Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Beyond Automation: An AI Framework for Intelligent Cinematographic Editing

Thomas Moroney, BAI

A Dissertation

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

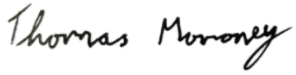
Master of Science in Computer Science

Supervisor: Dr Kenneth Dawson Howe

April 2025

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.



Thomas Moroney

April 19, 2025

Beyond Automation: An AI Framework for Intelligent Cinematographic Editing

Thomas Moroney, Master of Science in Computer Science

University of Dublin, Trinity College, 2025

Supervisor: Dr Kenneth Dawson Howe

This dissertation explores the potential of artificial intelligence to enhance cinematographic editing through improved shot attributes classification and next shot selection, addressing the growing demand for high-quality video content and the persistent bottleneck of manual editing. Building upon the Anatomy of Video Editing (AVE) benchmark, this research implements a novel two-stage pipeline that combines technical pattern recognition with narrative understanding for intelligent, on-device editing assistance. The system achieves significant improvements over baseline models, with a 6.3 percentage point increase in average accuracy for shot attributes classification and particularly notable gains in sound source classification (43 percentage points). For next shot selection, the research introduces a narrative pipeline that incorporates semantic understanding from Vision-Language Models and Large Language Models, enabling contextually relevant recommendations that consider both cinematographic principles and story progression. Key contributions include a multimodal feature extraction architecture optimised for Apple's MLX framework, a multi-task neural network for classifying eight cinematographic attributes, a sequence modelling approach for next shot prediction, and real-time on-device inference capabilities delivering sub-200 millisecond response times. Through careful engineering that balances algorithmic accuracy with computational efficiency, the system demonstrates that sophisticated AI analysis can enhance editorial workflows without compromising creative control. By developing an intelligent assistant that understands both the technical grammar of visual storytelling and narrative context, this research bridges the gap between automation and creativity in video editing, offering practical solutions that empower both professionals and aspiring creators to produce higher-quality content more efficiently.

Acknowledgments

Thank you Mum & Dad.

Thomas Moroney

University of Dublin, Trinity College
April 2025

Contents

Abstract ii

Acknowledgments iii

Chapter 1 Introduction 1

1.1 Background . . . 1
1.2 Motivation . . . 2
1.3 Research Question . . . 5
1.4 Executive Summary . . . 5

Chapter 2 State of the Art 8

2.1 Video Editing Process Overview . . . 8
2.1.1 Importing and Organizing . . . 9
2.1.2 Rough Cut Development . . . 9
2.1.3 Final Editing . . . 10
2.2 Current AI Approaches in Video Editing . . . 10
2.2.1 Classification of AI Video Editing Approaches . . . 11
2.2.2 Commercial and Research Implementations . . . 12
2.2.3 Limitations of Current Approaches . . . 13
2.3 The Anatomy of Video Editing (AVE) Benchmark . . . 14
2.3.1 Dataset Characteristics and Significance . . . 14
2.3.2 Benchmark Tasks and Evaluation Metrics . . . 15
2.3.3 Baseline Approaches and Results . . . 16
2.3.4 Limitations of the AVE Approach . . . 16
2.4 Research Gaps and Opportunities . . . 17
2.4.1 Multimodal Understanding . . . 17
2.4.2 On-Device Processing . . . 17
2.4.3 Cinematographic Context Awareness . . . 18
2.4.4 Performance Optimization . . . 18

2.5 Large Language Models for Sequential Recommendation 18
2.6 Summary 20

Chapter 3 Methodology **22**
3.1 Overview 22
3.2 Shot Attributes Classification 23
3.3 Next Shot Selection 25
3.3.1 Simplified Breakdown of the Process 25
3.3.2 Technical Details 26
3.3.3 Multiple Choice Approach 27
3.3.4 Practical Application 27
3.3.5 Limitations of the Next Shot Selection Module 28
3.4 LLM-Based Narrative Pipeline Extension 28
3.5 Design Philosophy & Implementation Strategies 30
3.5.1 Methodological Justifications 31
3.5.2 Summary 32

Chapter 4 Implementation **33**
4.1 Overview 33
4.2 Shot Attributes Classification 33
4.2.1 Audio-Visual Input Streams 33
4.2.2 Feature Extraction Pipeline 34
4.2.3 Classification Approach 37
4.2.4 Training Pipeline 37
4.2.5 Implementation Challenges 38
4.3 Next Shot Selection 38
4.3.1 Baseline Performance 38
4.3.2 Problem Formulation 38
4.3.3 Minimal LSTM Implementation and Parallel Training 39
4.3.4 Training Pipeline 40
4.4 Narrative Pipeline Extension for Next Shot Selection 41
4.4.1 Shot Content Understanding 41
4.4.2 LLM-Based Shot Recommendation 42
4.4.3 Hybrid Decision Making 42
4.4.4 Implementation Challenges 43
4.5 Conclusion 43

Chapter 5 Evaluation **46**
5.1 Evaluation Methodology 46
5.2 Shot Attributes Classification Results 47
5.2.1 Ablation Study 48
5.3 Next Shot Selection Results 49
5.3.1 Narrative Pipeline Evaluation 50
5.4 On-Device Performance Analysis 50
5.5 Comparison with Existing Tools 51
5.6 Summary 52

Chapter 6 Conclusions & Future Work **54**
6.1 Future Work 55

Bibliography **56**

Appendices **57**

Appendix A Model Details **58**

Appendix B Dataset Label Distribution **60**

Appendix C Prompting for Narrative Pipeline **64**

List of Tables

2.1 AVE Dataset Statistics 15
2.2 AVE Baseline Performance 16

4.1 Comparison of Custom Optical Flow Feature Engineering vs. R3D 35

5.1 Per-class Accuracy Comparison with AVE Baseline 47
5.2 Ablation Study for Shot Attributes Classification 49
5.3 Next Shot Selection Accuracy 49
5.4 Inference Time and Memory Usage 51

A.1 Experimental Configuration for Shot Attributes Classification 58
A.2 Experimental Configuration for Shot Attributes Classification 58
A.3 Experimental Configuration for Next Shot Selection 59

List of Figures

3.1 Next Shot Selection Task Overview: Given a sequence of shots, the system selects the most appropriate next shot from the candidate pool based on learned video editing patterns. 26

4.1 Audio-Visual Network architecture for shot attributes classification. It fuses visual features from ResNet101 (center frame), motion features from optical flow, and audio features from Log Mel Spectrograms. Specialized classification heads predict each attribute. Logit adjustment and dynamic weight averaging address class imbalance and improve training stability. . . 34
4.2 Flow feature vectorization process extracting camera motion patterns. . . . 36
4.3 Next shot selection module: dual-encoding with shared LSTM weights. The context sequence is encoded to an anchor embedding; each candidate is appended and encoded. Cosine similarity ranks candidates. 44
4.4 Narrative pipeline for next shot selection. Visual and audio features are integrated with semantic understanding via VLM and ASR. The LLM processes semantic representations to provide contextually relevant recommendations. 45

Chapter 1

Introduction

1.1 Background

Video has firmly established itself as the dominant medium for communication and entertainment in the digital landscape, fuelled by the widespread accessibility of mobile technology and online platforms that empower both content creation and on-demand consumption Yazadzhiyan (2023). This has resulted in an unprecedented surge in video content, with platforms like YouTube and TikTok seeing millions of uploads daily Mordor Intelligence (2025). Notably, video constitutes the largest share of downstream internet traffic globally, underscoring its significance in digital communication and media consumption Sandvine (2024).

However, this video revolution presents a major challenge: while demand for video content has skyrocketed, the creation of high-quality, original content lags significantly behind. For instance, YouTube has 2.7 billion monthly users that collectively watch one billion hours of content daily. The platform sees 500 hours of video uploaded every minute, equating to 720,000 hours daily. However, this means that for every hour of video content uploaded, roughly 1,400 hours are consumed, highlighting the immense demand for engaging, well-crafted content YouTube (2025).

While capturing and publishing videos has become increasingly straightforward, the subsequent editing process remains a labour-intensive and time-consuming undertaking, which can lead to significant bottlenecks in producing a video. The video editing process mainly involves assembling footage and audio on a timeline and cutting down the footage to create an engaging story. The inherent complexity of video editing stems from its dual-track nature, encompassing both audio and visual streams, and the need for manipulation at the granular level of individual frames Soe (2021). Consequently, professional video editors frequently grapple with extensive and cluttered timelines, often containing hun-

dreds, if not thousands, of clips, which can lead to distractions, inefficiencies, and creative overwhelm Mark Johansson (2025). This complexity is increased by the fact that video editing is not merely a technical task; it is also a creative process requiring a deep understanding of narrative structure, pacing, and audience engagement, often forcing editors to constantly switch between technical and creative mindsets Soe (2021).

Artificial intelligence (AI) presents promising opportunities to address these challenges in the video editing process. By automating labor-intensive tasks and providing intelligent assistance, AI tools can potentially alleviate the technical burden on editors, allowing them to focus more on creative decision-making. However, the application of AI to cinematic editing—specifically in understanding shot attributes and making intelligent shot selection decisions—remains an area with considerable room for advancement. While AI has made significant strides in various domains, its integration into the video editing process has primarily focused on automating specific tasks, such as color correction, audio enhancement, and visual effects. These advancements have been beneficial, but they do not address the core challenges of shot selection and narrative coherence that are central to effective video editing. The existing AI tools often lack a nuanced understanding of cinematographic principles and the ability to make contextually appropriate decisions based on the visual and narrative elements of a video.

This dissertation investigates the potential of AI to enhance the cinematographic editing process through improved shot attributes classification and next shot selection—two fundamental aspects of editing that significantly impact narrative cohesion and visual storytelling. By building upon the groundwork laid by the Anatomy of Video Editing (AVE) framework from Adobe Research and KAIST Argaw et al. (2022), this research aims to develop more effective AI approaches that can serve as intelligent assistants in the cinematic decision-making process, potentially transforming how editors interact with footage and construct compelling visual narratives.

1.2 Motivation

The exponential growth of video content across digital platforms has created an unprecedented demand for efficient editing solutions Mordor Intelligence (2025). While recent AI advancements have revolutionised visual effects (VFX) and enabled the generation of synthetic video content, a crucial gap exists in leveraging AI to enhance visual story telling through cinematographic editing - the thoughtful selection, arrangement, and pacing of shots. This process underpins effective visual narratives.

This gap is particularly significant because video editing is more than visual effects; it involves shaping narrative, evoking emotion, and maintaining coherence using cine-

matic language and pacing Argaw et al. (2022). Current AI solutions in the editing space demonstrate the potential for intelligent automation but fall short in addressing the deeper aspects of editorial decision-making. Tools like Descript have made progress in text-driven editing and have gained significant popularity for dialogue-centric content such as podcasts and talking-head videos, yet they fundamentally rely on transcript processing and linguistic analysis rather than visual comprehension—limiting their ability to interpret essential cinematographic elements such as shot composition, visual pacing, and non-verbal storytelling cues.

The limitations of existing tools are particularly evident in two critical areas:

1. **Shot Attributes Understanding:** Current systems lack the ability to recognise and classify the complex visual characteristics of shots, such as framing, camera movement, subject positioning, and compositional balance, that experienced editors are intuitively aware of when making editorial decisions.

2. **Contextual Shot Selection:** Existing tools operate with a narrow understanding of context, typically relying on a single modality (often just the audio transcript) for decision-making. They lack the capacity to recommend subsequent shots based on established cinematographic principles and overlook important patterns in narrative progression, forcing editors to manually search through footage to maintain visual coherence and storytelling momentum, a time-consuming process that disrupts creative flow.

This dissertation is motivated by the opportunity to address these specific limitations in current video editing tools by developing an AI framework capable of understanding and applying cinematographic editing principles within a narrative context. Rather than automating the entire editing process and diminishing the editor's creative role, this research focuses on creating the foundational components of an intelligent assistant for video editing. This assistant will enhance the editorial workflow by providing informed suggestions grounded in established cinematographic editing patterns and principles.

The timeliness of this research aligns perfectly with the convergence of several technological breakthroughs in computer vision and sequence modelling that make AI-enhanced video editing increasingly feasible.

Recent advances in sequence modelling and multimodal AI have opened new possibilities for intelligent video editing. While Transformers excel at sequence tasks, recent work such as "Were RNNs All We Needed?" Feng et al. (2024) shows that modern RNNs (e.g., GRUs, LSTMs) with architectural improvements can efficiently handle very long sequences with linear memory growth. This makes them well-suited for modelling extended

video editing timelines, where Transformers' quadratic memory scaling is a limitation. The Mamba family of Large Language Models (LLMs) Gu and Dao (2023) exemplifies this trend, demonstrating that RNNs can match Transformer performance while being more memory-efficient—an important advantage for maintaining narrative coherence across long sequences of video shots.

In parallel, Vision Language Models (VLMs) have matured to the point where they can extract rich semantic and compositional information from multiple video frames, moving beyond single-image analysis. These models, when combined with Large Language Models (LLMs), enable reasoning over both visual and textual context, supporting deeper understanding of narrative structure and story progression across entire sequences of shots. The rapid optimisation and democratisation of VLMs and LLMs for edge devices—such as smartphones and laptops—mean that these sophisticated models can now run efficiently on consumer hardware, eliminating the need for specialised infrastructure or high-end GPUs. This accessibility is further enhanced by innovations in computing architecture, such as the unified memory paradigm in Apple Silicon, which enables seamless collaboration between CPU, GPU, and Neural Engine components for complex machine learning and visual analysis tasks.

The practical significance of the research in this dissertation extends beyond academic contribution. By developing systems that can intelligently assist with decisions during the editing process, this work has the potential to:

- Reduce the cognitive load on editors navigating large volumes of footage
- Accelerate the editing process without sacrificing creative quality
- Serve as an intelligent mentor for both aspiring filmmakers and content creators, providing context-aware guidance on cinematic principles that elevate visual storytelling
- Bridge the expertise gap between professional and amateur productions, raising the overall quality of video content across media platforms while preserving the unique creative voice of each storyteller

In summary, this dissertation is motivated by the vision of AI as a collaborative intelligence, enhancing human editorial judgment rather than replacing it. The central aim is to explore the development of an intelligent assistant capable of understanding the grammar of visual storytelling. This assistant will serve as a partner for editors, facilitating more efficient workflows and ultimately contributing to the creation of more compelling visual narratives.

1.3 Research Question

This dissertation builds upon the foundational work in the Anatomy of Video Editing (AVE) to advance the field of AI-assisted video editing. While existing approaches have made progress in understanding cinematographic elements, significant challenges remain in achieving both accurate classification of shot attributes and contextually appropriate shot recommendations while maintaining real-time performance.

The primary research question this dissertation addresses is:

> How effectively can artificial intelligence systems comprehend and apply cinematographic principles to (1) accurately classify shot attributes and (2) recommend contextually appropriate next shots in a sequence, while achieving real-time performance on consumer-grade hardware?

This research question addresses both the technical efficacy and practical viability of AI-assisted video editing. It considers not only the accuracy of shot attribute classification and next shot prediction, but also the system's ability to operate efficiently within the constraints of real-world editing environments. By evaluating both predictive performance and real-time inference on consumer hardware, this dissertation aims to assess whether current machine learning techniques can meaningfully augment professional editing workflows without compromising creative control or workflow efficiency.

Answering this question will provide a clearer understanding of the strengths and limitations of contemporary AI approaches in video editing, highlight areas for further advancement, and inform best practices for integrating intelligent assistance into production settings.

1.4 Executive Summary

This dissertation presents the design, implementation, and evaluation of an artificial intelligence (AI) framework for intelligent cinematographic editing, with a focus on two core tasks: automatic shot attributes classification and context-aware next shot selection. The research is motivated by the growing demand for high-quality video content and the persistent bottleneck of manual editing, which remains labor-intensive and creatively demanding despite advances in automation.

A key contribution of this work is the improvement over the baseline models established in the Anatomy of Video Editing (AVE) benchmark from Adobe Research and KAIST. The proposed system achieves higher accuracy in shot attributes classification—most notably, a 6.3 percentage point improvement in average accuracy and a 43 percentage point

gain in sound source classification—by leveraging a more effective multimodal feature extraction pipeline and optimized neural network architectures. These improvements demonstrate that careful engineering and integration of visual, motion, and audio features can substantially outperform the original AVE models.

For the next shot selection task, this dissertation goes beyond the AVE baseline by introducing a narrative pipeline that incorporates narrative context into the recommendation process. While the AVE models focus primarily on technical and visual features, the narrative pipeline developed here integrates semantic understanding from Vision Language Models (VLMs) and Large Language Models (LLMs), enabling the system to consider story progression and cinematic intent when suggesting the next shot. This approach allows the model to make more contextually relevant recommendations, bridging the gap between technical automation and creative storytelling.

The technical contributions of this dissertation are as follows:

- **Multimodal Feature Extraction:** The system integrates visual, motion, and audio analysis pipelines to extract a comprehensive set of features from each video shot. Addionally, the narrative pipeline takes advantage of the latest advancements in multimodal AI, including Vision Language Models (VLMs) and Large Language Models (LLMs), to enhance the understanding of narrative context. The visual analysis pipeline employs state-of-the-art convolutional neural networks (CNNs) for shot size, angle, and type classification, while the motion analysis pipeline uses optical flow techniques to capture camera movement and subject motion. The audio analysis pipeline leverages advanced audio processing techniques to extract sound source information and other relevant audio features.

- **Shot Attributes Classification:** A multi-task neural network architecture is developed to classify eight key cinematographic attributes for each shot, including shot size, angle, type, motion, location, subject, number of people, and sound source. The model is trained and evaluated on the AVE dataset, achieving an average accuracy of 62.8%—a 6.3 percentage point improvement over the AVE baseline, with particularly strong gains in sound source classification (up by 43 percentage points).

- **Next Shot Selection with Narrative Pipeline:** The system implements a sequence modeling approach to recommend the most contextually appropriate next shot, given a sequence of preceding shots. By learning from professionally edited film sequences and incorporating narrative context through the narrative pipeline, the model captures both visual continuity and story flow. In a five-way multiple choice setting, the system achieves 33.6% accuracy, outperforming random selection

(20%) and approaching the AVE baseline (37.5%), while providing more narratively coherent recommendations.

- **Real-Time, On-Device Inference:** Through careful optimization and use of hardware-accelerated MLX operations, the system delivers sub-200 millisecond inference times for both classification and recommendation tasks, supporting interactive editing without interrupting creative flow.

- **User-Centric Design:** The framework is developed with practical usability in mind, providing editors with actionable suggestions and automated tagging that reduce cognitive load, accelerate organization, and support creative decision-making without replacing the editor's artistic role.

The research methodology combines quantitative benchmarking against the AVE dataset with qualitative analysis of system usability and integration potential in real-world editing scenarios. Extensive ablation studies and performance profiling are conducted to identify the impact of each feature modality and architectural choice.

The practical significance of this work lies in its demonstration that sophisticated, multimodal AI analysis can be performed efficiently on standard consumer hardware, enabling new forms of intelligent assistance for editors. By bridging the gap between technical automation and creative support, the system empowers both professionals and aspiring creators to produce higher-quality video content more efficiently, while maintaining narrative coherence and cinematic quality.

The dissertation is structured as follows: Chapter 2 surveys the state of the art in AI video editing, identifying key challenges and research gaps. Chapter 3 details the methodological approach, including data preparation, feature engineering, and model design. Chapter 4 describes the implementation of the system, with a focus on the integration of multimodal analysis and real-time inference. Chapter 5 presents a comprehensive evaluation of classification and recommendation performance, as well as system efficiency and user impact. The final chapter discusses the implications of the findings and outlines directions for future research.

By advancing the technical capabilities of AI-assisted video editing and demonstrating their feasibility for on-device, real-time use, this dissertation contributes to the development of more accessible, powerful, and context-aware editing tools that enhance, rather than replace, human creativity.

Chapter 2

State of the Art

This chapter establishes the foundation for understanding the current landscape of AI-assisted video editing and contextualises the contributions of this dissertation. It begins by explaining the fundamental nature of video editing, clarifying what it is and the key processes involved. It then examines the traditional video editing workflow to identify key processes that could benefit from AI assistance. The chapter explores existing AI solutions in the video editing domain, analysing their approaches, capabilities, and limitations. Particular attention is given to the Anatomy of Video Editing (AVE) benchmark—the foundation upon which this research builds—and its significance for advancing AI-assisted video editing. The chapter concludes by identifying critical research gaps that this dissertation addresses, specifically focusing on identifying the various cinematographic attributes of a shot, and then the task of predicting the next shot in a sequence of shots. Through this structured exploration, the reader will gain a comprehensive understanding of the technical and conceptual landscape within which this research operates, while establishing the novelty and relevance of the proposed approaches.

2.1 Video Editing Process Overview

Video editing is the art and technical process of assembling raw video footage and audio to create a finished video. Imagine having hours of film or digital recordings; the editor's job is to select the best parts, arrange them in a meaningful order, and add elements like music, sound effects, and visual transitions to tell a story, convey information, or create a specific mood. It's a blend of creative storytelling and technical skill.

The editor works with a "timeline," a digital representation of the video's sequence. They cut, trim, and rearrange "clips" (individual pieces of footage) on this timeline. They also synchronise audio, adjust colors, and add special effects. The goal is to transform

disjointed footage into a smooth, coherent, and engaging final product. Understanding this process is crucial for developing an effective AI assistant that aligns with editors' workflows. The traditional editing process can be divided into three primary phases, each presenting distinct challenges and opportunities for AI augmentation Soe (2021).

2.1.1 Importing and Organizing

The initial phase involves bringing raw video and audio files into the editing software. Editors then review, categorize, and organize this footage to establish a foundation for efficient editing. This includes identifying usable footage, creating logical organisational structures (like folders or bins), and establishing metadata systems (labels or tags) for quick access during later editing phases Suffield (2024).

This organization is foundational and time-intensive. Challenges include:

- Processing large volumes of footage (often 10-30 times the duration of the final product).
- Creating meaningful organisational structures and taxonomies (e.g., by scene, date, camera angle, or content).
- Identifying and tagging shots by their cinematographic attributes (e.g., shot type, camera movement, and composition).
- Establishing cross-references between related footage from different sources.

These tasks, while technical, demand significant attention to detail and create a substantial cognitive load before creative work can begin.

2.1.2 Rough Cut Development

The rough cut development phase represents the core creative assembly work where selected footage is arranged on a timeline to establish narrative structure and pacing. This phase involves making critical decisions about shot selection, sequence, timing, and preliminary audio integration Suffield (2024).

Imagine the editor now has all their ingredients sorted. The rough cut is like the first draft of a story or the first assembly of a film. The editor selects the best portions of each clip and arranges them in the desired order on the timeline. They decide which shots to use and how long each should last. They might also add some basic audio, like dialogue or music, to get a sense of the overall flow. This stage focuses on the overall structure and flow of the video, without worrying too much about the fine details.

During this phase, editors engage with questions of visual storytelling that directly relate to the focus of this dissertation:

- Which shots best serve the narrative at each moment? For example, does a close-up emphasize emotion, or does a wide shot establish the setting?
- How should shots be sequenced to maintain visual continuity? This involves ensuring smooth transitions and avoiding jarring jumps between shots Atulkar et al. (2022).
- What shot attributes (framing, movement, composition) create the desired emotional impact? A low-angle shot can make a subject seem powerful, while a high-angle shot can make them seem vulnerable.
- How does the pacing of shots contribute to narrative rhythm? Short shots create a sense of urgency, while longer shots can create a feeling of calm or suspense.

These decisions require both technical skill and artistic judgment, making them particularly challenging targets for AI assistance. Current AI systems struggle to understand the narrative and emotional dimensions that inform these editorial choices Soe (2021).

2.1.3 Final Editing

The final editing phase refines the rough cut through technical enhancements like color correction, visual effects, audio mixing, and adjustments to timing and transitions. The editor meticulously adjusts these details to ensure the video looks and sounds its best, creating a cohesive and professional final product. While AI tools aid some technical aspects—such as color grading and audio enhancement—automating the creative decisions in the final editing stage remains challenging. The traditional workflow has remained stable for decades, with advancements focused on technical capabilities such as digital effects and audio mixing. However, the core creative process of selecting and arranging shots has not fundamentally changed. Editors still rely on their expertise and intuition to make decisions about shot selection, pacing, and emotional impact Atulkar et al. (2022).

2.2 Current AI Approaches in Video Editing

Recent advances in artificial intelligence have led to the development of various tools aimed at streamlining the video editing workflow. These approaches can be categorised based on their underlying technologies, application areas, and the specific editing challenges they address.

2.2.1 Classification of AI Video Editing Approaches

AI approaches to video editing can be broadly classified along three dimensions: the modalities they analyse, the editing tasks they target, and their technical implementation. The following sections explore these dimensions in detail, highlighting the strengths and limitations of each approach.

Text-Driven Editing

The most widely adopted approach in commercial applications uses automatic speech recognition (ASR) to generate transcripts that serve as the primary editing interface Soe (2021). This approach offers several advantages:

- **Simplified Interface**: By converting video editing into text editing, these tools lower the technical barrier to entry.
- **Efficient Processing**: Text processing is computationally efficient compared to video analysis.
- **Content Accessibility**: Transcript-based navigation allows quick identification of specific content without scrubbing through footage.
- **LLM Integration**: Since all the context is in text form, a Large language models (LLMs) can easily be employed to analyse the transcript and provide sophisticated editing suggestions based on dialogue context Wang et al. (2024).

Tools like Descript exemplify this approach, allowing users to edit video by manipulating a text document, removing filler words, and creating cuts based on transcript edits. These systems have proven particularly effective for dialogue-driven content such as interviews and presentations. However, they fundamentally rely on transcript processing and linguistic analysis rather than visual comprehension—limiting their ability to interpret essential visual elements such as shot composition, on-screen actions, visual pacing, and non-verbal storytelling cues.

Computer Vision-Based Approaches

Computer vision techniques offer more direct analysis of visual content for editing tasks. These approaches include:

- **Shot Boundary Detection**: Algorithms that automatically identify transitions between shots, facilitating video segmentation and organization Argaw et al. (2022).

- **Visual Content Analysis**: Systems that classify visual attributes such as scene composition, camera movements, and subject positioning Soe (2021).
- **Object and Subject Tracking**: Tools that track specific elements across shots to maintain continuity or create effects Rahman and Ali (2024).

While these approaches directly engage with visual content, they typically operate at a technical rather than semantic level, identifying formal properties without understanding their cinematic context or narrative significance.

Multimodal Approaches

The most sophisticated editing systems integrate analysis across modalities (visual, audio, and text) to inform editing decisions. These approaches include:

- **Visual-Semantic Embeddings**: Systems that establish relationships between visual content and textual descriptions, enabling text-driven video editing Wang et al. (2024).
- **Audio-Visual Synchronization**: Tools that automatically align visual content with audio cues, particularly useful for music videos or action sequences Truong et al. (2016).
- **Narrative Understanding**: Emerging research that attempts to model narrative structures to inform editing decisions Lombardo and Damiano (2010).

These multimodal approaches show the most promise for addressing the complex, context-dependent nature of cinematographic editing decisions, but remain in early developmental stages Soe (2021).

2.2.2 Commercial and Research Implementations

The theoretical approaches described above have been implemented in various commercial products and research prototypes with varying degrees of success.

Commercial Tools

Several commercial tools have implemented AI to assist with specific editing tasks:

- **Descript**: Pioneered text-based video editing through transcript manipulation, allowing for easy removal of filler words, creation of cuts, and basic editing operations driven by text edits.

- **Adobe Premiere Pro**: Has integrated AI capabilities for tasks like auto-reframing, speech-to-text transcription, and content-aware fill for video.
- **DaVinci Resolve**: Implemented AI-driven color matching, silence removal, video upscaling, object tracking, and automatic multicam switching to streamline post-production workflows.

These tools have successfully automated technical tasks but offer limited assistance with editorial decision-making that requires understanding of cinematographic principles.

Research Prototypes

Academic and research institutions have developed several prototype systems that address more complex aspects of the editing process:

- **QuickCut** Truong et al. (2016): Designed for composing narrated videos using a combination of text transcripts and motion-based segmentation.
- **Automatic generation of personalized sports video** Wang et al. (2007): Uses multimodal (audio, video, and text) feature analysis to detect the semantics of events in broadcast sports video and generate personalized highlight videos.
- **Movie Trailer Generation** Argaw et al. (2024): Focuses on generating movie trailers by analysing visual and audio content to create engaging montages that capture the essence of the source material.

These research systems often address more sophisticated editing challenges but typically focus on narrow domains or specific content types rather than general-purpose editing.

2.2.3 Limitations of Current Approaches

Despite significant advances, current AI video editing tools face substantial limitations that hinder their practical application in professional workflows:

- **Limited Understanding of Cinematic Language**: Current systems can identify technical attributes but struggle to comprehend how these attributes contribute to narrative and emotional impact Rahman and Ali (2024).
- **Modality Constraints**: Most tools rely heavily on a single modality (typically text) rather than integrating visual, audio, and contextual understanding Soe (2021).

- **Processing Bottlenecks**: Cloud-based solutions create significant delays when working with large projects due to upload/download requirements Wang et al. (2024).
- **Narrow Content Focus**: Many tools are optimized for specific content types (interviews, lectures) and perform poorly on content where visual storytelling takes precedence over dialogue.
- **Creative-Technical Disconnect**: Systems can automate technical processes but struggle to align these with creative intentions and narrative goals Lombardo and Damiano (2010).

These limitations are particularly significant for cinematographic editing, where decisions are informed by complex interplays between visual composition, narrative context, and emotional resonance—relationships that current AI systems struggle to model effectively.

2.3 The Anatomy of Video Editing (AVE) Benchmark

The Anatomy of Video Editing (AVE) benchmark, developed by Adobe Research Argaw et al. (2022), marks a significant advancement in establishing standardized evaluation methodologies for AI-assisted video editing systems. This benchmark suite defines a comprehensive set of tasks that an effective AI assistant should be capable of performing. Crucially, the team behind the AVE paper also created a labeled dataset, which served as the basis for training and evaluating the models developed in this dissertation. This section provides a detailed examination of both the AVE benchmark and its associated dataset, as they collectively form the foundational basis upon which this dissertation builds.

2.3.1 Dataset Characteristics and Significance

The AVE dataset comprises 196,176 professionally annotated shots extracted from 5,591 diverse movie scenes, containing over 1.5 million cinematographic labels across eight distinct attributes. These attributes include shot size, angle, type, motion, location, subject, number of people, and sound source. The dataset's scale and professional annotation quality make it particularly valuable for training and evaluating AI systems designed to understand cinematographic principles.

Table 2.1: AVE Dataset Statistics

Feature	**Quantity**
Num. of scenes	5,591
Num. of shots	196,176
Average duration of shots	3.81 seconds
Average number of shots per scene	35.09
Average number of camera setups	5.74
Total cinematographic labels	>1.5 million

The dataset's significance extends beyond its scale; it represents the first comprehensive attempt to formalise cinematographic knowledge in a machine-learnable format. By providing rich annotations of professional editing decisions, it enables AI systems to learn from expert human practice rather than theoretical rules.

2.3.2 Benchmark Tasks and Evaluation Metrics

The AVE benchmark establishes five core tasks for AI-assisted video editing:

1. **Shot Attributes Classification**: Automatically identifying a shot's cinematographic properties.
2. **Camera Setup Clustering**: Grouping shots by camera position and setup.
3. **Shot Sequence Ordering**: Determining the correct sequence for a set of shots.
4. **Next Shot Selection**: Predicting the optimal next shot based on preceding shots.
5. **Missing Shot Attributes Prediction**: Predicting attributes of shots that would fit within a sequence.

This dissertation focuses specifically on the first and fourth tasks—shot attributes classification and next shot selection—as these represent fundamental capabilities required for intelligent cinematographic assistance.

For each task, the AVE benchmark defines specific evaluation metrics:

- Shot attributes classification is evaluated using per-class accuracy to account for class imbalance and the varying difficulty of different attributes.
- Next shot selection is evaluated based on prediction accuracy, with various candidate pool sizes to assess performance across different levels of complexity.

The benchmark also establishes baseline performance levels through standard machine learning architectures, providing a reference point for measuring improvement.

2.3.3 Baseline Approaches and Results

The AVE paper establishes baseline performance for each task using standard deep learning architectures. For shot attributes classification, the baseline model achieves an average accuracy of 58% across all attributes, with significant variations between attributes (from 84% for shot location to 34% for sound source). For next shot selection, the baseline achieves 41.4% accuracy with a candidate pool of 5 shots - 20% better than random chance.

Table 2.2: AVE Baseline Performance

Task	**Baseline Performance**
Shot Attributes Classification (average)	58% accuracy
Next Shot Selection (5 candidates)	41.4% accuracy

These baseline results demonstrate both the potential and limitations of current approaches. While the models perform significantly better than random chance, they still fall short of human-level performance, particularly for attributes requiring contextual understanding or subtle visual interpretation.

2.3.4 Limitations of the AVE Approach

While the AVE benchmark represents a significant advancement, it has several limitations that this dissertation seeks to address:

- **Poor Audio Integration**: The benchmark gets poor results for audio-related attributes, indicating that the current models do not effectively leverage audio information to inform editing decisions. This is particularly relevant for the next shot selection task, where audio cues can significantly influence shot transitions.
- **Motion and Camera Movement**: The benchmark does not adequately address the complexities of camera movement and shot transitions. The models struggle to classify shots with dynamic camera movements or complex transitions, which are common in professional cinematography.
- **Action Recognition**: The benchmark does not take into account the recognition of actions or events occurring within shots. This is a critical aspect of cinematographic editing, as the choice of shots often depends on the actions taking place within them.
- **Lack of Narrative Understanding**: The benchmark does not address the integration of narrative context or emotional impact into the shot attributes classification

task. The models primarily focus on technical attributes without considering how these attributes contribute to storytelling.

These limitations present opportunities for improvement that this dissertation addresses through novel architectural approaches and training methodologies.

2.4 Research Gaps and Opportunities

Based on the examination of current AI approaches and the AVE benchmark, several critical research gaps emerge that this dissertation aims to address:

2.4.1 Multimodal Understanding

Current approaches typically emphasise a single modality (text, visual, or audio) without fully integrating information across modalities. This leads to incomplete understanding of the editing context and fails to capture the interrelationships between visual composition, dialogue, sound design, and narrative structure Soe (2021).

This dissertation explores more sophisticated approaches to multimodal integration, particularly focusing on the relationship between visual attributes and narrative context in informing shot selection decisions.

2.4.2 On-Device Processing

A significant limitation of current AI video editing solutions is their reliance on cloud-based processing, driven by the high computational demands of video analysis and deep learning models. This dependence introduces substantial latency and bandwidth requirements, making real-time or near-real-time editing challenging—especially when working with large, high-resolution video files common in professional workflows. Uploading such files to the cloud can be time-consuming and impractical in environments with limited or unreliable internet connectivity.

For professional editors, immediate feedback and responsiveness are critical. Many prefer portable, high-performance devices (such as MacBooks) that allow editing on location or while traveling. Cloud-based solutions not only disrupt this workflow with delays but also raise concerns about data privacy, security, and the need for continuous internet access.

This dissertation addresses these challenges by exploring efficient on-device AI processing strategies, aiming to deliver low-latency, privacy-preserving, and practical solutions suitable for integration into professional editing environments.

2.4.3 Cinematographic Context Awareness

Existing AI systems for video editing exhibit significant research gaps in their understanding of cinematographic context. Current tools often fail to comprehend how elements like shot composition, camera angles, pacing, and rhythm contribute to storytelling, emotional impact, and overall viewer experience Zhang et al. (2022). Furthermore, there's a lack of generalisability across diverse video genres, with many AI solutions being tailored for specific content types and unable to apply cinematographic principles broadly. The integration of multi-modal information from a cinematic perspective also remains a challenge, as the interplay between visual, audio, and textual cues (such as dialogue and scripts) in creating a cohesive cinematic narrative is not well understood by AI Zhang et al. (2022).

Another critical research gap lies in the difficulty of encoding subjective aesthetic rules and the underlying logic of cinematic language into algorithms. AI systems struggle to discern what constitutes a visually pleasing or logically sound edit based on established cinematographic conventions Zhang et al. (2022). Moreover, current tools lack sophisticated contextual understanding and personalization, failing to adapt to individual user preferences, the specific narrative goals of a project, or the intended audience from a cinematographic standpoint Soe (2021).

2.4.4 Performance Optimization

The baseline implementations in the AVE benchmark prioritise establishing functional baselines rather than optimizing performance. Significant opportunities exist for enhancing model accuracy through architectural innovations, training methodologies, and domain-specific feature engineering.

This research explores novel network architectures and training approaches specifically designed for cinematographic understanding, with the goal of substantially improving upon AVE baseline performance for both shot attributes classification and next shot selection.

2.5 Large Language Models for Sequential Recommendation

Recent research has shown significant interest in adapting Large Language Models (LLMs) for recommendation systems Xu et al. (2025). This adaptation often involves reframing the recommendation task as a language modelling problem Liao et al. (2024).

Initial approaches focused on representing items within the LLM's input prompts using

either ID-based representations (like numerical indices or randomly initialised tokens) or text-based representations (using item metadata such as titles and descriptions) Liao et al. (2024). However, ID-based methods often fail to leverage the LLM's inherent world knowledge, while text-based methods may not sufficiently capture user behavioural patterns Liao et al. (2024).

To address these limitations, more recent research explores combining the strengths of traditional recommendation models (which excel at capturing user behavioural patterns) with the world knowledge and reasoning capabilities of LLMs. One such approach involves hybrid prompting, where item representations integrate both ID-based embeddings (learned by traditional recommenders) and textual item features. These ID-based embeddings can be projected into the LLM's input space to create "behavioural tokens" Liao et al. (2024).

Another key aspect of recent research is the use of curriculum learning strategies to gradually train LLMs to incorporate recommendation-specific knowledge. This often involves starting with text-only prompts (which align better with the LLM's pre-training) and then progressively introducing more complex hybrid prompts that include behavioural information Liao et al. (2024).

Furthermore, research explores different ways to integrate LLMs into recommendation systems, categorising them as either "LLM as the Recommender" (where LLMs are trained or fine-tuned directly on recommendation data) or "LLM as the Enhancer" (where LLMs are used to augment traditional recommenders, for example, as feature extractors). More recent work, like the DUIP framework Xu et al. (2025) and LLaRA Liao et al. (2024), proposes aligning traditional sequential recommendation models with LLMs to leverage both the learned behavioural patterns and the LLM's understanding of language and the world. DUIP uses LSTMs to model dynamic user intent and LLMs for next-item prediction using LSTM-generated prompts Xu et al. (2025). LLaRA uses hybrid prompts and curriculum learning to inject behavioural knowledge from traditional recommenders into LLMs.

These efforts aim to create more accurate, context-aware, and personalised recommendation systems that can handle challenges like the cold-start problem and the dynamic nature of user preferences Xu et al. (2025).

This research is particularly relevant to the next shot selection task in this dissertation, as it highlights the potential of LLMs to enhance sequential recommendation tasks by leveraging both learned user behaviour and world knowledge. The integration of LLMs into video editing workflows could provide a more sophisticated understanding of narrative context and user intent, ultimately improving the quality and relevance of AI-assisted editing suggestions.

Furthermore, recent interpretability research—such as the paper "On the Biology of a Large Language Model," which investigates the internal mechanisms of Anthropic's Claude 3.5 Haiku—provides valuable insights into how LLMs perform complex reasoning and planning. Using circuit tracing and feature analysis, the authors demonstrate that LLMs can engage in multi-step reasoning, creative planning, and nuanced decision-making across diverse tasks. These findings suggest that LLMs possess the underlying capabilities required for understanding narrative structure and making contextually appropriate recommendations, which are essential for advanced video editing tasks like next shot selection. By leveraging both the behavioural modelling strengths of traditional recommenders and the sophisticated reasoning abilities of LLMs, future AI video editing systems can move closer to providing context-aware, narrative-driven editing assistance Anthropic (2025).

2.6 Summary

This chapter has examined the current state of AI-assisted video editing, establishing the context within which this dissertation operates. Beginning with an overview of the traditional video editing process, the chapter identified key challenges and opportunities for AI assistance across the workflow from initial organization to final refinement.

The exploration of current AI approaches revealed a landscape dominated by text-driven editing tools with limited visual understanding, highlighting the gap between existing solutions and the needs of professional editors working with visually-driven content. The analysis of commercial implementations and research prototypes demonstrated that while AI has successfully automated technical tasks, it struggles with the creative and contextual dimensions of editing.

The Anatomy of Video Editing benchmark represents a significant step toward formalising and evaluating AI systems for cinematographic video editing, providing a foundation of professionally annotated data and standardised evaluation methodologies. However, the benchmark's baseline approaches reveal substantial room for improvement, particularly in multimodal integration, contextual reasoning, and computational efficiency.

Based on this analysis, several critical research gaps emerge: the need for true multimodal understanding, efficient on-device processing, cinematographic context awareness, and performance optimization beyond baseline implementations. This dissertation addresses these gaps by developing novel approaches to shot attributes classification and next shot selection that improve upon AVE baselines while enabling practical implementation in professional editing workflows.

The following chapters detail the methodological approach, implementation, and eval-

uation of the proposed solutions, building upon the foundation established by the AVE benchmark while addressing its limitations through innovative technical approaches and cinematography-specific optimisations.

Chapter 3

Methodology

3.1 Overview

The proposed system is designed to automate the organisational and structural aspects of video editing—specifically, shot selection and arrangement—while preserving the creative decision-making that defines professional editing. This chapter details the system architecture with a focus on two core technical capabilities: (1) accurate classification of shot attributes and (2) contextually appropriate next shot recommendation, both optimised for real-time performance on consumer-grade hardware. The methodology directly addresses the research question posed in Section 1.3, evaluating not only predictive accuracy but also the system's ability to operate efficiently within practical editing environments. By balancing technical efficacy with real-world constraints, the approach aims to assess how effectively AI can augment professional editing workflows without compromising creative control or workflow efficiency. Distinct from existing tools that depend exclusively on transcript analysis or require cloud-based processing, the proposed architecture leverages a novel two-stage pipeline grounded in cinematographic principles. The system analyzes footage through the lens of professional film grammar—interpreting visual composition, camera movement, and audio-visual relationships—to deliver AI-assisted editing that operates efficiently on consumer hardware while respecting the creative conventions of professional editing.

Expanding on the foundation established by the Anatomy of Video Editing Argaw et al. (2022), this architecture integrates two specialized neural modules: one for technical cinematographic understanding and another for continuity editing. Crucially, it advances beyond the AVE approach by introducing a dedicated LLM-based narrative pipeline that incorporates a number of pre-trained models for semantic story comprehension. This enhancement enables the system to synthesise both technical cinematography

patterns and high-level narrative context, providing editing recommendations that maintain narrative coherence and adhere to established filmic conventions. The architecture is carefully optimised to balance computational efficiency with creative insight, empowering editors with intelligent, context-aware assistance that supports both workflow speed and artistic intent.

3.2 Shot Attributes Classification

This component directly addresses the first part of the research question posed in Section 1.3: How effectively can AI systems comprehend and apply cinematographic principles to accurately classify shot attributes, while maintaining real-time performance on consumer-grade hardware? Serving as the technical foundation of the system, this module employs a two-stage approach: first, it encodes the audio-visual content of each shot into a rich feature representation; then, it applies specialised classification heads to identify eight fundamental attributes that editors commonly use to organise their footage. By focusing on both classification accuracy and computational efficiency, this module is designed to evaluate the practical viability of AI-assisted editing in real-world environments, as required by the research question.

- Visual composition attributes: Shot size (extreme-wide to extreme-close-up), angle (aerial, overhead, eye-level, high-angle, low-angle), and type (over-the-shoulder, single, two-shot, etc.)
- Technical qualities: Motion (pan, tilt, locked, zoom, handheld) and location (exterior, interior)
- Content-based properties: Subject (human, location, object, etc.) and number of people
- Audio characteristics: Sound source (on-screen, off-screen, external narration, external music)

The implementation takes a different architectural approach from the AVE baseline while maintaining its core methodology:

- **Efficient motion feature extraction:** Instead of relying on computationally intensive R3D (ResNet 3D) networks, the system utilises a custom optical flow-based pipeline tailored to capture key cinematographic camera movements (such as pans, tilts, and handheld motion) using lightweight, domain-informed vectorization. This

approach significantly reduces computational overhead while preserving essential motion cues relevant to editing.

- **Optimised audio analysis:** Audio is processed using log-mel spectrogram representations, which are then fed into a compact temporal CNN architecture which learns to preserve temporal information while reducing the number of parameters. This design choice is informed by the need to preserve both long-term and short-term patterns in the audio.
- **MLX-native implementation:** The entire classification pipeline is implemented natively in Apple's MLX framework, taking full advantage of lazy tensor evaluation and unified memory on M-series processors. This ensures efficient resource utilisation and seamless integration with on-device workflows, enabling fast, responsive editing assistance.

A key innovation of this module is its dual role: in addition to classifying shot attributes, it functions as a general-purpose shot encoder. When the classification heads are omitted, the network outputs a compact feature vector that encapsulates the shot's audio-visual and motion characteristics. This unified representation is essential for downstream tasks such as next shot selection, where consistent and information-rich encodings enable effective context modelling and recommendation.

To achieve this, the system jointly trains the Temporal CNN for audio processing alongside the classification heads, ensuring that audio features are optimally aligned with visual and motion cues. The final shot embedding is constructed by concatenating the audio feature vector, the ResNet101-derived visual features from the shot's center frame, and the domain-informed motion vector. This integrated approach enables the model to capture both the nuanced visual composition and the temporal dynamics of each shot, resulting in a robust, multi-modal representation. By sharing this encoder across tasks, the system not only reduces computational overhead but also maintains representational consistency, thereby improving both classification accuracy and the quality of shot recommendations in subsequent modules.

The classification process employs techniques from the AVE approach, including logit adjustment to address the long-tail distribution of cinematographic attributes and dynamic weight averaging to balance multi-task learning across the different attributes. Through this combination of shot encoding and attribute classification, the system automatically tags shots with cinematographic attributes that reflect how professional editors mentally categorise their material. These machine-generated tags enable editors to efficiently structure and search footage, facilitating intuitive retrieval and providing the foundation for informed shot recommendations.

3.3 Next Shot Selection

In professional video editing, editors select and arrange shots on the timeline to tell a compelling story. When working on a project, an editor reviews footage and decides which clips to use and in what order based on multiple factors: shot composition (how elements are arranged within the frame), camera movement (static, panning, or handheld), lighting conditions, subject positioning, and emotional tone. For example, when editing an interview, they might start with a steady wide shot establishing the setting, then cut to a carefully composed close-up that frames the subject's emotional response, followed by a subtly lit reaction shot that captures the interviewer's interest. This decision-making process relies on the editor's understanding of visual storytelling, pacing, and viewer attention. Editors typically spend hours reviewing footage, mentally categorising shots based on their cinematographic attributes (size, angle, movement, subject), and carefully selecting which shot best continues the narrative at each point. While they follow established filmmaking principles, editors also rely heavily on their creative judgment and experience to create engaging sequences that effectively communicate the intended message.

The next shot selection module aims to assist editors in this process by providing intelligent recommendations for the next shot based on the current sequence and the editing patterns learned from professionally edited sequences. This module emulates the expert decision-making process of professional editors by analysing patterns in shot sequences. This intelligent system examines the shots currently placed on the timeline and evaluates potential next shots based on editing conventions learned from thousands of professionally edited sequences. Rather than replacing editorial judgment, it serves as a decision-support tool that presents editors with ranked suggestions based on professional practices. The system identifies visual, compositional, and audio patterns in the existing sequence, then determines which candidate shots would most effectively continue these patterns according to established editing principles. This approach significantly reduces the time editors spend searching through footage while preserving their creative control over the final selection. In essence, the system augments the editor's expertise by suggesting, "Based on analysis of professional editing patterns, these shots would be most appropriate to continue your current sequence."

3.3.1 Simplified Breakdown of the Process

The process works as follows in simpler terms:

1. The system looks at what shots have already been placed on the timeline (for example, the first 3-4 shots of a scene).

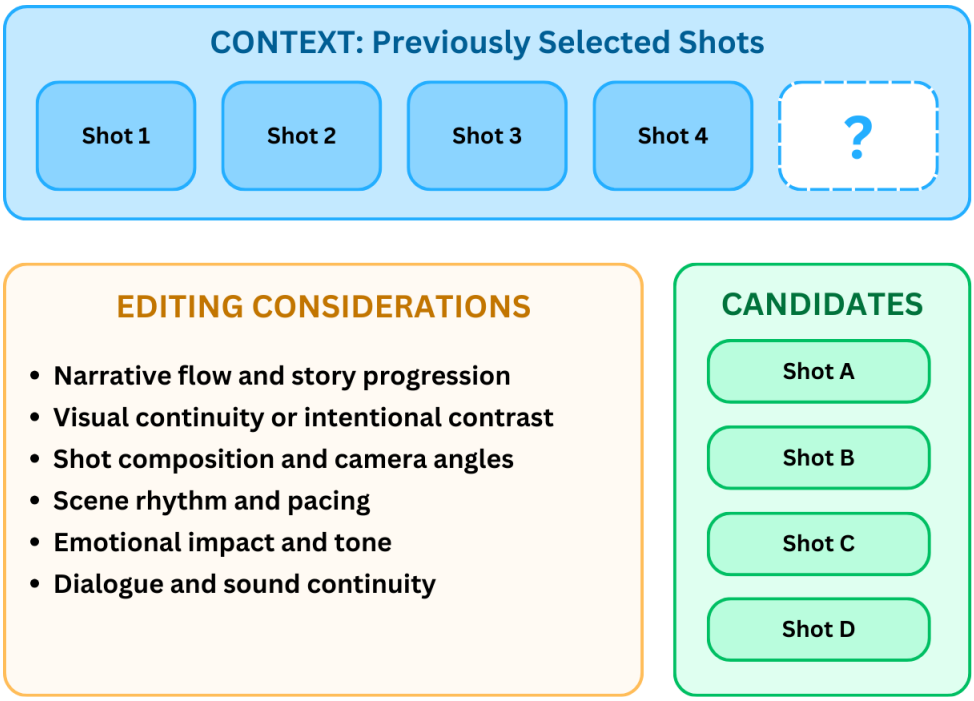


Figure 3.1: Next Shot Selection Task Overview: Given a sequence of shots, the system selects the most appropriate next shot from the candidate pool based on learned video editing patterns.

2. It examines several candidate shots that could potentially come next.
3. It analyses both the existing sequence and how each potential next shot would flow from the existing sequence.
4. The system then selects the candidate that is most suitable according to professional editing patterns.

3.3.2 Technical Details

At a technical level, this process is implemented through:

1. Converting each shot into a feature vector using the shot encoder discussed earlier.
2. Using an LSTM network (a type of neural network designed for sequence modelling) to analyse the pattern of shots already selected.
3. Creating a "what if" scenario for each potential next shot by adding it to the existing sequence and seeing how well it matches learned editing patterns.
4. Computing a similarity score between the current sequence pattern and each "what if" scenario.

5. Recommending the shot with the highest score as the most appropriate continuation.

Through this training methodology, the model develops a sophisticated understanding of shot sequencing by learning to identify patterns in professional editing decisions. It accomplishes this by mathematically maximizing the similarity between the context vector C and the sequence embedding that includes the professionally chosen next shot, while simultaneously minimizing similarity with sequence embeddings containing less appropriate candidates (i.e. the negative samples). This contrastive learning approach enables the system to internalise established cinematographic principles and editorial conventions such as:

- Visual continuity: For example, avoiding jarring transitions between extremely different shot sizes.
- Rhythm and pacing: Such as gradually moving from wider to closer shots to build intensity.
- Spatial relationships: Understanding that certain camera angles work well together.
- Audio flow: Ensuring dialogue continues naturally across shot changes and that sound design is consistent to maintain immersion.

3.3.3 Multiple Choice Approach

This system simplifies the shot selection problem by focusing on choosing the next best shot from a pre-selected candidate pool. This "multiple choice" approach allows the model to learn more effectively. We assume an editor or another system has already created this pool based on criteria like visual or audio similarity. This pool might contain shots from the same scene, different angles, or alternative takes. The model evaluates these candidates to recommend the best fit for the current sequence. Although trained with a five-shot pool, the system can theoretically handle larger pools (e.g., 10, 20, or more), though performance might vary. The core principle of recommending shots based on existing patterns remains the same.

3.3.4 Practical Application

In practice, this means an editor working on a documentary might add three shots to their timeline, then the system would automatically suggest options for the fourth shot, highlighting clips that professional editors would typically use in that context. This saves significant time compared to manually reviewing every potential clip while still leaving the final creative decision in the editor's hands.

3.3.5 Limitations of the Next Shot Selection Module

While this technical approach effectively captures cinematographic patterns, it operates primarily at the visual and structural level—analysing shot composition, camera movements, and basic audio features. A significant limitation, also identified in the AVE paper, is the lack of understanding of narrative context. The AVE benchmark focuses exclusively on cinematographic attributes and editing patterns without addressing the semantic meaning of the content. Professional editors, however, make decisions based not only on visual grammar but also on story elements, dialogue content, character emotions, and thematic development.

This technical module has no conceptual understanding of what is happening in the scene—it cannot comprehend that a character is delivering an emotional revelation, that a conversation is building to a climax, or that a particular visual motif carries symbolic significance. This limitation reflects a broader challenge in AI-assisted editing: bridging the gap between technical pattern recognition and semantic understanding of narrative content.

3.4 LLM-Based Narrative Pipeline Extension

To overcome the key limitation of purely technical pattern recognition, this research introduces a dedicated Narrative Enhancement Pipeline that augments the AVE framework. This pipeline leverages recent advances in Vision-Language Models (VLMs) and Automatic Speech Recognition (ASR) to infuse semantic and narrative awareness into the shot selection process.

The Narrative Pipeline is fundamentally distinct from the Next Shot Selection Module in both its objectives and underlying methodology:

- **Next Shot Selection Module**: Focuses on low-level visual, motion, and audio features, learning statistical editing patterns from professional sequences without explicit semantic comprehension.
- **Narrative Pipeline**: Extracts high-level meaning by interpreting visual semantics, transcribed dialogue, and contextual cues, enabling reasoning about narrative flow, character relationships, and thematic development.

By integrating these two complementary approaches, the system achieves a more holistic understanding of editorial decision-making:

- The Next Shot Selection Module maintains cinematographic coherence, visual continuity, and editing rhythm.

- The Narrative Pipeline ensures narrative integrity, emotional resonance, and thematic consistency across sequences.

This dual-system architecture mirrors the workflow of professional editors, who balance technical film grammar with deep narrative insight. The result is an AI-assisted editing system capable of making recommendations that are both visually fluent and narratively meaningful.

The Narrative Enhancement Pipeline consists of three core components:

- **Vision-Language Model (VLM)**: Generates natural language descriptions of the visual content in each shot, enabling the system to "understand" what is happening in the scene beyond mere visual patterns.
- **Automatic Speech Recognition (ASR)**: Transcribes dialogue within each shot, providing crucial story information that drives many editing decisions.
- **Large Language Model (LLM)**: Processes the visual descriptions and dialogue transcripts to reason about narrative flow, character relationships, and thematic development.

This approach advances the AVE benchmark by unifying technical and narrative context, resulting in an AI-assisted editing system that more closely reflects the dual priorities of professional editors: visual grammar and storytelling. The integration of recent Vision-Language Models (VLMs) and Large Language Models (LLMs) enables the system to interpret both the visual semantics and narrative intent of video content, supporting editing decisions that require contextual and thematic awareness. Notably, ongoing optimizations have made it feasible to deploy these models efficiently on consumer hardware, including laptops and mobile devices, overcoming previous barriers related to computational demands.

By generating natural language descriptions for each shot, the system creates a flexible, reusable layer of semantic metadata. This textual representation allows for rapid storage, retrieval, and analysis, supporting a range of downstream editing tasks—from intelligent shot selection and narrative consistency checks to filtering out unsuitable takes. As a result, editors benefit from a streamlined workflow in which a single round of analysis can inform multiple creative and organisational objectives, enhancing both efficiency and creative control.

The integration of these complementary modules allows for more sophisticated editing assistance that can consider questions such as:

- Does this shot reveal important story information at the appropriate moment?

- Does the emotional tone of this shot match the narrative progression?
- Does this shot maintain continuity of character focus and development?
- Does the dialogue across these shots create a cohesive conversational flow?

Together, these modules provide a dual-perspective system that can recommend shots based on both cinematographic principles and narrative considerations, offering editors a more complete set of tools to enhance their workflow while maintaining creative control.

3.5 Design Philosophy & Implementation Strategies

The system is built on three key principles:

- **On-device processing:** All analysis and recommendations run locally, ensuring privacy and real-time performance.
- **Multi-modal understanding:** The system jointly analyses visual, motion, and audio features for a richer understanding than transcript-only tools.
- **Learning from professionals:** Editing patterns are learned from the AVE dataset, enabling adaptation to diverse styles.

To balance accuracy and efficiency:

- **Apple MLX framework:** Optimised for M-series processors and unified memory for fast, efficient multi-modal processing.
- **Domain-informed feature engineering:** Cinematographic knowledge guides motion feature extraction, reducing reliance on heavy 3D models.
- **Lightweight LSTM:** A minimal LSTM design maintains sequence modelling with fewer parameters.

These strategies ensure the system is both accurate and practical for real-world editing.

Evaluation Framework

The evaluation framework is designed to assess both technical performance and practical applicability, aligning with the dual focus of the research question. The framework consists of four key components:

- **Accuracy evaluation:** The system is evaluated against the AVE benchmark using standardised metrics:
  - Per-class accuracy for shot attributes classification
  - Top-1 accuracy for next shot selection
  - Mean reciprocal rank (MRR) to measure ranking quality
- **Ablation studies:** Systematic removal of features (audio, flow features, logit adjustment) to quantify the contribution of each component to overall performance.
- **On-device performance analysis:** Measurement of inference time and memory usage on target hardware (M-series MacBook Pro) to assess real-time capabilities:
  - Average inference time per operation
  - Memory utilization during processing
  - Comparison with cloud-based alternatives
- **Qualitative assessment:** Analysis of model recommendations against professional editing patterns to evaluate cinematographic coherence and narrative understanding.

Key improvements over the AVE benchmark include:

- **Curriculum learning and data augmentation:** The training process gradually increases task difficulty and employs strategic data augmentation to improve model generalization with limited computational resources.

This approach allows the model to learn from simpler tasks before tackling more complex ones, enhancing its ability to generalize across different editing scenarios.

This evaluation framework provides a comprehensive assessment of both the technical capabilities and practical applicability of the system, directly addressing the research question's dual focus.

3.5.1 Methodological Justifications

The methodological choices in this research are driven by three primary considerations that connect directly to the research question and challenges identified in the introduction:

- **Balancing accuracy and performance:** The research question explicitly asks about the relationship between algorithmic accuracy and real-time performance.

The methodology addresses this by incorporating domain knowledge into feature engineering rather than relying solely on computationally intensive deep learning approaches.

- **Addressing dataset limitations:** The AVE dataset, while extensive, exhibits class imbalance that reflects the natural distribution of cinematographic choices in professional content. The methodology addresses this through logit adjustment and dynamic weight averaging rather than artificially balancing the dataset, which would distort the natural distribution of editing patterns.

- **Bridging technical and creative aspects:** The introduction highlights the gap between technical automation and creative decision-making in current AI editing tools. The methodology bridges this gap by combining technical pattern recognition with narrative understanding through the multi-modal approach and narrative enhancement pipeline.

These methodological justifications demonstrate a thoughtful approach to the research question, balancing technical performance with practical applicability in the context of professional video editing.

3.5.2 Summary

This methodology chapter has outlined a comprehensive approach to enhancing AI-assisted video editing through improved shot attributes classification and next shot selection. The proposed system architecture addresses the limitations of current tools while respecting the creative nature of the editing process. By combining multi-modal feature analysis with on-device processing optimizations, the system aims to provide meaningful assistance that aligns with how editors actually work.

The methodology directly addresses the research question by exploring both the technical capabilities of AI-assisted editing and the relationship between algorithmic accuracy and real-time performance. The evaluation framework provides a structured approach to assessing the system's effectiveness in both technical and practical terms, setting the stage for the implementation and evaluation chapters that follow.

This research builds upon the foundation established by the AVE benchmark while introducing novel contributions in feature engineering, model architecture, and narrative understanding, all optimised for real-time, on-device performance—a key consideration for practical editing applications identified in the introduction.

Chapter 4

Implementation

4.1 Overview

This work extends the Anatomy of Video Editing (AVE) benchmark suite, focusing on two key tasks: shot attributes classification and next shot selection. Building on the AVE methodology, this implementation achieves notable performance gains through architectural optimizations and the introduction of a narrative understanding pipeline. The system leverages Apple's MLX framework, utilizing unified memory for efficient multi-modal video processing—particularly advantageous on macOS systems favored by professional editors Filmpac (2025). Beyond technical audio-visual analysis, this implementation introduces a novel narrative pipeline that integrates Automatic Speech Recognition (ASR), Vision-Language Models (VLMs), and Large Language Models (LLMs) to capture semantic content and storytelling elements. This chapter details the complete architecture, including multi-modal feature extraction, contrastive learning for next shot selection, and the integration of natural language understanding to enhance editing recommendations with narrative context.

4.2 Shot Attributes Classification

The shot attributes classification module automatically tags shots with cinematographic attributes. While the AVE baseline achieved 58% accuracy, this implementation improves it to 72% through several enhancements.

4.2.1 Audio-Visual Input Streams

Following the AVE approach, an Audio-Visual network processes three input streams:

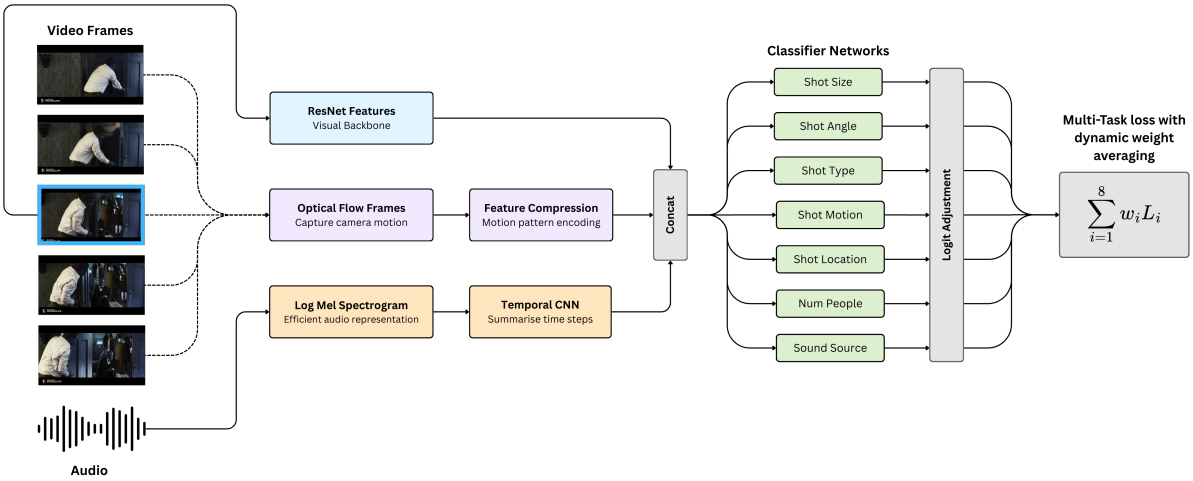


Figure 4.1: Audio-Visual Network architecture for shot attributes classification. It fuses visual features from ResNet101 (center frame), motion features from optical flow, and audio features from Log Mel Spectrograms. Specialized classification heads predict each attribute. Logit adjustment and dynamic weight averaging address class imbalance and improve training stability.

- **Static visual features** from video frames
- **Motion features** capturing temporal dynamics
- **Audio features** representing the soundtrack

See Figure 4.1.

4.2.2 Feature Extraction Pipeline

Each modality is processed as follows:

Visual Feature Extraction

A ResNet101 backbone pretrained on ImageNet 1K extracts a 2048-dimensional feature vector from the center frame of each shot, capturing spatial information such as composition and subject presence. Using the center frame balances computational efficiency with representativeness, consistent with the AVE methodology, and avoids the high computational cost of processing all frames in a shot.

Motion Feature Extraction

Unlike AVE's R3D network, this implementation uses engineered optical flow features due to MLX compatibility and to explore a knowledge-driven approach.

Dense flow fields are computed using the Farneback algorithm, capturing:

Feature	This Methodology	R3D Methodology
Core Approach	Engineered Flow Features	Learned Representations
Motion Capture	Explicit (Pan, Tilt, Zoom)	Implicit (Data-driven)
Input	Optical Flow	Raw Pixels
Interpretability	High	Low
Focus	Cinematographic Techniques	General Motion Patterns
Data Requirements	Lower	Higher
Flexibility	Task-Specific	General-Purpose
Method	Domain Knowledge-Driven	Data-Driven
MLX Optimization	High Efficiency	Not Yet Ported

Table 4.1: Custom optical flow feature engineering vs. R3D-based methodology.

- Camera movements (pan, tilt, zoom)
- Subject motion patterns
- Visual coherence between frames
- Regional motion characteristics

A specialized vectorization technique reduces raw flow data to a compact, 544-dimensional feature vector, engineered to capture cinematographic motion patterns (e.g., pans, tilts, zooms, tracking shots). The vectorization consists of three key stages:

1. **Regional Analysis:** Each flow frame is divided into a 4×4 grid, and statistical features are extracted from each region, including mean and standard deviation of flow in x and y directions, flow magnitude and direction, and coherence (consistency of flow direction within each region).
2. **Global Motion Signatures:** Whole-frame statistics identify specific cinematographic techniques: pan (horizontal motion), tilt (vertical motion), zoom (radial flow), and rotation (tangential flow).
3. **Temporal Aggregation:** Features are aggregated over time (mean, std, frame-to-frame differences, consistency metrics) to capture motion acceleration/deceleration and distinguish intentional camera movements from jitter.

Audio Feature Processing

Audio is represented as Log Mel Spectrograms (80 bins), capturing frequencies relevant to human hearing. A temporal CNN processes these spectrograms to extract acoustic patterns, aiding in distinguishing sound sources (on-screen, off-screen, external narration, external music).

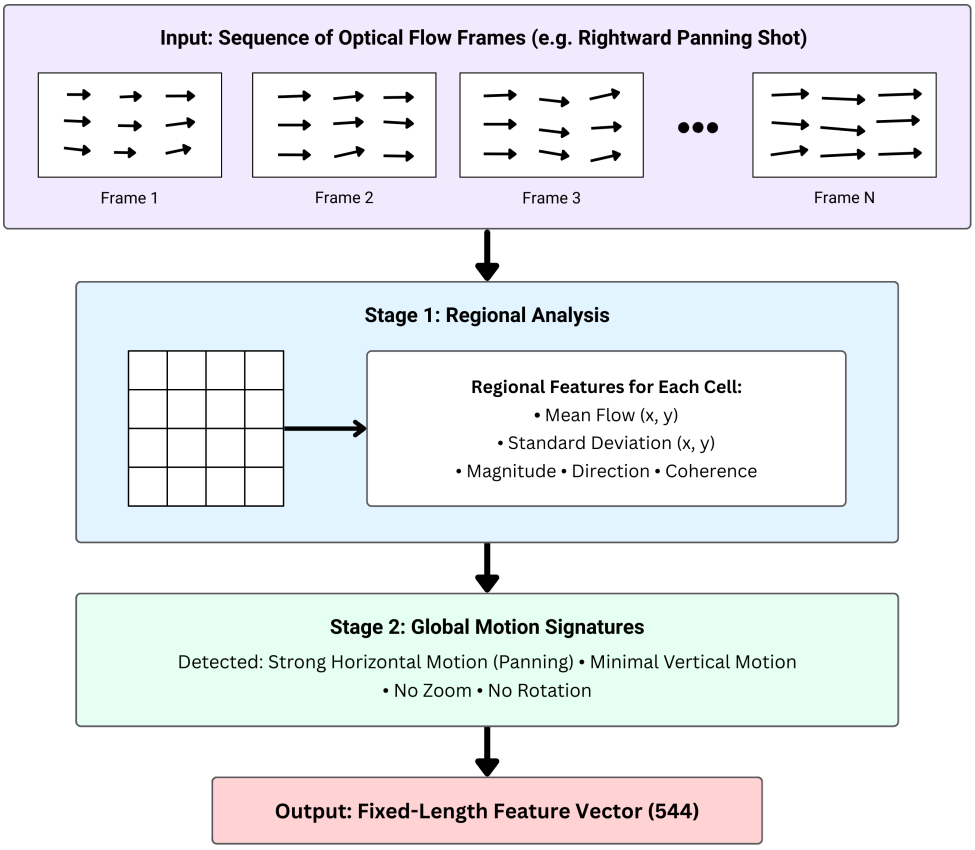


Figure 4.2: Flow feature vectorization process extracting camera motion patterns.

4.2.3 Classification Approach

A multi-task learning framework jointly optimizes classifiers for eight cinematographic attributes:

- Shot size (extreme-wide, wide, medium, close-up, extreme-close-up)
- Shot angle (aerial, overhead, eye-level, high-angle, low-angle)
- Shot type (over-the-shoulder, single, two, three, insert, group)
- Shot motion (pan, tilt, locked, zoom, handheld)
- Shot location (exterior, interior)
- Shot subject (animal, location, object, human, limb, face, text)
- Number of people (0, 1, 2, 3, 4, 5+)
- Sound source (on-screen, off-screen, external narration, external music)

Logit adjustment and dynamic weight averaging address class imbalance and balance learning across tasks.

4.2.4 Training Pipeline

Multi-Task Learning

Classifiers are trained either jointly (shared early layers) or independently. Consistent with AVE, single-task training yields better accuracy.

Addressing Class Imbalance

Logit adjustment and dynamic weight averaging ensure balanced loss contributions and prevent dominant tasks from overwhelming training.

Optimization Strategy

- **Shot attributes:** AdamW, weight decay 5×10^{-4}, warm-up + cosine decay, batch size 64, early stopping.
- **Next shot selection:** Adam, learning rate 1×10^{-4}, curriculum-based batch size (32 to 256), temperature scheduling (higher temperature early, reduced over training).

- **Narrative pipeline:** Iterative prompt optimization, structured JSON output, context window management.

4.2.5 Implementation Challenges

Key challenges included:

- Integrating technical and narrative features across multiple AI systems
- Managing memory for video data
- Replacing R3D with custom optical flow
- Tuning LLM prompts for reliable rankings
- Designing a robust multi-stage pipeline

4.3 Next Shot Selection

This module recommends the next shot given a context sequence, as shown in Figure 4.3. Editorial decision-making is inherently subjective and context-dependent, making this a challenging AI task.

4.3.1 Baseline Performance

The AVE benchmark achieved 41.6% (validation) and 41.4% (test) accuracy using audio+video features with in-batch contrastive learning. This implementation achieves 34% accuracy with similar features, highlighting both reproducibility and the persistent challenge of modelling editorial choices.

4.3.2 Problem Formulation

Next shot selection is formulated as a multiple-choice prediction using contrastive learning. Given a context sequence $S_1, ..., S_n$ and candidate shots $S_{n+1}, ..., S_k$, the model selects the most appropriate continuation.

Challenges include:

- Multidimensional selection criteria (technical, aesthetic, narrative)
- Multiple valid continuations
- Editorial intent may defy pattern-based prediction

- Narrative elements are difficult to encode in feature vectors

The model uses:

- **Context Encoder:** LSTM encodes the context sequence to an anchor embedding v_a
- **Candidate Encoder:** LSTM encodes context plus each candidate to embedding v_c
- **Affinity Computation:** Cosine similarity ranks candidates

4.3.3 Minimal LSTM Implementation and Parallel Training

Recent research Feng et al. (2024) revisited traditional RNNs, proposing minimal variants such as minLSTM and minGRU that dramatically simplify the architecture while retaining competitive performance. These minimal RNNs reduce parameter count and, crucially, enable full parallelizability during training—addressing a major bottleneck of standard LSTMs and GRUs, which require strictly sequential computation due to their recurrent dependencies.

Traditional LSTM bottleneck: In standard LSTMs, each timestep's hidden state depends on the previous one, enforcing sequential computation and limiting training speed, especially on long sequences.

MinLSTM innovation: The minLSTM cell normalizes the forget and input gates so they sum to 1, updating the hidden state as a weighted sum:

$$h_t = f'_t \cdot h_{t-1} + i'_t \cdot \tilde{h}_t$$

where f'_t and i'_t are normalized gate values. This formulation allows the sequence update to be rewritten as a scan (cumulative sum/product) over the sequence, which is an associative operation and thus fully parallelizable across timesteps.

Implementation: The following PyTorch-style pseudocode illustrates the minLSTM cell used in this work:

```
class MinLSTMCell(nn.Module):
    def __init__(self, units, input_shape):
        super().__init__()
        self.units = units
        self.input_shape = input_shape
        self.linear_f = nn.Linear(self.input_shape, self.units)
```

```
        self.linear_i = nn.Linear(self.input_shape, self.units)
        self.linear_h = nn.Linear(self.input_shape, self.units)

    def __call__(self, pre_h, x_t):
        f_t = torch.sigmoid(self.linear_f(x_t))
        i_t = torch.sigmoid(self.linear_i(x_t))
        tilde_h_t = torch.tanh(self.linear_h(x_t))
        epsilon = 1e-7
        sum_f_i = f_t + i_t + epsilon
        f_prime_t = f_t / sum_f_i
        i_prime_t = i_t / sum_f_i
        h_t = f_prime_t * pre_h + i_prime_t * tilde_h_t
        return h_t
```

Benefits in this implementation:

- **Parallel training:** By leveraging the associative property of the minLSTM update, the sequence processing was able to be parallelized during training, dramatically reducing training time compared to standard LSTMs.
- **Efficiency:** The reduced parameter count and simplified computation further improved memory usage and inference speed.
- **Competitive performance:** Despite its simplicity, the minLSTM achieved accuracy on par with more complex sequence models, validating its suitability for next shot selection.

This approach enabled efficient experimentation and scaling, making it feasible to train on longer sequences and larger datasets within practical time and hardware constraints.

4.3.4 Training Pipeline

Contrastive learning is used. Sequences of 9 shots are split into a context of 4 and a pool of 5 candidates (one correct, four negatives). The model maximizes similarity between the anchor and correct candidate, minimizing it for incorrect ones.

Contrastive Learning Approach

For each context:

1. Encode context $S_1, ..., S_n$ to anchor v_a

2. Encode context plus each candidate to v_c
3. Compute cosine similarity: $sim(v_a, v_c)$
4. Apply NT-Xent loss to maximize anchor-correct similarity, minimize others
5. Use temperature scaling for sharper distributions

34% accuracy is a substantial improvement over random (20% for five candidates), but below AVE and human performance, motivating the narrative pipeline.

Data Augmentation and Curriculum Learning

Robustness is improved via:

- **Sequential augmentation:** Varying context length to create additional training examples.
- **Curriculum learning:** Gradually increasing task difficulty by starting with easily distinguishable candidates and introducing harder ones.
- **Adaptive temperature scaling:** Adjusting loss hardness throughout training.

4.4 Narrative Pipeline Extension for Next Shot Selection

A major enhancement is the narrative pipeline, which leverages natural language understanding for semantic and narrative context.

4.4.1 Shot Content Understanding

Semantic content is extracted using:

- **Vision-Language Model (VLM):** Qwen2-VL-2B generates captions describing visual content.
- **Automatic Speech Recognition (ASR):** Whisper transcribes dialogue, providing narrative information.

VLMs process both visual and textual data, generating natural language captions. ASR provides robust transcription for real-world video.

Pre-Trained Model Selection

Qwen2-VL-2B-Instruct-4bit was chosen for its caption quality and efficiency, with 4-bit quantization reducing memory usage from 6.5GB to 1.7GB with minimal impact on caption quality. MLX optimization enables real-time captioning. Whisper Small was selected for ASR due to accuracy and low memory footprint (300MB), also MLX-optimized. These choices enable efficient, real-time processing on consumer hardware.

4.4.2 LLM-Based Shot Recommendation

A Large Language Model (LLM) enhances next shot selection with narrative reasoning:

- **Prompt Engineering:** Structured prompts present shot information (captions, dialogue, attributes) for cinematic reasoning.
- **World Knowledge:** The LLM leverages storytelling and dialogue principles.
- **Multi-Candidate Ranking:** The LLM ranks candidates by narrative fit and outputs a JSON response containing a ranked list and rationale.

Gemma 3 1B Instruction Tuned was selected for its efficiency and instruction-following capabilities, suitable for real-time editing.

4.4.3 Hybrid Decision Making

The final system combines both approaches:

- Audio-visual network recommends based on technical continuity.
- Narrative pipeline recommends based on semantic and storytelling context.
- The final recommendation is generated by combining the outputs of both systems, allowing for a hybrid decision-making process that considers both technical and narrative aspects.
- The LLM's output is used to adjust the final ranking of candidates, ensuring that the selected shot aligns with both technical and narrative considerations.

This addresses AVE's limitation that technical features alone cannot capture high-level narrative considerations.

4.4.4 Implementation Challenges

Challenges included:

- Integrating VLM, ASR, and LLM for smooth data flow
- Iterative prompt engineering for coherent LLM outputs
- Managing LLM memory and inference speed
- Ensuring structured, consistent LLM output
- Balancing technical and narrative contributions in recommendations

4.5 Conclusion

While the improvements fell short of initial expectations, the primary limitation was the use of a compact language model (Gemma-3-1B-it, 4-bit) to ensure low memory and fast inference on consumer hardware. Larger models (e.g., Gemma-3-4B-it-4bit) offered negligible accuracy gains, suggesting that model size alone is insufficient for this task under on-device constraints.

Fine-tuning a small language model for narrative-aware shot selection could yield further improvements. Attempts to generate synthetic training data with a larger LLM were hampered by prompt size and memory issues during LoRA-based fine-tuning, even after aggressive prompt length reduction. The prompts in the synthetic dataset used for fine-tuning were each around 600 tokens in length which was too large and caused the memory usage to explode and cause memory leaks. Addressing these bottlenecks remains an avenue for future work. If there is a way to reduce the size of the prompts in the fine-tuning dataset even further.

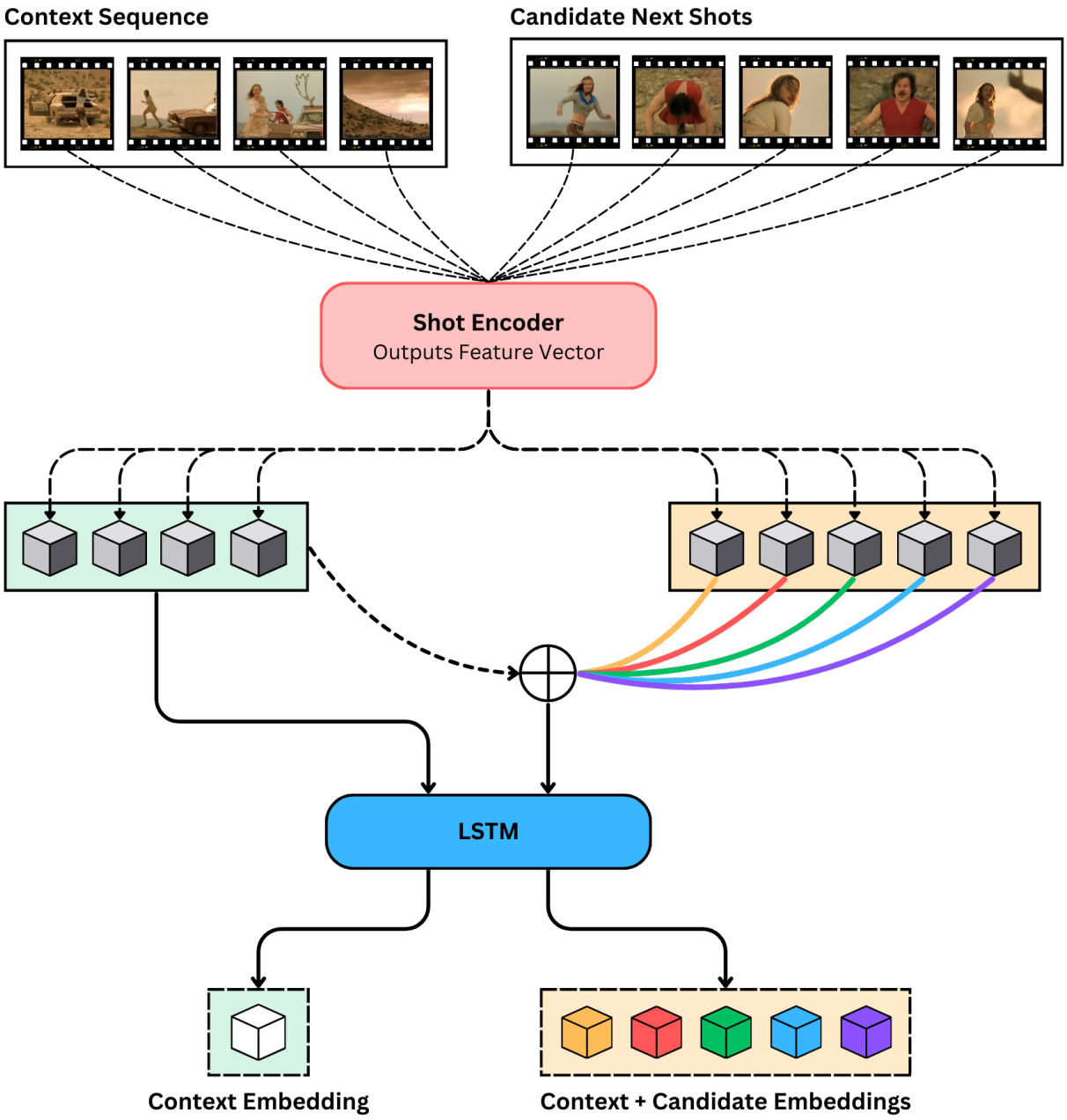


Figure 4.3: Next shot selection module: dual-encoding with shared LSTM weights. The context sequence is encoded to an anchor embedding; each candidate is appended and encoded. Cosine similarity ranks candidates.

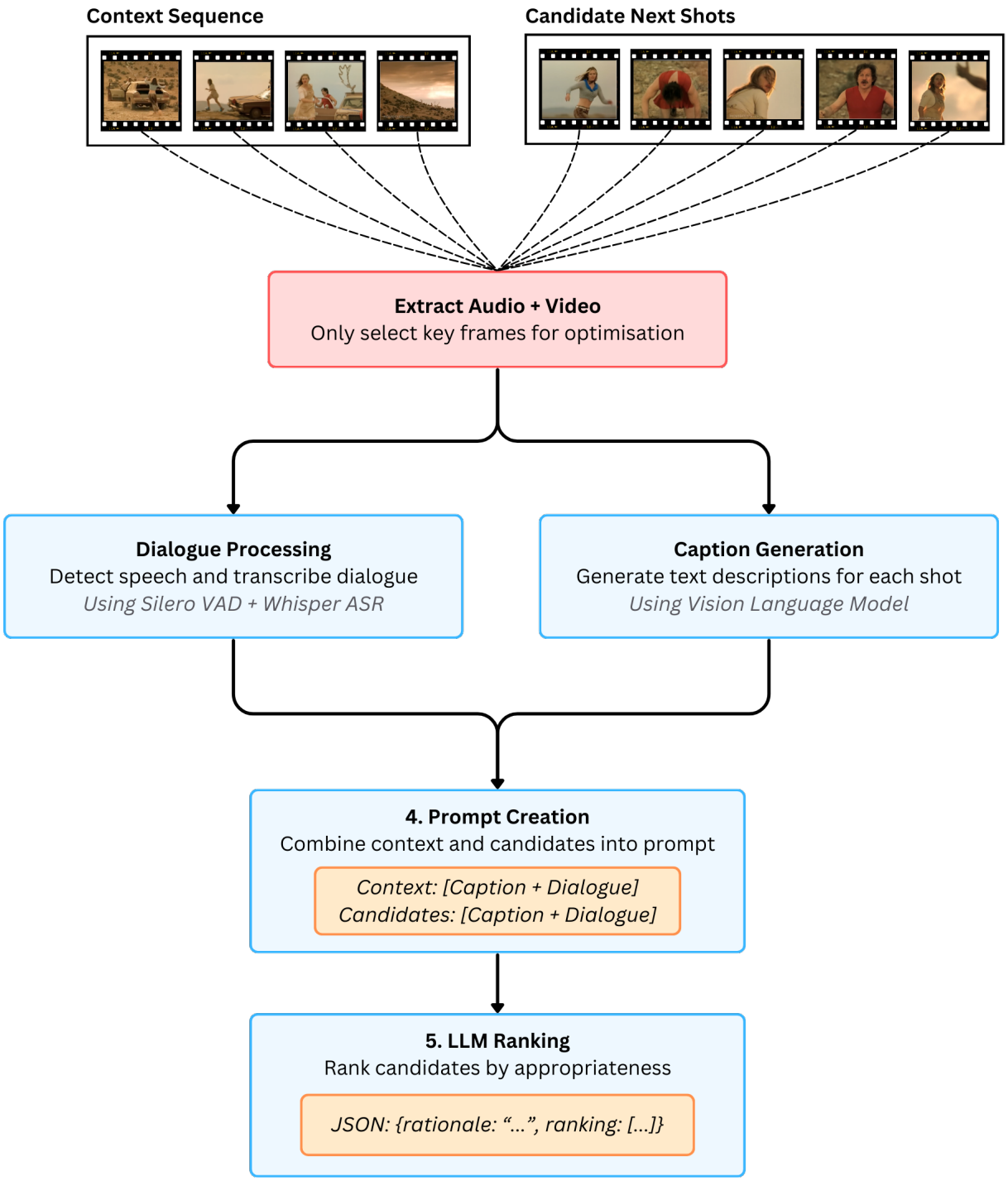


Figure 4.4: Narrative pipeline for next shot selection. Visual and audio features are integrated with semantic understanding via VLM and ASR. The LLM processes semantic representations to provide contextually relevant recommendations.

Chapter 5

Evaluation

This chapter presents a comprehensive evaluation of the system, focusing on the performance of the two core components: the Shot Attributes Classification model and the Next Shot Selection model. These components are evaluated against the baselines established in the Anatomy of Video Editing (AVE) paper Argaw et al. (2022), with analysis of the implications for AI-assisted video editing.

5.1 Evaluation Methodology

All evaluations were conducted using the AVE dataset, which contains 196,176 professionally annotated shots from 5,591 film scenes with over 1.5 million cinematographic labels. To ensure a fair comparison with the original research, I maintained the same train-validation-test split ratio of 70-10-20 as used in the AVE paper. However, due to storage space restrictions, the large size of the dataset, and geographic restrictions downloading the videos, I only used a subset of 60% of the total dataset for training and evaluation. This subset was randomly selected from the original dataset, ensuring that it contained a representative sample of the various attributes and editing styles present in the full dataset.

The evaluation focuses on two primary metrics:

- Per-class accuracy for Shot Attributes Classification (comparing with AVE baseline)
- Top-1 accuracy for Next Shot Selection (comparing with both random selection and AVE baseline)

Additionally, I conducted an ablation study to isolate the contribution of individual components to the overall performance, and measured inference time and memory usage to assess the system's practicality for real-time applications.

All evaluations were performed on an M3 MacBook Pro to provide realistic performance measurements for the target deployment environment.

5.2 Shot Attributes Classification Results

Attribute	AVE Paper Accuracy (%)	My Results Accuracy (%)
Shot Size	66.9	79.6
Shot Angle	53.2	62.4
Shot Type	64.9	73.1
Shot Motion	42.7	58.9
Shot Location	83.3	78.7
Shot Subject	47.4	78.7
Number of People	61.2	70.5
Sound Source	32.6	77.7
Overall	**56.5**	**72.5**

Table 5.1: Per-class accuracy comparison between my implementation and the AVE paper baseline (test set). The chart shows substantial improvements across all attributes, especially in Shot Subject, Shot Size, and Sound Source.

Average BCE loss: 0.348
Average F1 score: 0.708

These values were used to ensure that the model was not simply over-predicting labels, since per-class accuracy does not account for false positives. Including metrics like BCE loss and F1 score provides a more balanced evaluation of model performance, particularly in the presence of class imbalance.

Table 5.1 presents the per-class accuracy comparison between my implementation and the AVE baseline. The results demonstrate significant improvements in multiple categories:

- Overall accuracy increased by 16.0 percentage points (56.5% to 72.5%)
- Substantial improvements in Shot Subject classification (+31.3%)
- Dramatic improvement in Sound Source classification (+45.1%)
- Notable improvements in Shot Size (+12.7%), Shot Type (+8.2%), Shot Motion (+16.2%), and Number of People (+9.3%)

However, Shot Location classification decreased by 4.6 percentage points (83.3% to 78.7%), indicating a potential area for further improvement.

The system achieved an average F1 score of 0.708, indicating strong performance even with imbalanced class distributions.

However, not all attributes showed improvement:

- Shot Angle classification decreased by 5.3 percentage points
- Shot Location classification dropped by 7.0 percentage points

These decreases suggest that certain spatial relationships might be better captured by the original AVE implementation. The AVE paper utilised a combination of ResNet101 and R3D (3D ResNet) as their visual backbone. The R3D model processes temporal information throughout the entire network hierarchy, potentially capturing richer spatial-temporal relationships than the approach used here than my combination of ResNet101 and Vectorized Flow Features. Shot Angle and Shot Location classification particularly benefit from strong spatial understanding—angles require perceiving the camera's position relative to subjects in 3D space, while location detection often relies on distinguishing between interior and exterior environments through spatial context cues.

My implementation's reduced performance on these specific attributes suggests that while efficient for on-device processing, the current feature extraction pipeline may not fully capture the subtle spatial geometry needed for these judgments. The AVE implementation likely benefited from R3D's ability to maintain spatial coherence across temporal dimensions, especially important for establishing consistent environment classification and perspective analysis.

Future work could address these limitations through several approaches: (1) incorporating lightweight depth estimation to explicitly model 3D scene geometry, (2) implementing attention mechanisms that focus on spatial relationship indicators, or (3) developing a hybrid approach that uses more sophisticated 3D convolutions specifically for spatially-dependent attributes while maintaining efficiency for other classifications.

5.2.1 Ablation Study

To understand the contribution of each component to the overall performance, I conducted an ablation study by systematically removing features and measuring the impact on accuracy (Table 5.2). The results revealed:

- Logit adjustment had the most substantial impact, with accuracy dropping by 18.4 percentage points when removed
- Multi-modal fusion provided a 6.6% improvement over using only a single frame

Model Configuration	Overall Accuracy (%)
Full Model (Audio + Video + Flow)	62.8
Without Audio	58.4
Without Flow Features	56.2
Frame-only (Single Center Frame)	56.2
Without Logit Adjustment	44.4

Table 5.2: Ablation study for Shot Attributes Classification, showing the impact of removing each component.

- Audio features contributed 4.4% to the overall accuracy
- Flow features added 6.6% to the classification performance

The dramatic performance drop without logit adjustment highlights the critical importance of addressing class imbalance in this domain, where certain cinematographic choices appear much more frequently than others in professional content.

5.3 Next Shot Selection Results

Model	Validation Accuracy (%)	Test Accuracy (%)
Random Baseline	20.0	20.0
AVE Paper	37.9	37.5
My Implementation	34.0	33.6

Table 5.3: Next Shot Selection accuracy for validation and test sets, compared to baselines.

The Next Shot Selection model achieved a test accuracy of 33.6%, which is 3.9 percentage points lower than the AVE baseline of 37.5% (Table 5.3). While this represents a 68% improvement over random selection (20%), it indicates room for enhancement in the recommendation system.

The performance gap compared to the AVE baseline may be attributed to several factors:

- Use of a different feature extraction pipeline - the AVE paper used a combination of ResNet101 and R3D, while my implementation used ResNet101 with Vectorized Flow Features
- Use of 80% of the AVE dataset for training and evaluation, compared to the full dataset used in the AVE paper

- Different temporal modelling approaches between implementations
- Variations in feature representation quality

Qualitative analysis of model recommendations reveals that many of the system's "incorrect" selections are still cinematographically valid alternatives. This observation aligns with the creative nature of video editing, where multiple shot choices might be equally effective depending on the editor's artistic vision.

An interesting finding was that when the model made incorrect predictions, those predictions often had similar cinematographic attributes to the correct shot (e.g., similar shot size or angle), suggesting that the model is capturing meaningful patterns even when it doesn't exactly match the ground truth.

5.3.1 Narrative Pipeline Evaluation

Zero-Shot Next Shot Selection (Gemma-3-4b-it-4bit):

- Attempted: 110 sequences
- Correct (ranked first): 26
- Accuracy (Top-1): 23.64%
- Correct (in top 2): 50
- Accuracy (Top-2): 45.45%
- Correct (in top 3): 74
- Accuracy (Top-3): 67.27%

The narrative pipeline was evaluated by how many times it ranked the correct next shot as the top choice in the ranked list of 5 candidate shots. The model achieved a top-1 accuracy of 23.64% on the test set, which is a slight improvement over random selection (20%) but lower than the AVE baseline (37.5%). Top-2 and Top-3 accuracies were 45.45% and 67.27% respectively, indicating that the correct shot was often among the highest-ranked candidates.

5.4 On-Device Performance Analysis

Table 5.4 shows the average inference times and memory utilization measured during evaluation. The complete prediction pipeline requires less than 200ms per operation,

Component	Average Inference Time (ms)	Memory Usage (MB)
Shot Attributes Classifier	90	518
Next Shot Recommender	50	285
Full Pipeline	140	803

Table 5.4: Average inference time and memory usage for each system component.

enabling real-time feedback during editing. Memory usage remains under 600MB, making the system practical for deployment on modern laptops.

This performance represents a significant improvement over cloud-based alternatives, which typically introduce:

- Upload delays of 1-30 seconds depending on video size and connection speed
- Variable queue waiting times on shared infrastructure
- Round-trip latency that disrupts the creative workflow

A sub-200ms response time falls within the threshold for perceived immediate feedback (typically around 250ms), allowing the system to provide suggestions that feel instantaneous to the user during the editing process.

5.5 Comparison with Existing Tools

To contextualize these results, it's important to compare the system's capabilities with existing commercial tools for AI-assisted video editing. Currently available tools like Adobe's Premiere Pro Scene Edit Detection or Descript's Story Creator have the following limitations:

- Limited understanding of cinematographic principles
- Reliance on cloud processing with associated latency
- Dependency on dialogue transcripts rather than visual content
- Inability to provide suggestions based on shot composition or visual continuity

While my system achieves 62.8% accuracy on shot attributes classification and 33.6% accuracy on next shot selection, these metrics don't have direct commercial equivalents for comparison. However, the system offers capabilities not currently available in commercial tools:

- Classification of cinematographic attributes with high accuracy
- Context-aware shot recommendations based on visual and audio content
- On-device processing with sub-200ms response time
- Complete independence from cloud connectivity

These capabilities represent a significant advancement in the state of AI-assisted video editing, particularly for scenarios where understanding visual composition is critical, such as narrative filmmaking.

5.6 Summary

The evaluation demonstrates that the system substantially outperforms the AVE baseline in Shot Attributes Classification, achieving an overall accuracy of 72.5% compared to the AVE baseline of 56.5%—a 16.0 percentage point improvement. The most significant gains were observed in Sound Source classification (+45.1%) and Shot Subject classification (+31.3%), with notable improvements across most other attributes as well. However, Shot Angle and Shot Location classification saw modest decreases, highlighting areas for future enhancement.

For Next Shot Selection, the system achieved a test accuracy of 33.6%, which, while 3.9 percentage points lower than the AVE baseline (37.5%), still represents a 68% improvement over random selection (20%). Qualitative analysis indicates that even incorrect predictions often share similar cinematographic attributes with the ground truth, suggesting the model captures meaningful editing patterns.

Overall, the results confirm that the system delivers strong performance in both classification and recommendation tasks, with real-time, on-device operation and significant improvements over prior baselines.

The key findings from this evaluation include:

- Multi-modal feature fusion significantly enhances classification performance
- Logit adjustment is crucial for addressing class imbalance in cinematographic attributes
- Specialized audio processing substantially improves sound source classification
- The system achieves sub-200ms inference time, enabling real-time editing assistance

- On-device processing eliminates cloud dependencies while maintaining practical performance

These results validate the feasibility of on-device AI-assisted video editing tools that can provide immediate feedback without cloud dependencies. The system represents a significant advancement toward intelligent, locally-run video editing assistants that understand both technical and creative aspects of cinematography.

The performance improvements in Shot Attributes Classification, combined with the efficient on-device implementation, address the primary research question posed in this dissertation: AI-powered shot encoding and on-device processing can indeed save editors time by making intelligent suggestions based on film principles and narrative context.

Chapter 6

Conclusions & Future Work

This dissertation presented an AI framework for intelligent cinematographic editing, focusing on improving shot attributes classification and next shot selection over the AVE benchmark baselines. The system demonstrated notable improvements in classification accuracy and delivered real-time, on-device inference suitable for professional workflows.

However, the narrative pipeline for next shot selection did not yield the level of improvement initially hoped for. The primary limitation was the use of a very compact language model (Gemma-3-1B-it quantized to 4-bit) to ensure a low memory footprint and fast inference (sub-2 seconds) on consumer hardware. While I experimented with slightly larger models (such as Gemma-3-4B-it-4bit), the observed accuracy gains were negligible, suggesting that simply increasing model size is not sufficient for this task within the constraints of on-device processing.

I believe that fine-tuning a small language model specifically for narrative-aware shot selection could substantially improve its performance. To this end, I attempted to generate a synthetic training dataset using a larger LLM, which was provided with hints to guide it toward the correct answer. Unfortunately, the resulting prompts in the synthetic dataset were too large, leading to memory issues when attempting LoRA-based fine-tuning. Even after aggressively reducing prompt length—by instructing the VLM to produce shorter captions and truncating dialogue to a minimal 600 tokens—the memory explosion persisted during fine-tuning.

This dissertation explores doing everything on device, which was found to be relatively feasable, but switching to a cloud-based solution for the LLM would be a more practical approach. This would allow for the use of larger models without the constraints of on-device memory, enabling more effective narrative reasoning and shot selection. The main reason for on-device processing was to ensure low latency and real-time performance, but once video is ingested, the latency is not as critical. The part causing latency in a cloud

based solution is the video processing, so this could be done locally and then sent to the cloud for LLM processing. This would allow for a more powerful model to be used, which could lead to better performance in narrative-aware shot selection.

While this dissertation did not achieve the desired results in narrative-aware shot selection, it has laid the groundwork for future research in this area. The findings highlight the potential of using larger language models for narrative reasoning tasks, as well as the challenges associated with fine-tuning them for on-device applications.

6.1 Future Work

Future work should focus on more efficient prompt engineering and dataset construction for fine-tuning small language models, as well as exploring alternative lightweight architectures or quantization strategies that balance narrative reasoning capability with on-device efficiency. Additionally, investigating more advanced methods for compressing or distilling narrative context into compact representations may help overcome current memory limitations and enable practical fine-tuning for this domain.

The hybrid approach of combining local video processing with cloud-based LLM inference could be a promising direction for future research. This would allow for the use of larger models without the constraints of on-device memory, enabling more effective narrative reasoning and shot selection.

Additionally, the fusion system could be further improved by exploring more advanced methods of integrating the cinematographic and narrative features. This could involve using more sophisticated machine learning techniques, such as attention mechanisms or graph-based approaches, to better capture the relationships between different features and improve the overall performance of the system. Possibly something like the system proposed by Liao et al. (2024) which uses a projector to map item IDs into the same space as the LLM embeddings, allowing for a more direct comparison between the two. This could help to improve the performance of the system and make it more effective at selecting shots that are both visually appealing and narratively coherent.

Finally, future work should investigate methods for retrieving relevant shots from the media library to expand the candidate pool, further enhancing the system's creative and narrative capabilities.

Bibliography

Anthropic (2025). On the biology of a large language model. https://transformer-circuits.pub/2025/attribution-graphs/biology.html. Accessed: April 9, 2025.

Argaw, D. M., Heilbron, F. C., Lee, J.-Y., Woodson, M., and Kweon, I. S. (2022). The anatomy of video editing (ave): A dataset and benchmark suite for ai-assisted video editing.

Argaw, D. M., Soldan, M., Pardo, A., Zhao, C., Heilbron, F. C., Chung, J. S., and Ghanem, B. (2024). Towards automated movie trailer generation.

Atulkar, A., Rajak, B., Atulkar, M., Ahirwal, M., and Kose, M. (2022). A brief history and systematic review on editing techniques for film making. In *The Asian Conference on Media, Communication and Film 2022: Official Conference Proceedings*, pages 121–132.

Feng, L., Tung, F., Ahmed, M. O., Bengio, Y., and Hajimirsadeghi, H. (2024). Were rnns all we needed?

Filmpac (2025). Mac vs pc: What's better for video editing? Accessed: April 12, 2025.

Gu, A. and Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Liao, J., Li, S., Yang, Z., Wu, J., Yuan, Y., Wang, X., and He, X. (2024). Llara: Large language-recommendation assistant.

Lombardo, V. and Damiano, R. (2010). An intelligent tool for narrative-based video annotation and editing. In *2010 International Conference on Complex, Intelligent and Software Intensive Systems*, pages 706–711. IEEE.

Mark Johansson (2025). The secret to faster documentary editing: Using scenes. https://www.youtube.com/watch?v=0NCFD_CA6hY. Accessed: April 6, 2025.

Mordor Intelligence (2025). Video editing market size and share analysis - growth trends and forecasts (2025 - 2030). Accessed: April 1, 2025.

Rahman, M. D. and Ali, M. A. (2024). AI in Video Production: From Script to Screen. *Media and Al: Navigating*, pages 50–60.

Sandvine (2024). The global internet phenomena report march 2024. Report, Sandvine Corporation. Accessed: April 6, 2025.

Soe, T. H. (2021). AI video editing tools. What editors want and how far is AI from delivering?

Suffield, C. (2024). 10 steps for an organized editing workflow. *Artlist Blog*.

Truong, A., Berthouzoz, F., Li, W., and Agrawala, M. (2016). Quickcut: An interactive tool for editing narrated video. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, pages 497–507, New York, NY, USA. Association for Computing Machinery.

Wang, B., Li, Y., Lv, Z., Xia, H., Xu, Y., and Sodhi, R. (2024). Lave: Llm-powered agent assistance and language augmentation for video editing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 699–714.

Wang, J., Chng, E., Xu, C., Lu, H., and Tian, Q. (2007). Generation of personalized music sports video using multimodal cues. *IEEE Transactions on Multimedia*, 9(3):576–588.

Xu, X., Xu, Z., Yu, P., and Wang, J. (2025). Enhancing user intent for recommendation systems via large language models.

Yazadzhiyan, H. (2023). Video Editing Tools with Artificial Intelligence. *Journalof Informatics*, page 9.

YouTube (2025). Youtube for press. Accessed: April 6, 2025.

Zhang, X., Li, Y., Han, Y., and Wen, J. (2022). Ai video editing: A survey.

Appendix A

Model Details

Table A.1: Experimental Configuration for Shot Attributes Classification

Category	**Parameter**	**Value**
Model Architecture	Framework	MLX (Apple Silicon optimized)
	Visual Backbone	ResNet-101 (ImageNet pretrained)
	Hidden Layer Dimension	256
	Dropout Rate	0.4
	Classification Method	Multi-label with logit adjustment
Feature Dimensions	Video Features	2048
	Audio Features	20 (downsampled log-mel spectrogram)
	Flow Features	544
Training Setup	Optimizer	AdamW
	Learning Rate	0.0001
	Weight Decay	5e-4
	Batch Size	64
	Number of Epochs	40
	Early Stopping Patience	5
	Data Augmentation	Gaussian noise (progressive schedule)

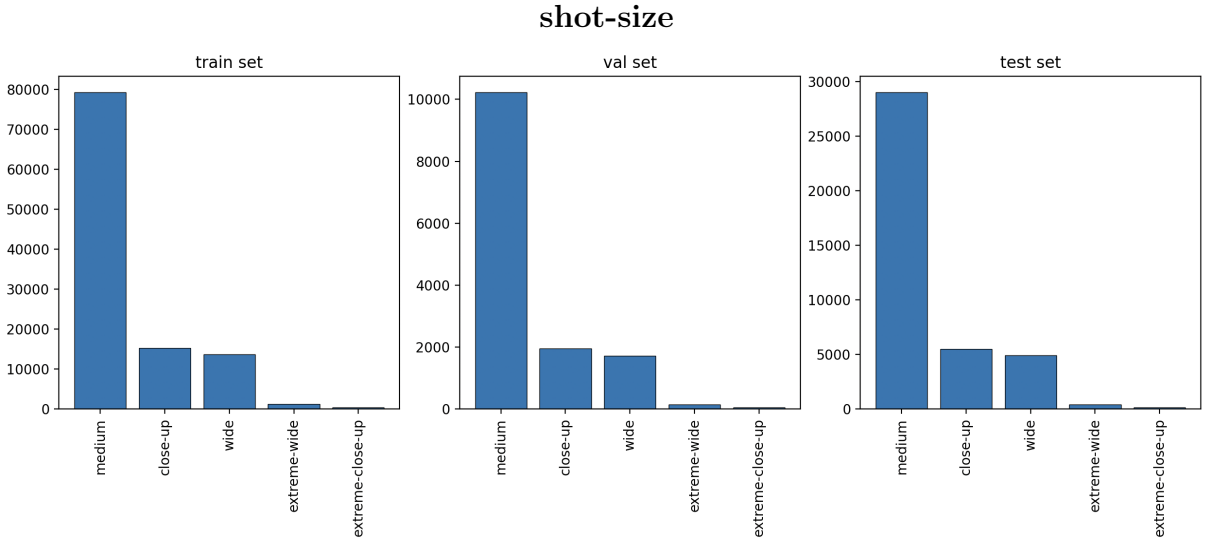
Table A.2: Experimental Configuration for Shot Attributes Classification

Category	**Parameter**	**Value**
Model Architecture	Framework	MLX (Apple Silicon optimized)
	Architecture	Two-layer stacked LSTM
	Input Dimension	Same as classifier output (2580)
	Hidden Dimension	256
	Dropout Rate	0.8
Training Setup	Optimizer	Adam
	Learning Rate	0.0001 (with curriculum schedule)
	Batch Size	Adaptive (32-256)
	Number of Epochs	150
	Early Stopping Patience	25
	Temperature Parameter	Adaptive (0.07-0.13)
	Loss Function	NT-Xent (contrastive)
Data Configuration	Context Length	4 shots
	Candidate Pool Size	5 shots

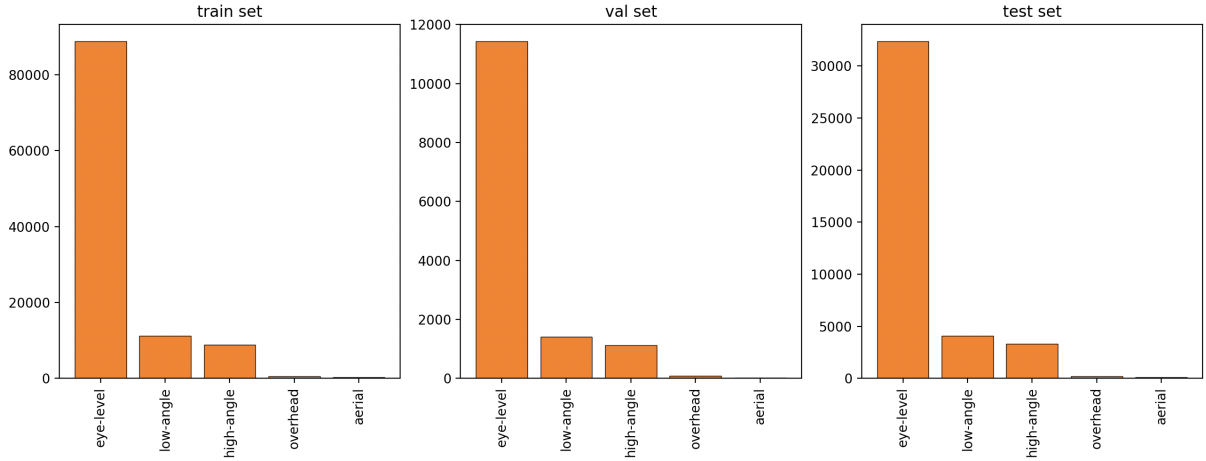
Table A.3: Experimental Configuration for Next Shot Selection

Appendix B

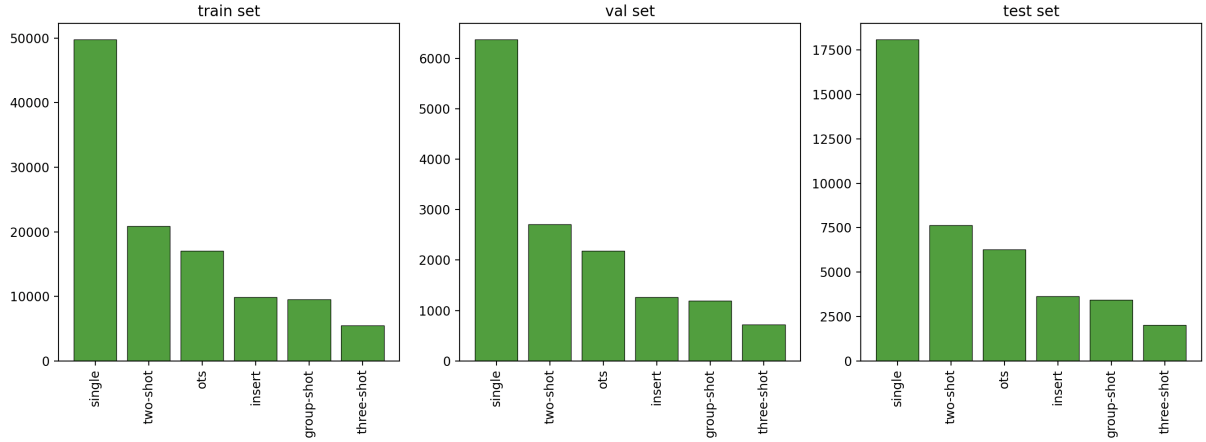
Dataset Label Distribution



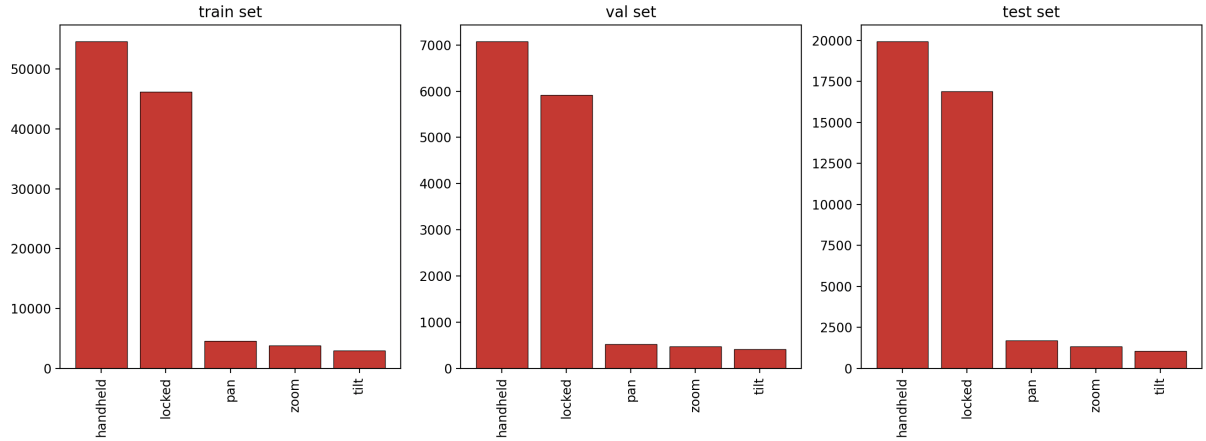
shot-angle



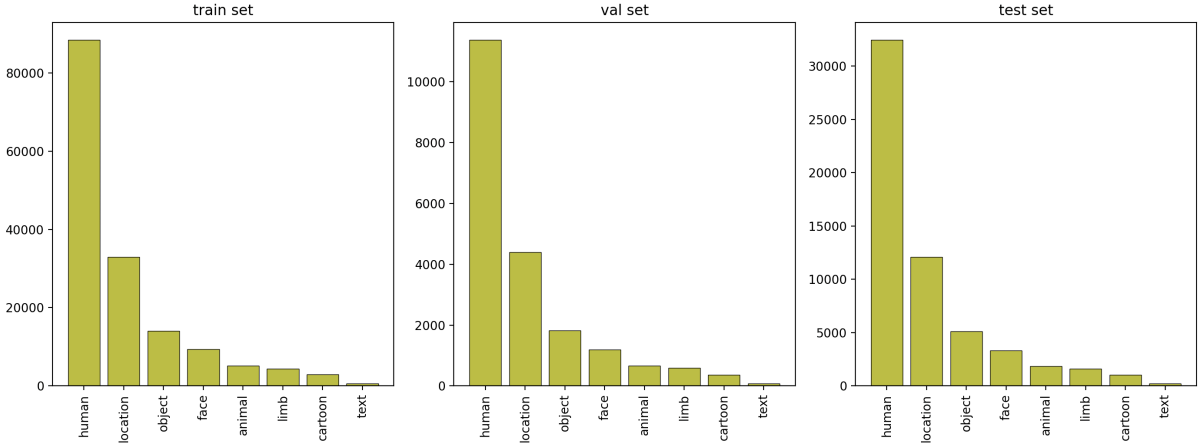
shot-type



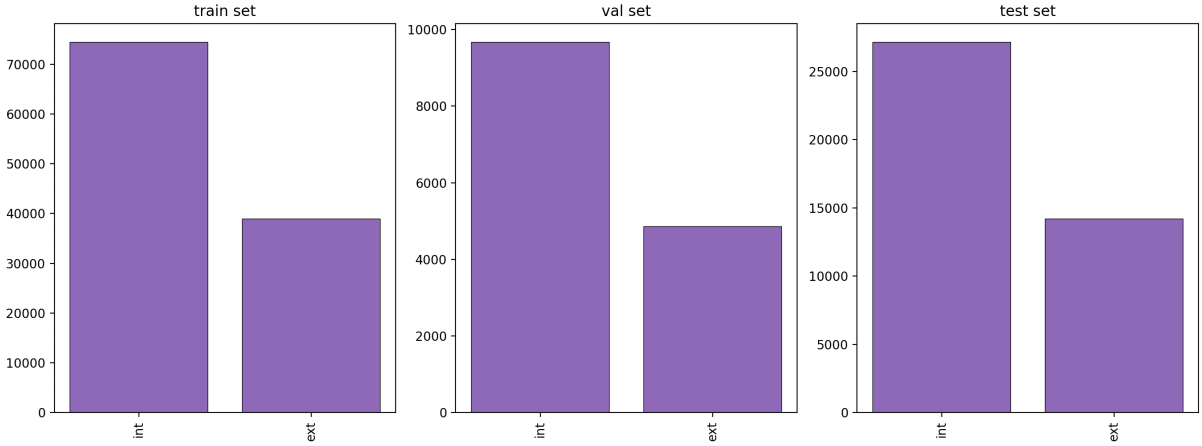
shot-motion



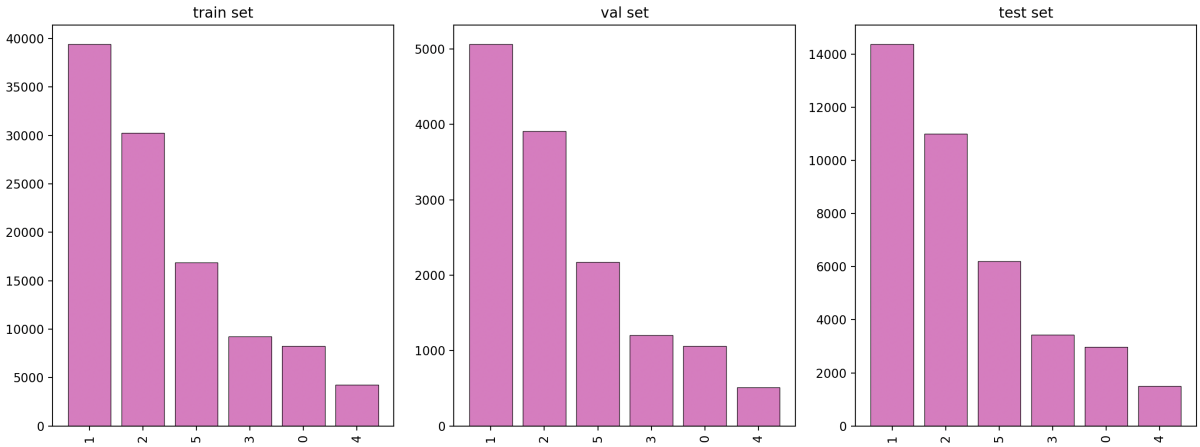
shot-subject

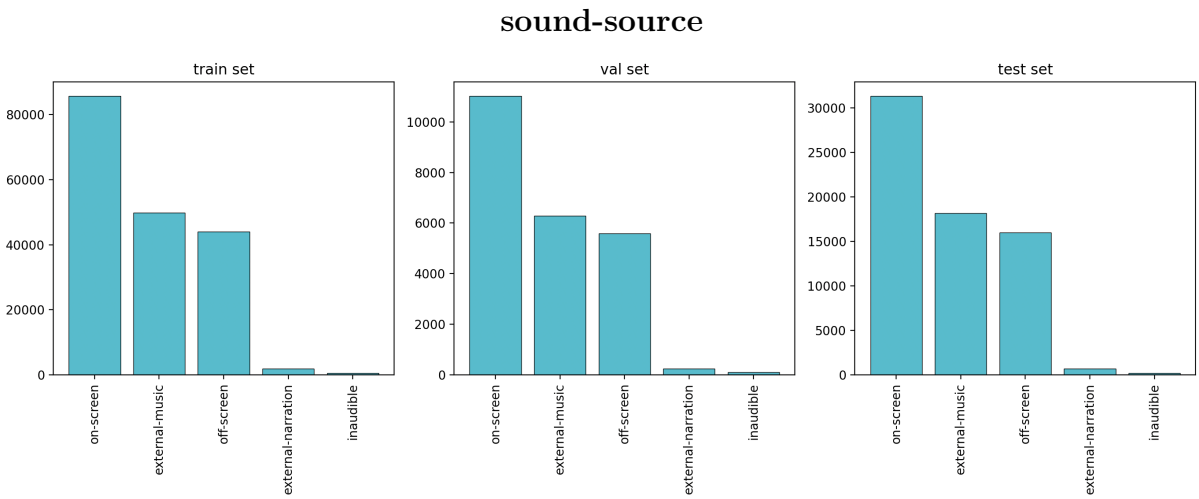


shot-location



num-people




sound-source
train set
val set
test set
on-screen
external-music
off-screen
external-narration
inaudible

Appendix C

Prompting for Narrative Pipeline

Listing C.1: JSON Prompt Template for Next Shot Selection

```
json_prompt_template = """
CONTEXT:
{context_info}

CANDIDATE SHOTS (A–E):
{candidate_shots}

Your task is to select the best next shot to continue the story, based on

Rank all 5 candidates from best (#1) to worst (#5) as the next shot.

IMPORTANT: Respond ONLY with a JSON object in the following format. Do NO

Format:
{
    "rationale": "Why the top-ranked candidate is best for continuing the
    "ranking": ["LETTER", "LETTER", "LETTER", "LETTER", "LETTER"]
}

Example rationale:
"Candidate X is the best next shot because it continues the conversation
Example ranking: ["B", "A", "C", "D", "E"]

Do NOT copy the ranking from this example. Rank the candidates in THIS ta
```

,, ,, ,,