

# More than a Moment: Towards Coherent Sequences of Audio Descriptions

Eshika Khandelwal<sup>1</sup>, Junyu Xie<sup>2</sup>, Tengda Han<sup>2</sup>, Max Bain<sup>2</sup>,

Arsha Nagrani<sup>2</sup>, Andrew Zisserman<sup>2</sup>, GüL Varol<sup>2,3</sup>, Makarand Tapaswi<sup>1</sup>

<sup>1</sup>CVIT, IIIT Hyderabad

<sup>2</sup>VGG, University of Oxford

<sup>3</sup>LIGM, École des Ponts, IP Paris, UGE, CNRS

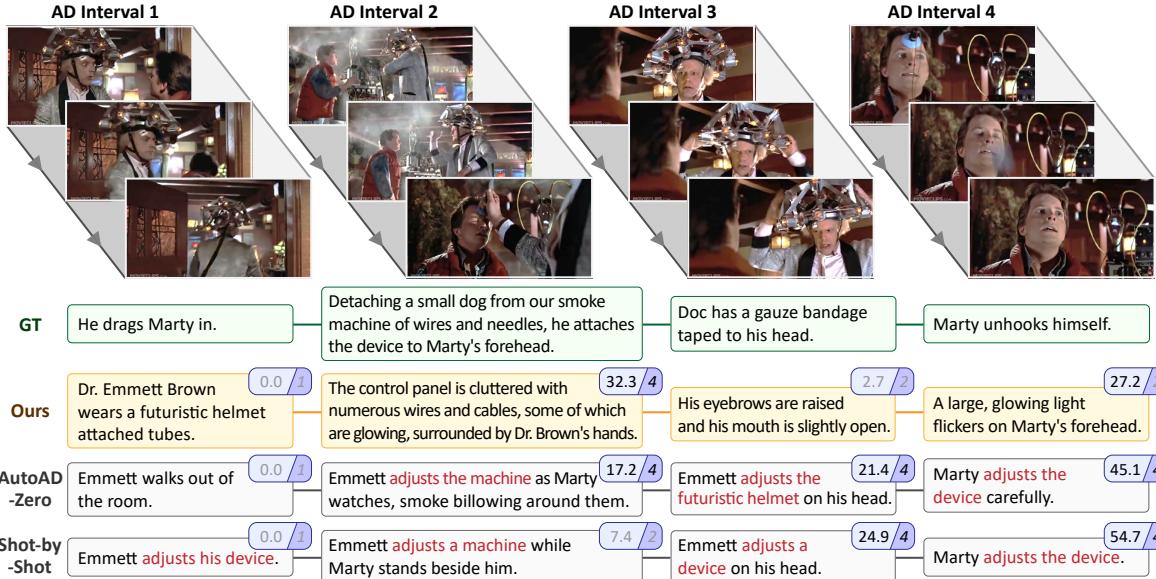


Figure 1: **Predicted ADs across the video** (i.e. a sequence of AD intervals). The results reported by per-AD evaluation metrics are shown on the top right of each prediction (left: CIDEr; right: LLM-AD-Eval, score 1-5), with low scores indicating poor performance coloured in grey. The repetitions across 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 0:22 – 1:05, that can be watched here: <https://www.youtube.com/watch?v=SR5BfQ4rEqQ&t=22s>.

## Abstract

Audio Descriptions (ADs) convey essential on-screen information, allowing visually impaired audiences to follow videos. To be effective, ADs must form a coherent sequence that helps listeners to 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.

## 1 Introduction

Audio Descriptions (ADs) help the visually impaired follow a movie or a 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 (Snyder, 2014), there is growing interest in automatic generation (Han et al., 2023b,a, 2024; Xie et al., 2024a; Fang et al., 2025; Park et al., 2025).

AD generation is historically treated as video captioning, i.e. a short description is generated independently for each predefined *AD interval* in the video (Rohrbach et al., 2017; Soldan et al., 2022; Han et al., 2024). However, ADs are a coherent sequence of descriptions that build a visual story and take the narrative forward. As seen in Fig. 1, this independent generation (e.g. AutoAD-Zero (Xie et al., 2024a), Shot-by-Shot (Xie et al., 2025)) results in repeating similar information, and failing to capture the narrative structure of the movie.

On the other hand, mainstream AD evaluations

are typically conducted on a *per-AD* basis, with CIDEr (Vedantam et al., 2015) and LLM-AD-Eval (Han et al., 2024) being widely adopted. As shown in Fig. 1, both metrics 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 ground truth, overlooking the fact that each time interval may encompass multiple valid descriptions.

We therefore posit that both AD generation and evaluation 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. 1).

In this work, 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 (Xie et al., 2024a), we first extract structured information from each trimmed clip. 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.

For evaluation, we move away from conventional metrics that compare ground truth (GT) and predicted ADs for a single interval. Instead, we adopt: (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. We evaluate sequence-level AD generation on CMD-AD and TV-AD videos and observe qualitative and quantitative improvements in generated ADs, further validated through user studies.

## 2 Related Work

**AD generation** aims to produce concise, coherent narrations of the salient visual content that complement auditory signals. Prior work falls into two categories: (i) *End-to-end models* (Han et al., 2023b,a, 2024; Wang et al., 2024a; Lin et al., 2024; Wang et al., 2025; Fang et al., 2025; Ye et al., 2025), which are fine-tuned on domain-specific AD datasets (Soldan et al., 2022; Han et al., 2023b; Xie et al., 2024b); (ii) *Training-free frameworks* (Zhang et al., 2024; Ye et al., 2024; Chu et al., 2024; Xie et al., 2024a; Park et al., 2025), which adopt multi-stage setups built on pre-trained Vision-Language Models (VLMs) and Large Language Models (LLMs).

While early methods generate ADs independently for each time segment, recent works

focus on maintaining narrative coherence and reducing redundancy across consecutive outputs. AutoAD-I (Han et al., 2023b) and UniAD (Wang et al., 2025) adopt recursive generation processes that condition on previous AD outputs. AutoAD-II (Han et al., 2023a) trains a localisation module to predict temporal segments for AD injection. DistinctAD (Fang et al., 2025) jointly processes adjacent AD clips and reduces redundancy via a Contextual Expectation-Maximisation Attention mechanism. Our work explores coherent AD generation in a training-free setup.

**AD evaluation.** Early AD evaluation adopts captioning metrics, including n-gram overlap metrics such as CIDEr (Vedantam et al., 2015), ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005), as well as semantic-oriented ones like SPICE (Anderson et al., 2016) and BERTScore (Zhang\* et al., 2020). More recent efforts have introduced AD-specific evaluations, such as retrieval-based (e.g. Recall@k/N (Han et al., 2023b)) and LLM-based assessments (Han et al., 2024; Zhang et al., 2024). Other metrics focus on specific aspects of AD quality, for example, CRITIC (Han et al., 2024) for character accuracy and “Action Score” (Xie et al., 2025) for action groundedness.

Beyond single-AD evaluation, few metrics assess coherence and redundancy across consecutive ADs. Lin et al. (2024) measure n-gram repetition using R@4, while Ye et al. (2025) introduce a redundancy-aware metric based on semantic similarity. Zhang et al. (2024) propose SegEval to score short AD windows using GPT-4 (OpenAI, 2024); while (Kala et al., 2025) proposes a question-answering based evaluation. In this work, we propose a suite of metrics for multi-AD evaluations, focusing on repetitions and overall storyline coherence.

## 3 Sequence-Level AD Generation

Given a sequence of predefined video intervals, our goal is to generate the corresponding AD texts. We introduce CoherentAD, a *training-free* method that generates multiple candidate ADs per interval and selects a coherent, non-redundant sequence of ADs. The method consists of three stages (see Fig. 2; ablated in Sec. B.1): (i) video interval description; (ii) multiple AD generation for each interval; and (iii) selection of the optimal AD sequence.

### 3.1 Video interval to summarised narrative

We describe the visual frames within each AD interval by first extracting structured textual descriptions of the visual content with a VLM, and then by

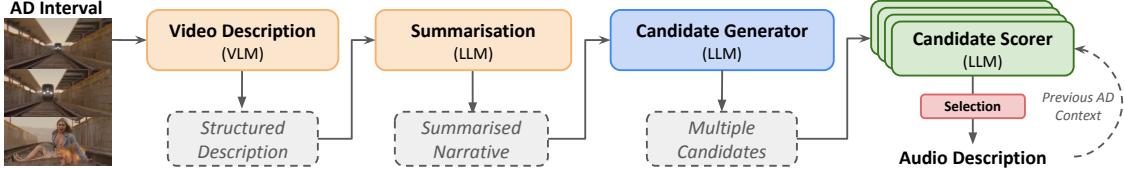


Figure 2: **Overview of our multi-stage AD generation pipeline CoherentAD.** For each AD interval, the VLM generates a structured description, which is then summarised. The summary is used to produce multiple candidate descriptions. Each candidate is scored by four independent LLM-based scorers that consider previous selections as context. The highest-scoring candidate is selected in an auto-regressive manner to form a coherent sequence.

summarising them with an LLM. Specifically, we first prompt a VLM using a three-part instruction (see Sec. A.1) to extract all visually relevant details in a structured format. Since later stages rely solely on text, any missed detail is unrecoverable, making it essential to capture a complete description at this stage. This description also supports diverse candidate generation later in the pipeline. Second, we use an LLM to rephrase the exhaustive structured visual description into a concise, yet complete paragraph  $P$  that preserves all relevant content. This serves as the basis for the next stage.

### 3.2 Multiple AD candidates per interval

Given the paragraph  $P$ , we prompt 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 the full 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 suitable for inclusion in the final sequence on its own. Not forcing a fixed number, but allowing less than  $m$  candidates avoids redundancy across candidates, while preventing fragmentation.

### 3.3 Coherent sequence selection

Finally, to form a sequence of AD predictions over an input video, we perform auto-regressive selection across intervals, conditioning on previously selected ADs. This selection process is guided by four scoring criteria, with each score assigned independently by a separate LLM scorer to ensure focused and unbiased judgment for each aspect. (i) *Adherence to AD guidelines* (Snyder, 2014) checks if the description is strictly grounded in what is visually perceptible (e.g. no inferences, speculation, or camera references) and focuses only on what a visually impaired viewer cannot directly access. (ii) *Redundancy* penalises repeated content from prior descriptions. (iii) *Story advancement* prioritises candidates that introduce new observable actions, interactions, or scene changes that move the narrative forward. (iv) *Counts of visual elements* tallies the number

of unique participants, actions, and salient visual details explicitly mentioned in the description, rewarding candidates that convey more information.

The final score for each candidate is computed as a weighted average of these four criteria, and the highest-scoring one is selected auto-regressively, conditioned on the previous  $r$  selected descriptions.

## 4 Sequence-Level AD Evaluation

Traditional AD evaluation independently compares each description to a single ground truth (GT) (Kala et al., 2025). 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 metrics. *StoryRecall* evaluates whether the generated sequence captures the same visual story as the reference GT sequence (Sec. 4.1), and *repetition* metrics measure redundancy. (Sec. 4.2).

### 4.1 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 (Kala et al., 2025), the full sequence collectively captures the core visual events, making it a reliable reference.

We concatenate GT and predicted ADs for each interval and compare resulting AD sequences using an LLM. A score from 1 to 5 reflects 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.

### 4.2 Repetition metrics

We also propose to explicitly monitor repetition in AD evaluation, as it reduces 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.

Method	Train Free	CMD-AD				TV-AD			
		SR ↑	Exact Repeat % ↓	Partial Repeat % ↓	SR ↑	Exact Repeat % ↓	Partial Repeat % ↓		
Ground Truth	-	-	( 0, 0, 0)	( 3.71, 4.42, 3.97)	-	( 0, 0, 0)	( 4.56, 4.38, 3.70)		
AutoAD-III (Han et al., 2024)	✗	2.11	(4.51, 2.81, 1.99)	(16.01, 11.90, 10.44)	-	-	-	-	-
UniAD (Wang et al., 2025)	✗	2.13	(4.15, 2.44, 2.10)	(16.21, 12.27, 11.57)	-	-	-	-	-
AutoAD-Zero† (Xie et al., 2024a)	✓	2.27	(0.19, 0.32, 0.30)	( 7.87, 8.13, 8.13)	1.69	(1.15, 1.23, 0.32)	(21.78, 17.47, 14.83)		
AutoAD-Zero	✓	2.27	(0.55, 0.31, 0.21)	(13.26, 10.62, 9.59)	1.79	(0.89, 0.18, 0.16)	(19.23, 14.90, 13.74)		
Shot-by-Shot (Xie et al., 2025)	✓	2.43	(0.42, 0.15, 0.12)	(19.06, 15.18, 13.56)	<b>1.84</b>	(0.73, 0.35, 0.16)	(21.19, 15.89, 13.46)		
CoherentAD (Ours)	✓	<b>2.63</b>	( 0, 0, 0)	( <b>5.21</b> , <b>4.45</b> , <b>4.18</b> )	1.83	( 0, 0, 0)	( <b>6.16</b> , <b>5.36</b> , <b>4.26</b> )		
w/o multiple candidates	✓	2.49	(0.04, 0, 0.02)	( 8.17, 6.78, 6.19)	-	-	-	-	-

Table 1: **Quantitative comparison on CMD-AD and TV-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. SR denotes StoryRecall.

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.

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 (Bird et al., 2009). As with exact repeats, we report three scores for the three offsets.

## 5 Experiments

**Datasets and details.** We evaluate on (i) CMD-AD (Han et al., 2024), with 7,316 ADs for 591 videos of 98 movies, and (ii) TV-AD (Xie et al., 2024a), with 2,983 ADs spanning 100 episodes across two TV series. We use Qwen2-VL-7B (Wang et al., 2024b) as the VLM and LLaMA3.1-Instruct-8B (Meta, 2024) as the LLM. We generate up to  $m=5$  candidates per interval and condition scoring on the  $r=3$  previously selected candidates. Additional details and ablations in Secs. A and B.

**Quantitative results.** Tab. 1 reports the performance on CMD-AD and TV-AD for both fine-tuned (top) and training-free (bottom) methods, where all prior works fall short on our sequence-level metrics. For instance, Shot-by-Shot (SbS) produces higher partial repetitions (19.06%) between consecutive predictions than the ground truth (3.71%). In contrast, CoherentAD achieves repetition scores that closely match GT (5.21%), with zero exact repeats. Our method prioritises sequence-level coherence by design, outperforming previous state-of-the-art on *StoryRecall* (2.63 vs 2.43) on CMD-AD, while achieving comparable performance for TV-AD.

Notably, the ablation without multiple candidate generation achieves significantly lower repetition

vs SbS (8.17% vs 19.06%). The latter uses concise VLM outputs highlighting the most prominent character, action, or interaction, often leading to redundant phrasing. Our setup aggregates wider VLM outputs, resulting in diverse and informative descriptions. Further, shifting to our multi-candidate setup increases StoryRecall (2.49 to 2.63) and decreases repetition (8.17% to 5.21%).

Finally, we also see good results on the ADQA benchmark (Kala et al., 2025) in Sec. C.

**Qualitative results** are shown in Fig. 1 and Sec. D. Repetitions are clearly observed in baseline predictions, while CoherentAD introduces distinct, visibly grounded details at each interval forming a coherent and non-redundant narrative.

## 6 Conclusion

We highlighted the limitations of current methods, producing repetitions and incoherence in sequential ADs. CoherentAD, our training-free approach attempted to address this through multiple coherence criteria, and posing the problem as sequence search among multiple candidates per AD interval.

**Acknowledgements.** This project was funded in part by the ANR project CorVis ANR-21-CE23-0003-01 and a research gift from Google. It was also supported by an SERB SRG/2023/002544 grant. The authors also thank Divy Kala for evaluating generated ADs on ADQA.

## Limitations

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.

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 post-processing. Editing or merging candidates could further enhance sequence-level fluency.

## References

- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*.
- Satanjeev Banerjee and Alon Lavie. 2005. 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*.
- Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media.
- Peng Chu, Jiang Wang, and Andre Abrantes. 2024. LLM-AD: Large language model based audio description system. *arXiv preprint arXiv:2405.00983*.
- Cursor. Cursor: Ai-powered code editor. <https://www.cursor.sh>.
- Bo Fang, Wenhao Wu, Qiangqiang Wu, Yuxin Song, and Antoni B. Chan. 2025. DistinctAD: Distinctive audio description generation in contexts. In *CVPR*.
- Tengda Han, Max Bain, Arsha Nagrani, Gü̈l Varol, Weidi Xie, and Andrew Zisserman. 2023a. AutoAD II: The sequel – who, when, and what in movie audio description. In *ICCV*.
- Tengda Han, Max Bain, Arsha Nagrani, Gü̈l Varol, Weidi Xie, and Andrew Zisserman. 2023b. Autoad: Movie description in context. In *CVPR*.
- Tengda Han, Max Bain, Arsha Nagrani, Gü̈l Varol, Weidi Xie, and Andrew Zisserman. 2024. AutoAD III: The prequel – back to the pixels. In *CVPR*.
- Divy Kala, Eshika Khandelwal, and Makarand Tapaswi. 2025. What You See is What You Ask: Evaluating audio descriptions. In *EMNLP*.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.
- Kevin Qinghong Lin, Pengchuan Zhang, Difei Gao, Xide Xia, Joya Chen, Ziteng Gao, Jinheng Xie, Xuhong Xiao, and Mike Zheng Shou. 2024. Learning video context as interleaved multimodal sequences. In *ECCV*.
- Meta. 2024. The Llama 3 herd of models. *arXiv preprint arXiv: 2407.21783*.
- OpenAI. Chatgpt: Language model for dialogue. <https://openai.com/chatgpt>.
- OpenAI. 2024. Gpt-4 technical report. *arXiv preprint arXiv: 2303.08774*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.
- Jaehyeong Park, Juncheol Ye, Seungkook Lee, Hyun W. Ka, and Dongsu Han. 2025. NarrAD: Automatic generation of audio descriptions for movies with rich narrative context. In *WACV*.
- Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *IJCV*.
- Joel Snyder. 2014. *The Visual Made Verbal: A Comprehensive Training Manual and Guide to the History and Applications of Audio Description*. Dog Ear Publishing, LLC.
- Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. 2022. MAD: A scalable dataset for language grounding in videos from movie audio descriptions. In *CVPR*.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *CVPR*.
- Hanlin Wang, Zhan Tong, Kecheng Zheng, Yujun Shen, and Limin Wang. 2025. Contextual AD narration with interleaved multimodal sequence. In *CVPR*.
- Jiayi Wang, Zihao Liu, and Xiaoyu Wu. 2024a. LoCo-MAD: Long-range context-enhanced model towards plot-centric movie audio description. In *ACCV*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Eshika Khandelwal, Gü̈l Varol, Weidi Xie, and Andrew Zisserman. 2025. Shot-by-shot: Film-grammar-aware training-free audio description generation. In *ICCV*.
- Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Gü̈l Varol, Weidi Xie, and Andrew Zisserman. 2024a. Autoad-zero: A training-free framework for zero-shot audio description. In *ACCV*.
- Junyu Xie, Weidi Xie, and Andrew Zisserman. 2024b. Appearance-based refinement for object-centric motion segmentation. In *ECCV*.

Junyu Xie, Chargin Yang, Weidi Xie, and Andrew Zisserman. 2024c. Moving object segmentation: All you need is sam (and flow). In *ACCV*.

Xiaojun Ye, Junhao Chen, Xiang Li, Haidong Xin, Chao Li, Sheng Zhou, and Jiajun Bu. 2024. MMAD: Multi-modal movie audio description. In *LREC-COLING*.

Xiaojun Ye, Chun Wang, Yiren Song, Sheng Zhou, Liangcheng Li, and Jiajun Bu. 2025. FocusedAD: Character-centric movie audio description. *arXiv preprint arXiv:2504.12157*.

Chaoyi Zhang, Kevin Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2024. MM-Narrator: Narrating long-form videos with multimodal in-context learning. In *CVPR*.

Tianyi Zhang\*, Varsha Kishore\*, Felix Wu\*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *ICLR*.

## Appendix

Sec. A provides additional implementation details of our multi-stage pipeline, Sec. B presents various additional ablations and experiments, and Sec. C evaluates CoherentAD on the new AD evaluation benchmark, ADQA, showing promising results, and Sec. D shows qualitative examples.

## A Implementation Details

### A.1 Video interval to summarised narrative

We complement Sec. 3.1 with additional details.

**Extracting structured visual descriptions.** We uniformly sample 16 frames per AD interval, following standard practice in Xie et al. (2024a, 2025). For character recognition, we adopt the method of Xie et al. (2024a), which overlays coloured circles around detected faces and uses them to provide character names as text prompts.

We found that when shown a sequence of sampled frames, the VLM often defaults to static, high-level descriptions (e.g. “a man is standing”) rather than capturing actions and changes unfolding over time. To address this, we prompt the VLM with a structured instruction that guides the model to extract all visually relevant details across three aspects—key actions, interactions, and environmental changes, respectively. The exact prompt is provided in Algorithm 1, which consists of:

- *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 clear and subtle), interactions, and any environmental changes;
- *Overall Summary* - a brief description of the primary event.

**Narrative summarisation.** We instruct the LLM to retain all meaningful visual content while rephrasing the dense structured description into a concise paragraph (using Algorithm 2). It explicitly discourages inference, dialogue, and speculation, resulting in a grounded description suitable for the candidate generation stage. To ensure adherence to the guidelines, the model drafts and iteratively refines the paragraph until it satisfies all specified constraints. The resulting output is compact and fluent, serving as the basis for the next stage.

Overall, this two-step design allows us to first extract all relevant visual details, and then organise

them into a concise, event-level narrative suitable for Audio Description.

### A.2 Multiple candidate generation

As explained in Sec. 3.2, we instruct the LLM to generate up to  $m$  candidate diverse ADs, each having a word limit of  $l_{\max}$ . Here,  $l_{\max}$  adapted from Xie et al. (2025), treated as target length rather than an upper bound, prompting each candidate to be as informative as possible. The prompt (Algorithm 3) encourages grouping related observations, resulting in each candidate conveying a complete and cohesive visual event.

### A.3 Criterion weights for sequence selection

In Sec. 3.3, all 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). Prompts for each criterion can be found in Algorithms 4 to 7. The weights were chosen to reflect the intended focus of the selection: the highest weight is given to whether the AD follows guidelines and advances the story, followed by redundancy, and the least to the number of participants, actions, and other details.

We experimented with ten additional configurations on CMD-AD by randomly varying each weight between 0.02 – 0.05 while preserving the relative ordering. The results remain consistent, with StoryRecall at  $2.60 \pm 0.02$  and partial repeats at  $(5.27 \pm 0.02, 4.41 \pm 0.02, 4.14 \pm 0.02)$ , indicating that small shifts in weights do not meaningfully affect performance.

### A.4 Model sizes and hardware setup

Our pipeline uses open-source models with 7B and 8B parameters: Qwen2-VL-7B (Wang et al., 2024b) for extracting structured visual descriptions, and LLaMA-3.1-Instruct-8B (Meta, 2024) for candidate generation and scoring. All experiments are run on NVIDIA A6000 GPUs, with a single run per setting; all reported values reflect these single runs.

### A.5 Computational cost

On average, processing a single AD interval takes approximately 22 seconds, with around 15 seconds spent on the VLM step (see Tab. A.2, for a detailed breakdown). Note, this process can also be parallelised across video segments.

We use a lightweight LLaMA3-8B model for efficient inference. Additionally, since AD generation is an offline process and not intended for real-time

Method Name	VLM: Video Desc.	LLM 1: Summ.	LLM 2: Candidate Gen.	LLM 3: Scorer	Story Recall ↑	Exact Repeat% ↓	Partial Repeat% ↓
Coherent AD	✓	✓	✓	✓	2.63 ( 0, 0, 0)	( 5.21, 4.45, 4.18)	
	✓	-	✓	✓	2.58 ( 0, 0, 0)	( 6.23, 5.32, 4.88)	
	✓	✓	✓	-	2.49 (0.04, 0, 0.02)	( 8.17, 6.78, 6.19)	
VLM only	✓	-	-	-	2.38 (1.37, 0.62, 0.41)	(15.31, 12.20, 10.92)	

Table A.1: **Ablating pipeline stages.** The full multi-stage pipeline, comprising video description (VLM), summarization (LLM 1), candidate generation (LLM 2), and scoring (LLM 3), achieves the best overall performance with highest story recall and lowest repetitions.

Step	Seconds
Structured descriptions	14.88
Summarisation	1.46
Multiple candidates	0.66
Calculating Counts	0.87
AD guidelines compliance	1.57
Story + Redundancy + recursively selecting	2.60
<b>Total Time per AD</b>	<b>22.05</b>

Table A.2: **Processing time per AD.** Breakdown of the average time (in seconds) taken by each stage in the pipeline, totalling 22.05 seconds per description.

use, the computational costs are incurred only once. Overall, the time and cost remains practical relative to typical movie production timelines and budgets.

## A.6 Use of AI Assistants

We use [Cursor](#) for code autocomplete, and [OpenAI](#) to test zero-shot prompts during development.

## B Additional Experiments

### B.1 Ablation of pipeline stages

While using a single VLM or combining all stages may seem simpler, we found it significantly reduced output quality. Each stage (event extraction, candidate generation, and scoring) involves complex, distinct instructions (see Algorithms 1 to 7), which a single model (of comparable size) struggles to handle reliably. Breaking the process into smaller tasks helps ensure that each instruction is properly followed.

In Tab. A.1, we ablate the multiple stages. First, we evaluated a baseline where the VLM directly generates a single AD (row 4). This performs noticeably worse, with a StoryRecall of 2.38 and substantially higher repetition, highlighting the limitations of relying solely on the VLM. To guide the VLM toward producing a single AD without overwhelming it with detailed instructions, we replaced the three-part storyboard-style VLM prompt (Algorithm 1) with a simplified version that retains the core AD constraints: one present-tense sentence describing the visible action or interaction, using

provided character names, with no inference or emotion, and within the word limit. Even then, the VLM often failed to comply with these constraints.

Second, we removed the summarisation step from our pipeline (row 2) while still generating five candidates and applying the same scoring mechanism. This also leads to a performance drop (StoryRecall: 2.58 vs. 2.63) and increased partial repeats. Without summarisation, the candidate descriptions are less diverse and often repeat the same visual event. The pipeline produces candidates containing actions that violate AD guidelines (such as “talking” or “conversing”) which are then discarded during scoring, reducing the pool of viable options. To accommodate the shift from a short paragraph to a longer and structured input, we adapted the prompt by retaining only the core AD constraints from Algorithm 3: up to five present-tense sentences, grounded in visible content, distinct, standalone, and within the word limit. This was necessary, as the original, more detailed prompt led to verbose outputs that exceeded AD timing constraints.

Finally, row 1 is our proposed multi-stage pipeline while row 3 without multiple candidate generation and scoring was included in Tab. 1. We observe that our multi-stage approach outperforms all other variants.

### B.2 Varying context during scoring

Tab. A.4 shows results for different values of  $r$ , the number of prior descriptions used in Sec. A.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 plateaus beyond  $r = 3$ . While  $r = 5$  yields the lowest overall repetition (5.13, 4.45, 4.10), it does not improve StoryRecall (2.61).  $r=3$  achieves the best balance with the highest StoryRecall (2.63) and low overall repetition (5.21, 4.45, 4.18). Exact repeats are zero in all cases.

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	✓	<b>25.0</b>	2.01	14.9	49.3	<b>63.2</b>	<b>86.9</b>
UniAD*	✓	21.8	<b>2.92</b>	14.3	47.4	63.0	86.6
AutoAD-Zero	✗	17.7	1.96	13.4	44.3	62.9	86.5
Q2VL	✗	-	-	<b>17.2</b>	<b>57.0</b>	51.2	70.4
CoherentAD(Ours)	✗	13.2	2.17	<b>15.2</b>	<b>50.3</b>	<b>64.0</b>	<b>88.0</b>
AV <sub>1</sub>	-	-	-	-	-	72.7	100
AV <sub>2</sub> (17)	-	-	-	30.2	100	75.0	103

Table A.3: Evaluation of CoherentAD on ADQA. Acronyms are as follows: Vis App: Visual Appreciation, Narr Und: Narrative Understanding. The old metrics that compare GT and predicted ADs one-to-one are C: CIDEr and LLMe: LLM-AD-eval (Han et al., 2024). The new metrics proposed in ADQA are CC: correct answer using context, and Ratio: Accuracy ratio.

<i>r</i>	StoryRecall $\uparrow$	Partial Repeat% $\downarrow$
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)
4	2.58	(5.20, 4.44, 4.13)
5	2.61	(5.13, 4.45, 4.10)

Table A.4: Varying the number of prior descriptions (*r*) during scoring. *r*=3 (default) gives the best StoryRecall and lowest overall repetition across positions.

## C ADQA Benchmark

In Tab. A.3, we evaluate CoherentAD on ADQA (Kala et al., 2025), a multiple-choice question-answering (MCQA) benchmark designed to assess ADs across two key dimensions: Visual Appreciation (VA) and Narrative Understanding (NU). The benchmark is motivated by the core purposes of ADs: (i) enabling BVI audiences to appreciate the visual elements that enrich their experience, and (ii) supporting narrative understanding by conveying essential visual plot points.

**Results.** Q2VL achieves the highest VA accuracy ratio by densely summarising 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 NU with a CC score (correct answers grounded in the provided ADs) 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.

## D Additional Qualitative Results

Fig. A.1 and Fig. A.2 provide additional visualisations comparing our and existing methods.

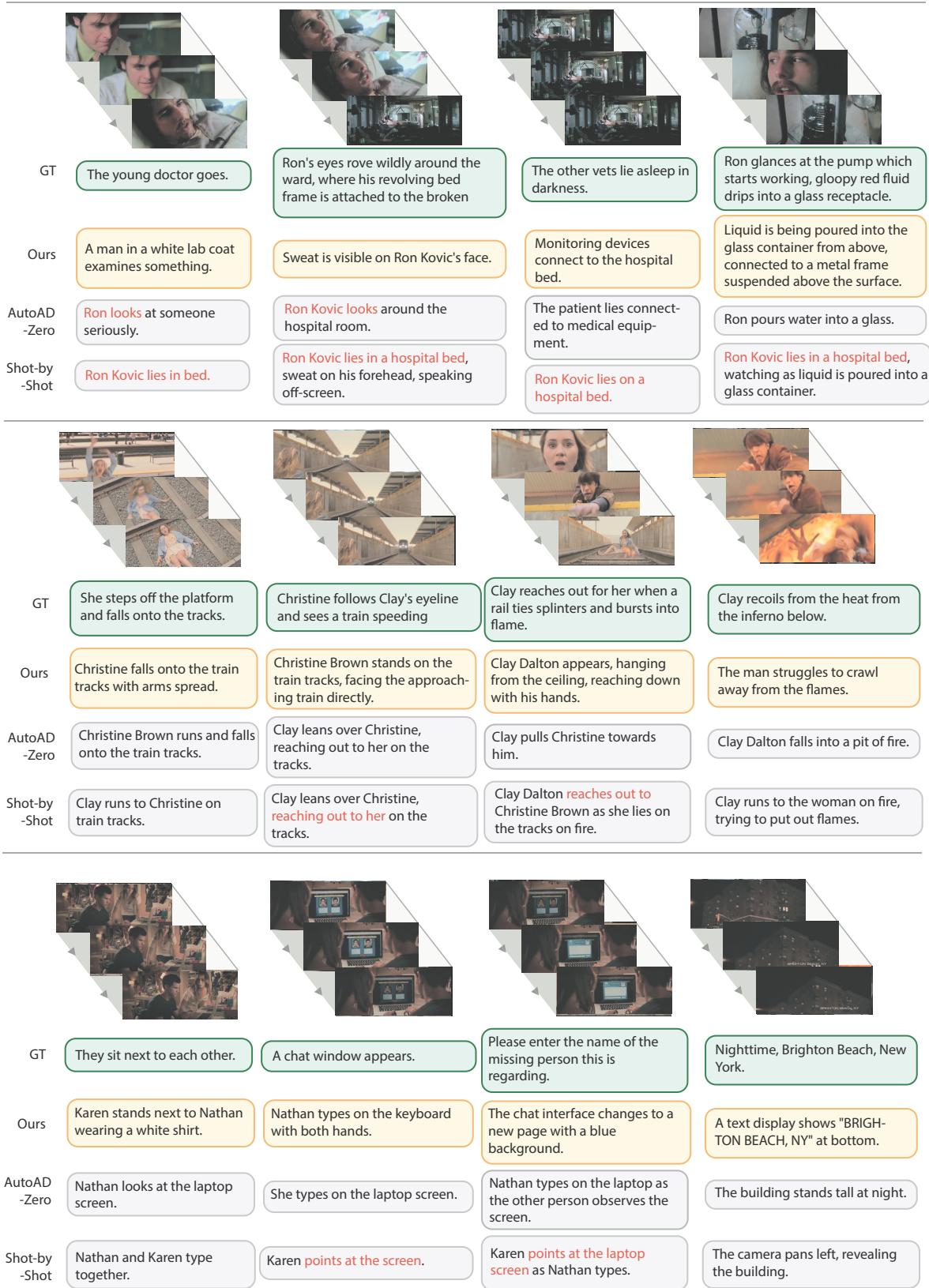


Figure A.1: Qualitative comparison showing GT, our outputs, AutoAD-Zero (Xie et al., 2024c) and Shot-by-Shot (Xie et al., 2025), with repetitions highlighted in red.

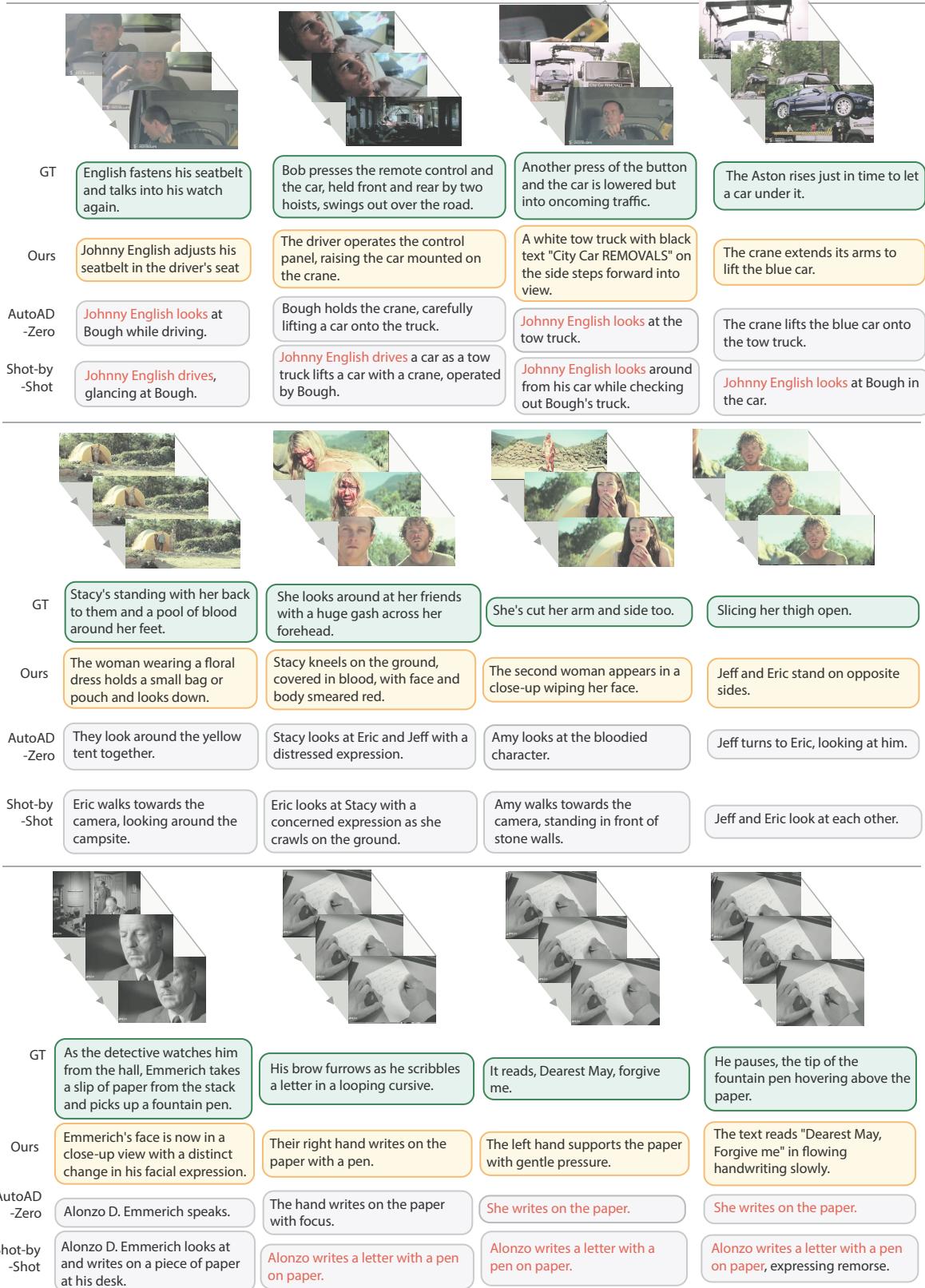


Figure A.2: Qualitative comparison showing GT, our outputs, AutoAD-Zero (Xie et al., 2024c) and Shot-by-Shot (Xie et al., 2025), with repetitions highlighted in red.

---

**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"\n{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"
    "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"
)
```

---