MM-Narrator: Audio Description Generation with Multimodal In-Context Learning (Supplementary)

https://MM-Narrator.github.io

In this supplementary, we present more details and discussions of AD generation with MM-Narrator (§A), our proposed complexity-based MM-ICL (§B), ablation studies of MM-ICL (§C) and AD evaluation with `SegEval` (§D). Next, we elaborate our implementation details (§E) and discuss the future work on both AD generation and evaluation (§F).

A. AD Generation

MM-Narrator builds prompts to query GPT-4 for recurrent AD generation, including the following elements: task-specific introduction $I_{\texttt{task}}$ and hint $H_{\texttt{task}}$, main query $q_{\texttt{main}}$, as well as a set of few-shot multimodal demonstrations $\mathcal{D}_{\texttt{ICL}}$ to conduct in-context learning. With a breakdown shown in Figure 7, we present the details as follows.

Querying with multimodal clues. Both the main query $q_{\texttt{main}}$ and the demonstration queries in $\mathcal{D}_{\texttt{ICL}}$ are formatted with the same query builder, which outputs AD query from multiple text-formed multimodal clues. These multimodal clues include visual captions ($\boldsymbol{x}_i^{cap}$) with successfully re-identified characters, recent context ADs ($\mathcal{M}_{\boldsymbol{short}}$) and character dialogues ($\boldsymbol{x}_{t\in T_{sub}}^{sub}$).

Prompting with MM-ICL. Each MM-ICL demonstration within $\mathcal{D}_{\texttt{ICL}}$, is composed of a pair $(\mathcal{Q}, \mathcal{A})$ or a tuple $(\mathcal{Q}, \mathcal{R}, \mathcal{A})$ when chain-of-thought (CoT) is adopted to generate the multimodal multi-step reasoning $\mathcal{R}$ that derives answer $\mathcal{A}$ from question $\mathcal{Q}$.

More qualitative results. Apart from Figure 1 and 4 in main paper, we show additional qualitative demonstrations of MM-Narrator on both MAD-eval-`Named` benchmark and other long-form videos (external to the MAD-eval dataset) as Figure 10 and Figure 11, respectively, in this supplementary.

B. Details of Complexity-based MM-ICL

Combining CoT with complexity-based ranking, our proposed complexity-based MM-ICL performs more favorably than classic ICL solutions. We reveal their details as follows.

Reasoning with CoT. We first employ GPT-4 to articulate the chain-of-thoughts (CoTs) as reasoning steps, denoted as $\mathcal{R}$, that assist in deriving the answer $\mathcal{A}$ from the question $\mathcal{Q}$. Practically, we found a CoT-specific constraint[1] helpful to derive reliable CoTs, ensuring a closed-loop reasoning to be inferred. Without this constraint, LLM might unexpectedly generate $\mathcal{R}$ followed by its own AD prediction, which are different from the human annotated $\mathcal{A}$.

Quantifying on atomic steps. Practically, we observe that raw steps decided by LLM itself, might not be a considerably consistent measurement among various examples. Take two demonstrations shown in Figure 8 as example: Steps 3 to 7 in *left example*, conduct reasoning over per-frame captions individually, which are equivalent to step 2 in *right example*, including several sub-steps in analysing the per-frame captions. To this end, following [19], we split $\mathcal{R}$ into *atomic steps* by newline char "\n", and propose using the number of atomic steps $N_{\texttt{atomic}}$ as our measurement of reasoning complexity.

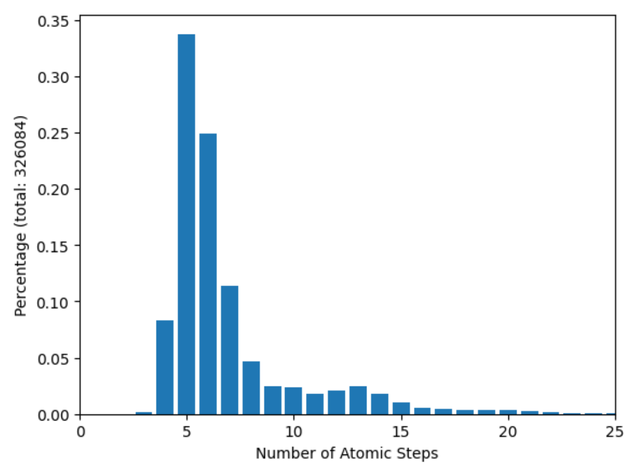

Figure 6. Distributions of multimodal MAD-v2-`Named` demonstrations over reasoning complexity, quantified by $N_{\texttt{atomic}}$.

Ranking by complexity. We propose to select the most intuitive examples to perform few-shot MM-ICL for improving AD generation. Here, we show the distributions of multimodal demonstrations over the complexity in Figure 6. Specifically, the 10% shortest examples lead to a simple demonstration pool $\mathcal{P}_{\texttt{simple}}$ with its maximum $N_{\texttt{atomic}}$ as 5, while the 10% longest ones result in another pool $\mathcal{P}_{\texttt{hard}}$ whose minimum $N_{\texttt{atomic}}$ equals to 12.

[1] CoT-specific constraint: "lets fill-in the REASONING process which derives the ANSWER from QUESTION."

C. More Ablations on MM-ICL

Table 4 in the main manuscript implies that complexity is a suitable criterion for selecting efficient ICL demonstrations to enhance AD generation. Here, we further discuss *three sub-questions* to elaborate a few in-depth ablation studies, as following:

Does CoT help? We propose to adopt CoT technique to obtain the intermediate reasoning steps $\mathcal{R}$ that help derive answer $\mathcal{A}$ from question $\mathcal{Q}$. This automatic process extends demonstration format from $(\mathcal{Q}, \mathcal{A})$ pairs to $(\mathcal{Q}, \mathcal{R}, \mathcal{A})$ tuples. As its consistent gains could be observed multiple times (R1 vs R2; R3 vs R4; C1 vs C2), adding multimodal multi-step reasoning $\mathcal{R}$ during MM-ICL could help MM-Narrator improve its multimodal reasoning capability to better incorporate multimodal inputs. Qualitative demonstrations of $\mathcal{R}$ are shown as Figure 8 in this supplementary.

Does complexity-based ranking help? We observed that conducting MM-ICL with the most intuitive examples benefits the overall performance (R4 vs C2), however, switching with the hardest ones which own the longest reasoning steps, MM-ICL actually leads to a decline in performance (R4 vs C3). These results indicate that more straightforward examples, quantified by the shortest number of reasoning steps, compile to a simpler yet more powerful subset MM-ICL demonstration pool for effective AD generation.

Does complexity-based MM-ICL work effectively? Combining CoT with complexity-based ranking, our proposed complexity-based MM-ICL (C2) performs more favorably than the random and similarity-based sampling approaches (R1 [13] and S1 [33]), which are classic solutions in choosing few-shot ICL examples. Besides, ours is easy-to-implement and explainable-to-human, avoiding the computation overhead of retrieval-based selection.

D. AD Evaluation with GPT-4

Suppose a few ADs form one segment `Seg`. For each `Seg`, our proposed `SegEval` evaluator takes into consideration an oracle context window `Ctx` of length W, to measure its AD quality with GPT-4. The details of `SegEval` prompt are shown as Figure 9. We elaborate each individual marking criteria as follows:

- ***originality***: Evaluates if the `Seg` is novel and non-repetitive, to enrich the watching experience of the visually impaired.
- ***consistency***: Checks if the generated `Seg` maintains a consistent tone or content throughout.
- ***coherence***: Determines whether `Seg` logically connects to the given `Ctx`. A coherent text flows smoothly and deepen the movie understanding for the visually impaired.
- ***diversity***: Focuses on the variety of `Seg` generated. A good model should produce varied outputs rather than repetitive or highly similar ones against the given `Ctx`.
- ***specificity***: Measures the level of detail in the generated `Seg`, assessing if it is sufficiently detailed and/or focused for the `Ctx`.

Noticeably, the first two marking aspects focus on text-level AD quality, which are context-free ($W = 0$) evaluation metrics, while the rest three metrics measure sequence-level AD generation, taking oracle context into consideration.

E. Implementation Details

Multimodal Experts. To obtain framewise image caption and people detection, we utilize vision experts publicly available via the Azure Cognitive Services APIs[2]. For speech recognition, we choose WhisperX [10] as our audio expert. To register and recall long-term visual memory for character re-identification purpose, we adopt CLIP-ViT-L14 [48] as our visual feature extractor, and query GPT-4 as our Person-NER tool with the following prefix: *"Extract the people names in the following text as a string splitted by '|' (return 'none' if none of names are recognized): "*.

Building MM-ICL Pool. We build the MM-ICL demonstrations for each sample in MAD-v2-`Named` split [22]. As the raw frames are not publicly available, we derive per-frame captions by inferring ClipCap [41] on the released CLIP-ViT features. Differing from the main query $q_{\texttt{main}}$, whose recent context ADs in $\mathcal{M}_{\texttt{short}}$ are recurrently generated by MM-Narrator, the queries in MM-ICL demonstrations $\mathcal{D}_{\texttt{ICL}}$ are instead built with human annotations as their recent context ADs. Additionally, we omit long-term visual memory retrievals when constructing MM-ICL demonstrations.

GPT-4 Error Handler. GPT-4 might inevitably return errors when the content filtering policies[3] are occasionally triggered in Azure OpenAI Service. Such cases account for a very small proportion (less than 0.1%), thus they would not largely affect the overall performance. To address them, we utilize ClipCap [41] as the error handler to output video caption as AD. Specifically, we inference ClipCap on the mean pooled feature among frames in each video clip.

Deployment on Long-form Videos. We utilize PySceneDetect [5] for scene detection, and based on that, we cut long-form videos into video clips for recurrent AD generation with MM-Narrator. We utilize Google Text-to-Speech (gTTS) [3] for voice-over audio creation, which narrates AD for each video clip.

[2] https://azure.microsoft.com/en-us/products/cognitive-services/vision-services

[3] https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/content-filter

Hyper-parameter Settings. Following [22], the number of frames N to be sampled per video clip is set to 8, while we utilize subtitles within a time window T_{sub} set to 0.25 minutes. Our short-term memory queue is maintained to contain K most recently predicted ADs with timestamps, where K = 7. The number of demonstrations C equals to 5, which are sampled for conducting MM-ICL. The API versions of GPT-4 and GPT-4V used in our experiments are 'gpt4-2023-03-15' and 'gpt4v-2023-08-01', respectively.

F. Future Works

AD Generation. In future developments in Audio Description (AD), a critical enhancement will be the integration of advanced audio-visual speaker and character identification, coupled with strategic timing for AD delivery. This direction involves not only recognizing who is speaking or present in a scene but also determining the most opportune moments to provide descriptions without interrupting critical dialogue or action. Additionally, the establishment of a much more comprehensive and reliable external character bank, facilitating retrieval-augmented generation, will further refine AD content, ensuring it is both contextually relevant and timely. These advances are poised to transform AD into a more coherent, immersive experience, significantly improving accessibility for visually impaired audiences.

AD Evaluation. In future work for AD evaluation, a crucial focus should be on enhancing the measurement of factuality, an aspect not adequately addressed by current evaluation criteria like `SegEval`. Given the limitation of traditional reference-based scores in precisely assessing the factual accuracy of AD content, employing AI models such as GPT-4V emerges as a promising solution. GPT-4V's advanced capabilities in understanding and contextualizing multimedia content could offer a more nuanced and accurate evaluation of AD factuality. This shift towards AI-driven, factuality-focused evaluation methods would not only provide a more comprehensive assessment of AD quality but also ensure that the generated descriptions are reliably accurate, ultimately benefiting visually impaired individuals with a more authentic storytelling experience.

Acknowledgment

We are deeply grateful to OpenAI for providing access to their exceptional tool [44, 45]. We also extend heartfelt thanks to our Microsoft colleagues for their insights, with special acknowledgment to Faisal Ahmed, Ehsan Azarnasab, and Lin Liang for their constructive feedback.

(A) AD Generation Prompt (Overview)

Task-specific Introduction

Task-specific Hint

Below are a few in-context examples of how you'd like to response for Audio Description Generation. Use these as a guide for the upcoming QUESTION:

------- Example Start ------------

Example 1: MM-ICL Demonstration

------- Example Split ------------

...... MM-ICL Demonstration

------- Example End ------------

Now, based on these examples, please DIRECTLY answer the following question for Audio Description Generation:

QUESTION: Main Query

ANSWER:

(B) Task-specific Introduction

Suppose you are an audio narrator who generates the audio descriptions (ADs) for blind people. However, instead of watching the videoclip, you will read the per-frame captions of the given videoclip, as well as the context ADs and/or subtitles before this videoclip. You must not narrate frame by frame. Instead, you should generate one-sentence brief AD covering the following videoclip. Note: the AD should be brief and concise and should not contain any redundant information. The length of the AD should match the length of the videoclip. Reminder: AD is not a video caption; thus, it must NOT contain words like "videoclip", "frames", "video". Specifically, the AD must NOT start with some common phrases like "This is a videoclip about ..." or "The videoclip describes/transitions/displays that ...".

(C) Task-specific Hint

(HINT: try to infer the character names from context ADs and subtitles, as well as the per-frame captions - where past memory might be recalled - for AD generation)

(D) MM-ICL Demonstration – *(Q, A)* pair or *(Q,R,A)* tuple

Q

*** QUESTION:**
[per-frame caption]
At T = 0.0s: car driving on a highway in bad weather.
At T = 0.2s: cars driving on the highway in bad weather.
At T = 0.4s: the car is driving on the road.
At T = 0.6s: the car is stuck in the middle of the road.
At T = 0.8s: cars driving on the highway in bad weather.
At T = 1.0s: cars driving on the road in bad weather conditions.
At T = 1.2s: cars driving on the road.
At T = 1.4s: cars driving on the road in bad weather conditions.

[Context (Previous) Audio Descriptions]
About 155.5 seconds ago: Deadpool sits on the railing of an overpass with a radio.
About 152.3 seconds ago: A small piece of tape on it reads, Wade.
About 149.5 seconds ago: He doodles a drawing of himself shooting a man in the head.
About 146.5 seconds ago: The man's speech bubble reads, hi, I'm Francis.
About 113.3 seconds ago: The caravan approaches on a freeway far below.
About 108.9 seconds ago: Deadpool drops down and crashes through an SUV sunroof.
About 58.1 seconds ago: Deadpool hits the driver on his wound and slams the gas pedal with his hand.

[Context (Previous) Subtitles]
About 14.6 seconds ago: Subtitle reads as "Shit."
About 13.2 seconds ago: Subtitle reads as "Did I leave this stove?"
About 10.3 seconds ago: Subtitle reads as "At normal speed, the chain whips forward, beheading the rider."
About 7.3 seconds ago: Subtitle reads as "Another body is thrown from the SUV and splats against the freeway sign."
About 3.3 seconds ago: Subtitle reads as "The rolling SUV skids to a halt, sideways on the bridge."

R

*** REASONING:**
1. Analyze the per-frame captions: The captions describe a car driving on a highway in bad weather, getting stuck in the middle of the road, and other cars driving in the same conditions.

2. Analyze the context audio descriptions: The previous audio descriptions mention Deadpool's actions, such as sitting on an overpass railing, doodling, and crashing into an SUV. There is no direct mention of other characters' names.

3. Analyze the context subtitles: The subtitles provide some dialogue and descriptions of action, such as a chain whipping forward, a body being thrown from the SUV, and the SUV skidding to a halt.

4. Infer the situation: Based on the information provided, it seems that Deadpool has caused an accident on the highway, and the car is now stuck in the middle of the road.

5. Generate the audio description: Considering the bad weather conditions and the car being stuck in the middle of the road, it is reasonable to assume that other cars would need to brake hard to avoid colliding with the stuck car. Therefore, the audio description would be "Other cars break hard before it."

A

*** ANSWER:** Other cars break hard before it.

(E) Main Query

[per-frame caption]
At T = 0.0s: a hand holding a yellow plastic container
At T = 0.3s: a hand holding a yellow plastic cup
At T = 0.9s: a hand holding a yellow plastic container
At T = 1.4s: a hand holding a yellow plastic container
At T = 1.9s: a hand holding a yellow plastic container *- who could be a character we've seen before, especially around the previous scene about 21.9s ago, whose AD reads as "Lisa excitedly opens a gift from George."*
At T = 2.4s: a hand holding a yellow plastic container
At T = 2.9s: a hand holding a yellow container *- who could be a character we've seen before, especially around the previous scene about 22.9s ago, whose AD reads as "Lisa excitedly opens a gift from George."*
At T = 3.4s: a hand holding a yellow container *- who could be a character we've seen before, especially around the previous scene about 23.4s ago, whose AD reads as "Lisa excitedly opens a gift from George."*

[Context (Previous) Audio Descriptions]
About 25.0 seconds ago: The man in a suit continues his conversation with Lisa as George watches.
About 21.4 seconds ago: Lisa excitedly opens a gift from George.
About 17.7 seconds ago: Lisa smiles as she discovers the contents of the gift from George.
About 12.4 seconds ago: George and Lisa engage in conversation at the table.
About 9.2 seconds ago: The man in a suit converses with Lisa, while George observes.
About 6.1 seconds ago: The man in a suit presents a gift to Lisa as George looks on.
About 4.4 seconds ago: Lisa receives a gift from the man in a suit while George watches, then she looks at a yellow cup.

[Context (Previous) Subtitles]
About 10.7 seconds ago: Subtitle reads as "Thanks for not rushing me."
About 7.2 seconds ago: Subtitle reads as "Yeah."

(F) Quantitative and Qualitative Analysis

How do you know (2010)
Start: 01:48:29 | End: 01:48:33

GT (via human annotation)
As Lisa examines it, her diamond watch glitters on her wrist.

PD (via MM-Narrator)
Lisa examines the yellow container from George.
R-L: 21.4 | **C:** 78.0 | **M:** 13.0 | **B1:** 16.1

Figure 7. A breakdown of the AD generation prompt constructed by MM-Narrator, including an (A) overview with ICL-specific instructions marked in green, (B) task-specific introduction $I_{\texttt{task}}$ and (C) hint $H_{\texttt{task}}$, a few multimodal ICL (MM-ICL) demonstrations $\mathcal{D}_{\texttt{ICL}}$ with an example shown as (D), and (E) the main query $q_{\texttt{main}}$ to be responded by GPT-4, with long-term visual memory marked in gray. Eventually, we show the corresponding (F) quantitative and qualitative analysis of the AD prediction via MM-Narrator against the human AD annotation. Zoom in for details.

(A) Prompting LLM to articulate CoTs as reasoning steps

You are an audio narrator who generates the audio descriptions (ADs) for blind people. However, instead of watching the videoclip, you will read the per-frame captions of the given videoclip, as well as the context ADs and/or subtitles before this videoclip.

Task-specific Hint

Now, given the below (**QUESTION**, **ANSWER**) pair example of AD generation, lets fill-in the **REASONING** process which derives the **ANSWER** from **Question**.

* **Question**: Question

* **Answer**: Answer

* **Reasoning**: **Lets think of step-by-step…**

(B) Two examples of deriving CoTs with LLM

[per-frame caption]
At T = 0.0s: close up of a burning firewood in a fireplace.
At T = 0.2s: a fire burns in a fireplace.
At T = 0.4s: slow motion of a burning flame in a metal bowl.
At T = 0.6s: a fire burns in slow motion.
At T = 0.8s: abstract background of a yellowish - orange background.
At T = 1.0s: person in a dark forest.
At T = 1.2s: a man is standing in front of a burning fireplace and is looking at the flame.
At T = 1.4s: slow motion of a girl running in the dark.

[Context (Previous) Audio Descriptions]
About 33.2 seconds ago: Another hand grabs Doris's wrist.
About 30.9 seconds ago: Lane's eyes clear and Debbie appears beside her.
About 27.3 seconds ago: The blonde girl stares intensely at the ghoul.
About 22.2 seconds ago: Doris flicks her eyes between the two friends.
About 12.2 seconds ago: Sarah emerges and throws the bundle into the raging furnace.
About 8.4 seconds ago: A shockwave throws Sarah back.
About 6.2 seconds ago: Doris' mouth hangs open and black smoke billows from it.

Lane throws the board into the furnace, surging flames throw her back against a wooden beam.

1. Analyze the per-frame captions and context audio descriptions to understand the scene and characters involved.
2. From the context audio descriptions, we can identify the characters as Doris, Lane, Debbie, and Sarah.
3. Now, let's focus on the per-frame captions to understand the actions happening in the scene.
4. At T = 0.0s to T = 0.8s, we see various descriptions of fire and flames, indicating that the scene is focused on a burning fireplace or furnace.
5. At T = 1.0s, there's a mention of a person in a dark forest, but it's not clear who this person is or how they relate to the scene.
6. At T = 1.2s, a man is standing in front of a burning fireplace, looking at the flame. This could be a character from the context audio descriptions, but we don't have enough information to determine who it is.
7. At T = 1.4s, there's a slow-motion scene of a girl running in the dark. This could be one of the female characters from the context audio descriptions, such as Lane, Debbie, or Sarah.
8. Considering the previous audio descriptions, we know that Sarah threw a bundle into the furnace, and a shockwave threw her back. This could be related to the current scene with the burning fireplace.
9. Based on the information available, we can infer that Lane is the one who throws the board into the furnace, as she is one of the characters mentioned in the context audio descriptions and is likely the girl running in the dark at T = 1.4s.
10. The surging flames and the shockwave mentioned in the context audio descriptions suggest that Lane is thrown back against a wooden beam after throwing the board into the furnace.

Thus, the audio description generated is: "Lane throws the board into the furnace, surging flames throw her back against a wooden beam."

[per-frame caption]
At T = 0.0s: person in a scene from the movie.
At T = 0.2s: a young woman is looking through a magnifying glass in a dark room.
At T = 0.4s: person in the dark : person.
At T = 0.6s: film character and person in the dark.
At T = 0.8s: the man in the dark.
At T = 1.0s: the man in the dark.
At T = 1.2s: the team in the library.
At T = 1.4s: the car is parked in the dark.

[Context (Previous) Audio Descriptions]
About 453.2 seconds ago: The handsome sailor sheepishly makes his way to the door.
About 448.9 seconds ago: Two dancers sandwich him between their bottoms and he bounces.
About 438.1 seconds ago: A director descends on a crane.
About 358.1 seconds ago: In the beach house living room...
About 128.2 seconds ago: Baird gazes off.
About 126.2 seconds ago: In an office, Eddie approaches a man wearing glasses.
About 9.5 seconds ago: He enters an editing room.

[Context (Previous) Subtitles]
About 13.1 seconds ago: Subtitle reads as "Why?"
About 11.9 seconds ago: Subtitle reads as "Never mind."
About 11.1 seconds ago: Subtitle reads as "Okay."
About 10.3 seconds ago: Subtitle reads as "Tell him I'll be there."
About 8.4 seconds ago: Subtitle reads as "Thank you, old lady."
About 7.6 seconds ago: Subtitle reads as "Oh, well, what is it?"
About 5.9 seconds ago: Subtitle reads as "Hello, Cece."
About 5.5 seconds ago: Subtitle reads as "I lost the thought we might go to later."
About 2.8 seconds ago: Subtitle reads as "Oh, hi, Eddie."
About 1.4 seconds ago: Subtitle reads as "Wanna lace up what you got on Barlowy Dance?"

It's up now.

1. Identify the character names and context from the previous audio descriptions and subtitles: We have Eddie, Cece, Baird, and a man wearing glasses.

2. Analyze the per-frame captions:
- At T = 0.0s: person in a scene from the movie.
- At T = 0.2s: a young woman is looking through a magnifying glass in a dark room.
- At T = 0.4s: person in the dark : person.
- At T = 0.6s: film character and person in the dark.
- At T = 0.8s: the man in the dark.
- At T = 1.0s: the man in the dark.
- At T = 1.2s: the team in the library.
- At T = 1.4s: the car is parked in the dark.

3. Generate the audio description based on the per-frame captions and context:
- "Eddie and Cece are in a dark room, where a young woman is looking through a magnifying glass. The man in the dark, possibly Baird or the man wearing glasses, is also present. The team gathers in the library, and a car is parked outside in the dark."

4. Final Answer: "It's up now."

Figure 8. Reasoning with CoT. (A) An overview of prompting LLM to articulate CoTs as multimodal multi-step reasoning $\mathcal{R}$ that derives the answer $\mathcal{A}$ from question $\mathcal{Q}$, with (B) two examples shown below. Zoom in for details.

(A) AD evaluation prompt with SegEval

Suppose you are a visually impaired person, and you will be "watching" a movie videoclip with audio description (AD). Here, you are requested to provide feedback (via reasoning and marking) on the performance of two AI assistants ("ASSISTANT1" and "ASSISTANT2") for automatic AD generation task:

Evaluation Steps:
1. you will be given <Context ADs>, <ASSISTANT1-output>, and <ASSISTANT2-output>, where <Context ADs> shows a few contextual human-annotated ADs but leaves "<PRESENT-SEGMENT>" empty to be filled with one or multiple AD(s) generated by two AI assistants (i.e., <ASSISTANT1-output> and <ASSISTANT2-output>).
2. you will read though the <Context ADs> and <ASSISTANT1-output> and <ASSISTANT2-output>, and then measure the AD generation quality of the two AI assistants in terms of Metric aspect.
3. you will complete the following five sections IN ORDER (namely, <Assistant1-Reasoning>, <Assistant2-Reasoning>, <Comparison-Reasoning>, <Assistant1-Score>, and <Assistant2-Score>).

HINT:
1. <Context ADs> will be used to provide the context of the movie scene. If it contains no valid ADs, it means that the current evaluation metric (Metric) will not take contextual information into account.
2. <Assistant1-Reasoning> and <Assistant2-Reasoning> will be used to record your reasoning and comments (with supporting evidence) on the Metric aspect of the ADs generated by two AI assistants, respectively;
3. <Comparison-Reasoning> will be used to record your feedback (with supporting evidence) for comparisons between the two AI assistants (with respect to the Metric aspect), which will be used to support the below two marking sections;
4. <Assistant1-Score> and <Assistant2-Score> will be used to record your AD generation Metric scores (from "1" to "5", where "1" indicates the worst and "10" indicates the excellent) of the two AI assistants, respectively.

Evaluation Criteria: Criteria

-------------Evaluation Starts---------------------

<Context ADs> Oracle Contexts

<ASSISTANT1-output> Current Segment *(Assistant 1)*

<ASSISTANT2-output> Current Segment *(Assistant 2)*

Please make sure you read and understand these instructions carefully, and complete the following five sections IN ORDER:
(1) firstly reason them individually within "<Assistant1-Reasoning>" and "<Assistant2-Reasoning>";
(2) secondly compare two assistants within "<Comparison-Reasoning>"; and
(3) finally mark them within "<Assistant1-Score>" and "<Assistant2-Score>"

(B) One example of **Diversity with L = 5 and W = 3**.

- Diversity: Focuses on the variety of <PRESENT-SEGMENT> generated. A good model should produce varied outputs rather than repetitive or highly similar ones against the given <Context ADs>.

About 57.0s ago: The Marines share a look.
About 43.0s ago: The staff sergeant wipes the wounded civilian's brow then sits back and unscrews a water bottle.
About 14.5s ago: Joe lifts his trembling hand.
About 11.2s ago: Doc eyes his superior.
About 8.9s ago: Nantz glances at the Marine and grips the wounded man's hand.
*** At present: ---PLACEHOLDER for "<PRESENT_SEGMENT>" to be generated by AI assistants below---**
About 62.0s later: On screen, a giant craft hovers.
About 75.0s later: Nantz meets her gaze and Santos nods.
About 81.0s later: They put Joe into the backroom.
About 83.5s later: Nantz compares a clock with his watch.
About 87.9s later: Harris checks his watch.

Now: The staff sergeant holds his gaze.
About 9.3s later: Nantz grabs his helmet.
About 12.9s later: Doc unfolds a foil blanket.
About 17.7s later: Now news footage.
About 34.0s later: Santos watches on.

Now: Cpl. Lee Imlay and Ssgt. Michael Nantz exchange concerned glances as they look down at something, surrounded by their fellow soldiers.
About 9.3s later: Nantz lies on the ground, tended to by Santos and fellow soldiers.
About 12.9s later: Santos tends to Nantz's wound as he lies on the floor, surrounded by concerned soldiers.
About 17.7s later: TSgt. Elena Santos stands silhouetted, raising her arms and pointing at a tree.
About 34.0s later: Santos and Mottola intently analyze screens, emotions running high.

<Assistant1-Reasoning>
Assistant1's output provides diverse descriptions of the scene, covering different actions and characters. The generated ADs include Nantz holding his gaze, grabbing his helmet, Doc unfolding a foil blanket, news footage, and Santos watching on. These descriptions are varied and not repetitive, showing a good level of diversity.

<Assistant2-Reasoning>
Assistant2's output also provides diverse descriptions of the scene, covering different actions and characters. The generated ADs include Nantz and Imlay exchanging concerned glances, Nantz lying on the ground being tended to, Santos tending to Nantz's wound, Santos raising her arms and pointing at a tree, and Santos and Mottola analyzing screens. These descriptions are varied and not repetitive, showing a good level of diversity.

<Comparison-Reasoning>
Both Assistant1 and Assistant2 provide diverse and varied descriptions of the scene, covering different actions and characters. Assistant1's output focuses more on Nantz's actions and reactions, while Assistant2's output provides more context about the situation and includes more characters. Assistant2's output seems to be slightly more diverse in terms of the actions and characters described, but both assistants perform well in terms of diversity.

<Assistant1-Score> 8

<Assistant2-Score> 9

Post-Processing (note: GPT-4 is agnostic to the output source during above marking)
Model Score: 1.125

(C) One example of **Originality with L = 5 and W = 1**.

- Originality: Evaluates if the <PRESENT-SEGMENT> is novel and non-repetitive, to enrich the watching experience of the visually impaired.

*** At present: ---PLACEHOLDER for "<PRESENT_SEGMENT>" to be generated by AI assistants below---**

Now: Children play and interact with a cat on the porch of a house.
About 14.2s later: The children gather on the porch, enjoying their time together.
About 14.9s later: The children sit together on the porch, petting the dog and chatting.
About 15.6s later: The children sit together on the porch, petting the dog and chatting.
About 18.2s later: The children gather around the dog in the yard.

Now: BO carefully pours the water into a metal dog bowl, and slides it across the grass to Houdini.
About 14.2s later: The Alsatian suddenly jerks his head up towards BO
About 14.9s later: BO jumps back in shock.
About 15.6s later: And MORGAN frowns.
About 18.2s later: BO stands.

<Assistant1-Reasoning>
The ADs generated by ASSISTANT1 are not very original. The descriptions are repetitive, especially the two instances of "The children sit together on the porch, petting the dog and chatting." This repetition does not enrich the watching experience for the visually impaired.

<Assistant2-Reasoning>
ASSISTANT2's ADs are more original and non-repetitive. Each description provides a unique action or event, such as BO pouring water into a dog bowl, the Alsatian jerking its head up, and BO jumping back in shock. These descriptions provide a richer watching experience for the visually impaired.

<Comparison-Reasoning>
Comparing the two AI assistants, ASSISTANT2's ADs are more original and non-repetitive than ASSISTANT1's. ASSISTANT1's descriptions are repetitive and do not provide a rich watching experience for the visually impaired. On the other hand, ASSISTANT2's descriptions are unique and provide a better understanding of the scene.

<Assistant1-Score> 2

<Assistant2-Score> 8

Post-Processing (note: GPT-4 is agnostic to the output source during above marking)
Model Score: 0.25

Figure 9. AD evaluation with `SegEval`. (A) An overview of prompting GPT-4 to evaluate AD generation quality, with (B) one *diversity* and (C) one *originality* examples shown below. Zoom in for details.

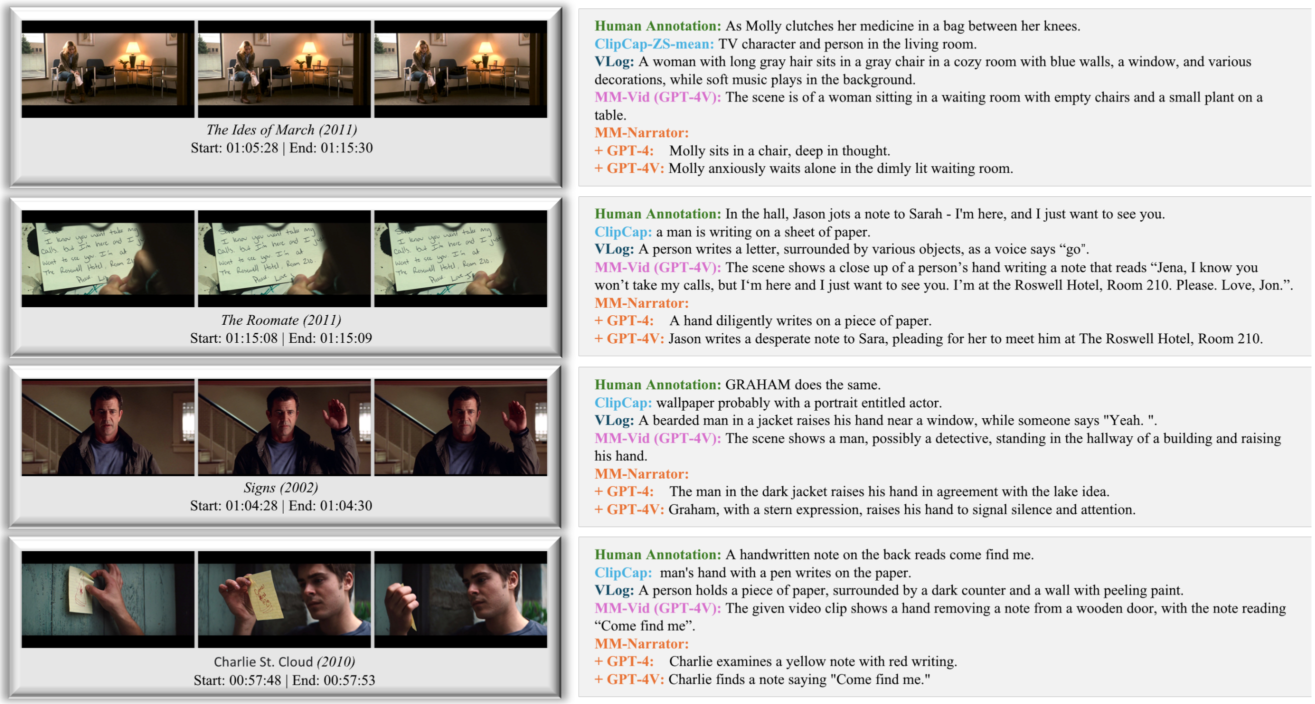

Figure 10. More qualitative comparisons on MAD-eval-`Named` benchmark. For example, in *The ides of March (2011)*, our method generates AD by conditioning on current video clip and the contextual information from timestamp 00:00:00 to 01:05:28. Zoom in for details.

Figure 11. More qualitative demonstrations of MM-Narrator on other long-form videos. For example, in *Inception (2010)*, our method generates AD by conditioning on current video clip and contextual information from timestamp 00:00:00 to 00:31:52. Zoom in for details.