in multimodal machine learning: Principles, challenges, and open questions," *ACM Computing Surveys*, vol. 56, no. 10, pp. 1–42, 2024.

[7] H. Ha, J. Ge, B. Feng, K. Ma, and G. Chakraborty, "Narrativetrack: Evaluating video language models beyond the frame," *arXiv preprint arXiv:2601.01095*, 2026.

[8] P. Antonenko, F. Paas, R. Grabner, and T. Van Gog, "Using electroencephalography to measure cognitive load," *Educational psychology review*, vol. 22, no. 4, pp. 425–438, 2010.

[9] IAB Tech Lab, "Content taxonomy," Interactive Advertising Bureau Tech Lab, Tech. Rep., 2023, version 3.0. [Online]. Available: https://iabtechlab.com/standards/content-taxonomy/

[10] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[11] Q. Zhang, Y. Wei, Z. Han, H. Fu, X. Peng, C. Deng, Q. Hu, C. Xu, J. Wen, D. Hu *et al.*, "Multimodal fusion on low-quality data: A comprehensive survey," *arXiv preprint arXiv:2404.18947*, 2024.

[12] T. Zhou, F. Porikli, D. J. Crandall, L. Van Gool, and W. Wang, "A survey on deep learning technique for video segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7099–7122, 2023.

[13] A. Yang, A. Nagrani, P. H. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, and C. Schmid, "Vid2seq: Large-scale pretraining of a visual language model for dense video captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 714–10 726.

[14] L. Ventura, A. Yang, C. Schmid, and G. Varol, "Chapter-llama: Efficient chaptering in hour-long videos with llms," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 18 947–18 958.

[15] R. E. Mayer, "Multimedia learning," in *Psychology of learning and motivation*. Elsevier, 2002, vol. 41, pp. 85–139.

[16] L. D. Nelson, T. Meyvis, and J. Galak, "Enhancing the television-viewing experience through commercial interruptions," *Journal of consumer research*, vol. 36, no. 2, pp. 160–172, 2009.

[17] S. Bellman, S. Treleaven-Hassard, J. A. Robinson, A. Rask, and D. Varan, "Getting the balance right," *Journal of Advertising*, vol. 41, no. 2, pp. 5–24, 2012.

[18] N. T. Nguyen, A. Zuniga, H. Lee, P. Hui, H. Flores, and P. Nurmi, "(m) ad to see me? intelligent advertisement placement: Balancing user annoyance and advertising effectiveness," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 2, pp. 1–26, 2020.

[19] U. Bulkan, T. Dagiuklas, and M. Iqbal, "Modelling quality of experience for online video advertisement insertion," *IEEE Transactions on Broadcasting*, vol. 66, no. 4, pp. 835–846, 2020.

[20] J. Brechman, S. Bellman, J. A. Robinson, A. Rask, and D. Varan, "Limited-interruption advertising in digital-video content: An analysis compares the effects of "midroll" versus "preroll" spots and clutter advertising," *Journal of advertising research*, vol. 56, no. 3, pp. 289–298, 2016.

[21] A. Kapoor, S. Narayanan, and A. Sharma, "Does emotional matching between video ads and content lead to better engagement: Evidence from a large-scale field experiment," *SSRN Electronic Journal*, 2022. [Online]. Available: https://ssrn.com/abstract=4131151

[22] F. Niu, N. Goela, A. Divakaran, and M. Abdel-Mottaleb, "Audio scene segmentation for video with generic content," in *Multimedia Content Access: Algorithms and Systems II*, vol. 6820. SPIE, 2008, pp. 307–311.

[23] S. Sen, R. Dewan, and V. M. Gupta, "Systems and method for dynamic insertion of advertisements," May 18 2021, uS Patent 11,010,946.

[24] T. J. Wen, C.-H. Chuan, W. S. Tsai, and J. Yang, "Decoding emotional (in) congruency: a computational approach toward ad placement on youtube," *Journal of interactive marketing*, vol. 57, no. 3, pp. 421–441, 2022.

[25] S. Germano, T.-L. Pham, and A. Mileo, "Web stream reasoning in practice: on the expressivity vs. scalability tradeoff," in *International Conference on Web Reasoning and Rule Systems*. Springer, 2015, pp. 105–112.

[26] W. D. Silva and A. Fernando, "Toward intelligent ad breaks: A survey and taxonomy of ai-driven ad placement in streaming media," *IEEE Access*, vol. 13, pp. 163 844–163 868, 2025.

[27] B. Gao, Y. Wang, H. Xie, Y. Hu, and Y. Hu, "Artificial intelligence in advertising: advancements, challenges, and ethical considerations in targeting, personalization, content creation, and ad optimization," *Sage Open*, vol. 13, no. 4, p. 21582440231210759, 2023.

[28] N. Devarajan, "Machine learning in targeted advertising: Examining its role, ethical implications, and impacts on consumer privacy," *Journal of Informatics Education and Research*, vol. 5, no. 1, pp. 1823–1832, 2025.

[29] X. Chen, T. V. Nguyen, Z. Shen, and M. Kankanhalli, "Livesense: Contextual advertising in live streaming videos," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 392–400.

[30] H. D. Ercan, N. S. Tanrıverdi, and N. Taşkın, "A systematic literature review for artificial intelligence in advertising," 2024. [Online]. Available: https://aisel.aisnet.org/confirm2024/1/

[31] C. Y. Joa, K. Kim, and L. Ha, "What makes people watch online in-stream video advertisements?" *Journal of Interactive Advertising*, vol. 18, no. 1, pp. 1–14, 2018.

[32] T. Prihatiningsih, R. Panudju, and I. J. Prasetyo, "Digital advertising trends and effectiveness in the modern era: A systematic literature review," *Golden Ratio of Marketing and Applied Psychology of Business*, vol. 5, no. 1, pp. 01–12, 2025.

[33] J. L. H. Frade, J. H. C. de Oliveira, and J. d. M. E. Giraldi, "Advertising in streaming video: An integrative literature review and research agenda," *Telecommunications Policy*, vol. 45, no. 9, p. 102186, 2021.

[34] J. Freeman, L. Wei, H. Yang, and F. Shen, "Does in-stream video advertising work? effects of position and congruence on consumer responses," *Journal of Promotion Management*, vol. 28, no. 5, pp. 515–536, 2022.

[35] Y. Jeong, "The impact of commercial break position on advertising effectiveness in different mood conditions," *Journal of Promotion Management*, vol. 17, no. 3, pp. 291–314, 2011.

[36] T.-K. Fan and C.-H. Chang, "Sentiment-oriented contextual advertising," *Knowledge and information systems*, vol. 23, pp. 321–344, 2010.

[37] T. J. Wen, C.-H. Chuan, J. Yang, and W. S. Tsai, "Predicting advertising persuasiveness: A decision tree method for understanding emotional (in) congruence of ad placement on youtube," *Journal of Current Issues & Research in Advertising*, vol. 43, no. 2, pp. 200–218, 2022.

[38] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cognitive science*, vol. 12, no. 2, pp. 257–285, 1988.

[39] Z. Liu, "A deep neural framework to detect individual advertisement (ad) from videos," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3578–3587.

[40] T. Akgul, S. Ozcan, and A. Iplik, "A cloud-based end-to-end server-side dynamic ad insertion platform for live content," in *Proceedings of the 11th ACM Multimedia Systems Conference*, 2020, pp. 361–364.

[41] C. Hackley, R. A. Tiwsakul, and L. Preuss, "An ethical evaluation of product placement: a deceptive practice?" *Business Ethics: A European Review*, vol. 17, no. 2, pp. 109–120, 2008.

[42] H. Anderson, "The guardian of the digital era: Assessing the impact and challenges of the children's online privacy protection act," *Law and Economy*, vol. 3, no. 2, pp. 6–10, 2024.

[43] V. Verdoodt, Y. Zhang, and E. Lievens, "Safeguarding the child's right to privacy and data protection in the european union and china: a tale of state duties and business responsibilities," *The International Journal of Human Rights*, vol. 28, no. 2, pp. 125–147, 2024.

[44] E. B. Khan, N. Tanveer, A. Shahid, M. J. Iqbal, H. A. Mirza, A. Javed, I. A. Qazi, and Z. A. Qazi, "Analyzing ad exposure and content in child-oriented videos on youtube," in *Proceedings of the ACM Web Conference 2024*, 2024, pp. 1215–1226.

[45] T. Medjkoune, O. Goga, and J. Senechal, "Marketing to children through online targeted advertising: Targeting mechanisms and legal aspects," in *Proceedings of the 2023 ACM SIGSAC conference on computer and communications Security*, 2023, pp. 180–194.

[46] S. H. Sengamedu, N. Sawant, and S. Wadhwa, "vadeo: video advertising system," in *Proceedings of the 15th ACM international conference on Multimedia*, 2007, pp. 455–456.

[47] T. Mei, L. Yang, X.-S. Hua, H. Wei, and S. Li, "Videosense: A contextual video advertising system," in *Proceedings of the 15th ACM international conference on Multimedia*, 2007, pp. 463–464.

[48] S. Chen, X. Nie, D. Fan, D. Zhang, V. Bhat, and R. Hamid, "Shot contrastive self-supervised learning for scene boundary detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9796–9805.

[49] Y. Yuan and J. Zhang, "Shot boundary detection using color clustering and attention mechanism," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 6, pp. 1–23, 2023.

[50] R. Liang, Q. Zhu, H. Wei, and S. Liao, "A video shot boundary detection approach based on cnn feature," in *2017 IEEE International Symposium on Multimedia (ISM)*, 2017, pp. 489–494.

[51] I. Bieda, A. Kisil, and T. Panchenko, "An approach to scene change detection," in *2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, vol. 1, 2021, pp. 489–493.

[52] T. Soucek and J. Lokoc, "Transnet v2: An effective deep network architecture for fast shot transition detection," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 11 218–11 221.

[53] M. Mancas and O. Le Meur, "Applications of saliency models," *From human attention to computational attention: A multidisciplinary approach*, pp. 331–377, 2016.

[54] E. Kim, S. Ratneshwar, and E. Thorson, "Why narrative ads work: An integrated process explanation," *Journal of Advertising*, vol. 46, no. 2, pp. 283–296, 2017.

[55] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor, "Learning video saliency from human gaze using candidate selection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1147–1154.

[56] M. Kümmerer, T. S. Wallis, and M. Bethge, "Deepgaze ii: Reading fixations from deep features trained on object recognition," *arXiv preprint arXiv:1610.01563*, 2016.

[57] F. Lu, Y. Lian, W. Sun, X. Min, and G. Zhai, "A study on the user viewing experience of implanted advertisement videos based on visual saliency," *Expert Systems with Applications*, p. 128499, 2025.

[58] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI'81: 7th international joint conference on Artificial intelligence*, vol. 2, 1981, pp. 674–679.

[59] H. E. Wong, O. Akar, E. A. Cuevas, I. Tabian, D. Ravichandran, I. Fu, and C. Carter, "Markerless augmented advertising for sports videos," in *Computer Vision–ACCV 2018 Workshops: 14th Asian Conference on Computer Vision,*

Perth, Australia, December 2–6, 2018, Revised Selected Papers 14. Springer, 2019, pp. 494–509.

[60] V. Efimova, L. Fedotov, V. Shalamov, and A. Filchenkov, "Advertisement replacement in video," in *Fourteenth International Conference on Machine Vision (ICMV 2021)*, vol. 12084. SPIE, 2022, pp. 232–240.

[61] I. Bacher, H. Javidnia, S. Dev, R. Agrahari, M. Hossari, M. Nicholson, C. Conran, J. Tang, P. Song, D. Corrigan *et al.*, "An advert creation system for 3d product placements," in *Machine Learning and Knowledge Discovery in Databases: Applied Data Science Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part IV*. Springer, 2021, pp. 224–239.

[62] J. Hu, G. Li, Z. Lu, J. Xiao, and R. Hong, "Videoader: a video advertising system based on intelligent analysis of visual content," in *Proceedings of the Third International Conference on Internet Multimedia Computing and Service*, 2011, pp. 30–33.

[63] H. Zhang, X. Cao, J. K. Ho, and T. W. Chow, "Object-level video advertising: an optimization framework," *IEEE Transactions on industrial informatics*, vol. 13, no. 2, pp. 520–531, 2016.

[64] C. Luo, Y. Peng, T. Zhu, and L. Li, "An optimization framework of video advertising: using deep learning algorithm based on global image information," *Cluster Computing*, vol. 22, no. Suppl 4, pp. 8939–8951, 2019.

[65] H. Zhang, Y. Ji, W. Huang, and L. Liu, "Sitcom-star-based clothing retrieval for video advertising: a deep learning framework," *Neural computing and applications*, vol. 31, pp. 7361–7380, 2019.

[66] R. Ferreira, K. Okada, M. Cristo, K. Berlt, E. S. de Moura, and D. Fernandes, "Scripts as source of information to contextual video advertising," in *Proceedings of the 18th Brazilian symposium on Multimedia and the web*, 2012, pp. 39–46.

[67] M. J. Welch, J. Cho, and W. Chang, "Generating advertising keywords from video content," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010, pp. 1421–1424.

[68] B.-J. Yi, J.-T. Lee, H.-W. Woo, and H.-C. Rim, "Contextual video advertising system using scene information inferred from video scripts," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010, pp. 771–772.

[69] R. De Michele and M. Furini, "Viewer-tailored advertising for video on demand platforms," in *2019 16th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2019, pp. 1–4.

[70] R. Tapu, B. Mocanu, and T. Zaharia, "Deep-ad: A multimodal temporal video segmentation framework for online video advertising," *IEEE Access*, vol. 8, pp. 99 582–99 597, 2020.

[71] B. Mocanu and R. Tapu, "Semanticad: A multimodal contextual advertisement framework for online video streaming platforms," *IEEE Access*, vol. 12, pp. 63 142–63 155, 2024.

[72] Z.-Q. Cheng, X. Wu, Y. Liu, and X.-S. Hua, "Video ecommerce++: Toward large scale online video advertising," *IEEE transactions on multimedia*, vol. 19, no. 6, pp. 1170–1183, 2017.

[73] M. Yuan, L. Zhang, Z. Wu, and D. Zheng, "High-quality activity-level video advertising," in *2020 IEEE/ACM 28th International Symposium on Quality of Service (IWQoS)*, 2020, pp. 1–10.

[74] A. Chaubey, A. Agarwaal, S. S. Roy, A. Agrawal, and S. Ghose, "Contextiq: A multimodal expert-based video retrieval system for contextual advertising," *arXiv preprint arXiv:2410.22233*, 2024.

[75] W. de Silva and A. Fernando, "Explainable video topics for content taxonomy: A multimodal retrieval approach to industry-compliant contextual advertising," *IEEE Access*, vol. 13, pp. 30 597–30 612, 2025.

[76] S. Ramagundam and N. Karne, "Decoding sentiments: An efficient trans-long short term memory to understand viewer reactions on ad-supported video contents," in *2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*, 2024, pp. 1–6.

[77] X. Zhuang, F. Liu, J. Hou, J. Hao, and X. Cai, "Transformer-based interactive multi-modal attention network for video sentiment detection," *Neural Processing Letters*, vol. 54, no. 3, pp. 1943–1960, 2022.

[78] Q. Qi, L. Lin, R. Zhang, and C. Xue, "Medt: Using multimodal encoding-decoding network as in transformer for multimodal sentiment analysis," *IEEE Access*, vol. 10, pp. 28 750–28 759, 2022.

[79] G. Wang, L. Zhuo, J. Li, D. Ren, and J. Zhang, "An efficient method of content-targeted online video advertising," *Journal of visual communication and image representation*, vol. 50, pp. 40–48, 2018.

[80] O. K. Chong, H.-N. Goh, and J. See, "What modality matters? exploiting highly relevant features for video advertisement insertion," in *2023 IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 3344–3348.

[81] T. Ye, Y. Liu, M. Yang, L. Zhang, and Z. Li, "Semantic fusion based graph network for video scene detection," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–8.

[82] A. Rao, L. Xu, Y. Xiong, G. Xu, Q. Huang, B. Zhou, and D. Lin, "A local-to-global approach to multi-modal movie scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 146–10 155.

[83] J. Guo, T. Mei, F. Liu, and X.-S. Hua, "Adon: an intelligent overlay video advertising system," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009, pp. 628–629.

[84] H. Mahmood, S. M. S. Islam, S. O. Gilani, and Y. Ayaz, "Dynamic saliency model inspired by middle temporal visual area: A spatio-temporal perspective," in *2018 Digital Image Computing: Techniques and Applications (DICTA)*, 2018, pp. 1–8.

[85] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections in 3d," in *ACM siggraph 2006 papers*, 2006, pp. 835–846.

[86] R. Madhok, S. Mujumdar, N. Gupta, and S. Mehta, "Semantic understanding for contextual in-video advertising," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[87] N. Kobets, T. Kovaliuk, G. Shekhet, and T. Tielysheva, "Analysis of streaming video content and generation relevant contextual advertising," in *CEUR Workshop Proceedings Vol-2604, Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020). Volume I: Main Conference Lviv, Ukraine, April 23-24, 2020.*, vol. 2604, no. 1. ceur-ws. org, 2020, pp. 829–843.

[88] X. Song, B. Xu, and Y.-G. Jiang, "Predicting content similarity via multimodal modeling for video-in-video advertising," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 569–581, 2021.

[89] G. Dobrița, S.-V. Oprea, and A. Bâra, "Federated learning's role in next-gen tv ad optimization." *Computers, Materials & Continua*, vol. 82, no. 1, 2025.

[90] A. Modi and H. R. Chowdhary, "The impact of predictive analytics on modern marketing strategies: A data-driven approach," *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 13, no. II, pp. 662–667, 2025. [Online]. Available: https://doi.org/10.22214/ijraset.2025.66929

[91] J. Ikeda, H. Seshime, X. Wang, and T. Yamasaki, "Predicting online video advertising effects with multimodal deep learning," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 2995–3002.

[92] E. Artioli, F. Tashtarian, and C. Timmerer, "Digitwise: Digital twin-based modeling of adaptive video streaming engagement," in *Proceedings of the 15th ACM Multimedia Systems Conference*, 2024, pp. 78–88.

[93] D. McDuff and J. Berger, "Do facial expressions predict ad sharing? a large-scale observational study," *arXiv preprint arXiv:1912.10311*, 2019.

[94] R. R. Sarukkai, "Real-time user modeling and prediction: examples from youtube," in *Proceedings of the 22nd International Conference on World Wide Web*, 2013, pp. 775–776.

[95] J. Yang, K. Wang, X. Peng, and Y. Qiao, "Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction," in *Proceedings of the 20th ACM international conference on multimodal interaction*, 2018, pp. 594–598.

[96] X. Tian, B. P. Nunes, Y. Liu, and R. Manrique, "Predicting student engagement using sequential ensemble model," *IEEE Transactions on Learning Technologies*, vol. 17, pp. 939–950, 2024.

[97] S. Wu, M.-A. Rizoiu, and L. Xie, "Beyond views: Measuring and predicting engagement in online videos," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, 2018.

[98] L. Stappen, A. Baird, M. Lienhart, A. Bätz, and B. Schuller, "An estimation of online video user engagement from features of time-and value-continuous, dimensional emotions," *Frontiers in Computer Science*, vol. 4, p. 773154, 2022.

[99] A. Dedieu, R. Mazumder, Z. Zhu, and H. Vahabi, "Hierarchical modeling and shrinkage for user session lengthprediction in media streaming," in *Proceedings of*

the 27th ACM International Conference on Information and Knowledge Management, 2018, pp. 607–616.

[100] S. Wassermann, N. Wehner, and P. Casas, "Machine learning models for youtube qoe and user engagement prediction in smartphones," *ACM SIGMETRICS Performance Evaluation Review*, vol. 46, no. 3, pp. 155–158, 2019.

[101] C. Qiao, J. Wang, Y. Wang, Y. Liu, and H. Tuo, "Understanding and improving user engagement in adaptive video streaming," in *2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS)*, 2021, pp. 1–10.

[102] P. Lebreton, K. Kawashima, K. Yamagishi, and J. Okamoto, "Study on viewing time with regards to quality factors in adaptive bitrate video streaming," in *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, 2018, pp. 1–6.

[103] P. Lebreton and K. Yamagishi, "Predicting user quitting ratio in adaptive bitrate video streaming," *IEEE Transactions on Multimedia*, vol. 23, pp. 4526–4540, 2021.

[104] R. Huber, B. Scheibehenne, A. Chapiro, S. Frey, and R. W. Sumner, "The influence of visual salience on video consumption behavior: A survival analysis approach," in *Proceedings of the ACM Web Science Conference*, 2015, pp. 1–2.

[105] E. Aguiar, S. Nagrecha, and N. V. Chawla, "Predicting online video engagement using clickstreams," in *2015 IEEE international conference on data science and advanced analytics (DSAA)*. IEEE, 2015, pp. 1–10.

[106] D. R. Cox, *Analysis of survival data*. Chapman and Hall/CRC, 2018.

[107] F. Qiu and Y. Cui, "A quantitative study of user satisfaction in online video streaming," in *2011 IEEE Consumer Communications and Networking Conference (CCNC)*, 2011, pp. 410–414.

[108] P. Lebreton and K. Yamagishi, "Study on user quitting in the puffer live tv video streaming service," in *2021 13th International Conference on Quality of Multimedia Experience (QoMEX)*, 2021, pp. 19–24.

[109] K. Kapoor, M. Sun, J. Srivastava, and T. Ye, "A hazard based approach to user return time prediction," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 1719–1728.

[110] H. Zhang, J. Yan, and Y. Zhang, "An attention-based deep network for ctr prediction," in *Proceedings of the 2020 12th international conference on machine learning and computing*, 2020, pp. 1–5.

[111] T. Mehta and G. Deshmukh, "Youtube ad view sentiment analysis using deep learning and machine learning," *arXiv preprint arXiv:2205.11082*, 2022.

[112] P. Rajaram, P. Manchanda, and E. Schwartz, "Finding the sweet spot: Ad scheduling on streaming media," *SSRN Electron. J.*, 2019.

[113] R. Gunasundari and C. Balakumar, "Improve accuracy in advertisement viewability prediction using plc and logistic regression," in *2023 International Conference on Emerging Research in Computational Science (ICERCS)*. IEEE, 2023, pp. 1–6.

[114] J. Liu and H. Yang, "Research on interactive advertising strategies using machine learning algorithms for user behavior analysis," in *Proceedings of the 2023 4th International Conference on Big Data Economy and Information Management*, 2023, pp. 181–184.

[115] X. Zhao, C. Gu, H. Zhang, X. Liu, X. Yang, and J. Tang, "Deep reinforcement learning for online advertising in recommender systems," *arXiv preprint arXiv:1909.03602*, 2019.

[116] X. Zhao, X. Zheng, X. Yang, X. Liu, and J. Tang, "Jointly learning to recommend and advertise," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 3319–3327.

[117] S. Ramagundam, "Machine learning algorithmic approaches to maximizing user engagement through ad placements," *International Journal of Scientific Research*

in Computer Science, Engineering and Information Technology, vol. 6, no. 3, pp. 1164–1174, 2020.

[118] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, pp. 235–256, 2002.

[119] D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen *et al.*, "A tutorial on thompson sampling," *Foundations and Trends® in Machine Learning*, vol. 11, no. 1, pp. 1–96, 2018.

[120] R. Jain, P. Nagrath, S. T. Raina, P. Prakash, and A. Thareja, "Ads optimization using reinforcement learning," in *Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020*. Springer, 2021, pp. 53–63.

[121] S. Zhang, "Utilizing reinforcement learning bandit algorithms in advertising optimization," *Highlights in Science, Engineering and Technology*, vol. 94, pp. 195–200, 2024.

[122] R. Reddy, S. Sharma, D. Reddy, and P. Singh, "Leveraging deep reinforcement learning and natural language processing for enhanced personalized video marketing strategies," *Journal of AI ML Research*, vol. 10, no. 8, 2021.

[123] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[124] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. Pmlr, 2018, pp. 1861–1870.

[125] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[126] X. Li, Y. Wang, J. Yu, X. Zeng, Y. Zhu, H. Huang, J. Gao, K. Li, Y. He, C. Wang *et al.*, "Videochat-flash: Hierarchical compression for long-context video modeling," *arXiv preprint arXiv:2501.00574*, 2024.

[127] M. Xu, M. Gao, Z. Gan, H.-Y. Chen, Z. Lai, H. Gang, K. Kang, and A. Dehghan, "Slowfast-llava: A strong training-free baseline for video large language models," *arXiv preprint arXiv:2407.15841*, 2024.

[128] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[129] M. T. Ribeiro, S. Singh, and C. Guestrin, """ why should i trust you?""" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[130] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in neural information processing systems*, vol. 12, 1999.

[131] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[132] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "A brief survey of deep reinforcement learning," *arXiv preprint arXiv:1708.05866*, 2017.

[133] G. Dulac-Arnold, D. Mankowitz, and T. Hester, "Challenges of real-world reinforcement learning," *arXiv preprint arXiv:1904.12901*, 2019.

[134] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.

[135] X. Zhao, L. Zhang, L. Xia, Z. Ding, D. Yin, and J. Tang, "Deep reinforcement learning for list-wise recommendations," *arXiv preprint arXiv:1801.00209*, 2017.

[136] A. U. Rehman, X. Shi, F. Ullah, Z. Wang, and C. Ma, "Measuring student attention based on eeg brain signals using deep reinforcement learning," *Expert Systems with Applications*, vol. 269, p. 126426, 2025.

[137] J. K. Chalaby, "The streaming industry and the platform economy: An analysis," *Media, Culture & Society*, vol. 46, no. 3, pp. 552–571, 2024.

[138] K. M. Miller and B. Skiera, "Economic consequences of online tracking restrictions: Evidence from cookies," *International journal of research in marketing*, vol. 41, no. 2, pp. 241–264, 2024.

[139] D. A. Cooper, T. Yalcin, C. Nistor, M. Macrini, and E. Pehlivan, "Privacy considerations for online advertising: A stakeholder's perspective to programmatic advertising," *Journal of Consumer Marketing*, vol. 40, no. 2, pp. 235–247, 2023.

[140] C. Rohrer and J. Boyd, "The rise of intrusive online advertising and the response of user experience research at yahoo!" in *CHI'04 extended Abstracts on human factors in Computing systems*, 2004, pp. 1085–1086.

[141] H. Li, S. M. Edwards, and J.-H. Lee, "Measuring the intrusiveness of advertisements: Scale development and validation," *Journal of advertising*, vol. 31, no. 2, pp. 37–47, 2002.

[142] M. Singh and R. Lamba, "Proposing contextually relevant advertisements for online videos," in *Machine Learning and Metaheuristics Algorithms, and Applications: First Symposium, SoMMA 2019, Trivandrum, India, December 18–21, 2019, Revised Selected Papers 1*. Springer, 2020, pp. 218–224.

[143] C. Xiang, T. V. Nguyen, and M. Kankanhalli, "Salad: A multimodal approach for contextual video advertising," in *2015 IEEE international symposium on multimedia (ISM)*. IEEE, 2015, pp. 211–216.

[144] K. Okada, E. S. de Moura, M. Cristo, D. Fernandes, M. A. Gonçalves, and K. Berlt, "Advertisement selection for online videos," in *Proceedings of the 18th Brazilian symposium on Multimedia and the web*, 2012, pp. 367–374.

[145] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4694–4702.

[146] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20–36.

[147] Q. Wu, A. Zhu, R. Cui, T. Wang, F. Hu, Y. Bao, and H. Snoussi, "Pose-guided inflated 3d convnet for action recognition in videos," *Signal Processing: Image Communication*, vol. 91, p. 116098, 2021.

[148] H. Tang, L. Ding, S. Wu, B. Ren, N. Sebe, and P. Rota, "Deep unsupervised key frame extraction for efficient video classification," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 3, pp. 1–17, 2023.

[149] K. Drossos, S. Adavanne, and T. Virtanen, "Automated audio captioning with recurrent neural networks," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 374–378.

[150] M. Wu, H. Dinkel, and K. Yu, "Audio caption: Listen and tell," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 830–834.

[151] D. Yu and L. Deng, *Automatic speech recognition*. Springer, 2016, vol. 1.

[152] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[153] Interactive Advertising Bureau, "Interactive advertising bureau taxonomies," https://github.com/InteractiveAdvertisingBureau/Taxonomies, 2023, accessed: December 9, 2023.

[154] J. Wood, P. Tan, W. Wang, and C. Arnold, "Source-lda: Enhancing probabilistic topic models using prior knowledge sources," in *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, 2017, pp. 411–422.

[155] J. H. Lau, D. Newman, S. Karimi, and T. Baldwin, "Best topic word selection for topic labelling," in *Coling 2010: Posters*, 2010, pp. 605–613.

[156] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[157] G. J. et al., "ultralytics/yolov5: v7.0 - yolov5 sota realtime instance segmentation," Oct. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7002879

[158] J. Xue and K. Eguchi, "Sequential bayesian nonparametric multimodal topic models for video data analysis," *IEICE TRANSACTIONS on Information and Systems*, vol. 101, no. 4, pp. 1079–1087, 2018.

[159] J. Thies, L. Stappen, G. Hagerer, B. W. Schuller, and G. Groh, "Graphtmt: Unsupervised graph-based topic modeling from video transcripts," in *2021 IEEE Seventh International Conference on Multimedia Big Data (BigMM)*, 2021, pp. 1–8.

[160] G. Rafiq, M. Rafiq, and G. S. Choi, "Video description: A comprehensive survey of deep learning approaches," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 13 293–13 372, 2023.

[161] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[162] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PmLR, 2020, pp. 1597–1607.

[163] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[164] H. Yang, D. Huang, B. Wen, J. Wu, H. Yao, Y. Jiang, X. Zhu, and Z. Yuan, "Self-supervised video representation learning with motion-aware masked autoencoders," *arXiv preprint arXiv:2210.04154*, 2022.

[165] B. Huang, Z. Zhao, G. Zhang, Y. Qiao, and L. Wang, "Mgmae: Motion guided masking for video masked autoencoding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 493–13 504.

[166] J. Hernandez, R. Villegas, and V. Ordonez, "Vic-mae: Self-supervised representation learning from images and video with contrastive masked autoencoders," in *European Conference on Computer Vision*. Springer, 2024, pp. 444–463.

[167] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7464–7473.

[168] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Icml*, vol. 2, no. 3, 2021, p. 4.

[169] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.

[170] J. Chen, H. Zhang, M. Gong, and Z. Gao, "Collaborative compensative transformer network for salient object detection," *Pattern Recognition*, vol. 154, p. 110600, 2024.

[171] Y. Wang, J. Li, H. Zhang, J. Zhang, F. Wan, A. Qiu, and Z. Gao, "Fedmdd: Multi-deliberation based calibration for federated long-tailed learning," *Knowledge-Based Systems*, p. 113741, 2025.

[172] V. Sharma, "Object detection and recognition using amazon rekognition with boto3," in *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2022, pp. 727–732.

[173] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[174] T. Jiao, C. Guo, X. Feng, Y. Chen, and J. Song, "A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications," *Computers, Materials and Continua*, vol. 80, no. 1, pp. 1–35, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1546221824005216

[175] B. V. Barde and A. M. Bainwad, "An overview of topic modeling methods and tools," in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2017, pp. 745–750.

[176] Sivic and Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proceedings ninth IEEE international conference on computer vision*. IEEE, 2003, pp. 1470–1477.

[177] R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, vol. 6, no. 1, 2015.

[178] L. Zhou, Z. Zhou, and D. Hu, "Scene classification using a multi-resolution bag-of-features model," *Pattern Recognition*, vol. 46, no. 1, pp. 424–433, 2013.

[179] F. B. Silva, R. d. O. Werneck, S. Goldenstein, S. Tabbone, and R. d. S. Torres, "Graph-based bag-of-words for classification," *Pattern Recognition*, vol. 74, pp. 266–285, 2018.

[180] D. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," *IEEE signal processing magazine*, vol. 27, no. 6, pp. 55–65, 2010.

[181] W. Wang, P. M. Barnaghi, and A. Bargiela, "Probabilistic topic models for learning terminological ontologies," *IEEE Transactions on knowledge and Data engineering*, vol. 22, no. 7, pp. 1028–1040, 2009.

[182] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, pp. 530–539.

[183] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, 2012, pp. 952–961.

[184] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[185] Z. Cao, S. Li, Y. Liu, W. Li, and H. Ji, "A novel neural topic model and its supervised extension," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.

[186] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/

[187] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[188] R. Egger and J. Yu, "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts," *Frontiers in sociology*, vol. 7, p. 886498, 2022.

[189] H. Rahimi, H. Naacke, C. Constantin, and B. Amann, "Antm: aligned neural topic models for exploring evolving topics," in *Transactions on Large-Scale Data-and Knowledge-Centered Systems LVI: Special Issue on Data Management-Principles, Technologies, and Applications*. Springer, 2024, pp. 76–97.

[190] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[191] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2013, pp. 160–172.

[192] L. McInnes, J. Healy, S. Astels *et al.*, "hdbscan: Hierarchical density based clustering." *J. Open Source Softw.*, vol. 2, no. 11, p. 205, 2017.

[193] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[194] M. Grootendorst, "Keybert: Minimal keyword extraction with bert," https://github.com/MaartenGr/KeyBERT, 2020.

[195] J. C. Van Gemert, C. J. Veenman, A. W. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 7, pp. 1271–1283, 2009.

[196] K. Van De Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1582–1596, 2009.

[197] D. Hall, D. Jurafsky, and C. D. Manning, "Studying the history of ideas using topic models," in *Proceedings of the 2008 conference on empirical methods in natural language processing*, 2008, pp. 363–371.

[198] S. Li, "Food.com recipes and user interactions," https://www.kaggle.com/datasets/shuyangli94/food-com-recipes-and-user-interactions, 2019, kaggle dataset, accessed: Apr. 10, 2023.

[199] M. Roccetti, G. Delnevo, L. Casini, and G. Cappiello, "Is bigger always better? a controversial journey to the center of machine learning design, with uses and misuses of big data for predicting water meter failures," *Journal of Big Data*, vol. 6, no. 1, pp. 1–23, 2019.

[200] C. Grece, "Trends in the vod market in eu28," *European Audiovisual Observatory*, 2021.

[201] S. Broughton Micova and S. Jacques, "Platform power in the video advertising ecosystem," *Broughton Micova, S. & Jacques, S.(2020). Platform power in the video advertising ecosystem. Internet Policy Review*, vol. 9, no. 4, 2020.

[202] Z. Liu, U. Iqbal, and N. Saxena, "Opted out, yet tracked: Are regulations enough to protect your privacy?" *arXiv preprint arXiv:2202.00885*, 2022.

[203] A. Turillazzi, M. Taddeo, L. Floridi, and F. Casolari, "The digital services act: an analysis of its ethical, legal, and social implications," *Law, Innovation and Technology*, vol. 15, no. 1, pp. 83–106, 2023.

[204] M. Veale and F. Z. Borgesius, "Adtech and real-time bidding under european data protection law," *German Law Journal*, vol. 23, no. 2, pp. 226–256, 2022.

[205] T. Mei, J. Guo, X.-S. Hua, and F. Liu, "Adon: Toward contextual overlay in-video advertising," *Multimedia systems*, vol. 16, pp. 335–344, 2010.

[206] T. Mei and X.-S. Hua, "Contextual internet multimedia advertising," *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1416–1433, 2010.

[207] T. Kozlova, "Efficiency of business and intercultural communication: Multilingual advertising discourse," in *III International Scientific Congress Society of Ambient Intelligence 2020 (ISC-SAI 2020)*. Atlantis Press, 2020, pp. 272–278.

[208] E. Häglund and J. Björklund, "Ai-driven contextual advertising: Toward relevant messaging without personal data," *Journal of Current Issues & Research in Advertising*, vol. 45, no. 3, pp. 301–319, 2024.

[209] Q. Yang, M. Ongpin, S. Nikolenko, A. Huang, and A. Farseev, "Against opacity: Explainable ai and large language models for effective digital advertising," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 9299–9305.

[210] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399–408.

[211] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[212] I. T. Lab, "Content taxonomy 3.0," 2021, accessed: 2024-07-19. [Online]. Available: https://iabtechlab.com/standards/content-taxonomy/

[213] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[214] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[215] H. Face, "Vit-gpt2 image captioning," 2023. [Online]. Available: https://huggingface.co/nlpconnect/vit-gpt2-image-captioning

[216] N. Reimers, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[217] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," *arXiv preprint arXiv:2007.01852*, 2020.

[218] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, A. Chen, A. Chowdhery, S. Narang, G. Mishra, A. W. Yu, X. Zhao, Y. Huang, H. Lee, D. Zhou, Q. V. Le, E. Chi, J. Dean, and J. Devlin, "Scaling instruction-finetuned language models," 2022. [Online]. Available: https://arxiv.org/abs/2210.11416

[219] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen *et al.*, "Palm 2 technical report," *arXiv preprint arXiv:2305.10403*, 2023.

[220] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey," *Multimedia tools and applications*, vol. 78, pp. 15 169–15 211, 2019.

[221] F. Bianchi, S. Terragni, D. Hovy, D. Nozza, and E. Fersini, "Cross-lingual contextualized topic models with zero-shot learning," *arXiv preprint arXiv:2004.07737*, 2020.

[222] N. Prakash, H. Wang, N. K. Hoang, M. S. Hee, and R. K.-W. Lee, "Promptmtopic: Unsupervised multimodal topic modeling of memes using large language models," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 621–631.

[223] J. Pfau, A. T. Young, J. Wei, M. L. Wei, and M. J. Keiser, "Robust semantic interpretability: Revisiting concept activation vectors," *arXiv preprint arXiv:2104.02768*, 2021.

[224] S. Cai, L. Qiu, X. Chen, Q. Zhang, and L. Chen, "Semantic-enhanced image clustering," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 6, 2023, pp. 6869–6878.

[225] X. Dong, J. Bao, T. Zhang, D. Chen, S. Gu, W. Zhang, L. Yuan, D. Chen, F. Wen, and N. Yu, "Clip itself is a strong fine-tuner: Achieving 85.7% and 88.0% top-1 accuracy with vit-b and vit-l on imagenet," *arXiv preprint arXiv:2212.06138*, 2022.

[226] E. Zosa and L. Pivovarova, "Multilingual and multimodal topic modelling with pretrained embeddings," *arXiv preprint arXiv:2211.08057*, 2022.

[227] J. Libovickỳ, R. Rosa, and A. Fraser, "How language-neutral is multilingual bert?" *arXiv preprint arXiv:1911.03310*, 2019.

[228] D. Kozlowski, C. Pradier, and P. Benz, "Generative ai for automatic topic labelling," *arXiv preprint arXiv:2408.07003*, 2024.

[229] Interactive Advertising Bureau, "Content taxonomy 3.1," 2024. [Online]. Available: https://iabtechlab.com/standards/content-taxonomy/

[230] FFmpeg Developers, "Ffmpeg," 2024. [Online]. Available: https://ffmpeg.org/

[231] OpenAI, "Whisper," 2024, accessed: 2024-07-19. [Online]. Available: https://github.com/openai/whisper

[232] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 478–493, 2020.

[233] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, 2020, pp. 1–6.

[234] J. Wu, Y. Liang, H. Akbari, Z. Wang, C. Yu *et al.*, "Scaling multimodal pretraining via cross-modality gradient harmonization," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 161–36 173, 2022.

[235] R. B. Pereira, A. Plastino, B. Zadrozny, and L. H. Merschmann, "Correlation analysis of performance measures for multi-label classification," *Information Processing & Management*, vol. 54, no. 3, pp. 359–369, 2018.

[236] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," *Advances in neural information processing systems*, vol. 36, pp. 11 809–11 822, 2023.

[237] B. Castellano, "Pyscenedetect: Video scene cut detection and analysis tool," 2025. [Online]. Available: https://github.com/Breakthrough/PySceneDetect

[238] M. A. Hearst, "Text tiling: Segmenting text into multi-paragraph subtopic passages," *Computational linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[239] M. Riedl and C. Biemann, "Topictiling: a text segmentation algorithm based on lda," in *Proceedings of ACL 2012 student research workshop*, 2012, pp. 37–42.

[240] M. Narasimhan, A. Rohrbach, and T. Darrell, "Clip-it! language-guided video summarization," *Advances in neural information processing systems*, vol. 34, pp. 13 988–14 000, 2021.

[241] U. Gargi, R. Kasturi, and S. Strayer, "Performance characterization of video-shot-change detection methods," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 1, pp. 1–13, 2000.

[242] J. S. Boreczky and L. A. Rowe, "Comparison of video shot boundary detection techniques," *Journal of Electronic Imaging*, vol. 5, no. 2, pp. 122–128, 1996.

[243] F. Y. Choi, "Advances in domain independent linear text segmentation," *arXiv preprint cs/0003083*, 2000.

[244] A. Yang, A. Nagrani, I. Laptev, J. Sivic, and C. Schmid, "Vidchapters-7m: Video chapters at scale," *Advances in Neural Information Processing Systems*, vol. 36, pp. 49 428–49 444, 2023.

[245] D. Bordwell and K. Thompson, *Film Art: An Introduction*, 10th ed. McGraw-Hill Education, 2013.

[246] M. Antoniak and D. Mimno, "Evaluating the stability of embedding-based word similarities," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 107–119, 2018.

[247] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," *arXiv preprint arXiv:1904.09751*, 2019.

[248] A. Solbiati, K. Heffernan, G. Damaskinos, S. Poddar, S. Modi, and J. Cali, "Unsupervised topic segmentation of meetings with bert embeddings," *arXiv preprint arXiv:2106.12978*, 2021.

[249] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *European conference on computer vision*. Springer, 2014, pp. 540–555.

[250] M. Afham, S. N. Shukla, O. Poursaeed, P. Zhang, A. Shah, and S. Lim, "Revisiting kernel temporal segmentation as an adaptive tokenizer for long-form video understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1189–1194.

[251] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," *arXiv preprint arXiv:2303.00747*, 2023.

[252] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière *et al.*, "Gemma 3 technical report," *arXiv preprint arXiv:2503.19786*, 2025.

[253] Amazon Web Services, "Amazon ec2 g6 instances," https://aws.amazon.com/ec2/instance-types/g6/, 2024, accessed: February 22, 2026.

[254] OpenAI, "Images and vision | openai api," https://platform.openai.com/docs/guides/vision, 2025, accessed: February 22, 2026. Official documentation on using vision capabilities and image generation in the OpenAI API.

[255] S. Y. Chen, G. Ghinea, and R. D. Macredie, "A cognitive approach to user perception of multimedia quality: An empirical investigation," *International Journal of Human-Computer Studies*, vol. 64, no. 12, pp. 1200–1213, 2006.

[256] J. A. McCabe, C. S. Banasik, M. G. Jackson, E. M. Postlethwait, A. Steitz, and A. R. Wenzel, "Exploring perceptions of cognitive load and mental fatigue in pandemic-era zoom classes." *Scholarship of Teaching and Learning in Psychology*, 2023.

[257] J. M. Zacks, N. K. Speer, K. M. Swallow, T. S. Braver, and J. R. Reynolds, "Event perception: a mind-brain perspective." *Psychological bulletin*, vol. 133, no. 2, p. 273, 2007.

[258] D. Liu, "The effects of segmentation on cognitive load, vocabulary learning and retention, and reading comprehension in a multimedia learning environment," *BMC psychology*, vol. 12, no. 1, p. 4, 2024.

[259] M. Ibrahim, P. D. Antonenko, C. M. Greenwood, and D. Wheeler, "Effects of segmenting, signalling, and weeding on learning from educational video," *Learning, media and technology*, vol. 37, no. 3, pp. 220–235, 2012.

[260] X. Wang, X. Li, Y. Wei, X. Song, Y. Song, X. Xia, F. Zeng, Z. Chen, L. Liu, G. Xu *et al.*, "From long videos to engaging clips: A human-inspired video editing framework with multimodal narrative understanding," *arXiv preprint arXiv:2507.02790*, 2025.

[261] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in Psychology*. Elsevier, 1988, vol. 52, pp. 139–183.

[262] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. M. Van Gerven, "Cognitive load measurement as a means to advance cognitive load theory," in *Cognitive Load Theory*. Routledge, 2016, pp. 63–71.

[263] M. P. Oppelt, A. Foltyn, J. Deuschel, N. R. Lang, N. Holzer, B. M. Eskofier, and S. H. Yang, "Adabase: a multimodal dataset for cognitive load estimation," *Sensors*, vol. 23, no. 1, p. 340, 2022.

[264] L. Lin and R. K. Atkinson, "Using animations and visual cueing to support learning of scientific concepts and processes," *Computers & Education*, vol. 56, no. 3, pp. 650–658, 2011.

[265] L. Nguyen-Phuoc, R. Gaboriau, D. Delacroix, and L. Navarro, "M&M: Multimodal-multitask model integrating audiovisual cues in cognitive load assessment," arXiv preprint arXiv:2403.09451, 2024.

[266] R. Sarailoo, K. Latifzadeh, S. H. Amiri, A. Bosaghzadeh, and R. Ebrahimpour, "Assessment of instantaneous cognitive load imposed by educational multime-

dia using electroencephalography signals," *Frontiers in Neuroscience*, vol. 16, p. 744737, 2022.

[267] P. J. Guo, J. Kim, and R. Rubin, "How video production affects student engagement: An empirical study of mooc videos," in *Proceedings of the First ACM Conference on Learning@Scale*, 2014, pp. 41–50.

[268] C. Lamontagne, S. Sénécal, M. Fredette, É. Labonté-LeMoyne, and P.-M. Léger, "The effect of the segmentation of video tutorials on user's training experience and performance," *Computers in Human Behavior Reports*, vol. 3, p. 100071, 2021.

[269] S. Gupta, P. Kumar, and R. Tekchandani, "A multimodal facial cues based engagement detection system in e-learning context using deep learning approach," *Multimedia Tools and Applications*, vol. 82, no. 18, pp. 28 589–28 615, 2023.

[270] K. Huttunen, H. Keränen, E. Väyrynen, R. Pääkkönen, and T. Leino, "Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights," *Applied ergonomics*, vol. 42, no. 2, pp. 348–357, 2011.

[271] A. Ramires, D. Cocharro, and M. E. P. Davies, "An audio-only method for advertisement detection in broadcast television content," *arXiv preprint arXiv:1811.02411*, 2018.

[272] T. O. Zander and C. Kothe, "Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general," *Journal of Neural Engineering*, vol. 8, no. 2, p. 025005, 2011.

[273] W. De Silva and A. Fernando, "Toward intelligent ad breaks: a survey and taxonomy of ai-driven ad placement in streaming media," *IEEE Access*, 2025.

[274] F. Paas, A. Renkl, and J. Sweller, "Cognitive load theory and instructional design: Recent developments," *Educational psychologist*, vol. 38, no. 1, pp. 1–4, 2003.

[275] A. Baddeley, "The episodic buffer: a new component of working memory?" *Trends in cognitive sciences*, vol. 4, no. 11, pp. 417–423, 2000.

[276] A. Berthold and A. Jameson, "Interpreting symptoms of cognitive load in speech input," in *UM99 user modeling: Proceedings of the seventh international conference.* Springer, 1999, pp. 235–244.

[277] J. E. Cutting, *Movies on our minds: The evolution of cinematic engagement.* Oxford University Press, 2021.

[278] ——, "Narrative theory and the dynamics of popular movies," *Psychonomic bulletin & review*, vol. 23, no. 6, pp. 1713–1743, 2016.

[279] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565.

[280] B. Park, J. L. Plass, and R. Brünken, "Cognitive and affective processes in multimedia learning," pp. 125–127, 2014.

[281] R. E. Mayer and R. Moreno, "Nine ways to reduce cognitive load in multimedia learning," *Educational psychologist*, vol. 38, no. 1, pp. 43–52, 2003.

[282] M. K. Pichora-Fuller, S. E. Kramer, M. A. Eckert, B. Edwards, B. W. Hornsby, L. E. Humes, U. Lemke, T. Lunner, M. Matthen, C. L. Mackersie *et al.*, "Hearing impairment and cognitive energy: The framework for understanding effortful listening (fuel)," *Ear and hearing*, vol. 37, pp. 5S–27S, 2016.

[283] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python." *SciPy*, vol. 2015, pp. 18–24, 2015.

[284] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[285] M. K. Pichora-Fuller and G. Singh, "Effects of age on auditory and cognitive processing: implications for hearing aid fitting and audiologic rehabilitation," *Trends in amplification*, vol. 10, no. 1, pp. 29–59, 2006.

[286] J. F. Walker and L. M. Archibald, "Articulation rate in preschool children: a 3-year longitudinal study," *International Journal of Language & Communication Disorders*, vol. 41, no. 5, pp. 541–565, 2006.

[287] J. Schilperoord, "On the cognitive status of pauses in discourse production," in *Contemporary tools and techniques for studying writing*. Springer, 2002, pp. 61–87.

[288] M. A. Khawaja, N. Ruiz, and F. Chen, "Potential speech features for cognitive load measurement," in *Proceedings of the 19th Australasian conference on computer-human interaction: Entertaining user interfaces*, 2007, pp. 57–60.

[289] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3-4, pp. 271–287, 2004.

[290] M. Barthel and S. Sauppe, "Speech planning at turn transitions in dialog is associated with increased processing load," *Cognitive Science*, vol. 43, no. 7, p. e12768, 2019.

[291] C. L. Mackersie and H. Cones, "Subjective and psychophysiological indexes of listening effort in a competing-talker task," *Journal of the American Academy of Audiology*, vol. 22, no. 02, pp. 113–122, 2011.

[292] J. P. Magliano and J. M. Zacks, "The impact of continuity editing in narrative film on event segmentation," *Cognitive science*, vol. 35, no. 8, pp. 1489–1517, 2011.

[293] T. J. Smith, "The attentional theory of cinematic continuity," *Projections*, vol. 6, no. 1, pp. 1–27, 2012.

[294] J. E. Cutting, J. E. DeLong, and K. L. Brunick, "Temporal fractals in movies and mind," *Cognitive Research: Principles and Implications*, vol. 3, no. 1, p. 8, 2018.

[295] J. E. Cutting, "The evolution of pace in popular movies," *Cognitive research: principles and implications*, vol. 1, no. 1, p. 30, 2016.

[296] X. J. Chai, N. Ofen, L. F. Jacobs, and J. D. Gabrieli, "Scene complexity: influence on perception, memory, and development in the medial temporal lobe," *Frontiers in human neuroscience*, vol. 4, p. 1021, 2010.

[297] S. H. Abdulhussain, A. R. Ramli, M. I. Saripan, B. M. Mahmmod, S. A. R. Al-Haddad, and W. A. Jassim, "Methods and challenges in shot boundary detection: a review," *Entropy*, vol. 20, no. 4, p. 214, 2018.

[298] L. C. Loschky, A. M. Larson, T. J. Smith, and J. P. Magliano, "The scene perception & event comprehension theory (spect) applied to visual narratives," *Topics in cognitive science*, vol. 12, no. 1, pp. 311–351, 2020.

[299] F. Nagle and N. Lavie, "Predicting human complexity perception of real-world scenes," *Royal Society Open Science*, vol. 7, no. 5, p. 191487, 05 2020. [Online]. Available: https://doi.org/10.1098/rsos.191487

[300] F. Bordes, R. Y. Pang, A. Ajay, A. C. Li, A. Bardes, S. Petryk, O. Mañas, Z. Lin, A. Mahmoud, B. Jayaraman *et al.*, "An introduction to vision-language modeling," *arXiv preprint arXiv:2405.17247*, 2024.

[301] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.