Seeing, Describing and Remembering: A Study on Audio Descriptions and Video Memorability

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computational Linguistics by Research

by

Eshika Khandelwal
2020114018
eshika.k@research.iiit.ac.in

Advisor: Dr. Makarand Tapaswi

International Institute of Information Technology Hyderabad
500 032, India

Jun 2025

International Institute of Information Technology Hyderabad
Hyderabad, India

CERTIFICATE

This is to certify that work presented in this thesis proposal titled ***Seeing, Describing and Remembering: A Study on Audio Descriptions and Video Memorability*** by *Eshika Khandelwal* has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Dr. Makarand Tapaswi

Acknowledgements

I feel incredibly lucky to have had the support, guidance, and encouragement of so many people throughout this journey. This thesis would not have been possible without them.

I would first like to thank my parents for supporting every decision I made and for always placing their steady trust in me. I am also thankful to my brother, Siddhesh. I often teased him about his achievements, not realising how challenging they were until I reached those points myself. I have been unknowingly following in his footsteps—just about seven years behind.

I owe my interest in research to Prof. Makarand Tapaswi. He always let me to pursue topics that excited me, even if it meant changing directions multiple times, to which I am grateful. I remain in awe of the way he thinks, especially his ability to notice nuances I would have completely missed on my own. He has always explained things with clarity and patience, and taught me to pay attention to the small details. From running careless experiments to (hopefully) adopting a more systematic and critical approach, a lot of how I work now is because of him.

Working with Prof. Vishnu Sreekumar and Prajneya Kumar was the most enjoyable part of this journey. Before this project, research felt like a pursuit of numbers and state of the art results. This work taught me how to face rejections, take pride in my work, and understand that research is much more than method and metrics. Prajneya, thank you for always being a source of motivation and someone I could talk to about all things research, even long after our project ended.

I am also thankful to Prof. Gül Varol for the opportunity to intern in Paris. Those six months gave me clarity about what I want to do next and helped me feel more certain about continuing in research.

To my college friends—Ameya, Anushree, Nandini, Pranali, Prajneya, Pranav, and Shreeya—thank you for patiently listening to my endless rants (often without context) and for putting up with me being constantly glued to my laptop. I promise you'll see less of that now (at least for a bit!). Your company gave me the much-needed breaks that kept me going through the deadlines. Special thanks to Pranav, whose support and patience brought me calm when nothing seemed to work. Thank you for always believing in me.

To my school friends—Amisha, Aryan, Sarthak, Shourya, and Vidhi—thank you for being a constant source of comfort and joy, from school days and entrance exams to the start of my research journey.

Finally, thank you to everyone who has contributed, big or small, to this rollercoaster of a journey. Your support has shaped this thesis in more ways than you know.

Abstract

This thesis investigates two aspects of how humans engage with visual content: how it is *described* and how it is *remembered*.

The first part focusses on *Audio Descriptions* (ADs), which convey essential on-screen information, allowing visually impaired audiences to follow videos. To be effective, ADs must form a coherent sequence that helps listeners visualise the unfolding scene, rather than describing isolated moments. However, most automatic methods generate each AD independently, often resulting in repetitive, incoherent descriptions. To address this, we propose a training-free method, CoherentAD, that first generates multiple candidate descriptions for each AD time interval, and then performs auto-regressive selection across the sequence to form a coherent and informative narrative. To evaluate AD sequences holistically, we introduce a sequence-level metric, StoryRecall, which measures how well the predicted ADs convey the ground truth narrative, alongside repetition metrics that capture the redundancy across consecutive AD outputs. Our method produces coherent AD sequences with enhanced narrative understanding, outperforming prior approaches that rely on independent generations.

The second part investigates *Video Memorability*. Understanding what makes a video memorable has important applications in advertising and education technology. Towards this goal, we investigate spatio-temporal attention mechanisms underlying video memorability. Different from previous works that fuse multiple features, we adopt a simple CNN+Transformer architecture that *enables analysis of spatio-temporal attention* while matching state-of-the-art (SoTA) performance on video memorability prediction. We compare model attention against human gaze fixations collected through a small-scale eye-tracking study where humans perform the video memory task. We uncover the following insights: (i) Quantitative saliency metrics show that our model, trained only to predict a memorability score, exhibits similar spatial attention patterns to human gaze, especially for more memorable videos. (ii) The model assigns greater importance to initial frames in a video, mimicking human attention patterns. (iii) Panoptic segmentation reveals that both (model and humans) assign a greater share of attention to *things* and less attention to *stuff* as compared to their occurrence probability.

Contents

List of Figures

List of Tables

List of Related Publications

[P1] Prajneya Kumar*, Eshika Khandelwal*, Makarand Tapaswi† and Vishnu Sreekumar†, **"Seeing Eye to AI: Comparing Human Gaze and Model Attention in Video Memorability"**, in proceedings of *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.
*† equal contribution

[P2] Eshika Khandelwal, Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Andrew Zisserman, Makarand Tapaswi and Gül Varol, **"More than a Moment: Towards Coherent Sequences of Audio Descriptions"**, *(Under Review)*

Related co-author publications:

[P3] Divy Kala, Eshika Khandelwal, and Makarand Tapaswi, **"What You See is What You Ask: Evaluating Audio Descriptions"**, *(Under Review)*

Chapter 1

Introduction

This thesis investigates two aspects of how humans engage with visual content: how scenes are *described* through narration, and how they are *remembered* after viewing. In the first part, we focus on making visual media more accessible through coherent, automatically generated *Audio Descriptions*. In the second, we explore *Video Memorability* by comparing model attention to human gaze during memory tasks.

Audio Descriptions. Imagine watching a suspenseful scene with your eyes closed. You hear a door creak, footsteps on a hard floor, then a long pause. You know something is happening, but without visuals, you can't make out what. Now add an Audio Description:

"A man enters the room and kneels beside a motionless body. He pulls a wallet from the jacket pocket and flips it open. Inside is a police badge."

Suddenly, the silence becomes meaningful. The scene isn't just quiet—it's a moment of discovery. Audio Descriptions (ADs) provide brief, carefully timed narrations that convey important visual information unavailable through sound (or dialogues) alone. They make visual media accessible to Blind and Visually Impaired (BVI) audiences by describing actions, gestures, facial expressions, scene changes, and on-screen text.

These descriptions are typically written to be neutral, concise, and delivered between lines of dialogue. While traditionally authored by professionals, there is increasing demand for automated AD generation. Most existing methods, however, generate one description per predefined time interval, without considering how these descriptions fit together as a sequence. This often leads to incoherent or redundant sequences. Similarly, current evaluation methods score ADs in isolation, missing how well they work as a complete narrative. In this thesis, we address both challenges by proposing a method for coherent AD sequence generation and introducing metrics that evaluate quality across the full sequence.

Video Memorability. Some videos linger in memory long after they are watched. A quintessential example is the *Glucon-D* advertisement[1], where an animated sun is portrayed sipping energy from

[1] `https://www.youtube.com/watch?v=0QKdXiP8rrY`

children through a straw, a metaphor for the exhaustion caused by summer heat. In contrast, slow pans over generic landscapes are rarely remembered.

Video memorability refers to the likelihood that a video will be remembered after viewing. Memorability is defined at the population level. For instance, if 1 in 10 viewers remembers a video, its memorability score is 0.1. Understanding what makes certain videos more memorable has significant implications in fields like advertising and education, where lasting impact and recall are crucial.

While previous research primarily focuses on predicting memorability scores based on video content, this thesis dives deeper. We aim to understand how models make these predictions—specifically, what they attend to in space and time. We analyze the spatio-temporal attention patterns of a CNN+Transformer model trained to predict video memorability and compare them to human gaze data collected from a memory experiment. This allows us to examine whether the model attends to the same visual cues as human viewers, and whether the alignment relates to the memorability scores.

1.1 Contribution

This thesis makes contributions in two areas:

Coherent Audio Descriptions (Generation and Evaluation).

- We propose CoherentAD, a training-free approach that promotes the generation of coherent and diverse visual descriptions across the video.
- To evaluate ADs at the sequence level, we introduce two new metrics that go beyond reference-based methods (e.g., CIDEr, LLM-AD-Eval) which assess descriptions in isolation:
  - *StoryRecall*, which measures how well the predicted sequence captures key visual and narrative details from the ground truth;
  - *Repetition*, which quantifies redundancy across consecutive ADs.
- We evaluate our method on the CMD-AD dataset and show improvements in both coherence and informativeness over prior baselines.

Video Memorability.

- We adopt a simple CNN+Transformer model that facilitates analysis of spatio-temporal attention mechanisms. Despite its simplicity, the model matches state-of-the-art performance on memorability prediction.
- We collect new eye-tracking data from human subjects performing a video memory task, following established experimental protocols [1, 2].
- Using panoptic segmentation and attention-weighted analysis, we show that both the model and humans vary their attention similarly across different visual categories, such as *things* and *stuff*.

- We find that the model, without any explicit temporal bias, learns to focus on early video frames—mirroring the natural temporal decay observed in human gaze patterns.

1.2 Organisation of the Thesis

The remainder of this thesis is organised as follows:

- **Chapter 2** reviews related work across both areas: Audio Descriptions (covering both generation and evaluation) and Video Memorability.
- **Chapter 3** presents our method for automatically generating coherent sequences of Audio Descriptions, along with qualitative comparisons with prior approaches.
- **Chapter 4** discusses the limitations of existing AD evaluation metrics and introduces our proposed sequence-level metrics designed to better assess coherence and redundancy.
- **Chapter 5** investigates video memorability. We analyse model attention patterns and compare them to human gaze behaviour using collected eye-tracking data.
- **Chapter 6** concludes the thesis and outlines future directions for both Audio Description and Video Memorability research.

Chapter 2

Literature Review

This chapter reviews the existing literature on Audio Descriptions and Memorability. Section 2.1 examines prior work on Audio Descriptions (ADs), including both supervised and training-free approaches to AD generation, as well as evaluation metrics developed to assess their quality. Section 2.2 surveys research on Memorability, drawing from both Cognitive Science and Computer Vision perspectives.

2.1 Audio Descriptions

ADs are verbal narrations that describe on-screen visual content to make media (like movies or TV shows) accessible to Blind and Visually Impaired (BVI) audiences. These narrations provide crucial visual context, typically inserted into natural pauses in dialogue or significant sound effects. This section surveys prior work on the automatic generation and evaluation of ADs, highlighting key approaches, innovations, and existing limitations.

2.1.1 Automatic Generation of Audio Descriptions

Prior work on automatic AD generation falls broadly into two categories: (i) *end-to-end models*, which are typically supervised and fine-tuned on specialized datasets, and (ii) *training-free frameworks*, which leverage the capabilities of pre-trained Vision-Language Models (VLMs) and Large Language Models (LLMs) within multi-stage pipelines.

2.1.1.1 End-to-end Models

Supervised methods [3–10] fine-tune end-to-end models on curated AD datasets such as MAD [11], CMD-AD [6], and TV-AD [12].

AutoAD-II [4] introduces a Flamingo-style multimodal cross-attention mechanism that incorporates visual features, cast metadata, and speech timing to generate character-aware ADs. It leverages an external Character Bank that dynamically links face embeddings to named entities, allowing the model to consistently refer to characters by name. UniAD [8] enhances understanding of complex scenes using interleaved multimodal instructions that jointly process visual and textual modalities. MovieSeq [7]

adopts a general-purpose multimodal transformer that represents video as an interleaved sequence of visual frames, subtitles, and plot elements. LoCo-MAD [5] explicitly models long-range temporal dependencies through a two-stage framework. It first encodes visual content using a compact representation learned via contrastive and generation objectives. It then applies a Dynamic Selection Module (DSM) to retrieve plot-relevant context for LLM-based generation. The DSM dynamically selects relevant subtitles and descriptions from earlier scenes using a sampler and contrastive loss, enabling better narrative consistency over long durations. FocusedAD [10] further advances narrative grounding by combining three modules: a Character Perception Module (CPM) for robust identity tracking across frames, a Dynamic Prior Module (DPM) for incorporating prior context, and a Focused Caption Module (FCM) that prioritizes narratively important content. This modular design ensures the model emphasises plot-relevant elements and tracks character identity without redundancy.

2.1.1.2 Training-free Frameworks

In contrast, training-free frameworks [12–17] leverage pre-trained VLMs and LLMs in multi-stage pipelines without task-specific fine-tuning. These approaches rely on prompting strategies and structured reasoning rather than supervised learning.

AutoAD-Zero [12] and Shot-by-Shot [16] implement explicit two-stage systems: a VLM first produces dense visual descriptions, which an LLM then compresses into concise ADs. Shot-by-Shot enhances this structure with cinematic priors, integrating film grammar cues such as shot boundaries, camera threads, and shot scale. LLM-AD leverages GPT-4V for zero-shot AD generation from visual and subtitle inputs. It uses natural language instructions to control output length and formatting, and integrates a character tracking module to ensure consistent naming across segments without additional training. MM-Narrator [13] tackles long-form video by incorporating a memory-augmented generation loop. It maintains short-term textual context (recent ADs) and long-term visual memory (for character re-identification), enabling scene-consistent narration across extended time spans. It constructs multimodal prompts for GPT-4 by integrating visual captions and dialogue with step-by-step reasoning in a Chain-of-Thought format. MMAD [14] enhances multimodal integration through dedicated modules: an audio-aware module captures environmental sound features, an actor-tracking module resolves character identity and ReID, and a contextual alignment module aligns non-dialogue scenes for AD insertion. NarrAD [17] incorporates movie scripts as an external source of narrative structure. It retrieves relevant script passages for each video segment, uses multimodal in-context prompting to generate candidate ADs, and then performs information curation to eliminate redundancy and meet timing constraints.

2.1.2 Coherence in Audio Descriptions

A common challenge in automatic AD generation is coherence. Models that generate ADs independently per time segment often produce repetitive or disjointed outputs. Ensuring continuity across descriptions is crucial for creating a smooth and comprehensible experience for BVI users.

Several supervised methods explicitly model temporal dependencies. AutoAD-I [3] and UniAD [8] implement recursive generation by conditioning each AD on previously generated outputs. AutoAD-II [4] adds a localisation module that predicts suitable time segments for inserting ADs. DistinctAD [9] jointly processes adjacent clips and employs a Contextual Expectation-Maximisation Attention (EMA) module to reduce redundancy across outputs. However, these methods rely on training with ground truth (GT) ADs. In contrast, our approach investigates sequence-level coherence and redundancy in a training-free setting

2.1.3 Evaluation Metrics for Audio Descriptions

Early evaluation of ADs adopted metrics from image captioning. N-gram-based metrics such as BLEU [18], ROUGE [19], METEOR [20], and CIDEr [21] measure lexical (n-gram) overlap between predicted and reference descriptions. Semantic similarity metrics like SPICE [22] and BERTScore [23] aim to capture meaning by comparing scene graphs or contextual embeddings.

To better reflect AD-specific quality, newer approaches propose targeted AD metrics. Retrieval-based evaluation (Recall@k/N [3]) assesses whether a predicted AD can retrieve its corresponding GT description from a pool of candidates. LLM-based evaluations [6, 13] use LLMs to judge predictions relative to reference ADs. CRITIC [6] promotes consistency in character identification, while the Action Score [16] evaluates whether visually grounded actions are correctly captured.

NarrAD [17] complements automatic metrics with human evaluations across four dimensions: *usefulness, specificity, likelihood of recommendation, and a comprehension test*. While such assessments provide valuable insights into user experience, they are resource-intensive and impractical to apply at scale.

Despite these advances, most metrics continue to evaluate individual ADs in isolation. MM-Narrator [13] introduces SegEval, which uses GPT-4 to assess AD quality over sliding windows of consecutive segments. However, because it performs relative comparisons between predicted ADs and the ground truth, SegEval may favor fluent but hallucinated predictions. Additionally, it is limited to fixed-size local windows and does not explicitly measure redundancy or global coherence across the entire sequence.

Our work moves beyond single-AD evaluation by introducing metrics designed for sequence-level analysis. We assess both local redundancy and global narrative coherence, enabling evaluation of ADs as cohesive sequences rather than isolated captions.

2.2 Memorability

Memorability refers to the likelihood that a visual stimulus will be remembered by viewers. In this section, we review how memorability has been studied in both cognitive science and computer vision.

2.2.1 Memorability in Cognitive Science

While human beings remember a huge amount of visual information, not all visual experiences are equal in our memory [24]. Some images are consistently better remembered across people, suggesting that memorability is observer-independent [25,26]. This makes algorithms suitable for predicting memorability [27]. Several factors such as scene semantics [24], object category [28], and visual saliency [28] correlate with memorability, yet considerable statistical variance in memorability scores remains unexplained [29]. Although image memorability has been studied extensively in cognitive science, videos have been used primarily in the study of event segmentation and to understand the neural processes underlying learning and memory [30, 31]. Observer-independent memorability of videos has received less attention in cognitive science compared to the work in computer vision.

2.2.2 Memorability in Computer Vision

The study of visual memorability in computer vision started with a focus on images [24,27]. Models such as *MemNet* were developed for image memorability prediction on large image datasets [27]. Improvements over the initial models involved incorporating attention mechanisms [32], image captioning modules [33], object and scene semantics [34], and aesthetic attributes [35]. The insights gained from these studies also led to the development of Generative Adversarial Networks (GAN) based models that can modify images to manipulate their memorability [36–38].

Video memorability has fewer works, typically evaluated on *VideoMem* [1] and *Memento10k* [2]. The semantic embeddings model of *VideoMem* [1] uses an image-captioning pipeline in conjunction with a 2-layer MLP for memorability prediction. *SemanticMemNet* [2] integrates visual cues with semantic information and decay patterns to predict memorability. Recent approaches involve multiple tiered representation structures, *M3S* [39], or use Large Language Models (LLMs) to generate textual descriptions that are then used to predict memorability scores [40]. In contrast, we adopt a simple CNN+Transformer attention-based model that matches SoTA, but also facilitates comparison between model attention and human gaze on semantic and temporal aspects of video memorability.

Chapter 3

Automatic Generation of Audio Descriptions

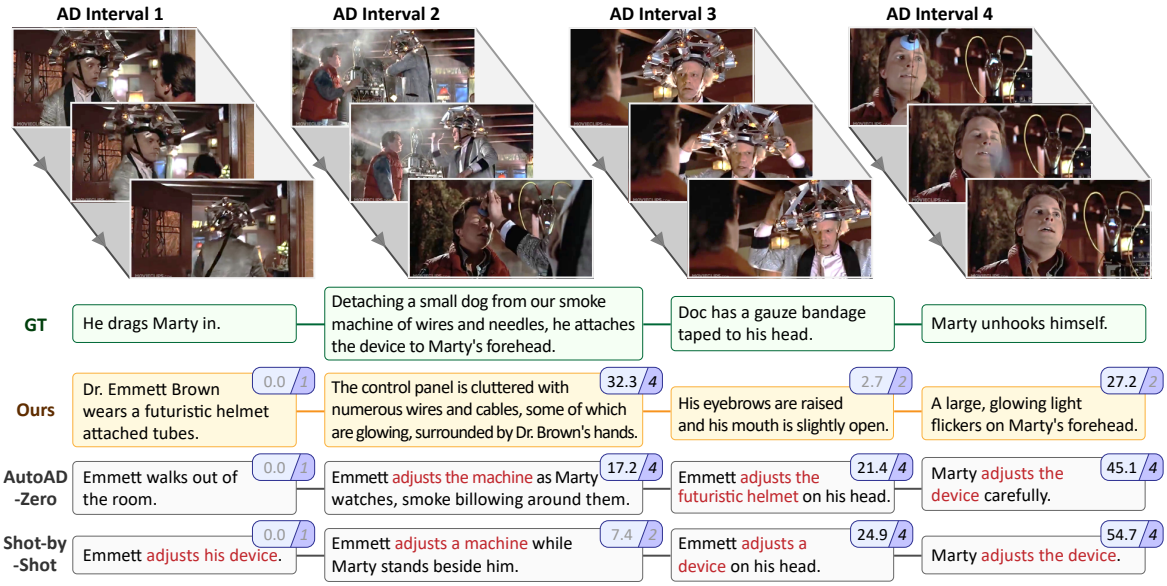


Figure 3.1 Predicted ADs across the video (i.e. a sequence of AD intervals). The results reported by per-AD evaluation metrics are shown on the top right of each prediction (left: CIDEr; right: LLM-AD-Eval, score 1-5), with low scores indicating poor performance coloured in grey. The repetitions across predictions are highlighted in red, where "adjust the device" is repeated multiple times. The video is sampled from the movie *Back to the Future*, corresponding to $0\colon 22 - 1\colon 05$, that can be watched here: `https://www.youtube.com/watch?v=SR5BfQ4rEqQ&t=22s`.

Audio Descriptions (ADs) help Blind and Visually Impaired (BVI) audiences follow a movie or any type of long video typically conveying a story. Narrated between dialogues, ADs often describe key visual elements of the scene, with emphasis on the setting, actions, and characters. While ADs are typically created by professionals [41], there is growing interest in generating ADs automatically [3, 4, 6, 9, 10, 12, 16, 17] to make content accessible.

AD generation is historically treated as video captioning, i.e. a short description is generated independently for each predefined *AD interval* in the video [6, 11, 42]. However, ADs are a coherent sequence of descriptions that build a visual story and take the narrative forward. As seen in Fig. 3.1,

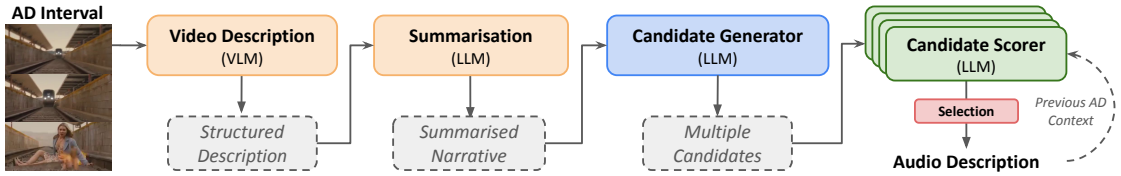


Figure 3.2 Overview of our multi-stage AD generation pipeline CoherentAD. For each AD interval, the VLM generates a structured description, which is then summarised. The summary is used to produce multiple candidate descriptions. Each candidate is scored by four independent LLM-based scorers that consider previous selections as context. The highest-scoring candidate is selected in an auto-regressive manner to form a coherent sequence.

this independent generation (e.g. AutoAD-Zero [12], Shot-by-Shot [16]) results in repeating similar information, and failing to capture the narrative structure of the movie.

We therefore posit that AD generation should be performed over a longer temporal extent, i.e. across the *video* that consists of a *sequence* of *AD intervals*. This is motivated by the subjective nature of ADs, where information is often distributed across multiple descriptions (see Fig. 3.1).

In this chapter, we propose a new training-free method, **CoherentAD** that encourages the generation of diverse visual descriptions across the video. Similar in spirit to AutoAD-Zero [12], we first extract structured information from each trimmed clip (the span of video corresponding to the AD interval). By contrast, we generate multiple AD-like candidate descriptions for each clip and auto-regressively choose one that would advance the narrative while also providing new visual details.

3.1 Training-Free Sequence-Level AD Generation

Given a sequence of predefined intervals in a video, our goal is to generate a corresponding set of ADs that together form a coherent narrative. To this end, we propose **CoherentAD**, a training-free pipeline that generates multiple candidate ADs for each interval and selects a coherent, non-redundant sequence from these candidates.

This section outlines the three components of our multi-stage pipeline. In Sec. 3.1.1, we extract structured descriptions for each interval and summarise them into concise, narratable paragraphs. In Sec. 3.1.2, we generate diverse candidate ADs from these summaries. Finally, in Sec. 3.1.3, we select an optimal sequence of ADs using a scoring mechanism that promotes coherence and informativeness. An overview of the full pipeline is shown in Fig. 3.2.

3.1.1 Video Clip to Summarised Narratives

We uniformly sample 16 frames per AD interval, following standard practice in prior work [6, 8, 12]. For character recognition, we adopt AutoADZero's [12] method of overlaying coloured circles on detected faces, which are then referenced in the text prompts.

We find that when a Vision-Langauge Model (VLM) is shown a sequence of sampled frames, it often defaults to static, high-level descriptions (e.g. "a man is standing"), failing to capture the progression of actions and visual changes over time. To address this, we introduce a two-stage process: first extracting a structured, detailed representation of the visual content, and then summarising it into a concise narrative suitable for AD generation.

3.1.1.1 Extracting Structured Visual Description

We prompt the VLM with a structured instruction that guides the model to extract all visually relevant details from the interval across three aspects—key actions, interactions, and environmental changes. The exact prompt is shown in Algorithm 1, which consists of three components:

- *Storyboard Description* - a step-by-step narration of events in order, treating the frames like a storyboard: a sequence of images that captures the key moments of a scene;
- *Character and Object Breakdown* - a list of all visible characters and objects, along with their observable actions (both prominent and subtle), interactions, and any environmental changes;
- *Overall Summary* - a brief description summarising the primary visual event.

This structured prompt produces a comprehensive representation of the interval. Since later stages rely solely on text, any missed detail at this stage is unrecoverable, making it essential to capture all relevant content. Thorough extraction at this stage also supports the generation of diverse and faithful AD candidates later in the pipeline.

3.1.1.2 Narrative Summarisation

While the structured visual description is exhaustive, it is fragmented and unsuitable for generating concise candidate ADs. We instruct the Large Language Model (LLM), using Algorithm 2, to rephrase this content into a single paragraph that retains all meaningful visual information. In doing so, it re-organises disjoint observations into a concise paragraph where visual details are presented in a more logically connected manner.

The prompt explicitly enforces compliance with AD guidelines by discouraging inference, dialogue, and speculation. To meet these constraints, we prompt the LLM to draft and iteratively refine the paragraph until all conditions are satisfied. The final output, denoted as paragraph P serves as the basis for the next stage.

3.1.2 Multiple Candidate Generation

Given paragraph P, we use an LLM to generate up to m diverse candidate ADs that are: (i) independent, (ii) concise, (iii) group related visual events, and (iv) collectively cover all content from the previous stage. Each candidate conveys a complete visual moment by grouping related elements (such

as actions, objects and context) into a compact description. This makes every candidate meaningful and self-sufficient, suitable for inclusion in the final sequence on its own.

We generate up to m candidates per interval, but stop early once all salient content in P has been covered. To avoid redundancy and prevent fragmentation (such as splitting a cohesive event across multiple descriptions) the prompt instructs the LLM to produce the smallest number of candidates needed to convey all relevant information. This constraint keeps the candidate pool focused, encourages grouping of related observations, and minimizes repetition across candidates. As a result, downstream selection becomes more tractable: given the previously selected description, the model chooses the most coherent next step from a concise and diverse set of alternatives.

Candidates are constrained by a word limit l_{max}, computed based on the AD interval's duration and a narration speed adapted from Shot-by-Shot [16].

See Algorithm 3 for the prompt used in this stage.

3.1.3 Coherent Sequence Selection

Finally, we select one candidate per interval to construct the full AD sequence. Each candidate is evaluated using four independent scoring criteria, with scores assigned separately by dedicated LLM prompts to ensure focused and unbiased assessment along each dimension. Each candidate is scored independently, with the prompt conditioned on the previous $r = 3$ selected candidates to ensure coherence with prior context.

The final score is computed as a weighted average of these four criteria, and the highest-scoring candidate is selected in an autoregressive manner, conditioned on the previous r selected descriptions.

The four evaluation criteria are:

1. **Adherence to AD guidelines** ensures the description is strictly grounded in what is visually perceivable (i.e., avoids inferences, speculation, or camera references) and focuses only on information inaccessible to blind viewers;

2. **Redundancy with prior descriptions** penalises repetition to avoid wasting narration time and to leave space for previously unmentioned content;

3. **Contribution to story advancement** prioritises candidates that introduce new observable actions, interactions, or scene changes that move the visual narrative forward;

4. **Counts of visual elements** measures the number of unique participants, actions, and salient visual details explicitly mentioned in the description, rewarding candidates that convey more information.

This scoring setup allows the system to reliably select a coherent, informative, and narratively rich description for each interval in the sequence. All individual scores are normalised to the range $[0, 1]$. We then apply the following weights: 0.40 for *Adherence to AD guidelines*, 0.25 for *Redundancy*, 0.40

for *Story Advancement*, and 0.29 for *Counts*. The "*Counts*" score is computed by combining sub-scores for Participants (0.13), Actions (0.11), and Salient Details (0.05). Each of these scores is computed using an independent LLM call with a dedicated prompt. Prompts for each criterion can be found in Algorithms 4 to 7.

3.2 Qualitative Results

Qualitative results are shown in Fig. 3.1 and further illustrated in Figs. 3.3 and 3.4, where repetitions in baseline predictions are clearly visible. In contrast, our method introduces distinct, visually grounded details at each interval. The selected descriptions collectively form a coherent and non-redundant narrative.

We present a quantitative evaluation of the pipeline in Sec. 4.3.

3.3 Implementation details.

All experiments use Qwen2-VL-7B [43] as the VLM and LLaMA3.1-Instruct-8B [44] as the LLM. We generate up to $m = 5$ candidates per interval and condition scoring on the $r = 3$ previously selected candidates. The word limit used in Sec. 3.1.2 is computed based on a narration speed of 0.275, adapted from Shot-by-Shot [16].

3.4 Challenges Faced

Generating a coherent sequence of ADs involves several challenges.

Temporal misalignment between ADs and video. ADs in movies are inserted between dialogues, not necessarily when the described visual event occurs. As a result, a given AD may refer to content that appears before or after the associated time span. In our experiments on CMD-AD [6], we find that 40% of GT ADs are not visually entailed by the corresponding trimmed video clip. This introduces a fundamental mismatch—models are expected to describe or evaluate content that may not even be visible within the predefined interval.

Limited coverage in VLM generations. Even among the 60% of ADs that are locally entailed, only a subset (1139 out of 4431) are captured in the VLM-generated structured descriptions (in Sec. 3.1.1). These outputs aim to exhaustively list visible entities and actions, but often miss key details that are central to the narrative. This reflects current limitations in visual parsing and grounding. While we expect improvements with the development of more robust VLMs, current models fall short of what is needed for high-quality AD generation.

Challenges in Sequence-Level Selection. Given five candidate descriptions per interval, the goal is to select the most coherent sequence. Beam search was initially considered to optimise selection across intervals, but this requires a reliable scoring function, something that existing metrics do not provide.

Metrics like *CIDEr* [21] and *BERTScore* [45] rely on one-to-one matches with a reference, which is problematic in itself due to the existence of multiple valid ADs per interval (discussed further in Chapter 4). Moreover, their behaviour is unreliable: *CIDEr* often spikes for candidates with a few matching words, even when the full description is suboptimal, and remains uniformly low otherwise. It may also assign low scores even when two descriptions convey the same meaning but use different surface forms—for instance, *"grabs a slice"* and *"grab slices"* fail to align completely despite having near-identical semantics. Such variation is often acceptable in ADs. Additionally, *BERTScore* fails to meaningfully differentiate between the five candidates, with very small score variation.

To go beyond reference matching, we explored using sentence-level *BERT Next Sentence Prediction (NSP)* scores [46] to evaluate transitions between descriptions. However, these scores cluster around 0 or 1, offering little discrimination due to their binary training objective.

In short, there was no meaningful way to compute either independent candidate scores or transition scores for use in beam search. In the absence of reliable scoring functions for beam search, we instead adopt a greedy, autoregressive selection strategy that chooses each candidate based on its compatibility with preceding context.

Supervised learning is not viable. Even when the ground-truth (GT) AD is among the candidates, existing metrics do not reliably rank it highest. This makes it impossible to assign consistent labels for supervised training. We also experimented with using LLM or VLM log-probabilities (the model's internal confidence scores for generating the sentence) for ranking, but both tend to favour generic, template-like outputs, lacking sensitivity to narrative quality or visual grounding. As a result, there was no usable signal to train a model to pick the best candidate.

Limitations of Direct LLM-Based Selection/Ranking We also tried prompting the LLM to choose the best candidate among the five directly, instead of scoring each one independently. However, this task proved too complex. It requires the model to reason over many overlapping descriptions, apply multiple criteria (coherence, salience, redundancy), and identify subtle differences. With multiple candidates per interval, the model's outputs became erratic and inconsistent—it failed to reliably compare and rank closely related descriptions. The LLM struggled to apply so many constraints across dense sets of alternatives. To make the task more manageable, we broke it down into simpler, well-defined criteria and scored each candidate independently.

These challenges motivate our design of a training-free, scoring-based approach.

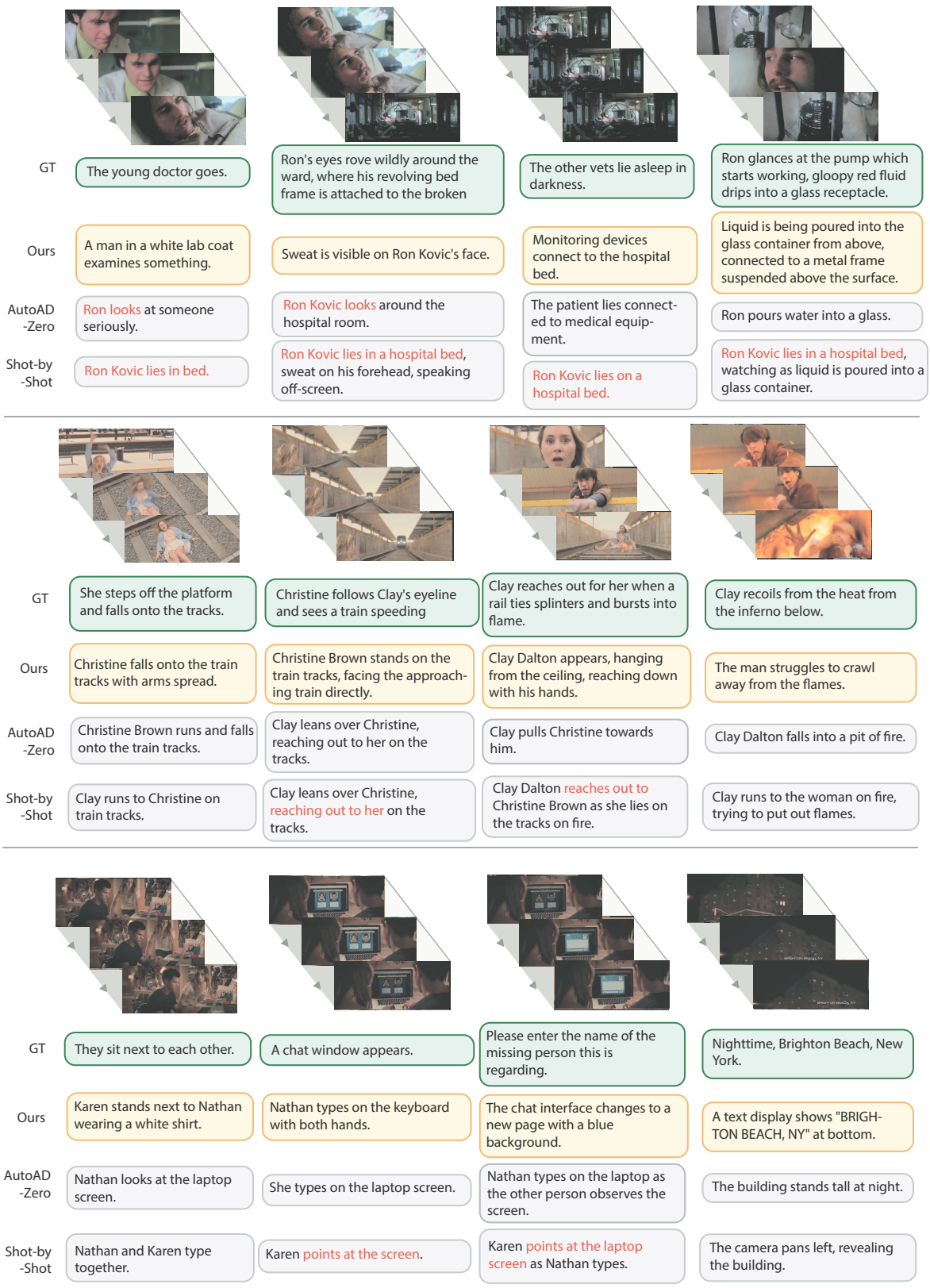


Figure 3.3 Qualitative comparison showing GT, our outputs, AutoAD-Zero [47] and Shot-by-Shot [16], with repetitions highlighted in red.

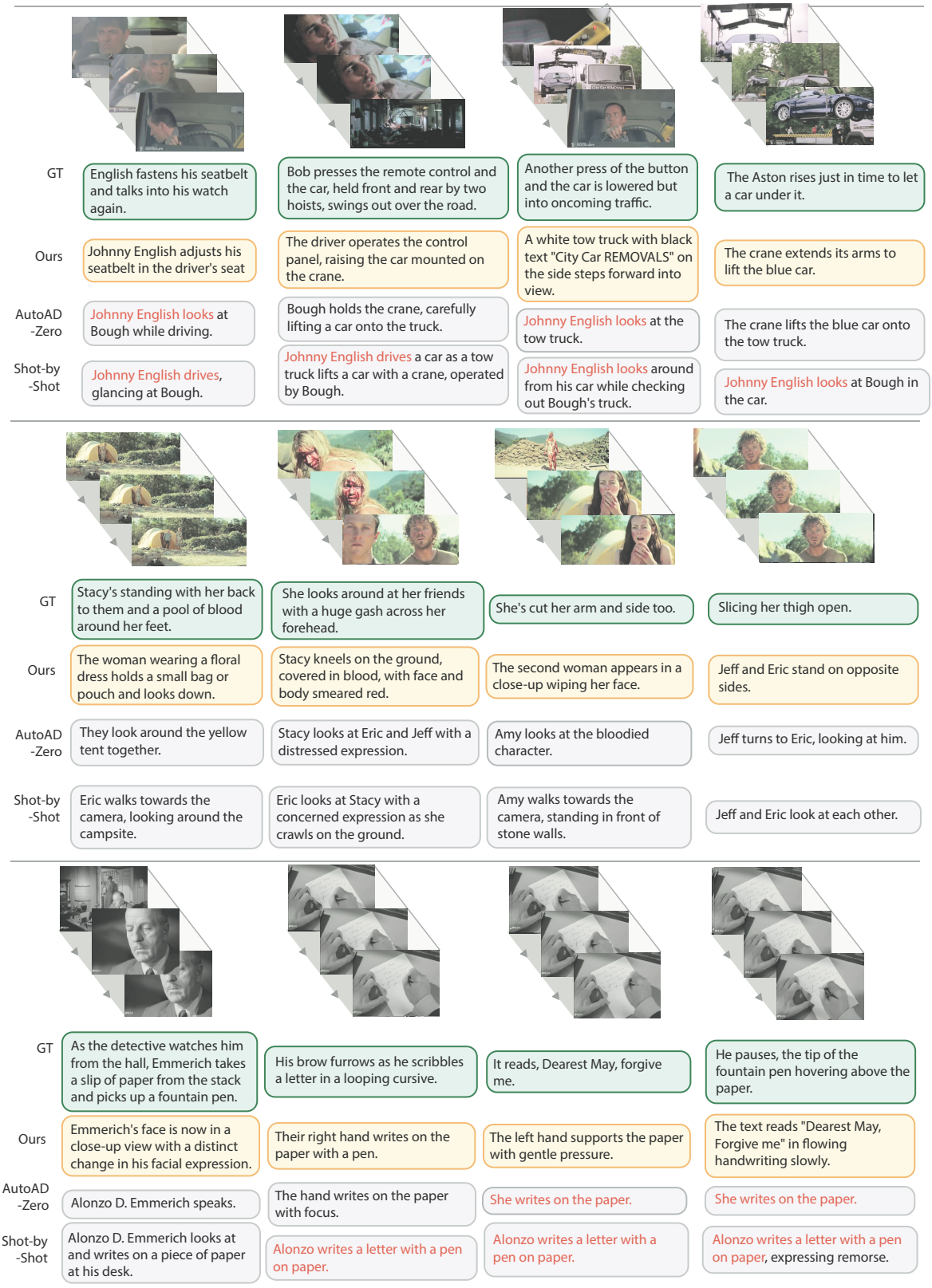


Figure 3.4 Qualitative comparison showing GT, our outputs, AutoAD-Zero [47] and Shot-by-Shot [16], with repetitions highlighted in red.

Algorithm 1 Stage I prompt for extracting structured visual descriptions.

```
user_prompt = (
    "Describe the video segment in detail using a three-part structure:\n"
    "1. **Storyboard Description**\n"
    "   - Describe the events in the order they happen, step by step.\n"
    "   - Use words like 'first,' 'next,' 'then,' 'finally' to show the timeline.\n"
    "   - Include clear descriptions of the location, background, objects, visible text, and any
    changes in the setting (e.g., lights switching on, doors opening, smoke appearing).\n"
    "   - Mention any uncertainty if something is unclear.\n\n"

    "2. **Character and Object Breakdown**\n"
    "{char_text}\n"
    "Break this section into these parts:\n"
    "   a. **Characters**\n"
    "      - List all visible people or animals (named or unnamed).\n"
    "      - For each character, provide two separate points:\n"
    "         i. **Visible Actions** - Big, clear movements (e.g., 'walks to the door,' 'sits on the
    chair'). If none are seen, write 'None.'\n"
    "         ii. **Subtle Actions** - Small, visible movements or facial changes (e.g., 'raises
    eyebrows,' 'nods'). If you believe the character seems to display an emotion (e.g., 'concerned,'
    'nervous,' 'relieved'), clearly state the physical cues you observed that led you to that
    interpretation. Avoid stating an emotion without citing the visible evidence. Be specific about
    gestures: say what the character does with which hand, where they point, etc. If none are seen,
    write 'None.'\n"
    "   b. **Character-Character Interactions**\n"
    "      - List any visible interactions between characters (e.g., touching, eye contact).\n"
    "   c. **Objects**\n"
    "      - List any important objects seen in the scene.\n"
    "      - Describe how they look, where they are, and any changes they go through.\n"
    "   d. **Character-Object Interactions**\n"
    "      - Describe how characters use or touch objects (e.g., 'opens drawer,' 'picks up phone').\n"
    "   e. **Changes in Environment**\n"
    "      - List any visible changes in the surroundings (e.g., light turns off, car drives in).\n\n"

    "3. **Overall Summary**\n"
    "   - Summarize, in 1-2 sentences, the main action or event that occurs in the clip.\n"
    "   - Mention who is primarily involved (characters and/or objects).\n\n"

    "Important rules:\n"
    "- Only describe what you can see clearly.\n"
    "- Don't guess what characters are thinking or feeling unless it's visible on their face or body.\n"
    "- Say if anything is unclear or hard to see.\n"
)
```

Algorithm 2 Stage I prompt for narrative summarisation.

```
system_prompt = (
    "You are a visual summarization expert trained to convert detailed scene descriptions into
    clean, narratable paragraphs. "
    "Your output is used as audio description for blind and visually impaired users, so it must:\n\n"
    "* Be strictly grounded in the input - only describe what is clearly visible in the source.\n"
    "* Use concrete physical verbs in present tense and active voice.\n"
    "* Avoid all emotions, sounds, dialogue, intentions, or inferences.\n"
    "* Convert static adjective phrases (like 'crossed arms') into dynamic verb phrases (like
    'crosses arms') only if those states appear in the source.\n"
    "* Use first names **only if** explicitly and visually grounded in the input. Never guess
    identities.\n\n"
    "Your paragraph should include all meaningful visual actions, facial or body movements, and
    environmental changes. "
    "Keep it crisp, literal, and compliant - no more, no less."
)

user_prompt = (
    "**TASK: Convert the input scene description into one concise, literal paragraph.**\n"
    "Describe only visible, physical events - include body actions, interactions with objects,
    posture or expression changes, and setting changes.\n\n"
    "**Guidelines**\n"
    "* Use present-tense, active voice.\n"
    "* Use concrete physical verbs (e.g., 'raises hand', 'steps forward').\n"
    "* Include subtle but visible actions (e.g., glances, nods, clenches fists).\n"
    "* Describe static elements only if they clarify the action.\n"
    "* Convert adjective states to verbs only if they appear in the input (e.g., 'crossed arms' ->
    'crosses arms').\n"
    "* Ensure all visual events are described once and only once.\n"
    "* Use first names of the characters only if present and clearly grounded. Otherwise, use
    specific roles or generic labels.\n"
    "* Never guess or add any action, name, or object not grounded in the source.\n"
    "**Strictly Remove**\n"
    "* Dialogue or speech verbs (speaks, talks, responds).\n"
    "* Emotions or mental states (nervous, concerned).\n"
    "* Sounds or spoken content (conversation, laughs, screams).\n"
    "* Uncertain or speculative language (e.g., 'seems to', 'possibly').\n"
    "* Camera or framing language (off-screen, towards the camera).\n"
    "* Visual markers like colored circles (red circle, green circle, blue circle, yellow circle).\n"

    "**Step 1 - Draft**\n"
    "Write a first-pass paragraph that follows all points under **Guidelines** and **Strictly
    Remove**.\n\n"

    "**Step 2 - Refine**\n"
    "Reread your draft paragraph and revise until it passes all four checks:\n\n"
    "1. **Forbidden Terms Check**\n"
    "   * Remove any mention of emotions, thoughts, sounds, dialogue, camera angles, or visual
    markers.\n"
    "   * Refer to the **Strictly Remove** section above.\n"
    "   * Examples to remove: 'nervous', 'appears to', 'camera pans', 'red circle', 'speaks',
    'indicating', 'green circle'.\n\n"
    "2. **Hallucination Check**\n"
    "   * Do not add any name, object, action, or interpretation that is not grounded in the
    input.\n"
    "   * If it's not clearly visible in the scene description, leave it out.\n\n"
    "3. **State-to-Verb Rewrite**\n"
    "   * Convert static descriptions to visible actions only when the original text implies
    motion.\n"
    "   * Example: 'with arms crossed' becomes 'crosses arms' (if and only if supported by
    input).\n\n"
    "4. **Coverage Check**\n"
    "   * Include every meaningful, visible action or interaction exactly once.\n"
    "   * Do not omit any relevant gestures, postures, or setting changes.\n\n"
    "Repeat this loop until the paragraph fully satisfies all checks and follows the **Guidelines**
    and **Strictly Remove** rules.\n\n"

    "**Output**\n"
    "Output only the final paragraph, nothing else.\n\n"
    "**Scene Description**\n"
    f"{text}\n"
)
```

Algorithm 3 Stage II prompt for multiple candidate generation.

```
system_prompt = (
    "You are a professional audio description writer.\n"
    "You convert summaries into concise, present-tense descriptions that cover only what can be
    physically visible on screen.\n"
    "You never speculate, interpret, or describe sound or speech.\n"
    "You group related visual details into single sentences when they belong to the same moment or
    subject.\n"
    "Your writing is literal, compact, and strictly grounded in visual facts.\n"
    "You stay within the word limit.\n"
    "Every word you use adds concrete visual value.\n"
    "You never include filler words or vague phrasing.\n"
)

user_prompt = (
    "You are given a paragraph that summarizes the visual content of a video clip. "
    "Your task is to convert this into up to 5 candidate audio descriptions (ADs) that strictly
    follow professional AD guidelines. "
    "Each candidate must be a complete sentence in present tense, describing only what can be
    physically visible. "
    "Do not infer or interpret anything beyond what can be explicitly seen.\n\n"

    "Follow these detailed instructions:\n"
    "1. Visual-Only Content:\n"
    "Describe only what can be physically visible: people, actions (both prominent and subtle),
    interactions, objects, spatial layout, and environmental context.\n"
    "Do NOT include:\n"
    "* Emotions or internal states\n"
    "* Intentions or speculation\n"
    "* Sounds or speech-related verbs\n"
    "* Any inferred meaning or visual interpretation\n"
    "* Camera or viewer references\n"
    "* Filler words or vague phrases\n"
    "* Any colored circles (they are not meaningful scene elements)\n\n"

    "2. Sentence Structure:\n"
    "* Use present tense only.\n"
    "* Each candidate must be complete and self-contained.\n"
    "* Keep sentences concise - no filler or padding.\n"
    f"* Each sentence should aim to be exactly {num_words} words, "
    "but only include words that convey clear visual information. Do not add words just to meet the
    target.\n\n"

    "3. Group Related Observations:\n"
    "Each sentence should describe a complete and coherent visual moment.\n"
    "Cluster visual details that naturally belong together, such as:\n"
    "* a person's action, posture, and gesture\n"
    "* a person's facial expression, gaze direction, and position in the scene\n"
    "* multiple people engaged in a single visible interaction\n"
    "* people jointly focused on a shared object or action\n"
    "* movement through a space with visible layout or surrounding elements\n"
    "* people and objects arranged in the same spatial scene\n"
    "* multiple characters described together by their clothing, positioning, or appearance\n"
    "* an object and its placement, motion, or use in the scene\n\n"
    "Avoid splitting visual details that form a single visual moment "
    "- even if they are brief or subtle. "
    "Only separate background elements if they clearly relate to different subjects or actions.\n"
    "If a gesture, facial movement, or small action is part of one visual act, group it with related
    observations.\n\n"

    "4. Naming Conventions:\n"
    "* Use first names for characters if available.\n"
    "* Do not use full names or titles - first names are enough.\n\n"

    "5. Language Precision:\n"
    "* Do not use vague, redundant, or filler phrases such as: 'is visible', 'can be seen', 'in the
    background'.\n"
    "* Prefer direct phrasing.\n"
    "* Avoid repetition across candidates.\n\n"

    "6. Candidate Count:\n"
    "* Generate the minimum number of candidates needed to fully cover the paragraph - up to 5.\n"
    "* If 2-4 are sufficient, stop there. Do not force 5.\n"
    f"* Remember, each candidate must aim to be {num_words} words.\n"

    "Format your output as a numbered list of 1-5 sentences, with no extra text.\n\n"

    "Here is the narrative paragraph:\n"
    f"{text} \n"

    "Now generate the candidate audio descriptions."
)
```

Algorithm 4 Stage III prompt for scoring adherence to AD guidelines

```
system_prompt = (
    "You are a precise and fair rule-checker. You only lower the score when a description clearly
    breaks one of the defined rules.\n"
    "Be especially careful to catch:\n"
    "* Any mention of inferred emotions or internal states (e.g., 'worried', 'nervous', 'concerned',
    'frustrated')\n"
    "* Any reference to speech or dialogue (e.g., 'talks', 'speaks', 'conversation')\n"
    "* Any reference to the camera perspective or the viewer (e.g., 'off-screen', 'away from the
    camera')\n"
    "* Any mention of coloured circles (e.g., red/green/blue/yellow circles)\n"
)


user_prompt = (
    "You are evaluating the following description for rule violations.\n\n"
    "Description:\n"
    f"{candidate}\n\n"
    "Check whether the description breaks any of the following rules:\n"
    "* Explicit mentions of emotion or internal state (e.g., 'nervous', 'worried', 'concerned',
    'frustrated')\n"
    "* Descriptions of speech or conversation (e.g., 'talks', 'speaks', 'discusses', conversation')\n"
    "* References to the camera perspective, screen, or the viewer (e.g., 'off-screen', 'in front of
    the camera', 'toward the camera')\n"
    "* Mentions of coloured circles (e.g., 'red circle', 'blue circle', 'green circle', 'yellow
    circle')\n\n"
    "Scoring:\n"
    "* 3 = Fully compliant - no rule violations\n"
    "* 2 = Partially compliant - minor violation, mostly adheres to rules.\n"
    "* 1 = Non-compliant - major rule violation(s)\n\n"
    "Important: Only give a score below 3 if the description clearly breaks one of the listed
    rules.\n\n"
    "Clarifications:\n"
    "* A minor violation means the description still conveys meaningful visual content on its own,
    despite the violation.\n"
    "* Do not penalize vague, brief, or underspecified descriptions.\n"
    "* Facial expressions (e.g., raising eyebrows), head movements (e.g., turning), and eye movements
    (e.g., looking around, glancing) are permitted unless they include clear emotion words (e.g.,
    'worried', 'nervous', 'concerned', 'frustrated').\n"
    "* Mouth movements (e.g., mouth opens) are fine unless they clearly imply speech.\n"
    "* Describing a camera or screen (e.g., TV, monitor) as an object is fine - only references to
    the camera's perspective or the viewer are violations.\n"
    "* Coloured circles are always violations.\n"
    "Examples:\n"
    "Score 3: 'The man furrows his brow and picks up a gun and a camera.' (no rule violations)\n"
    "Score 2: 'The woman in a red dress walks toward the door, looking tense.' ('looking tense'
    describes emotion/internal state - minor violation; the main action 'walking toward the door'
    still remains meaningful on its own)\n"
    "Score 1: 'The camera zooms in on a man with a red circle.' (references to camera and coloured
    circle - major violations)\n\n"
    "First, list any rule-relevant observations in 1-2 sentences. If a rule is broken, note whether
    the violation is minor or major. Then, assign a score from 1 to 3.\n"
    "Output your response in the following format:\n"
    "Observations: [...]\n"
    "Score: [1-3]"
)
```

Algorithm 5 Stage III prompt for scoring redundancy.

```
system_prompt = (
    "You are a redundancy evaluator.\n"
    "Your job is to judge how much new information a candidate description adds beyond the previous
    description(s).\n"
    "Do not go out of your way to find small or indirect overlaps "
    "- only count something as repeated if its meaning clearly matches what was already described.\n"
    "Default to a score of 3 unless the candidate clearly, without any assumptions, repeats an event
    (action, interaction, or visible state) that was already described.\n"
    "If no new event is introduced, assign a score of 1.\n"
)

user_prompt = (
    "You are checking how much new information the candidate description adds to the previous
    description(s).\n\n"
    "Compare the candidate with the previous description(s) and judge how much of the content is
    new.\n\n"
    "Previous description(s):\n"
    f"{current_desc}\n\n"
    "Candidate description:\n"
    f"{candidate}\n\n"
    "Assign a score from 1 to 3 based on the following criteria:\n"
    "* 1 = Almost all content is already stated - no new event is introduced.\n"
    "* 2 = Some content is new - a clearly repeated event is present, but a new event is also
    described.\n"
    "* 3 = Most of the content is new - no events are repeated.\n"
    "Clarifications:\n"
    "* People (including their appearance if unnamed), objects, or locations do not count as repeated
    content.\n"
    "* If the candidate continues a prior event, do not treat it as repetition if it adds clearly new
    and meaningful actions or visual developments.\n"
    "* Only assign a lower score if a clearly repeated event (action, interaction, or visible state)
    is present.\n"
    "* Do not go out of your way to find subtle or indirect overlaps - only count something as
    repeated if its meaning clearly matches what was already described.\n"
    "Examples:\n"
    "* Score 3:\n"
    "    Previous: 'A man in a red shirt is picking up a gun from the table.'\n"
    "    Candidate: 'A man in a red shirt is looking at a gun and smiling.'\n"
    "    (All events, looking at the gun, smiling, are new and not mentioned before.)\n\n"
    "* Score 2:\n"
    "    Previous: 'Jim walks toward the door.'\n"
    "    Candidate: 'Jim approaches the door and opens it with his left hand.'\n"
    "    (One event, approaching the door, is similar to walking toward it, but the second event,
    opening the door, is clearly new.)\n\n"
    "* Score 1:\n"
    "    Previous: 'The car with its headlights on drives forward through the intersection.'\n"
    "    Candidate: 'The car moves forward with its headlights on.'\n"
    "    (The same event, moving forward with headlights on, is repeated in different words. No new
    events are added.)\n\n"

    "Instructions:\n"
    "Write your observations in 1-2 sentences explaining how much of the candidate's content is
    new.\n"
    "Then assign a score from 1 to 3.\n\n"
    "Important: Default to a score of 3 unless the candidate clearly, without any assumptions,
    repeats an event already described. If no new event is introduced, assign a score of 1.\n"
    "Output format:\n"
    "Observations: [...]\n"
    "Score: [1-3]"
)
```

Algorithm 6 Stage III prompt for scoring story advancement.

```
system_prompt = (
    "You are a visual narrative progression evaluator.\n"
    "Your job is to judge how much a candidate description advances the scene beyond the previous
    description(s).\n"
    "Treat the previous description(s) as the current state of the scene.\n"
    "Focus on new actions, interactions, or changes that clearly affect what is happening in the
    scene.\n"
    "Minor movements or posture shifts (e.g., turning, walking, looking around) should only be scored
    higher if they cause a clear shift in focus, direction, or interaction.\n"
    "Descriptions of appearance, background, or static visual elements should receive the lowest
    score unless they visibly affect the scene.\n"
    "Base your evaluation only on what is explicitly stated. Do not infer intent, emotions, or
    consequences that are not shown.\n"
)

user_prompt = (
    "You assess whether a candidate description advances the visual narrative beyond the previous
    description(s).\n\n"
    "Carefully read the previous description(s) and the candidate.\n\n"
    "Previous description(s):\n"
    f"{current_desc}\n\n"
    "Candidate description:\n"
    f"{candidate}\n\n"
    "Evaluate what new visual information the candidate explicitly adds. "
    "Look for new actions, interactions, or visually meaningful changes.\n"
    "Scoring Criteria:\n"
    "* 5 = Major action, event, or change that clearly advances the scene.\n"
    "  Example: 'The man pulls the trigger, and the gun fires.' (Highly significant change in the
    scene)\n"
    "* 4 = Clear action, interaction, or change that adds meaningful development to the scene.\n"
    "  Example: 'She picks up the phone from the table.' (Initiates a new event)\n"
    "* 3 = Minor action or movement that slightly advances the scene by shifting focus, direction, or
    interaction.\n"
    "  Example: 'The boy steps away from the table.' (shift in position)\n"
    "* 2 = Minor gestures or visual details that add tone or context but do not affect what is
    happening in the scene.\n"
    "  Example: 'The woman sits at her desk.' (No change in the scene)\n"
    "* 1 = Static visual detail with no narrative impact.\n"
    "  Example: 'A lamp rests on the side table.' (No change in the scene)\n\n"
    "Important:\n"
    "* Score based on whether the candidate changes the current state of the scene.\n"
    "* Descriptions that visibly change the course of events or introduce new interactions should
    score higher.\n"
    "* Minor actions or gestures with no effect on others or the unfolding situation should score 2
    or lower.\n"
    "* Purely descriptive details about appearance, background, or already-known elements should
    score 1.\n"
    "* Onscreen text should be scored by its narrative impact. If it introduces new facts or reframes
    the scene, it may merit a 3-5.\n"
    "* Only use the information explicitly shown in the candidate. Do not assume anything beyond what
    is described.\n"
    "Describe what the candidate contributes to the ongoing scene in 1-2 sentences. Then assign a
    score from 1 to 5.\n"
    "Output Format:\n"
    "Observation: [...]\n"
    "Score: [1-5]"
)
```

Algorithm 7 Stage III prompt for counting visual elements.

```
system_prompt = (
    "You are an expert in structured scene parsing.\n"
    "Extract only explicit, observable, and non-redundant visual details from a description.\n"
    "Each item must be counted exactly once - no duplicates within or across categories.\n"
    "Only include elements that are clearly described and visually relevant to the described event.\n"
    "Participants and Actions must play a central role in the described event. "
    "If an entity or action is not clearly central, demote it to 'Other Details'.\n"
)

user_prompt = (
    "You are given a scene description. "
    "Extract and count only the most visually **salient** elements under the following two
    categories:\n\n"

    "----\n\n"

    "1. **Participants**"
    "* Include only people, animals, or objects that play a **visually central and narratively
    important role**.\n"
    "* Do **not** include someone just for being present or named. They must be doing something
    important, or something important must be happening to them. \n"
    "* Prioritize scenes with **multiple active entities** - especially if they are interacting
    meaningfully. \n"
    "* Ask: *Would this participant make the moment feel different if removed?*\n\n"

    "**Valid examples:**\n"
    "- woman covered in blood\n"
    "- man pointing a gun\n"
    "- child gripping a torn photo\n\n"

    "**Invalid examples (unless clearly emphasized):**\n"
    "- person walking\n"
    "- woman seated in the background\n"
    "- man standing\n\n"

    "----\n\n"

    "2. **Other Details**\n"
    "* Include only **striking descriptive elements** - things that change the tone, reveal something
    dramatic, or stand out visually.\n"
    "* Focus on things like blood, injuries, fire, smoke, damage, or strong emotional expressions.\n"
    "* Do **not** include ordinary background elements, red/green circles, or routine
    clothing/furniture unless the sentence highlights them as important.\n"
    "* Ask: *Would a blind viewer miss something essential if this detail were skipped?*\n\n"

    "**Valid examples:**\n"
    "- blood on the floor\n"
    "- shattered glass underfoot\n"
    "- smoke billowing from a doorway\n\n"

    "**Invalid examples (unless clearly emphasized):**\n"
    "- red circle, green circle\n"
    "- lamp, couch, hat visible in the background\n\n"

    "----\n\n"

    "**Important Guidelines**\n"
    "- Leave categories empty unless something clearly stands out.\n"
    "- Count only what is **explicitly stated**, not inferred.\n"
    "- Do **not** list anything generic or background unless the sentence signals its importance.\n"
    "- Each detail must be **distinct** and appear in only one category.\n\n"

    "**Output Format (strict):**\n"
    "Participants: <comma-separated list> - <count>\n"
    "Other Details: <comma-separated list> - <count>\n\n"

    "Now extract salient visual content from the following description:\n"

    "Description:\n"
    f"{candidate}\n"
)
```

Chapter 4

Holistic Evaluation of Audio Descriptions

To improve accessibility, there has been a growing interest in automatic Audio Description (AD) generation [3, 4, 6, 8, 9, 12, 15–17]. However, even after development of several new methods and metrics (Sec. 2.1.3), evaluation centres around comparing ground-truth (GT) ADs against predictions for each *trimmed clip*—few second video clips trimmed to the AD narration timestamps. This is a consequence of treating AD generation as a video clip captioning task.

As shown in Fig. 3.1, metrics such as CIDEr [21] and LLM-AD-Eval [6] produce highly correlated scores and often reward predictions that simply mention correct names or objects, while failing to penalise redundancy across outputs or capture the coherence of the overall narrative. Moreover, these metrics enforce matching against a single GT, overlooking the fact that each time interval may encompass multiple valid descriptions. AD evaluation is subjective in nature—two experts may write different ADs for the same video. In Sec. 4.1, we present a thorough analysis of this variability for a subset of movies (with two AD sources).

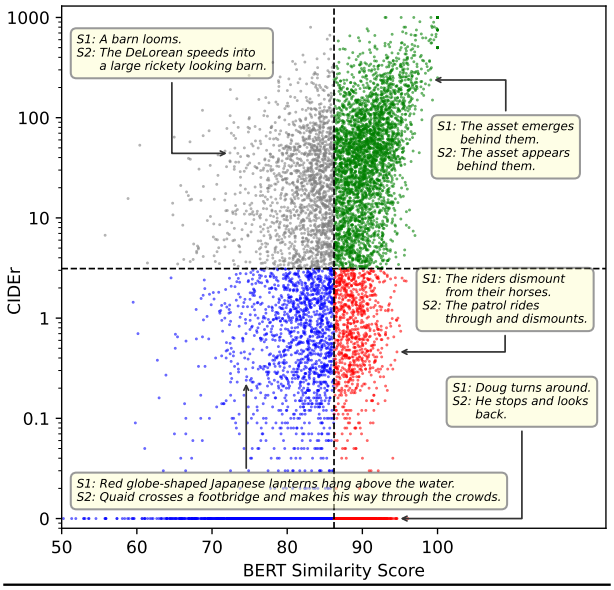
To overcome these issues, for evaluation, we move away from conventional metrics (e.g. CIDEr, LLM-AD-Eval) that compare GT to predicted ADs for a single interval. Instead, we adopt two metrics: (i) *StoryRecall* that captures whether visual details and narrative points mentioned in the GT are conveyed by the predictions; and (ii) *Repetition* metrics that assess the redundancy of generated ADs. In Sec. 4.3, we evaluate sequence-level AD generation on CMD-AD videos and observe qualitative and quantitative improvements in generated ADs. We additionally evaluate on *ADQA* in Sec. 4.4, a benchmark that poses multiple-choice questions to assess whether generated ADs enable blind or visually impaired (BVI) users to perceive visual details and follow the story.

4.1 Takeaways from Two AD Sources

Work under this section was done with Divy Kala, Research Assistant at the same group

The CMD-AD dataset relies on AudioVault [1] as the primary source of ADs. In this section, we study the variability across multiple professional Audio Description (AD) narrations for the same movie.

[1] See `https://audiovault.net/`.



B	C	P %	Reason / Implication
↑	↑	33.8	Same detail with similar words
↑	↓	16.3	Same detail with different words
↓	↓	33.8	Different details for the same clip
↓	↑	16.1	Matched words but poorer semantic match
-	0	26.6	Catastrophic failure, different words

Figure 4.1 BERTScore (B) vs. CIDEr (C) for *time-aligned ADs* of a movie. The quadrants and ↑ or ↓ labels are separated by median scores (B: 86.2, C: 3.1) and the proportion of samples in each quadrant is in P %. We summarize the reasons for these scores in the table. Best seen on screen in color. An interactive plotly chart is included in the supplement.

4.1.1 Aligning Multiple AD Sources

For a subset of movies in CMD-AD, we identify and analyze multiple AD narrations from AudioVault. Some are US vs. UK movie variants[2] while others are multiple narrators for the same movie.

AudioVault hosts movie audio files (including dialog, music, sound effects) on which AD narrations are overlaid. To align AD sources, we follow three steps: (i) obtain timestamped transcriptions using WhisperX [48]; (ii) classify each transcription as AD or dialog using an LLM; and (iii) align the two transcriptions using dynamic time warping [6] anchored via dialog that have strong matches. Apart from a few movies with missing scenes that are treated manually, the above process yields good alignments.

While this procedure is similar in spirit to AutoAD-3 [6], there are two important differences. (i) We observe that using an LLM results in better AD/dialog classification than identifying the narrator's voice. (ii) We align two transcribed sources containing dialog + AD, while AutoAD-3 aligns transcribed CMD videos (dialog) with AudioVault transcriptions (dialog + AD).

[2]In our analysis, the content across these variants is quite similar allowing us to use them for our work.

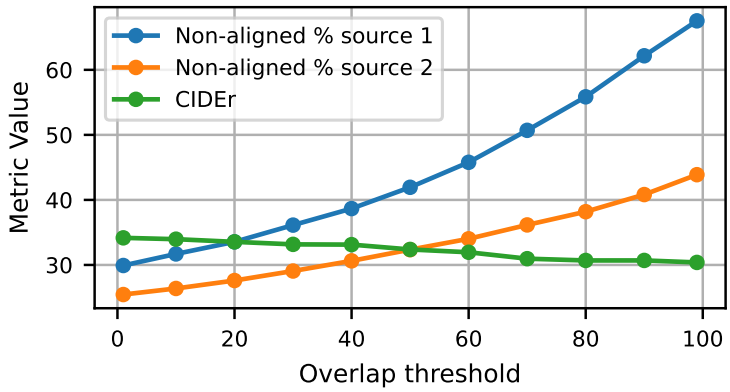


Figure 4.2 Impact of time duration overlap threshold on AD alignment on the two-source subset of CMD-AD movies. The fraction (%) of non-aligned ADs increase with threshold (expected). Interestingly, at low thresholds, 25-30% ADs are not aligned indicating that many ADs in one source are not present in the other. Additionally, CIDEr does not increase with better temporal overlap (high threshold) and is a brittle metric.

Aligned %	Overlap	BERTScore	CIDEr
60.7±16.7	85.6±16.8	85.3±6.4	37.3±97.6

Table 4.1 Results of aligning and mapping ADs between two sources for 17 movies of the CMD-AD dataset. Mean CIDEr of 37.3 is top-line *human performance*. About 39% ADs are unaligned (overlap < 0.5).

Our alignment process provides a timeline of sequentially aligned dialog from both sources, interspersed with ADs. Next, we greedily map ADs from both sources after compensating for the time difference (slope and offset). Specifically, we consider the duration of an AD in source 1, map it to source 2, and pair ADs based on the amount of overlapping duration While the decision to map hinges on an overlap threshold (0.5), Fig. 4.2 shows that even at low thresholds about 25-30% ADs remain *unaligned*.

4.1.2 Similarity between Aligned ADs

For a subset of CMD-AD movies with two AD sources, Tab. 4.1 shows the % of aligned ADs, and average similarity scores: BERTScore [45] and CIDEr [21]. Even for aligned ADs, we observe poor CIDEr scores highlighting the challenges of using N-gram based metrics. Next, in Fig. 4.2 we observe that the fraction of non-aligned ADs increase with the threshold (expected). Surprisingly, among aligned ADs, CIDEr stays constant and even reduces a bit. To investigate further, Fig. 4.1 compares BERTScore against CIDEr for mapped AD pairs. We see 5 important scenarios corresponding to low/high values of the similarity metrics. They highlight the subjective nature of ADs where aligned ADs may describe different details (33.8%), or the same detail, but using different words resulting in poor CIDEr scores (16.3%). Finally, about 27% of the pairs suffer failure and score 0 on CIDEr.

We summarize our *takeaways*: (i) Different AD narrators may describe events at different times resulting in a large proportion of unaligned ADs. (ii) Even among ADs aligned in time, narrators may pick different visual details appropriate for a coherent story, but resulting in low similarity scores. (iii) Different words may be used to describe the same visual detail. Here, N-gram metrics like CIDEr are a bad choice to evaluate ADs as they do not comprehend the semantic nature of ADs.

Overall, these findings advocate that comparing ADs in a one-to-one manner (evaluation like video captioning) is inappropriate as two professionals may often disagree. We need a better evaluation aimed at the heart of AD creation—do they help BVIs appreciate and understand the story.

4.2 Sequence-Level AD Evaluation

Traditional AD evaluation independently compares each description to a single ground truth (GT). This overlooks the fact that ADs are intended to form a coherent sequence that, among other goals, (i) conveys the story, and (ii) remains non-redundant. To address this, we propose two types of metrics: *StoryRecall* that evaluates whether the generated sequence captures the same visual story as the reference GT sequence (Sec. 4.2.1), and *repetition* metrics that measure redundancy (Sec. 4.2.2).

4.2.1 StoryRecall

To evaluate whether the predicted AD sequence captures the key events of the video, we assess how well they recover the storyline conveyed in the GT sequence. Although individual GT descriptions may not align one-to-one with the predicted ADs (due to subjective choices or differing focus, as discussed in Sec. 4.1), the full sequence collectively captures the core visual events, making it a reliable reference.

We construct the GT and predicted sequences by concatenating all ADs within each video. An LLM then compares the two sequences and assigns a score from 1 to 5, reflecting how much of the GT's visual content is conveyed. For example, a score of 5 indicates that the predicted sequence captures nearly all key actions, events, and visual details described in the GT. Note, during the comparison, paraphrasing and reordering are allowed, provided the core visual narrative remains intact. Extra information in the predicted sequence is not penalised. The final score is computed as the average across all videos.

4.2.2 Repetition metrics

Complementary to StoryRecall, we propose explicitly monitoring repetition in AD evaluation, as it can reduce the effectiveness of an AD sequence by wasting valuable narration time and limiting the inclusion of new information. Specifically, we consider two simple repetition measures.

First, we compute the number of *exact repetitions*. For each description, we compare it to the next three consecutive descriptions and check for exact string matches. We then report the proportion of ADs with exact matches across the entire dataset, yielding three percentages—one for each offset.

Method	Training-Free	StoryRecall ↑	Exact Repeat% ↓	Partial Repeat% ↓
Ground Truth	-	-	(0, 0, 0)	(3.71, 4.42, 3.97)
AutoAD-III [6]	✗	2.11	(4.51, 2.81, 1.99)	(16.01, 11.90, 10.44)
UniAD [8]	✗	2.13	(4.15, 2.44, 2.10)	(16.21, 12.27, 11.57)
AutoAD-Zero† [12]	✓	2.27	(0.19, 0.32, 0.30)	(7.87, 8.13, 8.13)
AutoAD-Zero [12]	✓	2.27	(0.55, 0.31, 0.21)	(13.26, 10.62, 9.59)
Shot-by-Shot [16]	✓	2.43	(0.42, 0.15, 0.12)	(19.06, 15.18, 13.56)
CoherentAD (Ours)	✓	**2.63**	(**0**, **0**, **0**)	(**5.21**, **4.45**, **4.18**)
w/o multiple candidates	✓	2.49	(0.04, **0**, 0.02)	(8.17, 6.78, 6.19)

Table 4.2 Quantitative comparison on CMD-AD. The first row indicates the inherent level of repetition in ground-truth descriptions, serving as a lower bound for repetition scores. † denotes the original AutoAD-Zero adopting the VideoLLaMA-7B VLM backbone in the first stage, while all other training-free methods (including *Ours*) employ Qwen2-VL-7B. The repetitions are reported against the three offsets.

Second, we capture *partial repetitions* through lexical overlap, by computing the intersection over union between descriptions. Specifically, we extract a set of tokens from each description by lowercasing the text, removing punctuation and English stopwords, and applying tokenisation using NLTK [49]. As with exact repetitions, we compute three overlap scores per video, corresponding to the three offsets.

4.3 Evaluating Sequences of Audio Descriptions

4.3.1 Dataset and metrics.

We evaluate our method CoherentAD (from Chapter 3) on the test set of CMD-AD [6], with $7,316$ ADs for 591 videos of 98 movies. We report the sequence-level metrics proposed in Sec. 4.2.

4.3.2 Quantitative results.

Tab. 4.2 reports the performance on the CMD-AD dataset for both fine-tuned (top) and training-free (bottom) methods, where all prior works fall short on our sequence-level metrics. For instance, [16] produce substantially higher partial repetitions (19.06%) between consecutive predictions than the ground truth (3.71%). In contrast, our method prioritises sequence-level coherence by design, significantly outperforming the previous state-of-the-art on *StoryRecall* (2.63 vs 2.43) and achieving repetition scores that closely match the ground truth, with zero exact repeats. This indicates that the overall sequence remains coherent and faithful to the original narrative.

4.3.3 Ablation without Multiple Candidates.

Notably, our ablation without multiple candidate generation (generating one AD per interval) achieves significantly lower repetition than [16] (8.17 vs 19.06). [16] rely on concise VLM outputs that highlight

r	StoryRecall ↑	Partial Repeat% ↓
1	2.56	(5.05, 4.71, 4.33)
2	2.60	(5.10, 4.28, 4.30)
3 (default)	2.63	(5.21, 4.45, 4.18)

Table 4.3 Varying the number of prior descriptions (r) during scoring. $r = 3$ (default) gives the best StoryRecall and lowest overall repetition across positions.

Method	Train	Old Metrics		Vis App		Narr Und	
		C	LLMe	CC	Ratio	CC	Ratio
Dialog only	-	-	-	10.0	33.1	58.9	81.0
AutoAD-III	✓	**25.0**	2.01	14.9	49.3	***63.2***	***86.9***
UniAD*	✓	21.8	**2.92**	14.3	47.4	63.0	86.6
AutoAD-Zero	✗	17.7	1.96	13.4	44.3	62.9	86.5
Q2VL	✗	-	-	**17.2**	**57.0**	51.2	70.4
CoherentAD(Ours)	✗	13.2	2.17	***15.2***	***50.34***	**64.0**	**88.0**
AV_1	-	-	-	-	-	72.7	100
AV_2 (17)	-	-	-	30.2	100	75.0	103

Table 4.4 Evaluation of various generated and human authored ADs on ADQA. Results are reported with dialog + AD as context. DistinctAD is with Llama. NarrAD is with curation. UniAD has some missing outputs. The "Train" column indicates if the method is trained (✓) or training-free (✗). Acronyms are as follows: Vis App: Visual Appreciation, Narr Und: Narrative Understanding. The metrics are C: CIDEr, LLMe: LLM-AD-eval [6], CC: correct answer using context, and Ratio: Accuracy ratio.

the most prominent character, action, or interaction, often leading to redundant phrasing and missing visual details. In contrast, our setup aggregates a wider range of outputs from the VLM, resulting in more diverse and informative descriptions. Even without leveraging neighbouring context, this setup achieves a slightly higher StoryRecall than [16] (2.49 vs 2.43). Further, shifting from single descriptions to our multi-candidate setup increases StoryRecall and decreases repetition. (8.17 vs 5.21).

4.3.4 Ablation on Varying Context Size.

Tab. 4.3 shows results for different values of r, the number of prior descriptions used in Sec. 3.1.3. Using $r = 1$ gives the lowest partial repetition at position 1 (5.05) but slightly higher repetition at later positions and lower StoryRecall (2.56). Performance improves with $r = 2$ (2.60), and $r = 3$ achieves the best balance with the highest StoryRecall (2.63) and lowest overall repetition (5.21, 4.45, 4.18). Exact repeats are zero in all cases.

4.4 ADQA: Evaluating ADs via Question Answering

The work in this section was done by Divy Kala.

Writing ADs is a complex task that requires professionals to identify the most relevant visual elements and describe them in a coherent and concise manner to fit within the gap between dialogs [50]. Two themes are central to ADs: (i) they help BVI audiences *appreciate* the visual elements enriching their experience, and (ii) they support *narrative understanding* by conveying essential visual plot points.

We posit that AD evaluations must capture both these aspects, and propose the ADQA benchmark, a multi-choice question-answering (MCQA) framework. For visual appreciation, questions are automatically generated from factual details in ground-truth ADs. For narrative understanding, questions are based on plot descriptions aligned to video segments using LLMs. In both cases, the questions have one correct answer among five options, and the LLM provides both the correct answer and a rationale, ensuring question quality through self-verification. This setup shifts the evaluation focus from surface similarity to functional utility: would the generated ADs help BVIs perceive key visual elements and follow the story?

In Tab. 4.4, we report performance using two metrics: **CC** (correct answers grounded in the provided ADs), and the **accuracy ratio**, which normalizes model performance relative to human-authored ADs (AV_1, AV_2, LSMDC). We evaluate various AD generation approaches. These include trained models (AutoAD-III [6], UniAD [8], DistinctAD [9]) and training-free baselines (AutoAD-Zero [12], NarrAD [17]). We also evaluate human-authored ADs (AV_1, AV_2, LSMDC) to establish upper bounds, and use dialog-only context as a lower-bound baseline. Finally, we include results from CoherentAD. To evaluate visual comprehension in VideoLMs, we use the summarised narratives produced by Qwen2VL [43] in Sec. 3.1.1, consisting of paragraph-length descriptions that are not bound by strict AD timing constraints.

Results. The dialog only context establishes a strong baseline for evaluation. Thankfully, all models outperform this on VA questions indicating that ADs are required to appreciate the rich visuals of a movie. However, the same does not hold for NU questions, where the gap between the dialog-only baseline and other methods is relatively small. This highlights the contribution of dialog to the narrative. The results indicate serious room for improvement to generate ADs that help drive the story. For the NU task, the results are a bit saturated and models give high scores on CMD-AD (88% Ratio).

Interestingly, AV_1 and AV_2 achieve similar scores on CMD-AD NU questions (72.7% and 75.0% CC) despite the latter covering only 17 of 98 movies. Q2VL achieves the highest accuracy ratio (57.0%) on CMD-AD VA questions hinting that many details are perceived and effective editing with training-free approaches is a good direction. Q2VL achieves this high Visual Appreciation (VA) accuracy ratio by densely summarizing everything visible in a scene. However, its outputs are paragraph-length and unconstrained by AD timing, making them unsuitable for real-world narration. In contrast, CoherentAD produces single-sentence descriptions that are concise enough to fit within AD intervals, yet still retains high VA performance (50.34)—second only to Q2VL. Given its real-world applicability and strong VA score, CoherentAD offers a more practical solution.

Despite scoring lower on conventional metrics like CIDEr (13.2) and LLM-AD-Eval (2.17), CoherentAD outperforms all prior methods on Narrative Understanding (NU) with a CC score of 64.0

and the highest accuracy ratio of 88.0. The lower NU scores of the summarised paragraphs, compared to our method, may be attributed to the LLM's difficulty in processing large dumps of information. This demonstrates that our training-free, sequence-level approach produces ADs that are more functionally useful—even when not favored by similarity-based metrics—highlighting the limitations of relying solely on CIDEr-style evaluations for assessing AD quality. Taken together, these results establish CoherentAD as the most effective method across both VA and NU.

Chapter 5

Comparing Human Gaze and Model Attention in Video Memorability

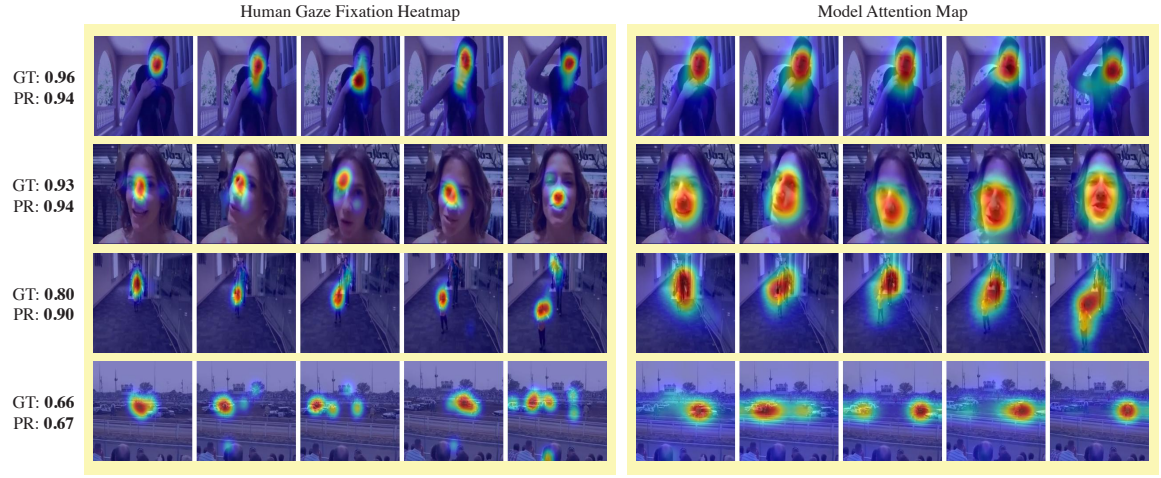


Figure 5.1 Comparing human gaze fixations (left) and model's attention maps (right) for 3 different videos (one per row). The memorability scores, ground-truth (GT) and model prediction (PR), are provided on the left. The heatmaps depict areas of high visual attention through warmer colors (red-yellow), indicating regions where human observers fixated (left) and model attended (right). The model's attention patterns are aligned with human gaze patterns, especially for more memorable videos. Samples from Memento10k [2].

This chapter was completed with the help of my co-advisor Prof. Vishnu Sreekumar, and co-author Prajneya Kumar. Please refer to our original paper for details.[1]

In 2018, Nike's "Dream Crazy" commercial featuring Colin Kaepernick captured nationwide attention in the US[2]. This advertisement was especially memorable because it was aired in the aftermath of Kaepernick's protests against race-based police brutality. While the context made this commercial memorable for US-based audiences, other types of commercials tend to be memorable in general. For example, a famous 2013 E-Trade Super Bowl commercial features a baby seated behind a stack of cash talking about investments and hidden fees[3]. This sort of ad is likely to be memorable regardless of

[1] https://arxiv.org/abs/2311.16484

[2] Dream Crazy https://www.youtube.com/watch?v=WW2yKSt2C_A

[3] E-Trade ad https://www.youtube.com/watch?v=EbnWbdR9wSY

cultural context due to several attention-grabbing features, notably, a baby talking in an adult voice and delivering investment advice. This latter type of memorability, thought to be consistent across individuals and cultures, has been extensively studied in both cognitive science and computer vision using images [24, 25] and words [51, 52]. In this work, we ask: what are the spatial, temporal, and semantic patterns of attention that are associated with video memorability? To answer this question, we train a CNN+Transformer model to predict human memorability of naturalistic videos, use self-attention scores to determine where the model *looks* across space and time, and collect human eye-tracking data to compare the model's attention against human fixations (Fig. 3.1).

Early work on image memorability reveals the importance of both object and scene categories in predicting memorability [24,28]. Semantic categories are also predictive of memorability across stimuli, including words [51,52] and indeed, prior work shows that context guides eye movements to task-relevant object locations [53]. Thus, we investigate what semantic categories in videos drive memorability. Video captioning approaches have been used in previous semantic analyses of video memorability [2, 54, 55]. However, to our knowledge, we are the first to present a detailed analysis of attention captured by different semantic categories when humans attempt to memorize videos and when a model is trained to predict these memorability scores. We apply panoptic segmentation [56] and adopt the COCO hierarchy [57] to distinguish between *things* (i.e. objects with well-defined shapes such as *person*) and *stuff* (i.e. amorphous background regions such as *sky*) in the video frames. Next, we compare pixel distributions weighted by model attention and human gaze and find that both the model and humans generally enhance attention to *things* and reduce attention to *stuff*. Furthermore, the model and humans agree on what specific *things* and *stuff* to emphasize or disregard. Overall, these results indicate that the model learns similar attentional strategies as humans *even though it is trained only to predict a memorability score*.

Beyond semantics, the time axis in videos begs an important question: how early does the model know about the memorability of a video? Human experiments using extremely fast presentation times reveal that image memorability differences can be observed in brain activity patterns as early as $400\,\mathrm{ms}$ [25, 58].

Therefore, it is possible that very early moments in a video are predictive of how memorable it will be. Furthermore, human attention tends to be highest at the beginning of an event and wanes over the course of the event [59]. Thus, video memorability scores may be influenced to a greater extent by the initial frames. Note that memorability scores are computed as a consensus across participants. Therefore, we expect the video frames that most people attend to in similar ways to drive the memorability scores. Despite having no intrinsic temporal bias, can models trained to predict memorability pick up on these human-like temporal attention patterns? To answer this question, we first analyse human-human gaze agreement in our videos and establish that different people are more likely to attend to similar regions in the initial frames. Next, summing over the model's spatial attention scores in a frame, we observe that the model indeed assigns greater importance to earlier frames within videos, thereby discovering a subtle temporal pattern in human behaviour.

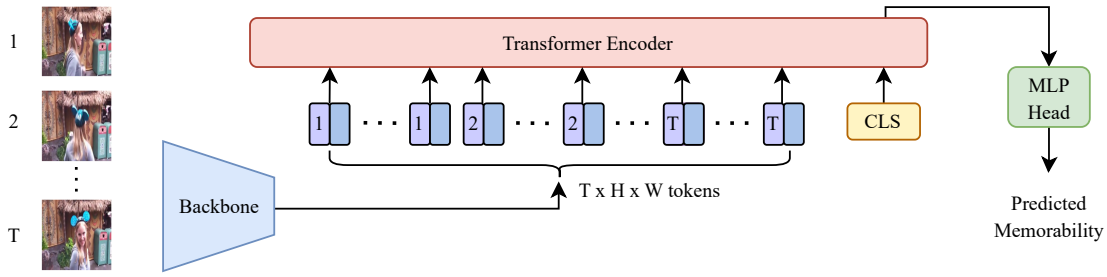


Figure 5.2 Model overview. T video frames are passed through an image backbone encoder to obtain spatio-temporal features $\mathbf{F} \in \mathbb{R}^{T \times H \times W \times D}$. Coupled with position embeddings, and after appending a CLS token, we pass them through a Transformer encoder with self-attention. A memorability score is calculated at the CLS representation with an MLP. *Attention scores between* CLS *and each token are used for downstream anaylsis*.

The video memorability literature [1, 39, 40] focuses on high prediction performance and lacks analysis of models' (dis)similarities to how humans view and remember videos. We aim to address this gap in this chapter.

Note, our work aims to highlight the similarities between *human fixations* when performing memorability experiments, and *model attention* when trained to predict memorability scores. A simple CNN+Transformer architecture enables this, matches SoTA, and has not been used in video memorability before.

5.1 Methods: Model and Human

We present two methods: (i) a CNN+Transformer model that predicts memorability scores using spatio-temporal attention; and (ii) an eye-tracking study to capture human gaze patterns during a memorability experiment.

5.1.1 Transformer-based Model

We begin by defining some notation. Our dataset consists of multiple videos with associated memorability scores, (V, m) pairs. Each video consists of multiple frames. We sub-sample T frames for memorability prediction and denote a video as $V = \{f_i\}_{i=1}^{T}$.

Our model consists of three parts: (i) a backbone image encoder Φ, (ii) a Transformer encoder that attends over spatio-temporal tokens extracted from T video frames, and (iii) a prediction head that estimates the memorability of a video (see Fig. 5.2).

1. Image encoder. Our goal is to employ a model that allows us to analyze the spatio-temporal attention over video frames. Thus, we consider CNN backbones such as ResNet-50 [60], trained with contrastive language-image pretraining (CLIP) [61]. We encode each video frame to obtain a space-aware repre-

sentation (from the conv5 layer):

$$\mathbf{f}_i = \Phi(f_i)\,,\ \text{where } \mathbf{f}_i \in \mathbb{R}^{H\times W\times D}\,, \forall i \in \{1,\ldots,T\}\,, \tag{5.1}$$

where $H \times W$ are height and width of the spatial resolution, and D is the dimensionality of the embeddings.

While previous works use multiple features: frames, flow, and video by [2]; low-, mid-, and high-level representations and a contextual similarity module by [39]; or a host of 10+ models fed to an LLM by [40], our model relies on a single semantic backbone (CLIP). Our simple approach enables the analysis of model's spatio-temporal attention maps through a comparison to human gaze.

2. Video encoder. We use a Transformer encoder [62] to capture attention across spatio-temporal tokens. First, we flatten and encode the image features using a linear layer $\mathbf{W}_d \in \mathbb{R}^{d\times D}$ to reduce dimensionality. Next, to each token, we add two types of position embeddings:

$$\mathbf{f}'_{ij} = \mathbf{W}_d\mathbf{f}_{ij} + \mathbf{E}^t_i + \mathbf{E}^s_j\,, \forall i \in \{1,\ldots,T\}, j \in \{1,\ldots,HW\}\,, \tag{5.2}$$

where $\mathbf{E}^t_i$ is the i^{th} row of the temporal embedding matrix (learnable or Fourier), and $\mathbf{E}^s_j$ is the j^{th} row of the spatial embedding matrix, and $\mathbf{f}_{ij} \in \mathbb{R}^D$ is the feature at frame i and spatial region j.

We prepend a CLS token (with learnable parameters $\mathbf{h}_{\mathsf{CLS}}$) to create a sequence of $1{+}THW$ tokens and post LayerNorm [63] feed this to a Transformer encoder (TE) of L layers with hidden dimension d:

$$[\tilde{\mathbf{h}}_{\mathsf{CLS}}, \tilde{\mathbf{f}}_{11}, \ldots, \tilde{\mathbf{f}}_{THW}] = \text{TE}([\mathbf{h}_{\mathsf{CLS}}, \mathbf{f}'_{11}, \ldots, \mathbf{f}'_{THW}])\,. \tag{5.3}$$

3. Predicting memorability. We pass the CLS token's contextualized representation to an MLP and predict the memorability score: $\hat{m} = \text{MLP}(\tilde{\mathbf{h}}_{\mathsf{CLS}})$.

Extracting attention scores. We extract the self-attention matrix from the multi-head attention module of the last layer of the TE. We mean pool over the heads and pick the row corresponding to the CLS token. Ignoring the self token, this attention vector $\alpha \in \mathbb{R}^{THW}, \sum \alpha = 1$, is used for further spatio-temporal analysis. We obtain an attention map of the size of the image by applying upscaling (pyramid expand) on the $H \times W$ attention scores of each frame.

Training and inference. Similar to previous work [2, 39] we use the MSE loss $\mathcal{L} = \|m - \hat{m}\|^2$ to train our model. We also considered the Spearman loss [39], but did not see significant performance gains. For most experiments, we freeze the backbone and rely on the strong semantic features extracted by CLIP pretraining.

5.1.2 Eyetracking Study: Capturing Gaze Patterns

The work in this subsection was done by Prajneya Kumar.

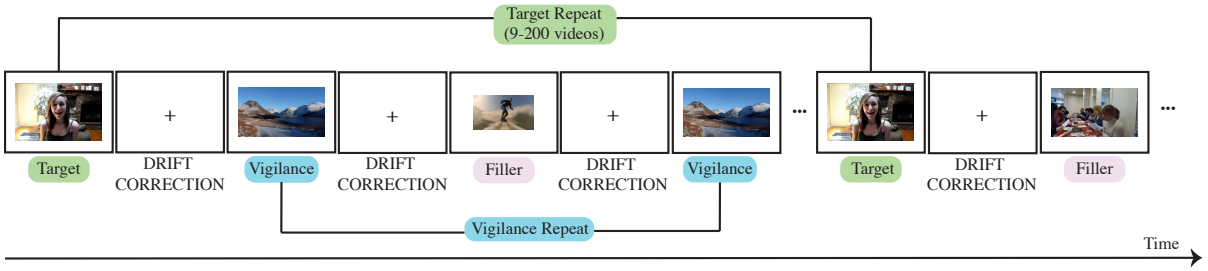


Figure 5.3 Design of eye-tracking experiment. A subject watches alternating videos and drift correction fixation crosses (typically between 0.5 s to 1 s). A vigilance video (one of 40) is repeated in a short interval of 2-3 videos to ensure that the subject is alert, while the target videos (one of 20) have a lag of at least 9 videos. Filler videos (80) are not repeated.

We collect eye-tracking data while participants view videos in a memory experiment. The setup (as showin in Fig. 5.3) follows the original video memorability experiments [1, 2], as we want the gaze patterns to accurately reflect the cognitive and visual processes involved in viewing and remembering videos.

Our study has 20 participants (9 females, 11 males, Age 22.15 ±0.52 (mean ±SEM)). Memento10K: 6 females, 4 males, Age 22.9 ±0.94. VideoMem: 3 females, 7 males, Age 21.4 ±0.37. We choose 140 unique videos each from both video datasets: *Memento10K* [2] and *Videomem* [1].

Further details on data collection and processing are available in the main paper.[4]

5.2 Datasets and Experimental Setup

Video memorability datasets. We perform experiments on two datasets: (i) **VideoMem** [1] consists of 10K, 7 second video clips, each associated with a memorability score. (ii) **Memento10K** [2], introduced as a dynamic video memorability dataset, contains human annotations at different viewing delays. This dataset consists of 10K clips, but they are shorter in duration (3 seconds).

Data splits. VideoMem has 7000 videos in the training set and 1000 in the validation set (MediaEval workshop [64]). Past works report results on the validation set as the test labels are not publicly available. Memento10k is split into 7000 videos for train and 1500 each for validation and test. We provided our model's outputs to the competition organizers and report results on the test set.

Memorability metrics. The memorability score associated with each video in the datasets captures the proportion of people in the original experiments who correctly recognized the video. We evaluate model's predictions relative to ground-truth (GT) memorability scores, using the Spearman rank correlation (RC ↑). Following previous works, we also report the mean squared error (MSE ↓) to measure the gap between GT and predictions.

[4] https://arxiv.org/abs/2311.16484

	Embedding					Memento10k (val)	
	CLIP	Time	Space	Sampling	Caption	RC ↑	MSE ↓
1	ST	F	-	Random	-	**0.706**	0.0061
2	T	F	-	Random	-	0.687	0.0062
3	ST	L	-	Random	-	0.696	0.0059
4	ST	F	1D	Random	-	*0.703*	*0.0057*
5	ST	F	2D	Random	-	0.701	**0.0056**
6	ST	F	-	Middle	-	*0.703*	0.0066
7	ST	F	-	Random	Orig.	**0.745**	**0.0050**
8	ST	F	-	Random	Pred.	*0.710*	*0.0056*

Table 5.1 Model ablations. Column 1 (C1) compares the impact of using spatio-temporal (ST) features versus temporal (T) features with global average pooling. C2 and C3 specify the types of temporal (L: learnable, F: Fourier) and spatial position embeddings used. C4 is the frame sampling method used during training. C5 indicates whether the video caption (Orig: original caption, Pred: predicted caption) is used in modeling. *Row 1 (R1) is chosen as the default configuration for further experiments* and represents the best vision-only model. R2-6 evaluate vision model choices: features, position-encodings, and frame sampling methods. R7 presents results with original captions (Orig.) as a part of the model and R8 aims to predict the captions on the fly. The best results in each section are in **bold**, with second-best in *italics*.

Implementation details. We break each video into T uniform segments and pick one frame at random from each segment during training - this acts as data augmentation [65]. For inference, we take the middle frame of the segment. T=5 works well for Memento10k (1.66fps) and T=7 for VideoMem (1fps). When not specified otherwise, we train our model with the Adam optimizer [66], learning rate 10^{-5}, and a step scheduler (for VideoMem only) with step size 10 epochs and multiplier 0.5.

5.3 Experiments: Memento

We begin with model ablation studies for Memento10k. VideoMem presents challenges due to data leakage, which are discussed in Sec. 5.4.

5.3.1 Ablation of Vision Models

Tab. 5.1 rows 1-6 show the results of various hyperparameters of the vision model evaluated on the validation set. Row 1 (R1) achieves best performance and is the *default configuration* for further experiments. Using spatio-temporal (ST, R1) image embeddings and not performing global average pooling (R2) shows a small improvement in RC. Similarly, using Fourier embeddings (R1) is better than learnable ones (R3), perhaps due to the small dataset size. Surprisingly, using spatial embeddings to identify the $H \times W$ tokens reduces performance (R1 vs. R4 or R5), perhaps due to the pyramidal

nature of the CNN representations. Finally, using random sampling during training (R1) instead of picking the middle frame of the segment (R6) results in a small increase. In general, the gap between all rows is small, indicating that results are not impacted strongly by hyperparameter changes. However, spatio-temporal (ST) CLIP embeddings are required to obtain spatio-temporal model attention maps.

5.3.2 Use of captions.

[2] introduced captions (descriptions) for the short videos in Memento10k as a way to emphasize semantic categories for predicting memorability. In this section, we explore two ways of incorporating caption information. First, in Sec. 5.3.2.1, we assume that the ground-truth caption is available and concatenate its token embeddings with the visual tokens during memorability prediction. Second, in Sec. 5.3.2.2, we investigate a joint learning setup where captions are not given but are predicted alongside the memorability score.

5.3.2.1 Assuming Caption is Available

When the caption is given, we first extract token-level representations through a BERT encoder and append them to the spatio-temporal video tokens for memorability prediction. **Text encoder.** We extract textual embeddings for the captions from the last hidden state of the BERT [46] model ψ:

$$\{\mathbf{g}_l\}_{l=1}^{N} = \psi(\{g_l\}_{l=1}^{N}), \tag{5.4}$$

where $\mathbf{g}_l \in \mathbb{R}^d$, N is the number of tokens, and d is the dimensionality of the embeddings, equal to the reduced dimensionality of images after the linear layer.

Changes to the video encoder. We append N text tokens to the THW visual tokens fed to the Transformer encoder. To distinguish between text and image, we append modality specific embeddings to both the visual (from Eq. 2) and text tokens. We also add position embeddings indicating order to the text tokens.

$$\mathbf{f}'_{ij} = \mathbf{W}^d \mathbf{f}_{ij} + \mathbf{E}^t_i + \mathbf{E}^s_j + \mathbf{E}^m_1, \tag{5.5}$$

$$\mathbf{g}'_l = \mathbf{g}_l + \mathbf{E}^c_l + \mathbf{E}^m_2, \tag{5.6}$$

where $i = [1, \ldots, T]$, $j = [1, \ldots, HW]$, $l = [1, \ldots, N]$, $\mathbf{E}^t_i$ is the i^{th} row of the temporal embedding matrix (learnable or Fourier) for images, $\mathbf{E}^c_l$ is the l^{th} row of the temporal embedding matrix for the caption, $\mathbf{E}^s_j$ is the j^{th} row of the spatial embedding matrix, and $\mathbf{E}^m_{[1,2]}$ are the modality embeddings, one for visual tokens, another for text.

We combine the CLS token (with learnable parameters $\mathbf{h}_{\mathsf{CLS}}$), image and text tokens to create a sequence of $1 + TWH + N$, apply LayerNorm, feed it to the TE.

$$[\tilde{\mathbf{h}}_{\mathsf{CLS}}, \tilde{\mathbf{f}}_{11}, \ldots, \tilde{\mathbf{f}}_{THW}, \tilde{\mathbf{g}}_1, \ldots, \tilde{\mathbf{g}}_N] = \text{TE}([\mathbf{h}_{\mathsf{CLS}}, \mathbf{f}'_{11}, \ldots, \mathbf{f}'_{THW}, \mathbf{g}'_1, \ldots, \mathbf{g}'_N]). \tag{5.7}$$

Methods	Caption	Memento10k Test RC	Test MSE	Val RC	Val MSE
SemanticMemNet ECCV20	No	0.659	-	-	-
M3-S CVPR23	No	-	-	0.670	0.0062
Ours (R1 Tab. 5.1)	No	**0.662**	0.0065	**0.706**	0.0061
SemanticMemNet ECCV20	Yes	0.663	-	-	-
Sharingan arXiv	Yes	-	-	0.72	-
Ours (R7 Tab. 5.1)	Yes	**0.713**	0.0050	**0.745**	0.0050

Table 5.2 Comparison against SoTA for video memorability. Baselines considered are SemanticMemNet [2], M3-S [39], and Sharingan [40]. Split-half human-human consistency RC for Memento10k is 0.73.

As before, $\tilde{\mathbf{h}}_{\mathsf{CLS}}$ is used to predict the memorability score.

We report results when using the ground-truth caption in Tab. 5.1, row 7 (w original captions as input). We see a 0.04 points increase in Spearman correlation (0.706 to 0.745).

5.3.2.2 Joint Prediction of Caption and Memorability

When the caption is not available, we consider predicting the caption along with the memorability scores. In particular, we adapt CLIPCap [67], a recent approach that connects CLIP visual features with the GPT-2 decoder using a Transformer mapping layer.

Specifically, we use a mapping network (a Transformer decoder) to convert the THW visual tokens at the output of the Transformer encoder $\tilde{\mathbf{f}}_{ij}$ to a set of P prefix tokens. The mapping network of L_D=6 layers consists of P query learnable tokens and uses visual inputs as memory, P=30. The outputs of this mapping network are fed as prefix tokens to the GPT-2, and captions are generated in an auto-regressive manner.

We train the model jointly, to predict both the memorability score (using L1 regression loss) and the caption (using cross-entropy loss). Results of this approach are presented in Tab. 5.1, row 3. A small increase of 0.004 is observed in the RC score (0.706 to 0.710).

5.3.3 SoTA comparison.

Comparison to state-of-the-art works on Memento10k with different setups (val or test split, with / without captions) is presented in Tab. 5.2. Note, our goal is to understand the attentional factors driving video memorability through a model that provides spatio-temporal attention. Nevertheless, our model with a single feature encoder (CLIP) achieves results comparable to SoTA (Memento10k: 0.706 val, 0.662 test). With captions, we obtain 0.713 (test). To interpret model performance reported as RC scores, we note that a model that performs well is expected to approach a human-human consistency RC of 0.73 for *Memento10K* [2].

Table 5.3 Results of transferring an image/video memorability model to images/videos. Datasets: LM: LaMem [27], M10k: Memento10k [2], VM: VideoMem [1], and FG: FIGRIM [68]. Training strategy: P for pretraining and F for fine-tuning. Results reported on validation set.

	Train on				LaMem		Memento10k		VideoMem		FIGRIM	
	LM	M10k	VM	FG	RC	MSE	RC	MSE	RC	MSE	RC	MSE
1	F	-	-	-	**0.729**	0.0074	0.526	0.0220	0.382	0.0233	0.647	0.0168
2	-	F	-	-	0.547	0.0273	*0.706*	0.0061	0.439	0.0165	0.351	0.0525
3	-	-	F	-	0.549	0.0147	0.525	0.0089	**0.513**	0.0060	0.501	0.0355
4	P	F	-	-	0.679	0.0161	**0.718**	0.0568	0.446	0.0144	0.634	0.0318
5	P	-	F	-	0.688	0.0090	0.459	0.0096	0.504	0.0059	0.627	0.0237
6	P	-	-	F	0.678	0.0113	0.507	0.0130	0.392	0.0191	**0.742**	0.0123
7	P	F	F	-	0.664	0.0135	0.689	0.0058	0.483	0.0062	0.626	0.0273

Furthermore, our model is trained only on the Memento10k training set, while all baselines train on a combination of image and video memorability datasets. For example, pretraining on LaMem [33] and fine-tuning on Memento10k improves performance from 0.706 to 0.715.

Note: All further analyses and experiments are conducted using the vision-only model, without incorporating captions.

5.3.4 Transferring from/to Image Memorability

For completeness, we present cross-domain transfer results of pretraining and fine-tuning our model on image or video memorability datasets and evaluation on all.

We evaluate on image memorability tasks by considering the image as a "video" of $T{=}1$. As seen in Tab. 5.3, on the LaMem dataset [27] we match SoTA results (0.720 RC [33]). On the FIGRIM dataset [68], we achieve results close to human performance (0.74 RC [68]). Previous studies [2] pretrain models on image memorability datasets and then fine-tune them for video memorability prediction. Tab. 5.3 R2 vs. R4 shows a small improvement in Memento10k RC score from 0.706 to 0.718 with LaMem pretraining. However other results do not improve. We also observe that training on one dataset and evaluating on another (rows 1-3) usually leads to significant degradation and is an important problem for future work.

5.4 Why is VideoMem challenging?

The RC scores on VideoMem [1] are significantly lower than on Memento10k, even with additional information like captions providing no improvement. EK add from supplement - Detailed results can be found in supplement In fact, most methods achieve RC greater than the human-human RC at 0.481, indicating that models have probably overfit to the dataset, especially as a held-out test set is not avail-

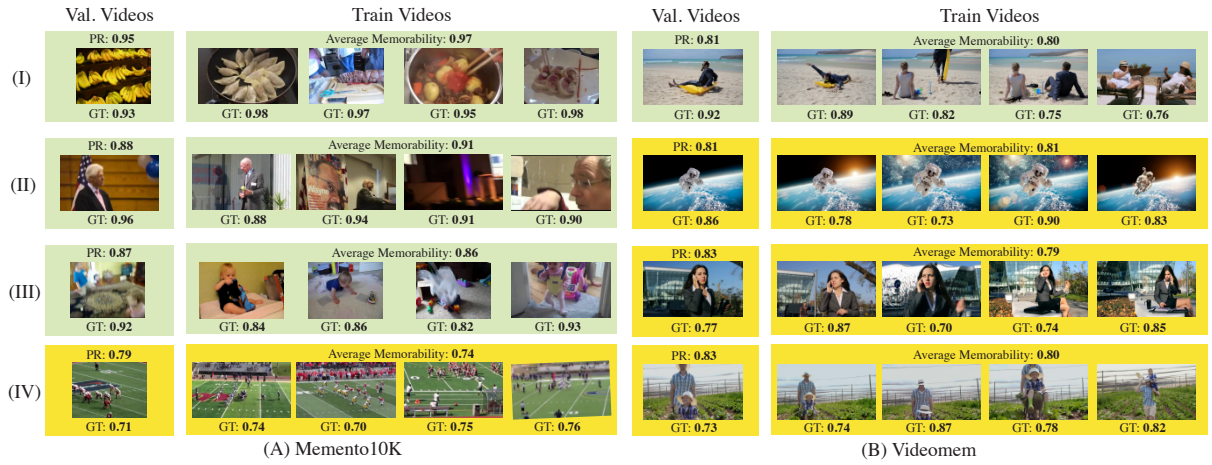


Figure 5.4 Nearest neighbor (NN) analysis for videos from Memento10K (left) and VideoMem (right). We illustrate four validation set videos and for each, four NN from the training set. We provide the GT memorability score (below), the predicted score on the val set (above), and the average of 4 NN scores from the training set. In B (right), multiple video clips with high visual similarity between train and validation sets are highlighted with a *yellow* background. Conversely, the green rows highlight clips that have similar content, but are likely from different source videos. We discuss how data leakage and variance in GT scores may adversely affect evaluation in Sec. 5.4.

able. As evidence, the code repository of a recent work, M3-S [39][5] shows that achieving a Spearman RC of 0.5158 is possible after using highly specific random seeds and hyperparameters.

5.4.1 Similar videos across splits.

We propose a nearest-neighbors (NN) analysis of representations and observe that improving results on VideoMem is challenging due to problems in *split creation*. We visualize the NN in the training set for each validation video based on $\hat{\mathbf{h}}_{\mathsf{CLS}}$, the representation before the MLP regressor. On VideoMem, from a random sample of 30 validation videos, 14 clips have visually identical NN in the training set. In contrast, on Memento10k, we are only able to find 1 clip among 30.

Fig. 5.4 displays a few videos illustrating this problem. On Memento10k (left), we see that NNs show semantic awareness and matching (I food, II speaker, IV sports field). On the other hand, on VideoMem, the NN are (probably) from the same long video. See right: I surfer, II astronaut, III news anchor, IV farmer. Given the identical visual stimuli, the model can do no better than predict the average memorability score of the NNs on the training set (which it does). *E.g.* in row II with the astronaut, PR=0.81 is equal to the average memorability, but is away from GT=0.86. In row IV farmer, PR=0.83 is close to the average 0.80, but away from GT=0.73. While using multiple feature backbones may help, this is not a satisfactory solution to a fundamental issue of data leakage across splits. To address this,

[5] `https://github.com/theodumont/modular-memorability`

		Embedding				Memento10k (val)		VideoMem (val)	
	CLIP	Time	Space	Sampling	Caption	RC ↑	MSE ↓	RC ↑	MSE ↓
1	Spatio-Temporal	Fourier	-	Random	-	**0.706**	0.0061	*0.513*	*0.0060*
2	Temporal	Fourier	-	Random	-	0.687	0.0062	0.508	0.0064
3	Spatio-Temporal	Learnable	-	Random	-	0.696	0.0059	0.502	0.0060
4	Spatio-Temporal	Fourier	1D	Random	-	*0.703*	*0.0057*	0.506	**0.0059**
5	Spatio-Temporal	Fourier	2D	Random	-	0.701	**0.0056**	0.505	*0.0060*
6	Spatio-Temporal	Fourier	-	Middle	-	*0.703*	0.0066	**0.515**	**0.0059**
7	Spatio-Temporal	Fourier	-	Random	Original	**0.745**	**0.0050**	0.505	0.0061
8	Spatio-Temporal	Fourier	-	Random	Predicted	*0.710*	*0.0056*	0.508	0.0061

Table 5.4 Model ablations. Column 1 (C1) compares the impact of using spatio-temporal features versus temporal features with global average pooling. C2 and C3 specify the types of temporal and spatial position embedding used. C4 is the frame sampling method used during training. C5 indicates whether the video caption is used in modeling. *Row 1 (R1) is chosen as the default configuration for further experiments* and represents the best vision-only model. R2-6 evaluate varying visual choices: features, position-encoding, and frame sampling methods. R7 presents results with original captions as a part of the model and R8 aims to predict the captions on the fly. The best results in each section are in **bold**, with second-best in *italics*.

we attempted to recreate the splits. However, as the original source video ids are unavailable, it is not easy to detect which video clips belong to the source video.

5.4.2 Implications for data collection.

We encourage researchers to analyze new datasets before they are released. Information about the video source and split creation process are crucial aspects for any dataset. Additionally, memorability scores are a measure of consensus among viewers and are therefore closely tied to the number of viewers per video. While LaMem averages 80 scores per image, Memento10K has over 90 annotations per video, Videomem averages 38 annotations per video, much smaller than the others. This variance in GT scores is also observed in Fig. 5.4 (B-II), videos of the same astronaut have GT scores varying from 0.73 to 0.90, making learning difficult.

5.4.3 Video Memorability prediction for Videomem

Expanding on the model ablations for Memento10k in Sec. 5.3.3, Tab. 5.4 shows results for VideoMem, which generally follows similar trends, with Row 1 (R1) achieving the best results. However, random sampling during training does not improve performance and including or predicting captions has no impact, perhaps due to the noise in the captions.

SoTA comparisons are shown in Tab. 5.5. As the test set memorability scores (labels) for VideoMem are not available, no previous work apart from the creators of the dataset have evaluated on a held-out

Methods	Caption	Memento10k (test)		VideoMem (test)		Memento10k (val)		VideoMem (val)	
		RC	MSE	RC	MSE	RC	MSE	RC	MSE
VideoMem ICCV19	No	-	-	0.494	-	-	-	0.503	-
SemanticMemNet ECCV20	No	0.659	-	-	-	-	-	0.555	-
M3-S CVPR23	No	-	-	-	-	0.670	0.0062	0.563	0.0046
Ours (R1 Tab. 5.4)	No	0.662	0.0065	-	-	0.706	0.0061	0.513	0.0060
SemanticMemNet	Yes	0.663	-	-	-	-	-	0.556	-
Sharingan arXiv	Yes	-	-	-	-	0.72	-	0.6	-
Ours (R7 Tab. 5.4)	Yes	0.713	0.0050	-	-	0.745	0.0050	0.505	0.0061

Table 5.5 Comparison against SoTA for video memorability on both test and validation sets for Memento10k and VideoMem. Baselines considered are VideoMem [1], SemanticMemNet [2], M3-S [39], and Sharingan [40]. Human-human split-half consistency scores are 0.73 for Memento10k and 0.481 for VideoMem.

test set. Instead, all approaches likely overfit on the validation set with RC scores much higher than the human-human consistency RC at 0.481. Our scores are lower than other SoTA methods, likely due to the challenges discussed in Sec. 5.4. However, we suspect that other models that leverage multiple modalities are strongly overfitting on this dataset.

5.5 Comparing Model Attention and Human Gaze

The following work on human gaze fixation, up to Sec. 5.5.1, was carried out by Prajneya Kumar.

Setup. To compare the human gaze fixation density maps and model-generated attention maps, we first min-max normalize them to $[0, 1]$. Next, we compute multiple popular metrics [6] in saliency evaluation [69]: AUC-Judd [70], Normalized Scanpath Saliency (NSS) [71], Linear Correlation Coefficient (CC) [72], and Kullback-Leibler Divergence (KLD) [73, 74].

We split participants into two random groups and for a given video, compute agreement between the two groups using the saliency metrics. These human-human (H-H) agreement scores are averaged over 10 random split iterations and then across videos. H-H scores act as a ceiling against which our model-human (M-H) agreement scores are compared. To obtain chance-level performance, we compute H-H agreement scores but now with shuffled videos (H-H Shuff.).

Results. While Chapter 5 shows qualitative results of human gaze and model attention, Tab. 5.6 indicates that there is a high degree of M-H similarity across both datasets. We observe that metrics (AUC-J, CC) often approach the H-H scores, and importantly, significantly improve over random chance (H-H Shuff.). In Fig. 5.6, we plot AUC-Judd and NSS against GT memorability bins and observe that the similarity between model attention and human gaze maps increases with GT memorability scores in

[6] We compute all metrics following the methods used by `https://github.com/imatge-upc/saliency-2019-SalBCE/blob/master/src/evaluation/metrics_functions.py`

Metrics	Memento10k			VideoMem		
	M-H	H-H	H-H Shuff.	M-H	H-H	H-H Shuff.
AUC-J ↑	0.89 ±0.007	0.90 ±0.001	0.70 ±0.002	0.89 ±0.007	0.80 ±0.002	0.55 ±0.001
AUC-P ↑	82.91 ±1.65	-	-	88.88 ±1.29	-	-
NSS ↑	1.95 ±0.074	3.07 ±0.024	0.84 ±0.022	2.00 ±0.068	3.12 ±0.023	0.23 ±0.012
CC ↑	0.46 ±0.014	0.49 ±0.003	0.16 ±0.003	0.27 ±0.007	0.27 ±0.018	0.03 ±0.001
KLD ↓	1.48 ±0.035	2.17 ±0.023	4.61 ±0.022	2.65 ±0.020	4.02 ±0.018	6.49 ±0.013

Table 5.6 Comparing gaze fixation maps against model's attention map via different metrics, along with human-human split-half reliability scores over 10 iterations. ↑ (↓) indicates higher (lower) is better. M-H: Model-human; H-H: Human-human; and H-H Shuff.: Human-Human_shuffled (random performance).

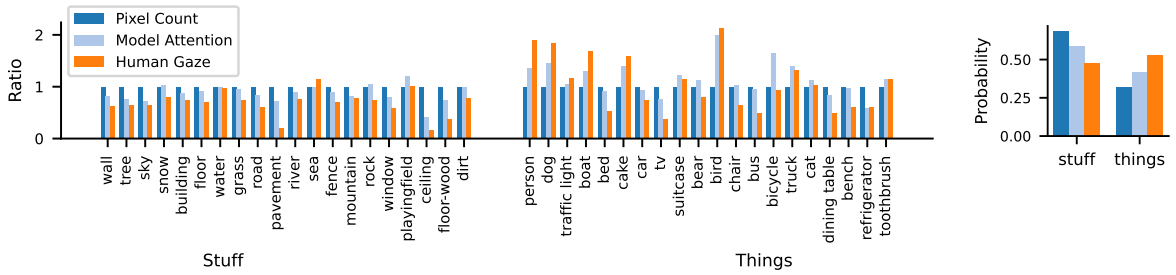


Figure 5.5 Analysis of panoptic segmentation for the most common 40 classes (20 stuff, 20 things). Left shows normalized pixel counts (blue), model attention-weighted counts (light blue), and human gaze-weighted counts (orange). Both, model and humans, show lower affinity for stuff classes and higher for thing classes, indicating their importance in memorability. Right Pixel counts are accumulated across stuff and thing classes, highlighting the above trend clearly. Best viewed on screen with zoom.

both datasets. This suggests that highly memorable videos have clear regions of focus for both humans and the model.

Center bias. Among metrics, we also considered the shuffled AUC (sAUC) [75], but it tends to unjustly penalize valid central predictions [76]. Therefore, we introduce a metric to measure relative similarity, *AUC-Percentile*. For a given video, we compare the true AUC-Judd between model attention and human gaze against a distribution of AUC-Judd values calculated by comparing model attention from that video and human gaze from other randomly selected videos. The percentile of the true AUC-Judd score within the distribution of random AUC-Judd scores estimates the probability that the true score is video-specific and is not obtained by chance or due to center bias. For instance, a model driven purely by center-bias (using a 2D Gaussian, σ=10% of the scene height [77]) yields an average AUC-Percentile score of 76.17 ±2.62 on Memento10K and 68.47 ±2.82 for VideoMem. Results in Tab. 5.6 show that our model's AUC-P scores at 82.91 ±1.65 and 88.88 ±1.29 exceed these center-bias-driven AUC-P scores.

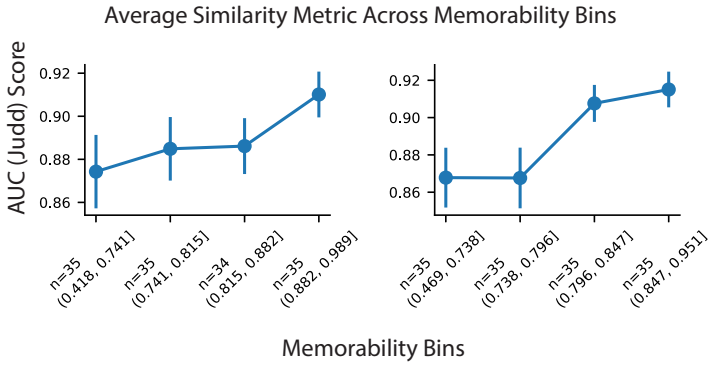


Figure 5.6 Gaze vs. attention similarity metrics with AUC-Judd scores on the Y-axis and Ground Truth on the X-Axis. Left: Memento10k, Right: VideoMem. Error bars depict SEMs.

Another approach to rule out the possibility that the high M-H similarity is due to center bias involves a direct comparison between the performance of the previously explained Gaussian-based center bias model [77] and our proposed gaze prediction model. We use the Gaussian to simulate central fixation and calculate median AUC-J score across frames per video. Compared to the Gaussian, our model is better aligned with human fixations across videos on both datasets, Memento10K ($p = 0.003$) and VideoMem ($p = 5.80 \times 10^{-12}$).

5.5.1 Panoptic Segmentation

We extract panoptic segmentation labels from MaskFormer [56], a SoTA model for segmentation, on the T selected video frames (see Fig. 5.7 for examples). We use the COCO-stuff hierarchy [57] to classify labels as *stuff* or *things*. We create three sets of counts: (i) *Pixel Count* sums the number of pixels attributed to each label across frames and videos (normalized by the total number of pixels in the frame). (ii) *Model Attention* weighted counts multiply the attention map with segmentation masks of each category, summing across frames and videos. (iii) *Human Gaze* weighted counts are similar and multiply gaze fixation densities with segmentation masks.

Stuff vs. things classes. We consider the most prevalent *stuff* and *things* labels (20 each) across the 140 videos of the eye-tracking dataset and observe that attention increases/decreases relative to normalized pixel counts in similar ways for models and humans (Fig. 5.5 left). Specifically, we observe a tendency for decreased attention to *stuff* and increased attention to *things*, which is clear in the cumulative distributions (Fig. 5.5 right).

Pixel count, human gaze, and model attention across all labels. In Fig. 5.8, we show the distributions for all stuff and things labels. Row 1 is the probability distribution of pixel counts and gaze/attention weighted counts for stuff labels (plotted in semilog scale). In row 2, we normalize these counts by the pixel count (blue), highlighting dynamic stuff labels such as *light, food, platform* receiving higher attention weighted scores, while other mundane labels such as *wall, sky, road* receiving lower scores.

A similar analysis is shown for *things* in rows 3 and 4. Here too, we observe that daily objects such as *bed, car, toilet* receive less human and model attention to account for memorability, while dynamic or interesting objects such as *person, dog, bird, wine glass, banana* (among others) receive higher attention. This confirms that not all objects are interesting.

Note, while this analysis is also subject to accuracy of Maskformer [56] (the panoptic segmentation approach), qualitatively, we find this to be quite reliable as seen in Fig. 5.7.

5.5.2 Temporal Attention

We first analyze whether humans look at similar regions across frames of a video and find that they are more consistent in the initial frames of the video as compared to later frames, see Fig. 5.9 (blue). However, it is possible that this result is driven by center bias if most videos have salient central regions at the start. To rule this out, we identify a subset of videos that have off-center salient regions in the initial frames.[7] Fig. 5.9 (green) shows us that there is stronger consensus across participants for the off-centered videos, and this too goes down as the video progresses.

Next, to ascertain whether our model displays similar temporal patterns of attention, we compute attention scores as $\alpha \in \mathbb{R}^{T \times HW}$ and sum over the spatial dimensions to obtain temporal attention, $\alpha_T \in \mathbb{R}^T$. As visualized in Fig. 5.10 left, our model preferentially attends to the initial frames of the video sequence, without any architectural bias towards this. We further rule out two possibilities: (i) reversing the frames (and preserving the same temporal position embeddings), we observe that the model still gives more attention to early frames (now appearing at the end, Fig. 5.10 middle); (ii) computing optical flow magnitude [79] per frame, averaged across all pixels, we find that motion is strongest around the middle (Fig. 5.10 right) and cannot be the reason for increased attention to early frames.

Therefore, we conclude that our model, only trained to predict memorability scores, has learned to attend to the visual information that most participants look at earlier on in the videos.

[7] We adopt DeepGaze [78] and compute saliency maps for T video frames. Next, we compute a distance between the predicted saliency map and a center bias, modeled as a Gaussian, and sort the videos in decreasing distance. For this analysis, we consider 25^{th} percentile most off-centered videos for Memento10k and VideoMem separately.

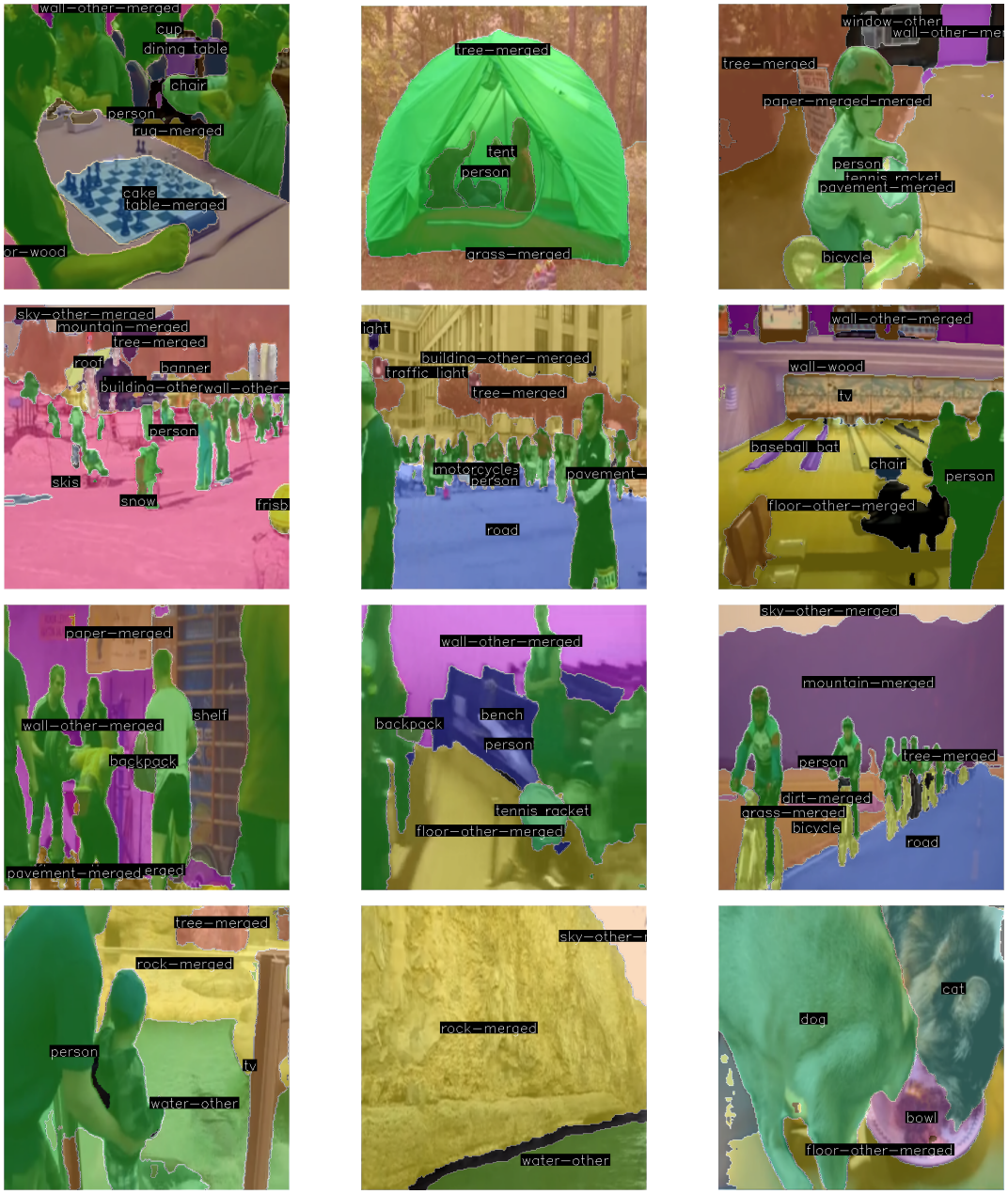


Figure 5.7 Visualisation of panoptic segmentation predictions on Memento10k dataset.

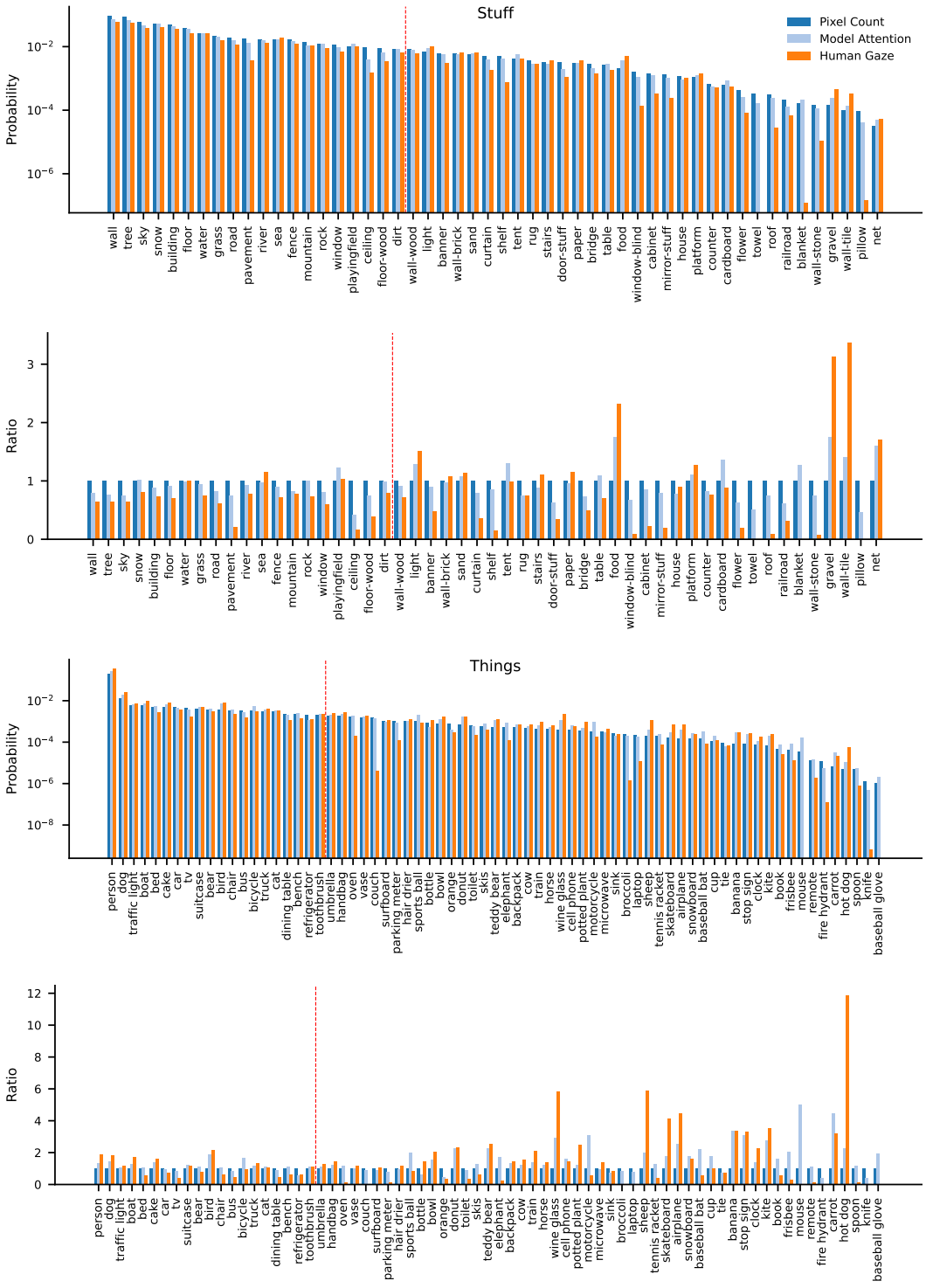


Figure 5.8 Analysis of panoptic segmentation results. The vertical red line marks the top-20 labels within these categories. **First** and **Third**: Raw, attention-, and gaze-weighted pixel probabilities for *stuff* and *things*, respectively (plotted in semilog scale); **Second** and **Fourth**: Highlights how model attention-weighted and human gaze-weighted pixel counts are higher or lower relative to normalized raw pixel counts for *stuff* and *things*.

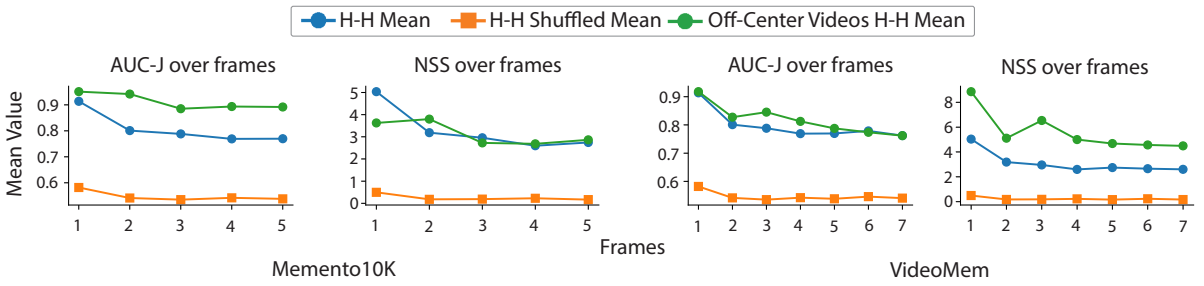


Figure 5.9 Framewise split-half AUC-J and NSS scores for Memento10K (left) and VideoMem (right). The x-axis shows sub-sampled frames at $T{=}5$ for Memento10K and $T{=}7$ for VideoMem. The blue line (H-H) indicates the framewise alignment between gaze patterns, averaged over all 140 videos. The green line captures framewise alignment averaged over 35/140 videos that have most off-center saliency in the initial frames. The orange line represents H-H shuffled, mean alignment when gaze patterns are compared across random videos.

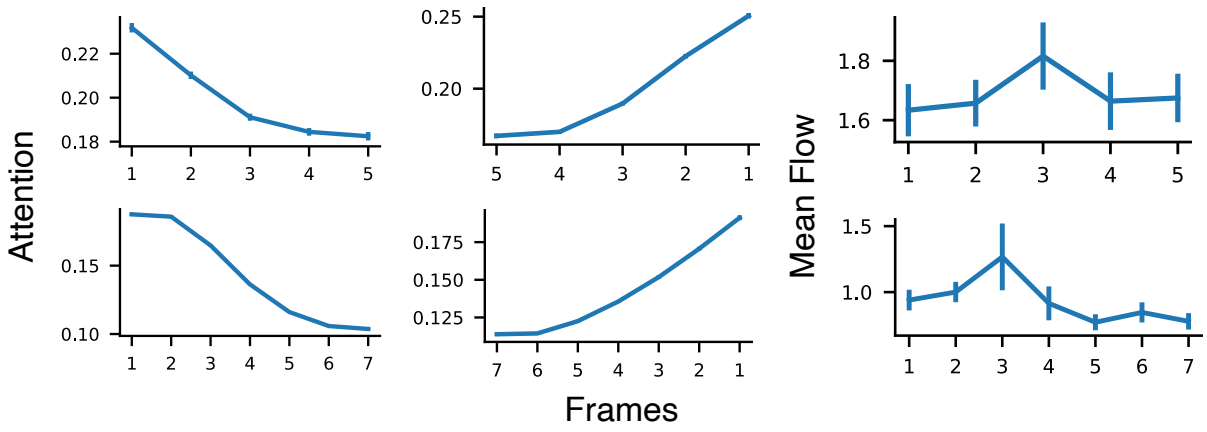


Figure 5.10 Left: Distribution of temporal attention across video frames in normal order, showing peak at the early frames. Middle: Distribution of temporal attention across video frames in *reversed* order as a control to rule out position bias. Right: Mean optical flow magnitude across frames to rule out motion as a bias for the stronger temporal attention at the beginning. The x-axis indicates the number of sub-sampled frames; $T{=}5$ for Memento10K (top) and $T{=}7$ for VideoMem (bottom).

Chapter 6

Conclusion and Future Directions

6.1 Summary

This thesis examined two facets of human engagement with visual content—how it is described and how it is remembered.

In the first part, we proposed a training-free method, CoherentAD, for generating coherent sequences of Audio Descriptions (ADs) by formulating the task as a sequence-level search over multiple candidates per interval. Our approach explicitly incorporates coherence criteria during selection, moving beyond the common practice of generating each AD independently. We also highlighted the limitations of existing metrics, which fail to capture repetition and other forms of incoherence across consecutive ADs. To address this gap, we introduced novel sequence-level evaluation metrics—StoryRecall and repetition scores—that better reflect the quality of AD sequences.

In the second part, we adopted a simple CNN+Transformer model that not only matches SoTA in predicting video memorability scores, but also enables exploring the underlying spatio-temporal attention mechanisms. Furthermore, we collected human gaze data to compare against model attention and observed that the model and humans look at similar regions. We also discovered novel semantic attention patterns relevant for video memorability. On the temporal dimension, the model exhibited strong preference for early frames of the videos, mimicking temporal patterns in human attention. We also analysed a widely used video memorability dataset, identifying several critical issues that researchers must consider when constructing new datasets.

6.2 Limitations and Future Work

Audio Descriptions. Similar to most prior work, our method relies on reference-provided temporal intervals and does not address the problem of AD localisation, i.e. predicting *when* an AD should be placed. These intervals are assumed as input to the pipeline. A promising direction for future work is to develop methods that jointly localise and generate ADs, potentially using audio cues to identify dialogue-gaps automatically.

Our approach also does not incorporate neighbouring context during generation or selection. However, ADs are not required to match the exact interval boundaries and can refer to nearby events. Leveraging neighbouring context could help fill in gaps and improve coherence.

Finally, we select the best candidate per interval without any post-processing. Editing or merging candidates could further enhance sequence-level fluency.

A useful direction for future work is to conduct user studies with BVI audiences to assess how well our metric aligns with human preferences.

Video Memorability. The current datasets have 10k videos each. A model trained on them may not generalize well to any video from the internet, especially in specific domains where the visual stimuli are typically similar across all clips, e.g. identifying memorable parts from a lecture video. Additionally, the model processes extracted frames rather than full videos, which may result in the loss of important details for memorability and could affect comparison with human data, where viewers see the entire video.

Bibliography

[1] R. Cohendet, C.-H. Demarty, N. Duong, and M. Engilberge, "VideoMem: Constructing, Analyzing, Predicting Short-term and Long-term Video Memorability," in *ICCV*, 2019.

[2] A. Newman, C. Fosco, V. Casser, A. Lee, B. McNamara, and A. Oliva, "Multimodal Memorability: Modeling Effects of Semantics and Decay on Video Memorability," in *ECCV*, 2020.

[3] T. Han, M. Bain, A. Nagrani, G. Varol, W. Xie, and A. Zisserman, "Autoad: Movie description in context," in *CVPR*, 2023.

[4] ——, "AutoAD II: The sequel – who, when, and what in movie audio description," in *ICCV*, 2023.

[5] J. Wang, Z. Liu, and X. Wu, "LoCo-MAD: Long-range context-enhanced model towards plot-centric movie audio description," in *ACCV*, 2024.

[6] T. Han, M. Bain, A. Nagrani, G. Varol, W. Xie, and A. Zisserman, "AutoAD III: The prequel – back to the pixels," in *CVPR*, 2024.

[7] K. Q. Lin, P. Zhang, D. Gao, X. Xia, J. Chen, Z. Gao, J. Xie, X. Xiao, and M. Z. Shou, "Learning video context as interleaved multimodal sequences," in *ECCV*, 2024.

[8] H. Wang, Z. Tong, K. Zheng, Y. Shen, and L. Wang, "Contextual AD narration with interleaved multimodal sequence," in *CVPR*, 2025.

[9] B. Fang, W. Wu, Q. Wu, Y. Song, and A. B. Chan, "DistinctAD: Distinctive audio description generation in contexts," in *CVPR*, 2025.

[10] X. Ye, C. Wang, Y. Song, S. Zhou, L. Li, and J. Bu, "FocusedAD: Character-centric movie audio description," *arXiv preprint arXiv:2504.12157*, 2025.

[11] M. Soldan, A. Pardo, J. L. Alcázar, F. Caba, C. Zhao, S. Giancola, and B. Ghanem, "MAD: A scalable dataset for language grounding in videos from movie audio descriptions," in *CVPR*, 2022.

[12] J. Xie, T. Han, M. Bain, A. Nagrani, G. Varol, W. Xie, and A. Zisserman, "Autoad-zero: A training-free framework for zero-shot audio description," in *ACCV*, 2024.

[13] C. Zhang, K. Lin, Z. Yang, J. Wang, L. Li, C.-C. Lin, Z. Liu, and L. Wang, "MM-Narrator: Narrating long-form videos with multimodal in-context learning," in *CVPR*, 2024.

[14] X. Ye, J. Chen, X. Li, H. Xin, C. Li, S. Zhou, and J. Bu, "MMAD: Multi-modal movie audio description," in *LREC-COLING*, 2024.

[15] P. Chu, J. Wang, and A. Abrantes, "LLM-AD: Large language model based audio description system," *arXiv preprint arXiv:2405.00983*, 2024.

[16] J. Xie, T. Han, M. Bain, A. Nagrani, E. Khandelwal, G. Varol, W. Xie, and A. Zisserman, "Shot-by-Shot: Film-grammar-aware training-free audio description generation," *arXiv preprint arXiv:2504.01020*, 2025.

[17] J. Park, J. Ye, S. Lee, H. W. Ka, and D. Han, "NarrAD: Automatic generation of audio descriptions for movies with rich narrative context," in *WACV*, 2025.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002.

[19] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004.

[20] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.

[21] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *CVPR*, 2015.

[22] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *ECCV*, 2016.

[23] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *ICLR*, 2020.

[24] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?" in *CVPR*, 2011.

[25] W. A. Bainbridge, "Chapter One - Memorability: How what we see influences what we remember," in *Psychology of Learning and Motivation*, ser. Knowledge and Vision, K. D. Federmeier and D. M. Beck, Eds. Academic Press, Jan. 2019, vol. 70, pp. 1–27.

[26] ——, "The memorability of people: Intrinsic memorability across transformations of a person's face," *Journal of Experimental Psychology. Learning, Memory, and Cognition*, vol. 43, no. 5, pp. 706–716, May 2017.

[27] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, "Understanding and Predicting Image Memorability at a Large Scale," in *ICCV*, 2015.

[28] R. Dubey, J. Peterson, A. Khosla, M.-H. Yang, and B. Ghanem, "What Makes an Object Memorable?" in *ICCV*, 12 2015.

[29] N. C. Rust and V. Mehrpour, "Understanding image memorability," *Trends in cognitive sciences*, vol. 24, no. 7, pp. 557–568, Jul. 2020.

[30] C. M. Bird, J. L. Keidel, L. P. Ing, A. J. Horner, and N. Burgess, "Consolidation of Complex Events via Reinstatement in Posterior Cingulate Cortex," *Journal of Neuroscience*, vol. 35, no. 43, pp. 14 426–14 434, Oct. 2015.

[31] C. Baldassano, J. Chen, A. Zadbood, J. W. Pillow, U. Hasson, and K. A. Norman, "Discovering Event Structure in Continuous Narrative Perception and Memory," *Neuron*, vol. 95, no. 3, pp. 709–721.e5, Aug. 2017.

[32] J. Fajtl, V. Argyriou, D. Monekosso, and P. Remagnino, "AMNet: Memorability Estimation with Attention," in *CVPR*, Jun. 2018, pp. 6363–6372.

[33] H. Squalli-Houssaini, N. Duong, M. Gwenaelle, and C.-H. Demarty, "Deep learning for predicting image memorability," in *ICASSP*, 2018.

[34] S. Perera, A. Tal, and L. Zelnik-Manor, "Is Image Memorability Prediction Solved?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Jun. 2019, pp. 800–808.

[35] T. Zhu, F. Zhu, H. Zhu, and L. Li, "Aesthetics-Assisted Multi-task Learning with Attention for Image Memorability Prediction," in *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2020.

[36] L. Goetschalckx, A. Andonian, A. Oliva, and P. Isola, "GANalyze: Toward Visual Definitions of Cognitive Image Properties," in *ICCV*, 2019.

[37] O. Sidorov, "Changing the image memorability: From basic photo editing to gans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[38] C. Kyle-Davidson, A. G. Bors, and K. K. Evans, "Generating memorable images based on human visual memory schemas," *arXiv preprint arXiv:2005.02969*, 2020.

[39] T. Dumont, J. Hevia, and C. L. Fosco, "Modular memorability: Tiered representations for video memorability prediction," in *CVPR*, 2023.

[40] H. S. I, S. Singh, Y. K. Singla, A. Bhattacharyya, V. Baths, C. Chen, R. Ratn Shah, and B. Krishnamurthy, "Long-Term Memorability On Advertisements," *arXiv:2309.00378v1*, 2023.

[41] J. Snyder, *The Visual Made Verbal: A Comprehensive Training Manual and Guide to the History and Applications of Audio Description*. Dog Ear Publishing, LLC, 2014.

[42] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, "Movie description," *IJCV*, 2017.

[43] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[44] Meta, "The Llama 3 herd of models," *arXiv preprint arXiv: 2407.21783*, 2024.

[45] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," in *ICLR*, 2020.

[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019.

[47] J. Xie, C. Yang, W. Xie, and A. Zisserman, "Moving object segmentation: All you need is sam (and flow)," in *ACCV*, 2024.

[48] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio," in *Interspeech*, 2023.

[49] S. Bird, E. Loper, and E. Klein, *Natural Language Processing with Python*. O'Reilly Media, 2009.

[50] A. Pavel, G. Reyes, and J. P. Bigham, "Rescribe: Authoring and Automatically Editing Audio Descriptions," in *ACM Symposium on User Interface Software and Technology (UIST)*, 2020.

[51] A. Aka, S. Bhatia, and J. McCoy, "Semantic determinants of memorability," *Cognition*, vol. 239, p. 105497, Oct. 2023.

[52] C. R. Madan, "Exploring word memorability: How well do different word properties explain item free-recall probability?" *Psychonomic Bulletin & Review*, vol. 28, no. 2, pp. 583–595, Apr. 2021.

[53] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search," *Psychological review*, vol. 113, 2006.

[54] R. Cohendet, K. Yadati, Q. Ngoc, and C.-H. Demarty, "Annotating, understanding, and predicting long-term video memorability," in *ICMR '18: 2018 International Conference on Multimedia Retrieval*, 2018, pp. 178–186.

[55] S. Shekhar, D. Singal, H. Singh, M. Kedia, and A. Shetty, "Show and recall: Learning what makes videos memorable," in *ICCV Workshops*, 2017, pp. 2730–2739.

[56] B. Cheng, A. G. Schwing, and A. Kirillov, "Per-Pixel Classification is Not All You Need for Semantic Segmentation," 2021.

[57] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *CVPR*, 2018, pp. 1209–1218.

[58] S.-M. Khaligh-Razavi, W. A. Bainbridge, D. Pantazis, and A. Oliva, "From what we perceive to what we remember: Characterizing representational dynamics of visual memorability," *bioRxiv doi: 10.1101/049700*, 2016.

[59] J. E. Kosie and D. Baldwin, "Attentional profiles linked to event segmentation are robust to missing information," *Cognitive Research: Principles and Implications*, vol. 4, no. 1, p. 8, Mar. 2019.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, 2016.

[61] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *arXiv: 2103.00020*, 2021.

[62] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," 2017.

[63] L. J. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," *arXiv: 1607.06450*, 2016.

[64] L. Sweeney, M. G. Constantin, C.-H. Demarty, C. Fosco, A. G. S. de Herrera, S. Halder, G. Healy, B. Ionescu, A. Matran-Fernandez, A. F. Smeaton *et al.*, "Overview of the mediaeval 2022 predicting video memorability task," *arXiv preprint arXiv:2212.06516*, 2022.

[65] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal Relational Reasoning in Videos," in *ECCV*, 2018.

[66] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *ICLR*, 2015.

[67] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," *arXiv preprint arXiv:2111.09734*, 2021.

[68] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva, "Intrinsic and extrinsic effects on image memorability," *Vision research*, vol. 116, pp. 165–178, 2015.

[69] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 3, pp. 740–757, 2018.

[70] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *ICCV*. IEEE, 2009, pp. 2106–2113.

[71] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 55–69, 2012.

[72] N. Ouerhani, H. Hügli, R. Müri, and R. Wartburg, "Empirical validation of the saliency-based model of visual attention," *Electronic Letters on Computer Vision and Image Analysis*, 2004.

[73] B. Tatler, R. Baddeley, and I. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vision research*, 2005.

[74] U. Rajashekar, L. Cormack, and A. Bovik, "Point of gaze analysis reveals visual search strategies," *Proceedings of SPIE - The International Society for Optical Engineering*, 2004.

[75] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of vision*, vol. 8, no. 7, pp. 32–32, 2008.

[76] S. Jia and N. Bruce, "Revisiting Saliency Metrics: Farthest-Neighbor Area Under Curve," in *CVPR*, 2020.

[77] M. Lyu, K. W. Choe, O. Kardan, H. Kotabe, J. Henderson, and M. Berman, "Overt attentional correlates of memorability of scene images and their relationships to scene semantics," *Journal of Vision*, vol. 20, 2020.

[78] M. Kümmerer, T. Wallis, L. Gatys, and M. Bethge, "Understanding Low- and High-Level Contributions to Fixation Prediction," in *ICCV*, 2017.

[79] Z. Teed and J. Deng, "RAFT: recurrent all-pairs field transforms for optical flow," *CoRR*, vol. abs/2003.12039, 2020.