What You See is What You Ask: Evaluating Audio Descriptions

Divy Kala, **Eshika Khandelwal,** **Makarand Tapaswi**
CVIT, IIIT Hyderabad, India
https://katha-ai.github.io/projects/adqa/

Abstract

Audio descriptions (ADs) narrate important visual details in movies, enabling Blind and Low Vision (BLV) users to *understand narratives* and *appreciate visual details*. Existing works in automatic AD generation mostly focus on few-second trimmed clips, and evaluate them by comparing against a single ground-truth reference AD. However, writing ADs is inherently subjective. Through alignment and analysis of two independent AD tracks for the same movies, we quantify the subjectivity in *when* and *whether* to describe, and *what* and *how* to highlight. Thus, we show that working with trimmed clips is inadequate. We propose *ADQA*, a QA benchmark that evaluates ADs at the level of few-minute long, coherent video segments, testing whether they would help BLV users *understand the story* and *appreciate visual details*. ADQA features visual appreciation (VA) questions about visual facts and narrative understanding (NU) questions based on the plot. Through ADQA, we show that current AD generation methods lag far behind human-authored ADs. We conclude with several recommendations for future work and introduce a public leaderboard for benchmarking.

1 Introduction

> *"It's just that my eyes don't work. My brain is perfectly intact. Let me think for myself."*
>
> – anonymous AD consumer

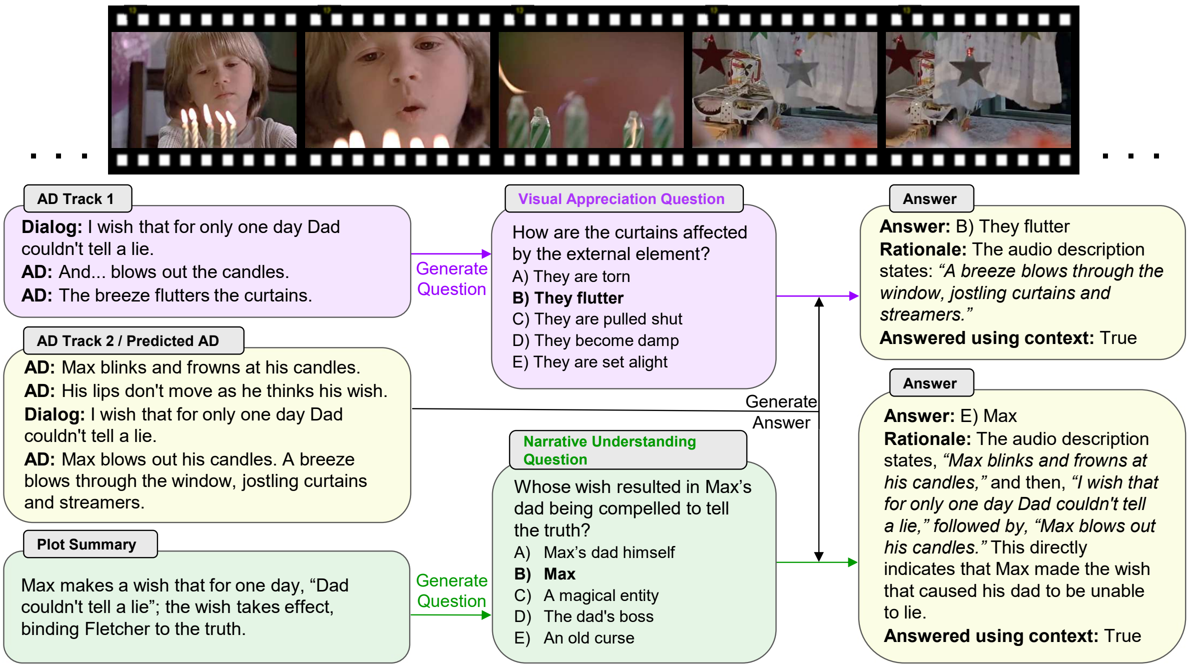
Blind and Low Vision (BLV) individuals watch movies and TV shows with assistance from audio descriptions (ADs). Writing ADs is a complex task that requires experts to identify the most relevant visual elements and describe them in a coherent and concise manner to fit within the gap between dialogs (Pavel et al., 2020). A seminal book on this topic, The Visual Made Verbal (Snyder, 2014), states the oft-referenced "first rule of description" as "What You See Is What You Say." It also emphasizes two central goals of AD: enabling BLV users to (i) *appreciate* the visual richness of a scene; and (ii) *understand* the narrative by providing important (visual) plot points.

In addition to these broad goals, several guidelines (*e.g.* (YouDescribe, 2013)) provide concrete recommendations: (i) focus on visual elements, avoid audible content (unless audio source is ambiguous *e.g.* mixer grinder), (ii) stick to visual facts and do not opine or interpret (*e.g.* beautiful woman), (iii) read on-screen text when relevant to the story (*e.g.* two years later), (iv) provide information just in time, neither too early nor late, (v) balance the amount of information and speed of narration, and (vi) match vocabulary to the material and be concise.

To make content more accessible, there is a rising interest in automatic AD generation (Han et al., 2023b, 2024; Xie et al., 2024; Gao et al., 2024; Chu et al., 2024; Park et al., 2025; Wang et al., 2025). However, even after development of several new methods and metrics (Sec. 2), evaluation treats ADs as isolated captions. Typically, each predicted AD is compared to a single ground-truth reference for each **trimmed clip**—few second video clips trimmed to the duration where the ground-truth ADs are spoken. A challenge with this evaluation is the subjective nature of ADs, *i.e.* two experts may write different ADs for the same video (Fig. 1). For a subset of movies that have two independently narrated AD tracks, we align them and quantify how this subjectivity manifests in Sec. 3. Working with trimmed clips also does not check whether generated ADs would help BLV users appreciate the visual richness of the media or better understand the unfolding story.

To address these gaps, we posit that AD evaluations need to be performed for longer video segments and check whether they assist *visual appreciation* and *narrative understanding*. We propose the

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 23496–23518
November 4-9, 2025 ©2025 Association for Computational Linguistics



AD Track 1

Dialog: I wish that for only one day Dad couldn't tell a lie.
AD: And... blows out the candles.
AD: The breeze flutters the curtains.

AD Track 2 / Predicted AD

AD: Max blinks and frowns at his candles.
AD: His lips don't move as he thinks his wish.
Dialog: I wish that for only one day Dad couldn't tell a lie.
AD: Max blows out his candles. A breeze blows through the window, jostling curtains and streamers.

Plot Summary

Max makes a wish that for one day, "Dad couldn't tell a lie"; the wish takes effect, binding Fletcher to the truth.

Visual Appreciation Question

How are the curtains affected by the external element?
A) They are torn
B) They flutter
C) They are pulled shut
D) They become damp
E) They are set alight

Narrative Understanding Question

Whose wish resulted in Max's dad being compelled to tell the truth?
A) Max's dad himself
B) Max
C) A magical entity
D) The dad's boss
E) An old curse

Answer

Answer: B) They flutter
Rationale: The audio description states: *"A breeze blows through the window, jostling curtains and streamers."*
Answered using context: True

Answer

Answer: E) Max
Rationale: The audio description states, *"Max blinks and frowns at his candles,"* and then, *"I wish that for only one day Dad couldn't tell a lie,"* followed by, *"Max blows out his candles."* This directly indicates that Max made the wish that caused his dad to be unable to lie.
Answered using context: True

Figure 1: We present ADQA's question generation and answering framework. A small part of a video from the film *Liar Liar* from CMD-AD (Han et al., 2024) is shown. AD Track 1 and 2 from AudioVault show the dialogs and ADs describing the video in different ways. The plot summary is taken from CMD (Bain et al., 2020). AD Track 1 is used to create **Visual Appreciation** questions, whereas the plot summary is used to create **Narrative Understanding** questions. LLMs are prompted to answer both question types using the AD track under evaluation, here, AD Track 2. The video can be watched here: `https://youtu.be/IsBB4i4k2PM`.

Audio Description Question Answering (ADQA) benchmark (Fig. 1), a multiple-choice QA framework (Sec. 4) that evaluates ADs of few-minute long videos. **ADQA evaluates generated ADs by mixing them with dialog and using them as context for an LLM to answer questions.** In fact, barring sound effects and speaker identity, dialog and AD are the primary modalities for BLV users.

We create two kinds of QAs. For visual appreciation, we create questions based on ground-truth ADs that ask about specific visual details of a scene. Answering these requires AD models to pick the most prominent visual story element and describe it correctly. For narrative understanding, we create questions based on the story plot. Answering these requires generating ADs that advance the storyline together with dialog.

In summary, our contributions are: (i) ADQA, a new benchmark to evaluate AD generation methods by focusing on a central tenet: would the generated ADs help BLV users *appreciate* and *understand* the story?; (ii) a first detailed analysis of the subjective nature of ADs; (iii) comparative evaluation of AD generation methods through ADQA, hosted as a public leaderboard; and (iv) concrete recommendations for future of AD generation.

2 Related Work

A comprehensive review of AD generation methods, including advances in VLMs/LLMs, is provided by Gao et al. (2024). Below we discuss related datasets, methods, and metrics.

AD datasets. LSMDC (Rohrbach et al., 2017) contains professionally written ADs manually aligned with full-length movies to create trimmed clips. MAD (Soldan et al., 2022) uses AudioVault[1] for training and LSMDC for evaluation. CMD-AD (Han et al., 2024) also uses AudioVault and pairs ADs with videos from the CMD dataset (Bain et al., 2020), while TV-AD (Xie et al., 2024) features ADs for TV shows. ADQA, our benchmark, leverages the 2 most commonly used datasets: MAD-eval and CMD-AD.

Generation methods are of two types: fine-tuned and zero-shot. Fine-tuned approaches adapt pre-trained models with light-weight modules and add character identity (Han et al., 2023b,a, 2024; Wang et al., 2025; Ye et al., 2025, 2024). Among them, DistinctAD (Fang et al., 2025) penalizes repeated content to reduce redundancy and DanteAD (De-

---

[1] AudioVault is a non-profit community-driven repository of movie ADs. See `https://audiovault.net`

ganutti et al., 2025) improves long-term context using transformers. Fill-in (Park et al., 2020) and MiCAP (Raajesh et al., 2024) address identity grounding without explicitly naming characters and building consistent identity clusters across clips. Among the zero-shot methods AutoAD-Zero (Xie et al., 2024) and LLM-AD (Chu et al., 2024) prompt vision-language models (VLMs) with frames and character cues, NarrAD (Park et al., 2025) uses scripts, and MMNarrator (Zhang et al., 2024) proposes a memory-augmentation to handle long video context. We evaluate some of these models on our benchmark.

AD evaluation. Since writing ADs is a subjective process, a large variety of metrics have been proposed over the years. Classically, ADs are evaluated using CIDEr (Rohrbach et al., 2017; Soldan et al., 2022) which compares n-grams between the generated and reference ADs. However, CIDEr penalizes linguistic diversity when only a single reference AD is available, the typical case for AD evaluation (Vedantam et al., 2015).

AutoAD II introduces an alternative to one-to-one AD matching, R@k/N, which rewards semantic relevance. AutoAD III proposes CRITIC for character names and LLM-AD-Eval for sentence-pair semantic scoring using LLMs. Shot-by-Shot (Xie et al., 2025) introduces the Action Score to evaluate action coverage. MMNarrator (Zhang et al., 2024) proposes SegEval to assess both textual properties (originality, consistency) and sequence-level attributes (coherence, diversity, specificity). However, like reference-based methods, it assumes that system-generated ADs and reference ADs highlight the same visual details, which is not always true (see Sec. 3). NarrAD (Park et al., 2025) introduces human-centered metrics like usefulness and recommendability via user studies, which, while highly effective, are not scalable. Different from above, our work focuses on evaluating ADs through a QA framework to check whether they would help BLV users appreciate and understand the story.

Finally, CinePile (Rawal et al., 2024) uses ADs to generate QAs for long-form video comprehension. We differ in multiple ways: our aim is to evaluate generated ADs, appreciation questions often stem from a single AD (trimmed clips), and narrative questions are created based on plots.

3 Takeaways from Two AD Tracks

While MAD-train, CMD-AD, and TV-AD have relied on AudioVault as the primary source of ADs, they all use a single AD track.

3.1 Aligning Multiple AD Tracks

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

AudioVault hosts complete movie audio tracks comprising dialog, AD, music, and sound effects. To align AD tracks, we follow three steps: (i) obtain timestamped transcriptions using WhisperX (Bain et al., 2023); (ii) classify each transcription sentence as AD or dialog using LLMs (App. F); and (iii) align the two transcriptions using dynamic time warping (Han et al., 2024) 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 (Han et al., 2024), there are two important differences. (i) We observe that using an LLM results in much better AD/dialog classification than identifying the narrator's voice. (ii) We align two transcribed tracks containing dialog + AD, while AutoAD-3 aligns transcribed CMD videos (dialog) with AudioVault transcriptions (dialog + AD).

Our alignment process provides a timeline of sequentially aligned dialog from both tracks, interspersed with non-aligned ADs. We use this dialog alignment to compute a linear transformation (slope and offset) that maps the timestamps from track 1 to track 2 and apply the transformation. We then create a mapping between ADs from the two tracks if their durations have an overlap score (computed in App. B, step 4a) higher than a chosen threshold (default: 50%). Any ADs for which we cannot find a mapping are marked as *non-aligned*. Details of the full alignment procedure, including additional mapping steps, are provided in App. B.

Fig. 2 shows that increasing the overlap threshold increases the number of non-aligned AD pairs, as expected. Yet, even at a very low threshold (1%), about 25–30% of ADs still remain *non-aligned*. This suggests that some ADs inherently lack counterparts in the other track, fundamentally challenging the validity of sentence-pair evaluation.

---

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

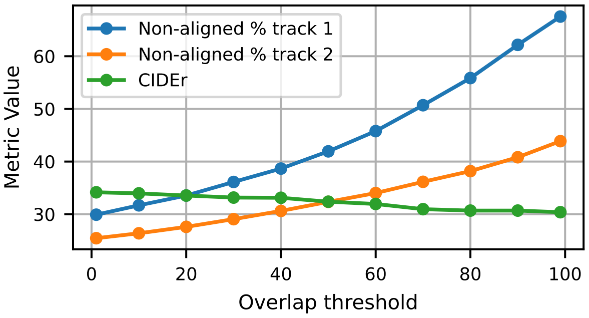


Figure 2: Impact of overlap threshold on AD alignment on the two-track subset of CMD-AD movies. The fraction (%) of non-aligned ADs increases with threshold (expected). Interestingly, even at low thresholds, 25-30% ADs are not aligned indicating that many ADs in one track are not present in the other. Additionally, CIDEr does not increase with better temporal overlap (high threshold) suggesting that even well-aligned ADs often differ substantially in wording.

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

Table 1: Results of aligning and mapping ADs between two sources for 17 movies of the CMD-AD dataset. CIDEr (37.3) is low when comparing the aligned ADs (60.7%) that have a high average overlap (85.6%).

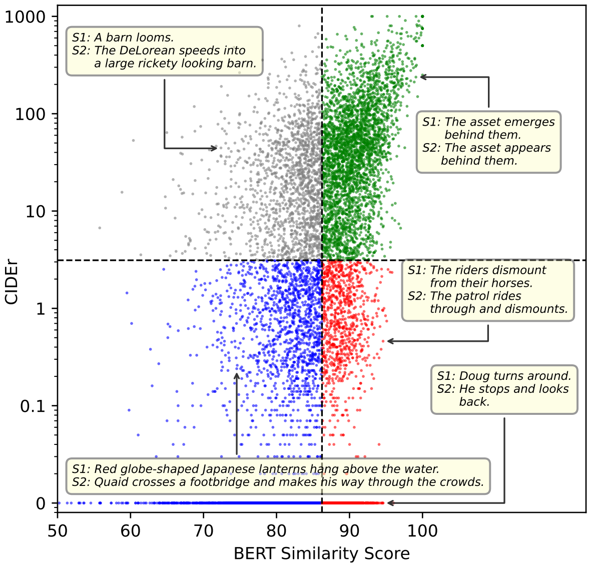
3.2 Similarity between Aligned ADs

For a subset of 17 CMD-AD movies with two AD tracks, Tab. 1 shows the % of aligned ADs, the average overlap %, and the average similarity scores between the aligned ADs: BERT cosine similarity (Devlin et al., 2019) and CIDEr (Vedantam et al., 2015). Even for aligned ADs with high overlap, we observe poor CIDEr scores highlighting the challenges of using n-gram based metrics.

Next, Fig. 2 shows that CIDEr stays constant with increasing overlap threshold, and even reduces a bit at higher thresholds. This indicates that even with very high temporal overlap, aligned AD pairs may still use different words.

To further investigate the nature of aligned AD pairs, Fig. 3 compares their BERT similarity against CIDEr. We see five important scenarios corresponding to low/high values of the similarity metrics. They highlight the subjective nature of ADs: where aligned ADs can describe different details (33.8%, quadrant III), or the same detail, but using different words resulting in poor CIDEr scores (16.3%, quadrant IV). Finally, about 27% pairs fail catastrophically with 0 CIDEr.

We summarize our *takeaways*: (i) Different AD experts may subjectively choose *when* to describe an event or *whether* to describe it at all, resulting in a large proportion of non-aligned ADs. (ii) Even among ADs aligned in time, experts may differ in *what* visual details to highlight for a coherent story, resulting in low semantic similarity scores. (iii) Finally, experts may differ in *how* they phrase the same detail, resulting in low scores from n-gram-based metrics like CIDEr, especially when evaluated against a single reference.



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

Figure 3: BERT Similarity (B) *vs*. CIDEr (C) for *time-aligned ADs* from two AD tracks on 17 movies from the CMD-AD dataset. 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.

Overall, these findings suggest that evaluating ADs in a strict one-to-one manner, as in video captioning, is unsuitable, since experts may disagree at multiple levels. We need a better evaluation aimed at the heart of AD creation—do they help BLV users appreciate and understand the story.

4 ADQA Benchmark for Evaluating ADs

Our goal is to evaluate ADs for their ability to help BLV users appreciate and understand the story. Given the subjective nature of the task and challenges of sentence-pair metrics (Sec. 3), we move from few-second long trimmed clips to a **video segment**—a few minute interval from the movie showcasing a coherent story segment (*e.g*. the 2-3 minute

Youtube videos used in CMD-AD) (Sec. 4.1). Next, we pose AD evaluation as a QA task. We measure the quality of predicted ADs (and thereby the generation method) based on our ability to correctly answer questions using those ADs. Sec. 4.2 presents the multiple-choice QA (MCQA) creation strategy and Sec. 4.3 the quantitative details of ADQA.

4.1 From Trimmed Clips to Video Segments

Traditionally, AD datasets are built by extracting trimmed clips corresponding to the duration where AD narration occurs, using transcription timestamps. However, AD narrations aren't strictly synchronized with the visual events being described; narration can anticipate or lag behind the visuals. While LSMDC (Rohrbach et al., 2017) adjusts the timestamps manually to ensure that the input videos contain the visual details, recent works such as MAD-train or CMD-AD move away from this tedious process. As a result, generated ADs often score poorly in 1-to-1 comparisons on trimmed clips, since they may lack the visual details described in the ground-truth ADs. Instead, we consider longer video segments (typically 2-3 minutes, which may include multiple ADs) as the evaluation unit. This shift not only provides more context for generating ADs, but also encourages producing coherent, non-redundant descriptions rather than isolated captions. For brevity, we will refer to trimmed clips as *clips* and longer video segments with potentially multiple ADs as *videos*.

CMD-AD is a special case as the videos are obtained from a YouTube channel that shares famous or important plot moments of a movie. These videos are 140 s on average and can be directly treated as video segments. The videos are also accompanied by plot-like descriptions (used in text-to-video retrieval by Bain et al. (2020)) that are appropriate for creating narrative understanding questions.

On the other hand, MAD provides ADs for the entire movie. Inspired by work on aligning plots and movies (Tapaswi et al., 2015), we use LLMs to align plot synopses sentences with dialog + AD transcriptions (referred as *script*) of the entire movie, thereby identifying *scene* boundaries. In the prompt to the LLM (see App. F), we observe the following clauses to be especially important: (i) merge consecutive *script* scenes when they align to long plot sentences describing multiple events; (ii) ensure that details in the plot are present in or can be inferred from the script; and (iii) not require every scene to be aligned to some plot sentence. This creates video segments of about 116 s that are accompanied by plot sentences. The first part of Tab. 2 presents some statistics for CMD-AD and MAD-eval, highlighting the number of video segments and those with an aligned plot sentence.

4.2 Creating QAs

We prompt Gemini-2.5-Pro to generate all questions in ADQA. Similar to video QA benchmarks (Tapaswi et al., 2016; Lei et al., 2018), our questions have 1 correct answer among 5 options.

Visual appreciation MCQAs. We obtain appreciation QAs by prompting the LLM to come up with questions based on the factual details conveyed in ADs (typically a single AD). While the input context includes dialog + ADs of a video segment, we prompt the model to (i) not create questions based on dialog; and (ii) not have questions that are answerable by looking at other questions (or answer options) of the same video. Along with question generation, the LLM is prompted to create multiple plausible answer options. We notice that an advanced model such as Gemini-2.5-Pro performs much better on the question generation task than smaller/faster models such as Gemini-2.0-Flash. An example provided in the prompt is: Given the AD: "A green truck speeds through the highway crossing a yellow barrier", multiple questions can be created: "What vehicle is seen on the highway?", "What is the color of the vehicle on the highway?", or "What color is the barrier on the highway?".

Narrative understanding MCQAs. Next, we create narrative understanding questions based on the plot descriptions accompanying the video segment. As indicated before, CMD-AD videos are accompanied by a plot-like description, while MAD videos are aligned with plot synopsis sentences. As these descriptions present the narrative, creating factual questions based on them results in narrative understanding MCQAs. We discourage generation of ambiguous questions that are not highlighted in the plot. In the prompt, we present the following example: For a plot description: "The shining spaceship lands on a strange planet covered in glowing blue plants and mist.", questions such as "Where does the spaceship land?" (A: on a strange planet), or "What makes the planet unusual?" (A: the planet is covered in glowing blue plants and mist) are considered good. On the other hand, questions such as "What did the spaceship do?" (it could have

	CMD-AD Full	CMD-AD TwoTrack	MAD-eval Full
#Movies	98	17	10
#VideoSegments (VS)	591	112	551
#VideoSegments w/ plot	591	112	338
Duration (s)	140	142	116
#ADs	7316	1484	6331
#VisualAppreciation Q	17595	2705	15441
#Vis App Q per AD	2.41	1.82	2.44
#NarrativeUnderstanding Q	3128	585	1962
#Nar Und Q per VS	5.29	5.22	5.80

Table 2: ADQA benchmark in numbers.

done several different things apart from landing) are marked ambiguous.

Answer rationale during question generation. As seen in the question generation prompts (App. F), beyond the question and answer options, the LLM is prompted to also select the correct answer and provide a rationale for the same. The rationale for appreciation QAs starts with "As specified in the AD, ...", indicating that the model should point to the specific AD(s) used to create the question. Similarly, the rationale for narrative understanding QAs is encouraged to refer to plot-like description. Beyond MCQA generation, asking the LLM to pick the correct answer and provide a rationale acts as a self-verification check resulting in higher quality QAs. We also rely on rationales during AD evaluation, as explained in the next section.

In summary, for each video segment, we generate: (i) Visual Appreciation (VA) question(s); (ii) Narrative Understanding (NU) question(s); (iii) correct answer and wrong options for each question; and (iv) a rationale for the correct answer referring to the dialog + AD script or plot.

4.3 ADQA Analysis

Tab. 2 shares some numbers of the ADQA benchmark. For CMD-AD, we present numbers on the full *evaluation* set (98 movies) used to evaluate AD generation approaches. As we obtain ADs from two tracks for 17 movies, we also report numbers for this *TwoTrack subset*. For MAD-eval, we obtain two AD tracks for all 10 movies: the original LSMDC ADs and different ADs sourced from AudioVault.

ADQA features multiple appreciation questions per AD ($\sim$2.4) and several understanding questions for each video segment ($\sim$5.5). While some example QAs from our benchmark are presented in Fig. 1, several more questions, answer options, and answer rationales are shared in App. D. We make *public* the questions, answers, and correct choices for 5 movies from CMD-AD and 1 movie from MAD-eval. Questions from the remaining movies are kept *private* for evaluation on the online leaderboard. More details are presented in App. C.

5 Experiments

We present the evaluation setup, followed by a user study assessing LLM reliability, and end with a comparison of various AD generation methods.

5.1 Answering Setup and Metrics

Base setup. We use a common prompt and LLM (Gemini 2.0 Flash) to perform all evaluations in our work. The setup involves providing all questions (and corresponding answer options) of a video segment, followed by a specified context. The LLM is prompted to answer each question independently. The key idea is to evaluate ADs generated using different methods by feeding them (separately) as context to the LLM.

Rationale-based answering. We find that LLMs trained on world knowledge know details of popular movies. This means that they are able to answer questions from prior knowledge or common sense rather than using the provided context. To evaluate whether the LLM answers based on the provided context, we generate three outputs: (i) the predicted answer among 5 choices; (ii) a descriptive rationale for choosing the answer; and (iii) a binary label indicating whether the rationale suggests that the answer was derived from context.

For completeness, we report three values: (i) CA: Correct Answer; (ii) AC: Answer uses Context (*i.e.* the binary label is `True`); and (iii) CC: Correct answer *and* uses Context. A prediction is considered correct only if the chosen answer option is correct *and* the binary label confirms context-based reasoning. Thus, CC is our primary metric of interest and can be considered as ADQA's *accuracy*.

Types of context. We evaluate five different types of context to understand how different information affects accuracy. They are: (i) no context, (ii) only movie name, (iii) only dialog, (iv) only ADs, and (v) dialog + ADs. These ablations allow us to quantify the contribution of each context type to the accuracy.

Accuracy ratio. As studied in Sec. 3, AD creation is inherently subjective and even human-generated

	Vis App	Narr Und
Valid options w/o context	95.3%	95.3%
Question clear and answerable	96.9%	98.4%
Answer rationale from context	93.7%	98.4%
Chosen answer from rationale	92.2%	96.9%

Table 3: 8 participants judged validity of generated questions, and checked whether the chosen answer and its rationale were well grounded.

ADs may not fully agree with each other. We therefore do not expect any method to score 100% accuracy on ADQA. Instead, we define accuracy ratio to measure how much of the gap between human performance and only dialog is closed by a model:

$$\text{Accuracy Ratio}_m = \frac{\text{CC}_m - \text{CC}_{\text{dialog}}}{\text{CC}_h - \text{CC}_{\text{dialog}}}, \quad (1)$$

where CC_m corresponds to accuracy of method m (the method under evaluation), $\text{CC}_{\text{dialog}}$ is the accuracy using only dialog context, and CC_h is computed by running the same LLM answering pipeline, but using human-authored AD tracks (e.g., from AudioVault) instead. Here, CC_h acts as topline *human performance* for AD generation.

5.2 Assessing LLM Reliability

To address potential unreliability in automatically generated questions and answers, we conducted a human evaluation with 8 participants covering both question generation and answering stages. The participants were unpaid volunteers from within our research group, with age ranging from 20-35 years. The study did not gather personally identifiable information. Volunteers were notified that research publications would present only the average scores calculated from their ratings.

We sampled 32 CMD-AD videos (16 each for Visual Appreciation and Narrative Understanding), yielding 128 questions correctly answered (CC) using track 2's AD and dialog. Depending on the task, users were shown: (i) the question generation context; (ii) the generated question and options; (iii) the answering context; and (iv) the chosen answer, and its rationale.

Each user then judged, with binary responses, whether (i) the answer options were valid and meaningful without the question generation context, (ii) the question was clear and answerable given the question generation context, (iii) the answering rationale was grounded in the answering context, and (iv) the chosen answer followed from the rationale.

Table 3 summarizes the results. The high scores across all criteria support the reliability of using LLMs to generate and answer ADQA questions. Issues flagged were generally very minor; for instance, including "Verbal" in the answer options for the question "Who ultimately executed the task that Keaton hesitated to perform?" was seen as odd by one user, who may have confused the name of one of the main characters in the movie—"Verbal", with the common noun (more examples in Tab. 9).

5.3 ADQA: Context Ablation

We present the impact of different context inputs with human narrated AD tracks in Tab. 4. The results are reported on all data (private + public).

For Visual Appreciation (VA), we use AudioVault track 1 to create questions, while for answering we use AudioVault track 2 (for CMD-AD) or LSMDC (for MAD-eval). Narrative Understanding (NU) questions are created using plot descriptions. For answering NU, we consider both AudioVault tracks for CMD-AD; and AudioVault and LSMDC tracks for MAD-eval.

Prior knowledge. LLMs are able to answer a substantial number of questions correctly (CA) with no context (NoCtx), highlighting that *LLMs have extensive prior knowledge and ability to guess based on common sense*. Interestingly, including the movie name (MN) does not improve results.

Impact of rationale. We first examine how generating rationales affects the LLM's ability to ground its answers in context. For VA questions on CMD-AD, the baseline accuracy without context (NoCtx) is 55.9% CA. Adding dialog + AD reduces CA to 41.4%, and CC further drops to 30.2%, indicating the LLM struggles when forced to justify answers using context. For MAD-eval, CC is 59.0% and hints at a stronger agreement between AV_1 and LSMDC. Across both datasets, this confirms that *requiring rationales compels the LLM to ground answers in context* rather than exclusively use prior knowledge.

Context utilization. We analyze how different contexts (dialog, AD, or both) are used by the LLM when answering ADQA questions. Across both datasets and question types, the *share of answers actually grounded in context (AC) rises as richer context is provided*: dialog alone yields low AC for VA (19.6%, 24.7%) since ADs provide most visual details; adding ADs raises AC substantially (60.2%, 75.9%); dialog+AD improves it further (62.6%,

	Dataset	#M	Source		NoCtx	MN	Dialog-only			AD-only			Dialog + AD		
			Q	A	CA	CA	CA	AC	CC	CA	AC	CC	CA	AC	CC
Visual Appreciation	CMD-AD	17	AV$_1$	AV$_2$	55.9	55.9	44.0	19.6	09.7	37.6	60.2	26.5	41.4	62.6	30.2
	MAD-eval	10	AV$_1$	L	46.9	47.0	41.1	24.7	11.7	67.7	75.9	58.5	67.0	76.8	59.0
Narrative Understanding	CMD-AD	17	plot	AV$_1$	60.2	59.1	70.3	81.9	61.9	70.9	73.5	56.4	76.7	92.0	73.0
	CMD-AD	17	plot	AV$_2$	-	-	69.2	80.8	60.5	68.9	69.6	51.8	77.8	92.0	75.0
	MAD-eval	10	plot	AV$_1$	52.1	52.7	64.1	68.1	50.1	62.0	58.7	42.5	74.9	86.4	68.4
	MAD-eval	10	plot	L	-	-	62.4	66.8	48.7	60.0	54.8	38.6	71.6	84.2	64.4

Table 4: Ablation results for different types of context. All sources are human authored ADs. Source acronyms are AV$_1$: AudioVault source 1, AV$_2$: AudioVault source 2, L: LSMDC, and plot: narrative description accompanying or aligned to the video segments. Other acronyms include #M: number of movies, NoCtx: no context, and MN: context is movie name. The metrics are CA: correct answer, AC: answer uses context, CC: correct answer using context.

76.8%). For NU, dialog alone already offers strong narrative cues (81.9%, 68.1% AC), while AD-only context is less helpful (73.5%, 58.7% AC). Yet, combining dialog+AD further boosts grounding (92.0%, 86.4% AC).

Impact of context type. Across both datasets and question types, we see a consistent trend: *enriching the answering context improves accuracy (CC)*. For VA, dialog-only context leads to lower CC (9.7% CMD-AD) vs. AD-only (26.5%), highlighting that ADs convey essential visual details missing from dialog. For NU, dialog only (61.9% CC, CMD-AD) outperforms AD-only (56.4% CC) emphasizing the importance of dialog in driving the story. However, the best results consistently come from dialog + AD with 30.2% CC for VA and 73.0% CC for NU on CMD-AD. MAD-eval shows similar trends, dialog + AD performs best. These experiments demonstrate that, across both datasets, *ADs significantly enhance visual appreciation* and strongly *complement dialog for improved narrative understanding*. They also validate that **ADQA is well-positioned to evaluate the richness and relevance of generated ADs**.

5.4 Evaluating AD Generation Methods

Methods. We report results for various methods in Tab. 5. We evaluate multiple fine-tuned (AutoAD-3 (Han et al., 2024), UniAD (Wang et al., 2025), DistinctAD (Fang et al., 2025)) and zero-shot (AutoAD-0 (Xie et al., 2024), NarrAD (Park et al., 2025)) AD generation methods. We also evaluate dense descriptions obtained from state-of-the-art VideoLMs (Qwen2VL (Wang et al., 2024)), prompted to generate paragraph-length rich descriptions relevant for AD generation (disregarding the temporal constraints required for ADs). Comparison against ground-truth AD sources (AudioVault AV$_1$, AV$_2$, and LSMDC) provide upper bounds for model performance and test the limits of what is realistically achievable on ADQA. A lower bound is established with context as dialog-only.

Setup. For CMD-AD, we evaluate generated ADs against questions obtained for the 93 private movies. AV$_2$ used in answering is an exception as it only has 17 movies. For MAD-eval, we evaluate all methods against the 9 private movies. The CC$_h$ in Accuracy Ratio (Eq. (1)) is determined as follows. For VA questions, we use AV$_2$ for CMD-AD and LSMDC for MAD-eval. For NU questions, we use AV$_1$ for CMD-AD and average accuracy of AV$_1$ and LSMDC for MAD-eval.

5.4.1 Results

Human ADs outperform all models across both question types and datasets. On CMD-AD NU, AV$_1$ (72.8%) and AV$_2$ (75.0%) achieve similar CC scores, further validating the benchmark. On MAD-eval NU, LSMDC (65.2%) and AV$_1$ (69.5%) reach similar CC scores, suggesting that community-driven AudioVault ADs match professional quality.

Dialog-only Baseline. For narrative understanding, dialog-only provides a strong baseline for CMD-AD (CC 59.1%), which all models surpass. In contrast, for MAD-eval, dialog-only achieves CC 50.3%, and ADs from most models do not surpass this score. NarrAD that uses scripts stands out with CC 52.4%. For visual appreciation, dialog-only is a good baseline for VA (CC 9.8% on CMD-AD, 11.8% on MAD-eval) and all models are able to outperform it.

Movie Scripts. NarrAD, which benefits from movie scripts as input, achieves the highest MAD-eval scores (VA CC 22.3%, NU CC 52.4%). However, as such scripts are not always available, we do not include NarrAD in further comparisons.

Best Performers. Looking at CC, AutoAD-III

	Method	Train	Old Metrics		**Vis App**		**Narr Und**	
			C	LLMe	CC	Ratio	CC	Ratio
CMD-AD	Dialog-only	-	-	-	9.8	0	59.1	0
	AutoAD-III	✓	**25.0**	2.01	14.7	24.0	63.5	**32.1**
	UniAD*	✓	21.8	**2.92**	14.1	21.0	63.2	29.9
	AutoAD-Zero	✗	17.7	1.96	13.2	16.7	63.2	29.9
	Q2VL	✗	-	-	17.1	**35.8**	51.5	-55.4
	AV_1	-	-	-	-	-	72.8	100
	AV_2 (17)	-	-	-	30.2	100	75.0	116
MAD-eval	Dialog-only	-	-	-	11.8	0	50.3	0
	AutoAD-III	✓	24.0	2.20	14.5	5.7	43.7	-38.7
	DistinctAD	✓	27.3	2.27	13.5	3.6	37.8	-73.3
	UniAD*	✓	**28.2**	2.46	15.8	8.5	44.0	-37.0
	AutoAD-Zero	✗	22.4	2.20	13.9	4.5	44.4	-34.6
	NarrAD	✗	26.4	**2.64**	22.3	**22.3**	52.4	**12.3**
	AV_1	-	-	-	-	-	69.5	112
	LSMDC	-	-	-	58.9	100	65.2	87.4

Table 5: Evaluation of various generated and human-authored ADs on ADQA private set. Results are reported with dialog + AD as context. DistinctAD uses Llama. NarrAD uses curated ADs. UniAD has some missing outputs. The "Train" column indicates if the method is fine-tuned (✓) or zero-shot (✗). Acronyms: Vis App: Visual Appreciation, Narr Und: Narrative Understanding. The metrics are C: CIDEr, LLMe: LLM-AD-eval (Han et al., 2024), CC: Correct answer using Context, and Ratio: Accuracy ratio.

achieves highest score on CMD-AD (VA 14.7%, NU 63.5%). This may be as AutoAD-III is fine-tuned on CMD-AD-train. On MAD-eval, UniAD leads in VA (27.1%), while AutoAD-Zero leads in NU (44.4%), with UniAD a close second (44.0%). DistinctAD, despite aiming for diversity, performs worse (MAD-eval VA 13.7%, NU 37.6%).

Accuracy ratio highlights the wide gap between human top-line performance (second AD track) and model generated ADs. On VA, MAD-eval (UniAD 8.5%) seems harder than CMD-AD (AutoAD-III 24.0%). Furthermore, on MAD-eval for the NU task, generated ADs (except NarrAD) contribute negatively in assisting narrative understanding.

Dense Descriptions. Q2VL achieves the highest CMD-AD VA CC (17.1%). This suggests that VA questions that ask about specific visual details, benefit from richly detailed captions. In contrast, it performs poorly on NU (CC 51.5%), likely because NU requires broader coherence, and an overload of dense information may hinder comprehension.

Evaluation server. We host ADQA as an evaluation server with a leaderboard where participants upload their generated ADs for evaluation on the private set. This ensures fair comparisons with the same answering prompt and LLM across all methods. For completeness, evaluation on the public set is reported in App. C.

6 Recommendations for Future Work

While automatic AD generation has made progress, a significant gap remains in supporting visual appreciation and narrative understanding for BLV users. We outline research directions to bridge this gap:

1. *From clips to videos.* Generation and evaluation should move to the video level as considering ADs in isolation will not result in coherent descriptions necessary for appreciation and understanding.

2. *Focus on narratives.* Current models fail to connect individual events into a story, often producing repetitive ADs. Combining ADs with dialog and training models to answer ADQA-like questions may help.

3. *VLMs hold promise.* Dense Q2VL results suggest VLMs can extract rich details. Future work may continue to focus on distilling these into concise ADs that fit naturally within dialog.

4. *Scripts, with caution.* NarrAD results show that scripts provide useful cues, but they aren't always available or aligned with the final cut. Relying on scripts may limit real-world adoption.

5. *More hands-on evaluation.* Overreliance on LLMs risks leakage of prior knowledge or ungrounded rationales. A holistic evaluation that follows the AD guidelines is needed.

7 Conclusion

We proposed ADQA, a new evaluation paradigm for automatic AD generation methods that addresses two themes central to ADs, whether they help BLV users with *visual appreciation* and *narrative understanding*. To analyze the subjectivity of ADs, we aligned and compared two human-narrated AD tracks, revealing issues with current video captioning-like setups, semantic similarity based metrics like BERT similarity, and n-gram based metrics like CIDEr. The second track also provided a top-line human performance on ADQA. Evaluation of current AD generation methods showed a large gap to human-authored ADs. We also provided several recommendations for future work based on our findings.

Acknowledgments. This project was supported by funding from SERB SRG/2023/002544 and an Adobe Research India gift. We thank volunteers for participating in the user study and all AD researchers that provided their model outputs.

8 Limitations

While ADQA introduces a narrative-aligned framework for evaluating ADs, we discuss some limitations:

1. Temporal feasibility not evaluated. We do not currently assess whether model-generated ADs are too long to fit naturally between dialog segments. A naive solution would be to cap the AD length based on available dialog-free intervals (*e.g.* assuming 160 words-per-minute (Snyder, 2014)). However, this may encourage models to saturate silent gaps with verbose descriptions, potentially overwhelming the viewer and detracting from the cinematic experience.

2. Variability in LLM outputs. Language model outputs are inherently non-deterministic, introducing small fluctuations in evaluation scores across runs. While this randomness is small, it raises some reproducibility concerns. We plan to manage this by running evaluations through a server where participants can upload their generated AD predictions with a rate limit.

3. Prior knowledge leakage. Despite best efforts to mitigate answering via prior knowledge, LLMs possess extensive information about popular movies. However, as the same model is used to assess all automatic AD generation methods and human-authored ADs, the relative scores remain valid, even if absolute performance may be inflated.

References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. In *Interspeech*.

Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. 2020. Condensed Movies: Story Based Retrieval with Contextual Embeddings. In *Asian Conference on Computer Vision (ACCV)*.

Peng Chu, Jiang Wang, and Andre Abrantes. 2024. LLM-AD: Large Language Model based Audio Description System. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Google Deepmind. 2024. Gemini-2.0-Flash. `https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/`. Last accessed 2025-05-17.

Google Deepmind. 2025. Gemini-2.5-Pro. `https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/`. Last accessed 2025-05-17.

Adrienne Deganutti, Simon Hadfield, and Andrew Gilbert. 2025. DANTE-AD: Dual-Vision Attention Network for Long-Term Audio Description. In *CVPR Workshop on AI for Content Creation*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of Association of Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Bo Fang, Wenhao Wu, Qiangqiang Wu, Yuxin Song, and Antoni B. Chan. 2025. DistinctAD: Distinctive Audio Description Generation in Contexts. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yingqiang Gao, Lukas Fischer, Alexa Lintner, and Sarah Ebling. 2024. Audio Description Generation in the Era of LLMs and VLMs: A Review of Transferable Generative AI Technologies. *arXiv preprint arXiv:2410.08860*.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2023a. AutoAD II: The Sequel – Who, When, and What in Movie Audio Description. In *International Conference on Computer Vision (ICCV)*.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2023b. AutoAD: Movie Description in Context. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2024. AutoAD III: The Prequel – Back to the Pixels. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. TVQA: Localized, Compositional Video Question Answering. In *Empirical Methods in Natural Language Processing (EMNLP)*.

OpenAI. 2022. ChatGPT. `https://openai.com/index/chatgpt/`. Last accessed 2025-05-17.

Jae Sung Park, Trevor Darrell, and Anna Rohrbach. 2020. Identity-Aware Multi-Sentence Video Description. In *European Conference on Computer Vision (ECCV)*.

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 *Winter Conference on Applications of Computer Vision (WACV)*.

Amy Pavel, Gabriel Reyes, and Jeffrey P. Bigham. 2020. Rescribe: Authoring and Automatically Editing Audio Descriptions. In *ACM Symposium on User Interface Software and Technology (UIST)*.

Haran Raajesh, Naveen Reddy Desanur, Zeeshan Khan, and Makarand Tapaswi. 2024. MICap: A Unified Model for Identity-Aware Movie Descriptions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. 2024. CinePile: A Long Video Question Answering Dataset and Benchmark. *arXiv preprint arXiv:2405.08813*.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie Description. *International Journal of Computer Vision (IJCV)*, 123:94–120.

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Makarand Tapaswi, Martin Baeuml, and Rainer Stiefelhagen. 2015. Aligning Plot Synopses to Videos for Story-based Retrieval. *International Journal of Multimedia Information Retrieval (IJMIR)*, 4:3–16.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding Stories in Movies through Question-Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hanlin Wang, Zhan Tong, Kecheng Zheng, Yujun Shen, and Limin Wang. 2025. Contextual AD Narration with Interleaved Multimodal Sequence. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

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. 2024. 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, Gül Varol, Weidi Xie, and Andrew Zisserman. 2025. Shot-by-Shot: Film-Grammar-Aware Training-Free Audio Description Generation. In *International Conference on Computer Vision (ICCV)*.

Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2024. AutoAD-Zero: A Training-Free Framework for Zero-Shot Audio Description. In *Asian Conference on Computer Vision (ACCV)*.

Xiaojun Ye, Junhao Chen, Xiang Li, Haidong Xin, Chao Li, Sheng Zhou, and Jiajun Bu. 2024. MMAD: Multi-Modal Movie Audio Description. In *International Conference on Computational Linguistics, Language Resources and Evaluation (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*.

YouDescribe. 2013. Audio Description Basics for Beginners: A do's and don'ts video tutorial. `https://www.youtube.com/playlist?list=PLNJrbI_nyy9uzywoJfyDRoeKA1SaIEFJ7`. YouTube playlist. Last accessed 2025-05-17.

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*.

A Responsible NLP Details

Model Size and Computational Budget: For computing sentence level similarity, we use *bert-base-uncased*, a 110M parameter model trained on English language. The bulk of work was done using Gemini APIs, with an approximate cost of generating questions < \$100 and evaluation of each AD generation method at < \$5 per experiment.

License or Terms of use for Artifacts: Movie video clips from CMD are available on YouTube for public access. AudioVault is a non-commercial entity that hosts audio-only movie files for the benefit of Blind and Low Vision individuals. We use AudioVault to access audio descriptions.

We made use of AI assistants such as ChatGPT (OpenAI, 2022) to help with coding. Commercial AI models Gemini were used for experiments and data creation. We use Gemini-2.5-Pro (Deepmind, 2025) for generating all questions in the dataset, and use Gemini-2.0-Flash (Deepmind, 2024) for answering the questions. Gemini API Terms of Service allows the API for commercial or research use.

Miscellaneous: Movie data may contain offensive or explicit content. All experimental results are presented for a single run.

B Details of the AD Mapping Process

Post-alignment of AD + dialog transcriptions, we map ADs from T_1 (track 1) and T_2 (track 2) using the process defined below:

1. Consider an AD A_i^1 in T_1 with duration d_i^1.
2. Identify the slope relevant to this AD (for some movies the offset changes across the movie due to censored scenes).
3. Predict the time duration $\hat{d}_i^2$ in T_2 for this AD using the slope and offset, and add a 1 s buffer on each side.
4. Get a list of ADs in T_2 having overlap with $\hat{d}_i^2$. For each AD, A_j^2 with duration d_j^2:
   (a) Compute overlap score $O(A_i^1, A_j^2)$ as:

$$O(A_i^1, A_j^2) = \frac{\cap(\hat{d}_i^2, d_j^2)}{\min(\hat{d}_i^2, d_j^2)}. \qquad (2)$$

   (b) If $O > 50\%$, create a mapping.
5. Repeat above process (steps 1-5) for all ADs A_i^1 in T_1.
6. Also repeat above process (steps 1-5) for all ADs A_j^2 in T_2.
   (a) If A_j^2 in T_2 fails to find a match and was not already mapped in steps 1-4, then we consider A_j^2 as non-aligned.
7. All ADs A_i^1 in T_1 without a mapping to any AD in T_2 are also considered non-aligned.

	Model	All		Private		Public	
		VA	NU	VA	NU	VA	NU
CMD-AD	Dialogue-only	10.0	58.9	9.8	59.1	13.8	54.9
	AutoAD-0	13.4	62.9	13.2	63.2	15.5	56.9
	AutoAD-III	14.9	63.2	14.7	63.5	18.0	56.9
	Qwen2	17.2	51.2	17.1	51.5	17.9	45.8
	UniAD*	14.3	63.0	14.1	63.2	17.8	60.1
	AV_1 (GT)	–	72.7	–	72.8	–	71.9
	AV_2 (17)	30.2	75.0	30.2	75.0	–	–
MAD-eval	Dialogue-only	11.7	48.7	11.8	50.3	11.0	33.7
	AutoAD-0	14.3	44.2	13.9	44.4	19.1	42.1
	AutoAD-III	14.7	42.9	14.5	43.7	17.5	34.7
	DistinctAD	13.7	37.6	13.5	37.8	16.1	35.3
	NarrAD	22.7	52.7	22.3	52.4	26.7	54.7
	UniAD*	16.0	43.4	15.8	44.0	18.5	38.4
	AV_1	–	68.4	–	69.5	–	57.9
	LSMDC	59.0	64.4	58.9	65.2	60.0	57.4

Table 6: Model performance on subsets of the benchmark. Columns report results on the full set (**All**), the hidden leaderboard set (**Private**), and the publicly released set (**Public**). **VA** = Visual Appreciation, **NU** = Narrative Understanding. For CMD-AD, $n = \{98, 93, 5\}$ correspond to All/Private/Public respectively; for MAD-eval, $n = \{10, 9, 1\}$. Note that AV_2 has no movies in the public set and hence no public results are reported. Scores are Correct answer using Context (CC) (↑).

C Benchmark

C.1 Public and Private Subsets

To facilitate understanding of the benchmark, we publicly release all questions and answers from 5 movies in CMD and 1 movie in MAD-eval. These movies were selected to be among the most representative of the full benchmark, capturing both model ranking and absolute performance. The rest of the 93 CMD movies and 9 MAD-eval movies form the private testing set to be used for evaluation in the leaderboard. Tab. 6 shows model outputs on the entire dataset and public and private sets.

C.2 Leaderboard

An online leaderboard can be accessed through the project website `https://katha-ai.github.io/projects/adqa/`. Researchers can submit their model generated ADs for evaluation on the private testing set with rate limits to prevent overfitting.

D Qualitative

Examples from ADQA. Tab. 7 presents several questions from the ADQA benchmark along with multiple choice options, the correct answer, and the rationale for the answer. We observe that the rationale often refers to specific parts of the AD (for appreciation QAs) or the plot-like descriptions (for understanding QAs).

Generated ADs and Answering ADQA. Tab. 8 presents some examples of LLM based answering for different input contexts. We see that dialog + AD results in meaningful rationales indicating that the model is able to evaluate the quality of AD provided as context.

E User Study

We ask users to rate generated questions and answering capabilities of the LLM on 4 criteria. The results are shared in Tab. 3.

Options valid without question generation context. You are given the generated question and answer options. Mark 1 if the options make sense, given the question. Otherwise mark 0. As long as the options make sense and are not wildly out of distribution you can mark 1.

- Example where 0 should be marked:
    - Question: How many coconuts were there on the tree?
    - Options A) 3 B) 2 C) 238746 (too large)
- Example where 1 should be marked:
    - Question: How many coconuts were there on the tree?
    - Options A) 3 B) 2 C) 7 D) 10

Question clear and answerable from question generation context. You are given the question context, generated question, answer options and correct answer. Mark 1 if question is clear and answerable from the question generation context, otherwise 0.

- Context:
    - Dialogue: The bear is going to attack us!
    - AD: The bear saunters away.
- Example where 0 should be marked:
    - Question: What color was the bear?
    - Options: A) Brown B) Black C) White
    - Reason: The correct answer is not derivable, given the context.
- Example where 1 should be marked:
    - Question: What action did the bear take?
    - Options: A) The bear attacked B) The bear stood on its hind legs C) The bear sauntered away.
    - Reason: The correct answer is derivable, given the context.

Answer rationale is derived from answering context. You are given the QA, answer rationale, and the answering context. Mark 1 if the answer rationale is correctly derived from the answering context, otherwise 0.

- Answering context:
    - Audio Description: Mark runs through a series of long hallways.
    - Audio Description: He stops before a large door.
- Example where 1 should be marked:
    - Answer rationale: The audio description clearly states that Mark stops before a large door
- Example where 0 should be marked:
    - Answer rationale: Although the color of the hallway is not explicitly stated, In Severence, hallways are famously white.

Answer is derived from Answer rationale. You are given QA, the model prediction, and the answer rationale. Do not worry about whether the answer rationale is based on the context. Mark 1 if the correct/generated answer is derived from the answer rationale, otherwise 0.

- Example where 1 should be marked:
    - Answer rationale: Although the color of the hallway is not explicitly stated, In Severence, hallways are famously white.
    - Correct/generated answer: White
- Example where 0 should be marked:
    - Answer rationale: Although the height of Burj Khalifa is not stated, skyscrapers are generally 100 meters.
    - Correct/generated answer: 830m

F Prompts

We provide multiple prompts used throughout this work. When not mentioned otherwise, we use Gemini 2.5 Pro (`gemini-2.5-pro-preview-03-25`) for this task.

1. Fig. 4 presents the prompt used to classify a WhisperX (Bain et al., 2023) transcription into AD or dialog.
2. Fig. 5 presents the prompt used to align the plot synopses sentences with a dialog + AD movie "script" (not the real script).

3. Fig. 6 presents the prompt used to generate *visual appreciation* MCQAs in our ADQA benchmark.
4. Fig. 7 presents the prompt used to generate *narrative understanding* MCQAs for the CMD-AD dataset.
5. Fig. 8 presents the prompt used to generate *narrative understanding* MCQAs for the MAD-eval dataset.
6. Fig. 9 presents the prompt used to answer questions from ADQA based on dialog and/or AD inputs. All answering is performed using Gemini-2.0-Flash.

#	Question	Answer Options	Rationale
Visual Appreciation			
1	How many companions accompany Biff?	A) One B) Two **C) Three** D) Four E) Five	As specified in the audio description, Biff has three buddies with him.
2	What physical characteristics are noted about Biff?	A) Short and slim with long hair B) Average height and build with curly hair **C) Tall and muscular with short hair** D) Short and stocky with blond hair E) Tall and thin with slicked-back hair	As specified in the audio description, Biff is tall and muscular and wears his hair cut short.
3	What is the location of the vehicle Biff and his buddies enter?	A) In the cafe's parking lot B) Down the street C) In an alleyway **D) Parked outside the cafe** E) Across the town square	As specified in the audio description, the black convertible is parked outside the cafe.
4	What type of eyewear is another of Biff's buddies wearing?	A) Sunglasses B) Reading glasses C) Safety goggles **D) 3D glasses** E) A monocle	As specified in the audio description, another buddy has 3D glasses.
5	What simultaneous action do Marty and the nearby kid perform?	A) They both sigh **B) They both turn their heads** C) They both check their watches D) They both take a drink E) They both look down	As specified in the audio description, both boys turn their heads.
Narrative Understanding			
6	What distressing situation involving his young father does Marty witness at this location after they meet?	A) His father failing an important test B) His father being rejected for a date **C) His father being subjected to bullying** D) His father getting into a car accident E) His father being scolded by his parents	The description indicates Marty ’watches him get bullied by Biff.’
7	Who is primarily responsible for the mistreatment directed towards Marty's young father during the scene at the diner?	A) Strickland, the school principal B) Marty, through an accidental intervention **C) Biff Tannen** D) A group of his father's classmates E) The diner's cook	The description explicitly states that Marty's father is bullied ’by Biff.’ Biff Tannen is the character's full name.
8	Whom does Marty specifically meet at the diner before witnessing an incident involving Biff?	A) His young mother, Lorraine **B) His young father, George** C) A younger version of Doc Brown D) His older brother, Dave, as a teen E) Goldie Wilson, the future mayor	The description states, ’Marty meets his young father at the diner...’. George is the name of Marty's father.
9	What is Marty's primary role concerning the confrontation involving his young father and Biff at the diner?	A) He directly intervenes to help his father. B) He actively encourages Biff. **C) He observes the situation unfold.** D) He calls for help from others in the diner. E) He tries to create a distraction.	The description states, ’Marty ... watches him get bullied by Biff,’ indicating his role as an observer of the event.
10	To what year is Marty transported after accelerating the DeLorean under duress?	A) 1885 **B) 1955** C) 1985 D) 2015 E) An unknown future year.	The description specifies that Marty "goes back in time to 1955."

Table 7: Example multiple-choice question-answers for **Visual Appreciation** (top) and **Narrative Understanding** (bottom) from ADQA. The correct answer is highlighted in bold.

Context Type	Answer Options	Rationale	AC
Question 1: How many companions accompany Biff?			
No Context	A) One B) Two **C) Three** D) Four E) Five	Bullies are often accompanied by multiple companions for intimidation purposes	-
Dialog	A) One **C) Two** *C) Three* D) Four E) Five	The dialogue references Biff and 'those boys', suggesting more than one companion, however without additional information on the exact number of companions accompanying Biff.	✗
Dialog + AD	A) One B) Two **C) Three** D) Four E) Five	The audio description says, "The young Beth and three cronies leave, jumping into an open-top car parked outside."	✓
Question 7: Who is primarily responsible for the mistreatment directed towards Marty's young father during the scene at the diner?			
No Context	A) Strickland, the school principal B) Marty, through an accidental intervention **C) Biff Tannen** D) A group of his father's classmates E) The diner's cook	In Back to the Future, Biff Tannen and his gang are the primary antagonists and bullies. It is Biff who instigates and leads the mistreatment of Marty's father, George McFly, at the diner.	-
Dialog	A) Strickland, the school principal B) Marty, through an accidental intervention **C) Biff Tannen** D) A group of his father's classmates E) The diner's cook	The dialogues "What are you looking at, butthead?", "Hey, Biff, get a hold of this guy's life preserver.", "Dork thinks he's gonna drown.", "So, uh, how about my homework, McFly?", "Oh, McFly, your shoe's untied.", "Don't be so gullible, McFly.", "I don't want to see you in here again." all indicate that Biff Tannen is mistreating George McFly.	✓
Dialog + AD	A) Strickland, the school principal B) Marty, through an accidental intervention **C) Biff Tannen** D) A group of his father's classmates E) The diner's cook	The audio descriptions and dialogue clearly indicate that Biff Tannen is the primary instigator of the mistreatment towards George McFly. Biff is shown interacting with George in a demeaning way, demanding his homework, tripping him, and generally bullying him, as stated by "Dialogue: Well, now, of course not, Biff.", "Dialogue: So, uh, how about my homework, McFly?", "Dialogue: Oh, McFly, your shoe's untied." and "Dialogue: Don't be so gullible, McFly."	✓

Table 8: Answering questions 1 and 7 (from Tab. 7) using the second AD source for example question from visual appreciation (top) and narrative understanding (bottom). We show the question and answering using different context types (col 1). Ground-truth correct answer is highlighted in green while the model predicted answer is in bold (red when wrong, green when correct). We see that the LLM is able to answer the question even without any context based on common sense reasoning. However, our process to check whether the context was used to create the rationale helps us assess whether the answer was generated using context (last column, AC).

Question	**Options**	**Response**	**Potential reason for marking False**
In this situation, who ultimately executed the task that Keaton hesitated to perform?	A) The individual eventually complied B) Another, unnamed character **C) Verbal** D) Keaton, after a delay E) The opportunity was lost	False	Likely confused by “Verbal” as a name.
Whose marriage to Greg is contingent upon the outcome of Jack’s lie detector test?	A) Greg’s sister B) A distant relative of Jack **C) Jack’s daughter** D) A mutual friend E) Jack’s ex-wife	False	Participant may have assumed the option "Greg’s sister" as unlikely.
How does Edward first observe the scene after the shooting?	A) By opening a door slightly B) By looking through a window **C) By peeping around the corner** D) By using a periscope E) By stepping out fully	False	Unknown.
What type of vehicle does Nikki observe arriving?	A) Bourne’s motorcycle B) A police car C) An ambulance **D) A van** E) The asset’s hatchback	False	"Asset’s hatchback" may have seemed strange without context about "The Asset" — a character in the movie.
What action does the pursuing individual take towards Bourne?	A) Shouts warnings B) Fires a weapon C) Rides a motorcycle **D) Runs** E) Throws an object	False	Unknown.
What object does Bourne discharge towards the van?	A) A bullet B) A flare **C) A gas canister** D) A grappling hook E) Water from the hose	False	Likely found the verb “discharge” odd for all options, one normally shoots a bullet, throws a canister, or fires a flare.

Table 9: Subset of user study questions where participants judged whether the options were valid given only the question and multiple-choice options (without question generation context). The correct answer is highlighted in **green and bold**. The final column lists potential reasons for incorrect judgments (hypothetical, since we did not collect explicit justifications).

You are an expert in analyzing movie scripts. You will be given a list of sentences that appear sequentially in a movie. For each sentence, classify it as either:

- "dialogue" — if it's a spoken line by a character. Music, background chatter, or anything that is not an audio description should also be classified as dialogue
- "AD" — if it's an audio description of what is happening on screen.

Further description:
Dialogue: Spoken lines from a movie, typically involving characters talking.
Example characteristics:

- Use of first-person pronouns like "I," "me," or "my."
- Often conversational or emotional in tone. Could be a command, exclamation, rambling, etc.
- Examples:
  - "You fought a bear? Are you insane?"
  - "It was either me or him. And honestly, I think I was more scared than he was."
  - "What even possessed you to go into the forest alone?"
  - "Get up"
  - "Move"

Audio Description (AD): Sentences that narrate the visual elements of a movie, intended for blind or visually impaired viewers.
Example characteristics:

- Is a narrator describing the scene visually.
- Descriptive and neutral tone.
- Often focuses on actions, settings, or appearances.
- Usually starts describing a scene by setting up the environment like "Outside", "Downstairs", "In a sunny afternoon outside", "Out in the snow", etc.
- No first-person perspective or conversational cues.
- Examples:
  - "A dense Russian forest, snow falls steadily, blanketing the ground. A man steps forward, his breath visible in the icy air."
  - "The man lunges at the bear with a crude spear, but the bear swats it aside effortlessly."
  - "He gets up"

Instructions:

1. For *every* input sentence, return exactly one classification: either "{dialogue_tag}" for dialogue or "{ad_tag}" for Audio Description.
2. Do *not* skip any inputs, even if they are very short or ambiguous.
3. Match the output count to the input count. If n sentences are given, return exactly n outputs.
4. Do *not* add any commentary, explanation, or extra lines. Just one output per sentence: "{dialogue_tag}" or "{ad_tag}".
5. Use context between sentences if helpful, since these sentences are sequential from a movie.
6. Some of the movies may be rated for adult audiences and might contain explicit sentences. This makes no difference; the classification should be done regardless just as for any other sentence. Be careful not to include any unsafe or overly sexual content.
7. Some sentences might be a mix of AD and dialogue due to transcription errors. These sentences should be labelled "{ad_tag}" if the audio description part is more prominent in the sentence, otherwise "{dialogue_tag}".

Input format:

1. {sentence1}
2. {sentence2}
3. {sentence3}
4. ...

Output format:

1. {classification1}
2. {classification2}
3. {classification3}
4. ...

Here is the input: {input}

Figure 4: Prompt to classify transcriptions into "dialogue" or "AD".

You are a movie editing AI assistant who's job is to segment the movie script into distinct scenes. Each scene is a self contained logical segment of the movie. You will be provided with two inputs: Movie script and Plot synopsis.

Movie script format:

```
Line <number>
<start time in hh:mm:ss.ss> --> <end time in hh:mm:ss.ss>
<Sentence type Dialogue or Audio Description>: <sentence>
```

Plot synopsis format: <plot synopsis paragraph>

Instructions:

- Segment the script into logical scenes, each spanning approximately few minutes of screen time (based on timestamps or logical transitions in dialogue and descriptions).
- For each scene, list the index range of script lines (e.g., 1–10, 11–18, etc.).
- For each scene, identify which sentence(s) from the plot synopsis match the scene's events, if any. If a scene doesn't match any part of the synopsis, note that no match was found.
- Use timestamps, audio descriptions, and dialogue shifts to define scene boundaries.
- If two script segments are logically distinct (e.g., a sudden change in location or topic), treat them as separate scenes.
- Pay special attention to changes in scenes described in audio descriptions.
- If a plot synopsis line spans across multiple consecutive scenes, then merge the scenes into one.
- Every detail in the plot synopsis should be explainable from the scene. If some detail exists in another consecutive scene, then merge the scenes.
- A scene may have one, multiple, or no corresponding synopsis lines.
- Every plot line must be associated to some scene, and each line can only be associated to at most 1 scene.

Output format:

```
[
  (<Line number of scene start>, <Line number of scene end>, <Plot synopsis sentence(s) that correspond to the scene OR None>),
  (<Line number of scene start>, <Line number of scene end>, <Plot synopsis sentence(s) that correspond to the scene OR None>),
   ...
]
```

Input:
Movie Script: {movie_script}
Plot synopsis: {plot_synopsis}
Output:

Figure 5: Prompt used to align the plot synopses sentences with a dialog + AD movie "script" (not the real script).

You are given a movie scene in text form, which consists of dialogues and audio descriptions. Your task is to generate questions exclusively based on the audio descriptions, ignoring the dialogues and only using them for context.
Every audio description sentence has to be used to construct 1 or more questions asking about direct facts. The questions must ask about every factual detail about the audio description sentence.

Examples: An AD sentence such as "A green truck speeds through the highway crossing a yellow barrier" can become multiple questions such as "What vehicle is seen on the highway?", "What is the color of the vehicle going on the highway?", "What can be said about the speed of the vehicle on the highway?", "What does the vehicle cross on the highway?", "What color is the barrier on the highway?", etc. Audio descriptions that are used to establish the scene such as "Outside", "Later that night", "Inside the home", "Now, inside" can be converted to questions about the scene "Where is the scene taking place?", "What time of day is the scene taking place?", etc.

Question and Answer Format:
Questions: Limited to one or two lines, formulated to be insightful and not overtly indicative of the answer. Avoid using overly descriptive language that could hint at the correct answer. If there are no good questions to be generated, return an empty json.
Answers: Five options per question, formatted as "- A), - B), - C), - D), and - E)", concise and reflective of the question's depth.
Answer Key: Specify the correct answer clearly with the formatting, "Correct Answer:", in the line following all the answer options.
Rationale: Write a rationale explaining the correctness of the "Answer Key" based on the scene's context in the next line.
The response should be in json format without any extra comments.

Very Important Rule: Make sure none of the question is answerable by looking at other questions and their options.

Output format:
Return json formatted text. Example:

```
[
    {
        "question": "question text1",
        "options": ["A) answer key 1", "B) answer key 2", "C) answer key 3", "D) answer key 4", "E) answer key 5"],
        "correct_answer": "E) answer key 5",
        "rationale": "As specified in the audio description, <rationale>",
    },
    {
        "question": "question text1",
        "options": ["A) answer key 1", "B) answer key 2", "C) answer key 3", "D) answer key 4", "E) answer key 5"],
        "correct_answer": "A) answer key 1",
        "rationale": "As specified in the audio description, <rationale>",
    }
]
```

Figure 6: Prompt used to generate **visual appreciation** MCQAs in our ADQA benchmark.

You are a teacher who's job is to create questions out of a 1 line description of a clip from a movie to test narrative understanding of the students. The questions must ask about factual details related to the description. The description is a 1 line summary, and the students are expected to answer the questions having watched the movie, without seeing the description.

Examples:
A description such as "The shining spaceship lands on a strange planet covered in glowing blue plants and mist." can be converted into many questions such as , "Where does the spaceship land?" (Answer: On a strange planet), "What makes the planet unusual?" (Answer: The planet is covered in glowing blue plants and mist)
A description such as "Mark waits alone by the lake after missing the last boat home." can be converted into many questions such as "Who is Mark with waiting by the lake?" (Answer: Mark is alone), "Why is Mark waiting by the lake?" (Answer: Mark missed the last boat home)
Remember that there may be many things happening in the clip from the movie, and the 1 line summary may choose to not highlight them. This may lead to ambiguous questions which should be avoided.

Example of ambiguous question: Question such as "What did the spaceship do?" is ambiguous given the description "The shining spaceship lands on a strange planet covered in glowing blue plants and mist.", because the spaceship might have done many things in the clip that were not described in the summary. The students will not know which action the question is asking for out of the many actions the spaceship performed in the clip.
Question such as "What is the spaceship described as?" is ambiguous given the description "The shining spaceship lands on a strange planet covered in glowing blue plants and mist.", because the word "shining" used to describe the spaceship might exist only in the 1 line summary (to which the students don't have access to), and the spaceship "shining" may have not been the most prominent feature of the spaceship in the movie clip and so the students may consider the question ambiguous.

Question and Answer Format:

Questions: Limited to one or two lines, formulated to be insightful and not overtly indicative of the answer. Avoid using overly descriptive language that could hint at the correct answer.
Answers: Five options per question, formatted as "- A), - B), - C), - D), and - E)", concise and reflective of the question's depth.
Answer Key: Specify the correct answer clearly with the formatting, "Correct Answer:", in the line following all the answer options.
Rationale: Write a rationale explaining the correctness of the "Answer Key" based on the scene's description.

Output format: Return json formatted text. Example:

```
[
    {
        "question": "question text1",
        "options": ["A) answer key 1", "B) answer key 2", "C) answer key 3", "D) answer key 4", "E) answer key 5"],
        "correct_answer": "E) answer key 5",
        "rationale": "<rationale>",
    },
    {
        "question": "question text1",
        "options": ["A) answer key 1", "B) answer key 2", "C) answer key 3", "D) answer key 4", "E) answer key 5"],
        "correct_answer": "A) answer key 1",
        "rationale": "<rationale>",
    }
]
```

Description: {description}

Figure 7: Prompt used to generate **narrative understanding** MCQAs for the CMD-AD dataset.

You are a teacher who's job is to create questions from plot summary of a clip from a movie to test narrative understanding of the students. The questions must ask about factual details related to the plot. The students are expected to answer the questions having watched the movie clip, without seeing the plot summary.

Examples:
- A summary such as "The shining spaceship lands on a strange planet covered in glowing blue plants and mist." can be converted into many questions such as , "Where does the spaceship land?" (Answer: On a strange planet), "What makes the planet unusual?" (Answer: The planet is covered in glowing blue plants and mist)
- A summary such as "Mark waits alone by the lake after missing the last boat home." can be converted into many questions such as "Who is Mark with waiting by the lake?" (Answer: Mark is alone), "Why is Mark waiting by the lake?" (Answer: Mark missed the last boat home)

Remember that there may be many things happening in the clip from the movie, and the summary may choose to not highlight them. This may lead to ambiguous questions which should be avoided.

Example of ambiguous question:
- Question such as "What did the spaceship do?" is ambiguous given the summary "The shining spaceship lands on a strange planet covered in glowing blue plants and mist.", because the spaceship might have done many things in the clip that were not described in the summary. The students will not know which action the question is asking for out of the many actions the spaceship performed in the clip.
- Question such as "What is the spaceship described as?" is ambiguous given the description "The shining spaceship lands on a strange planet covered in glowing blue plants and mist.", because the word "shining" used to describe the spaceship might exist only in the plot summary (to which the students don't have access to), and the spaceship "shining" may have not been the most prominent feature of the spaceship in the movie clip and so the students may consider the question ambiguous.

Question and Answer Format:

Questions: Limited to one or two lines, formulated to be insightful and not overtly indicative of the answer. Avoid using overly descriptive language that could hint at the correct answer.
Answers: Five options per question, formatted as "- A), - B), - C), - D), and - E)", concise and reflective of the question's depth.
Answer Key: Specify the correct answer clearly with the formatting, "Correct Answer:", in the line following all the answer options.
Rationale: Write a rationale explaining the correctness of the "Answer Key" based on the scene's description.

Output format: Return json formatted text. Example:

```
[
    {
        "question": "question text1",
        "options": ["A) answer key 1", "B) answer key 2", "C) answer key 3", "D) answer key 4", "E) answer key 5"],
        "correct_answer": "E) answer key 5",
        "rationale": "<rationale>",
    },
    {
        "question": "question text1",
        "options": ["A) answer key 1", "B) answer key 2", "C) answer key 3", "D) answer key 4", "E) answer key 5"],
        "correct_answer": "A) answer key 1",
        "rationale": "<rationale>",
    }
]
```

Plot summary {summary}

Figure 8: Prompt used to generate **narrative understanding** MCQAs for the MAD-eval dataset.

A series of questions and their options are given below.
`{questions_with_choices}`

Provide 1 answer to each of the questions based on the following `{context_type}`.
If `{context_type}` are not available, then they will not be provided. Also come up with rationale for the answers, quoting the specific (one or more) `{context_type}` used for answering the question.
If the rationale suggests that the question is answered by directly using `{context_type}`, then the boolean variable `{answered_from_var_name}` should be "True".
Otherwise, if the rationale suggests that the question is answered by not using `{context_type}`, but by prior knowledge or by common sense, then the variable `{answered_from_var_name}` should be "False".
Always answer `{answered_from_var_name}` as either "True" with T upper case and "rue" lower case OR "False" with F upper case and "alse" lower case.

`{context_type}`: `{context}`

Instructions

1. Every question has to be answered.
2. There should be 1 and only 1 answer to each question. If no answer is known, take an educated guess. Do not answer the same question more than once.
3. All questions should be answered independently, i.e., you may not use other questions and their options to answer any question.
4. Answer only as in the output format provided:

Output format (substitute the <...> with appropriate values):

```
[
    {
        "answer": "<answer>" ,
        "rationale": "<rationale>" ,
        "{answered_from_var_name}": "<{answered_from_var_name}>" ,
    },
    ...
]
```

Figure 9: Prompt used to answer questions from ADQA based on dialog and/or AD inputs.