MM-Narrator: Narrating Long-form Videos with Multimodal In-Context Learning

Chaoyi Zhang[♠*] Kevin Lin[♣] Zhengyuan Yang[♣] Jianfeng Wang[♣] Linjie Li[♣]
Chung-Ching Lin[♣] Zicheng Liu[♦] Lijuan Wang[♣]
[♠]University of Sydney [♣]Microsoft [♦]Advanced Micro Devices
https://MM-Narrator.github.io

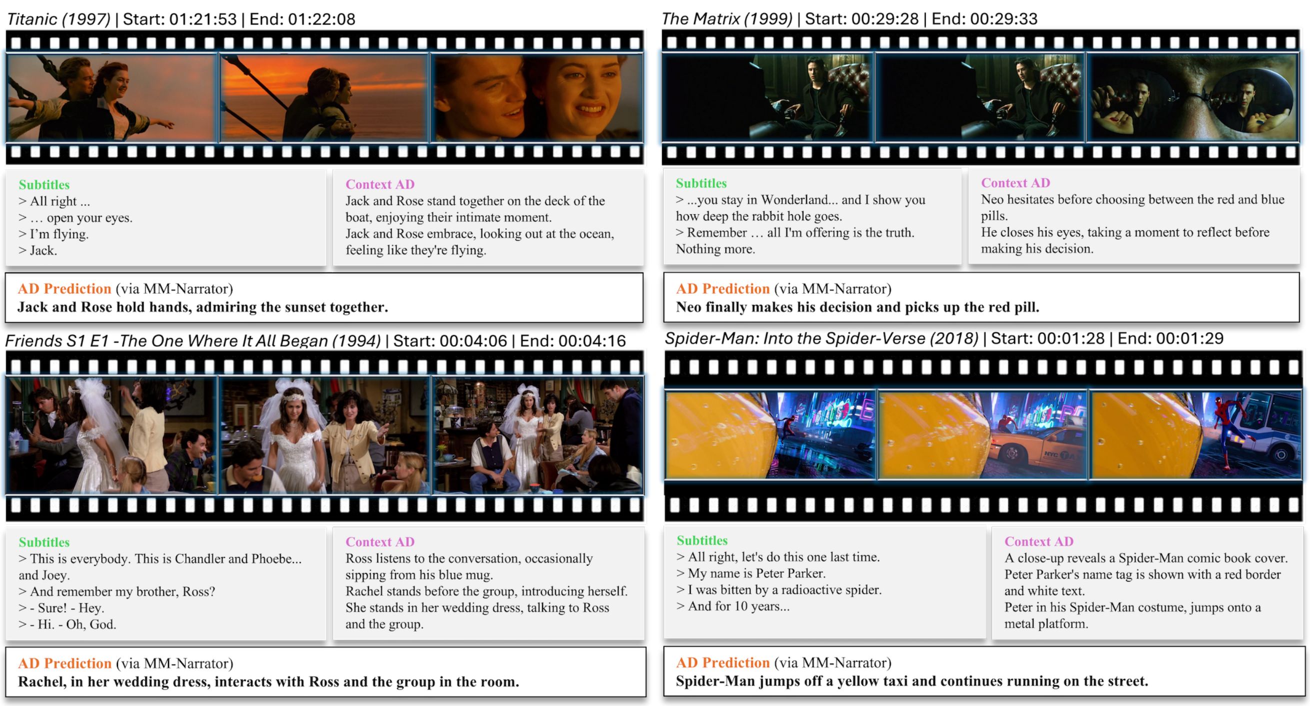

Figure 1. We present MM-Narrator, a training-free framework towards automatic audio description (AD) generation for long-form videos via iterations: for each scene, it perceives multimodal inputs (i.e., seeing visual frames and hearing character dialogues), recalls the context AD depicting past scenes, and infers AD prediction for the current scene. Zoom in for details.

Abstract

We present MM-Narrator, a novel system leveraging GPT-4 with multimodal in-context learning for the generation of audio descriptions (AD). Unlike previous methods that primarily focused on downstream fine-tuning with short video clips, MM-Narrator excels in generating precise audio descriptions for videos of extensive lengths, even beyond hours, in an autoregressive manner. This capability is made possible by the proposed memory-augmented generation process, which effectively utilizes both the short-term textual context and long-term visual memory through an efficient register-and-recall mechanism. These contextual memories compile pertinent past information, including storylines and character identities, ensuring an accurate tracking and depicting of story-coherent and character-centric audio descriptions. Maintaining the training-free design of MM-Narrator, we further propose a complexity-based demonstration selection strategy to largely enhance its multi-step reasoning capability via few-shot multimodal in-context learning (MM-ICL). Experimental results on MAD-eval dataset demonstrate that MM-Narrator consistently outperforms both the existing fine-tuning-based approaches and LLM-based approaches in most scenarios,

[*]Work done during internship at Microsoft.

as measured by standard evaluation metrics. Additionally, we introduce the first segment-based evaluator for recurrent text generation. Empowered by GPT-4, this evaluator comprehensively reasons and marks AD generation performance in various extendable dimensions.

1. Introduction

Audio Description (AD) is an essential task that transforms visual content into spoken narratives [1], primarily assisting visual impairments in accessing video content. Given its evident importance, the notable expectations for AD to fulfill include complementing the existing audio dialogue, enhancing viewer understanding, and avoiding overlap with the original audio. This process involves identifying not just who is present in the scene and what actions are taking place, but also precisely how and when the actions occur. Additionally, AD should capture subtle nuances and visual cues across different scenes, adding layers of complexity to its generation.

In addition to aiding visually impaired audiences, AD also enhances media comprehension for autistic individuals, supports eyes-free activities, facilitates child language development, and mitigates inattentional blindness for sighted users [25, 47]. However, traditional human-annotated AD, while detailed, incurs significant costs and often suffers from inconsistencies due to low inter-annotator agreement [21], highlighting the need for automatic AD generation systems. Furthermore, AD serves as an emerging testbed for benchmarking the capabilities of LLM/LMM systems in long-form multimodal reasoning [21, 22, 30], towards next-level advanced video understanding.

In this paper, we present MM-Narrator, a multimodal AD narrator, to effectively leverage multimodal clues, including visual, textual, and auditory elements, to enable comprehensive perception and reasoning. In particular, MM-Narrator distinguishes itself by naturally identifying characters through their dialogues, in contrast to existing methods that may underutilize subtitles [21, 22].

Apart from an intricate multimodal understanding of the video content, generating story-coherent AD for long-form videos also relies on an accurate tracking and depicting of character-centric evolving storylines over extended durations, even spanning hours. This differs AD generation from conventional dense video captioning [24, 28, 61, 65]: Unlike mere frame-by-frame scene description, AD should weave a coherent narrative, utilizing characters as pivotal elements to maintain an uninterrupted storytelling flow [1]. To achieve contextual understanding, we propose to leverage both short-term and long-term memories to assist MM-Narrator in its recurrent AD generation process. Specifically, short-term textual memory sets the stage for generating coherent narrations, whereas long-term visual memory aids in character re-identification during long-form videos, especially for scenes lacking dialogue.

As a GPT-4 empowered multimodal agent, MM-Narrator could further benefit from multimodal in-context learning (MM-ICL) via our proposed complexity-based multimodal demonstration selection. With complexity defined with the chain-of-thought (CoT) technique [63], MM-Narrator could efficiently form and learn from a smaller candidate pool of multimodal demonstrations, effectively improving its multimodal reasoning capability in a few-shot approach. This proposed complexity-based selection surpasses both random sampling and similarity-based retrieval, which are classic ICL solutions in choosing few-shot examples.

In summary, our contributions are four-folds: *(1)* We present MM-Narrator, an automatic AD narrator for long-form videos that can perceive multimodal inputs, recall past memories, and prompt GPT-4 to produce story-coherent and character-centric AD. *(2)* We propose a complexity-based multimodal in-context learning (MM-ICL) to further boost its AD generation performance with few-shot examples, offering new insights into the question *"what makes good ICL examples?"* under complex text generation scenarios with multimodal reasoning needed. *(3)* Our training-free MM-Narrator outperforms both fine-tuning-based SOTAs and LLM/LMM baselines, including GPT-4V, in most classic captioning metrics. *(4)* Furthermore, we introduce the first GPT-4 based evaluator for recurrent text generation, measuring more comprehensive AD generation qualities at both text-level and sequence-level. Results suggest that MM-Narrator generates AD comparable to human annotations across several considered aspects.

2. Related Work

Audio Description (AD) offers verbal narration of key visual elements in videos [1], enriching the viewing experience for individuals who are blind or have low vision. AD differs from video captioning [7, 24, 28, 31, 61, 65], which solely describes the visual content of a given video clip. Instead, AD generation considers multiple modalities, aiming to generate coherent narratives of storylines, characters, and actions in a way that complements the regular audio track. Initial studies [50, 51, 53, 57] concentrated on developing audio segmentation and transcription system to collect high-quality video datasets with temporally aligned ADs. These foundational efforts pave the way for more advanced explorations in LSMDC [51]. Recent research [22] has ventured into training transformer models equipped with a frozen LLM. Researchers also incorporate an external character bank [21] to enhance the accuracy of AD generation. Different from prior works [21, 22] that rely on downstream fine-tuning, our proposed MM-Narrator generates accurate ADs in a training-free manner.

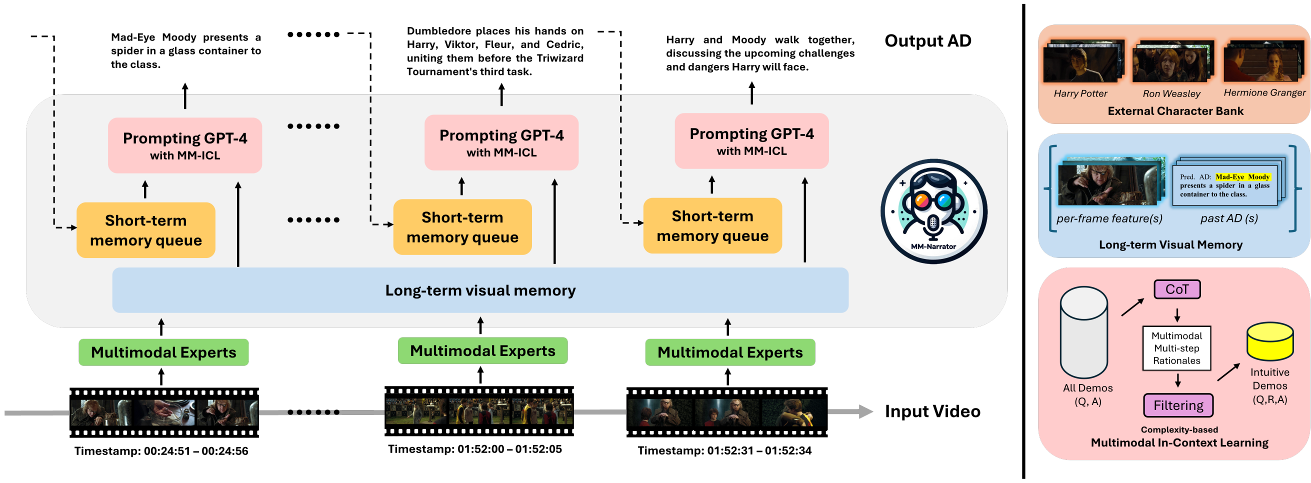

Figure 2. MM-Narrator generates AD sequence for long-form videos via iterations.

LLM for Video Understanding. The remarkable success of Large Language Models (LLMs) [8, 14, 15, 17, 44, 58] has sparked increasing interest in their application to video understanding. Recent works [6, 12, 27, 30, 37, 54] generally fall into two main categories: (*i*) visual instruction tuning, and (*ii*) prompting LLMs. The first approach [27, 32, 37, 38] typically fine-tunes an LLM-based model. This involves integrating the pre-trained LLMs and additional trainable networks. The second category [4] involves prompting LLMs to invoke specialized expert tools, transforming the input video into a textual document, which then serves as input to the LLMs for reasoning [6, 12, 30]. However, this strategy may not be effective for processing lengthy or speech-dense videos, as the LLMs often face challenges with excessive token lengths. Different from prior work, we propose to leverage short-term textual memory and long-term visual memory with a register-and-recall mechanism, to effectively generate ADs for long-form videos.

In-Context Learning (ICL) [13, 34, 36, 39, 40], as a new paradigm, allows LLMs to learn from a few examples without needing parameter updates via downstream fine-tuning. This learning-from-analogy strategy [16] augments original query question with a context formed by natural language demonstrations. Existing studies highlight that the success of ICL largely depends on the selection of effective demonstrations. One common solution [33, 52] is to form the ICL prompt with closest neighbors, which are retrieved with highest similarity to the query embedding. Other query-based metrics are also explored in finding supportive ICL examples on the basis of query content, such as mutual information [55, 56] and perplexity [20]. Although prior works have demonstrated their superiority in text classification tasks or open-domain QA [16], they have not explored ICL on complex text generation tasks under multimodal scenarios. In this work, we propose to quantify the demonstration complexity as the number of reasoning steps in chain-of-thoughts (CoTs) [63, 68], and select the most intuitive examples to improve AD generation with few-shot MM-ICL.

3. Method

Given a long-form video $\mathcal{V}$, consisting of multiple video clips $\{\boldsymbol{v}_t\}$, MM-Narrator generates an AD sequence $\{\mathcal{T}_t\}$ in an autoregressive manner, as shown in Figure 2. We first present MM-Narrator, a multimodal narrator that conducts recurrent AD generation via prompting GPT-4 (§3.1). Building upon MM-Narrator, we propose the complexity-based MM-ICL to further enhance its multimodal reasoning capabilities through intuitive few-shot demonstrations (§3.2). Notably, the entire MM-Narrator framework operates in a training-free manner.

3.1. Recurrent AD Narrator

At each iteration of scanning through a specific long-form video, MM-Narrator utilizes multimodal experts for perception, recalls past memories in both short-term and long-term contexts, and prompts LLM to generate an audio description. We describe each step as below.

Multimodal perceptions. We employ specialized vision and audio expert models to extract multimodal information from the input video clip. These off-the-shelf multimodal models are employed as integral tools within our MM-Narrator framework. We denote a video clip consisting of N frames with timestamp t as $\boldsymbol{v}_t = \{\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_N\}$. We deploy vision experts [48, 60, 64] to gather visual perceptions, which involves obtaining per-frame visual features and text-formed outputs. Specifically, for each frame $\mathcal{I}_i$, we collect CLIP-ViT features $\boldsymbol{x}_i^{CLIP}$, image captions $\boldsymbol{x}_i^{cap}$, and people detections $\boldsymbol{x}_i^{det}$. Alongside these crucial visuals, we observe the spoken dialogues play a profound role, which is underutilized in existing approaches [21, 22]. The spoken dialogues not only offer information complementary to the visuals, but also primarily serve as the only access to identify characters with their names when no external video metadata is given. To be specific, we concatenate the subtitles within a certain time window T_{sub} as $\boldsymbol{x}^{sub}_{t \in T_{sub}}$. These subtitles can be sourced from the Internet or generated through automated speech recognition (ASR) as an audio expert [10]. To summarize, for a given video clip $\boldsymbol{v}_t$, the multimodal experts produce a comprehensive tuple of perception clues $\mathcal{X}_t = (\{\boldsymbol{x}^{CLIP}_i\}, \{\boldsymbol{x}^{cap}_i\}, \{\boldsymbol{x}^{det}_i\}, \boldsymbol{x}^{sub}_{t \in T_{sub}})$, where $\{\boldsymbol{x}_{\dot{i}}\}$ denotes the per-frame outputs. Among these, $\{\boldsymbol{x}^{cap}_i\}$ and $\boldsymbol{x}^{sub}_{t \in T_{sub}}$ are directly used in constructing LLM prompts, while the others facilitate the proposed register-and-recall mechanism for long-term character re-identification.

Short-term memory queue. To equip MM-Narrator with contextual understanding for coherent AD generation, we maintain a short-term memory queue $\mathcal{M}_{\boldsymbol{short}} = \{\mathcal{T}_{t-K}, ..., \mathcal{T}_{t-1}\}$ to contain the K most recently predicted ADs with timestamps. The short-term memory queue will be updated over time during inference. This lightweight textual queue is instrumental in creating story-coherent AD narrations, enabling visually impaired audiences to follow the storytelling more intuitively.

Long-term visual memory. To endow MM-Narrator with the ability to recall characters identified in previous video clips, we construct a frame-level character re-identification visual bank. This visual bank, designed for long-term use, is operated by a register-and-recall mechanism as follows: (1) we register $\boldsymbol{x}^{CLIP}_j$ as the visual signature for each globally-indexed frame $\mathcal{I}_j$ in all previous video clips $\mathcal{I}_j \in \{v_1, v_2, ..., v_{t-1}\}$, and (2) for each current frame $\mathcal{I}_i$, we first filter-out the invalid matches resulting in nonpositive cosine similarity $\texttt{Sim}_{cos}(\boldsymbol{x}^{CLIP}_i, \boldsymbol{x}^{CLIP}_j)$, and then retrieve the past predicted AD which owns the highest similarity to the current visual signature $\boldsymbol{x}^{CLIP}_i$. For simplicity, this mechanism is activated only when a single individual is detected in a frame (i.e., $|\boldsymbol{x}^{det}_{\cdot}| = 1$), typically in close-up shots of the character, making frame-level CLIP-ViT features [48] compatible for character re-identification. Given any AD that covers multiple frames, this frame-level visual retriever supports the MM-Narrator in re-identifying multiple characters appearing in the video clip. Additionally, the retrieval candidate pool is refined to include only past predicted ADs where person named entities are recognized through a Named Entity Recognition (NER) tool [18]. This strategy focuses MM-Narrator on the main characters who contribute to the past storyline.

Prompting LLM for AD generation. Gathering all aforementioned text-formed outputs, MM-Narrator builds prompts to query GPT-4 for recurrent AD generation. Specifically, the input prompt contains the following elements: task introduction, visual captions ($\boldsymbol{x}^{cap}_i$) with successfully re-identified characters, recent context ADs ($\mathcal{M}_{\boldsymbol{short}}$) and character dialogues ($\boldsymbol{x}^{sub}_{t \in T_{sub}}$). Noticeably, we also found that adding task-specific hints into the prompt could empirically benefit overall AD generation, which we attribute as an explicit attention guidance via prompt engineering. A breakdown of our AD generation prompt constructed by MM-Narrator, is provided in the supplementary (Figure 7).

3.2. Multimodal In-Context Learning

In this section, we further extend MM-Narrator with multimodal in-context learning (MM-ICL) on few-shot examples. Our exploration begins by examining two primary methods of demonstration selection: random and similarity-based approaches. We then critically evaluate the question, "*What makes for effective ICL examples?*" and propose a complexity-based MM-ICL approach to improve the multimodal reasoning capability with the most intuitive multimodal demonstrations.

Random MM-ICL. Firstly, we build an in-context learning (ICL) demonstration pool, denoted as $\mathcal{P}$, from the training dataset. Each demonstration within the pool is composed of a pair $(\mathcal{Q}, \mathcal{A})$, where $\mathcal{Q}$ represents the text-formed *question* created using multimodal experts, and $\mathcal{A}$ is the corresponding ground-truth AD, serving as the *answer*. Then, for each test query q, we randomly sample C demonstrations from $\mathcal{P}$ to facilitate the ICL process.

Furthermore, we are further interested in two essential questions: *"What makes good examples for AD generation?"* and *"How to find and use them for ICL?"*

Similarity-based MM-ICL. A common approach, as suggested in existing literature [33], is to identify "good examples" based on similarity, employing a k-NN algorithm to select examples that exhibit the highest k similarity between the embeddings of $\mathcal{Q}_i$ and the test query q. This solution expects to find supportive examples to benefit few-shot performance via a "soft-copy" mechanism [23, 43], which is often used in text classification tasks, such as sentiment analysis, or relatively-simple text generation task such as open-domain QA [16].

However, we empirically find that this similarity-based approach does not manage to enhance the ICL capability for AD generation, regardless of whether the retrieved examples are presented in descending order [33] or ascending order [16]. We hypothesize that for complex text generation tasks such as AD generation, which requires multimodal perception and reasoning, similarity or relevance may not be the most suitable criteria for identifying effective ICL examples for improving overall performance.

Complexity-based MM-ICL. Our empirical analysis reveals that not all questions are equally challenging, in terms of the complexity of multimodal fusion. Take Figure 1 as an example: when comparing *Titanic (1997)* to *Spider Man (2018)*, the latter presents a more complex case. It requires

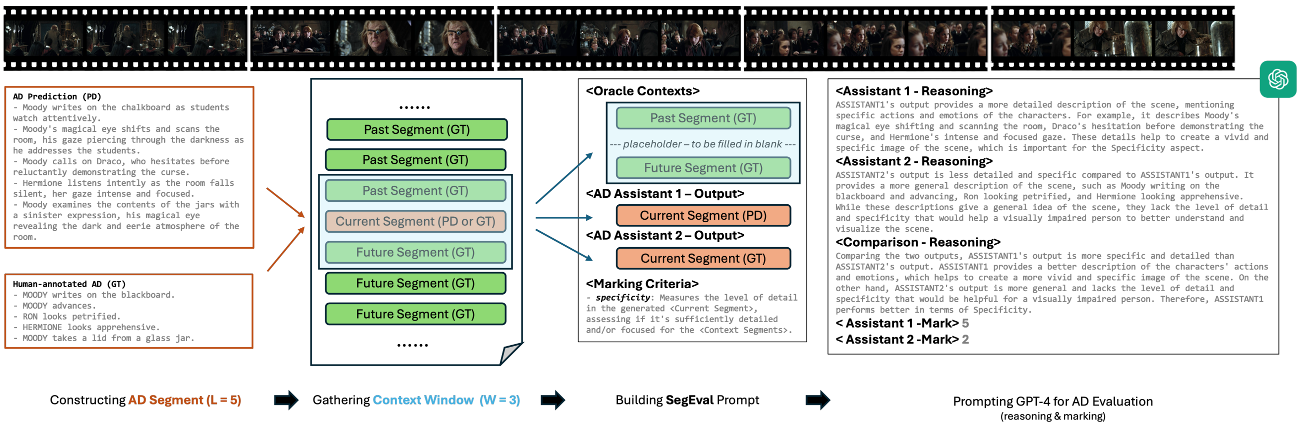

Figure 3. Our proposed `SegEval` evaluator to measure recurrent text generation quality with GPT-4 under customized marking criteria. Noticeably, GPT-4 is agnostic to the source of each assistant output (i.e., which `Seg` is GT or PD), and it would measure `Seg` quality taking oracle contexts into consideration. Take the response shown above as example, its corresponding re-scaled r is 2.25. Zoom in for details.

the inference that "Peter and Spider Man are the same character", a deduction drawn from context AD and subtitles, alongside describing his actions from visual frames, enriched by contextual understanding from the context AD.

This observation led us to hypothesize that complexity could be a more suitable metric for identifying effective ICL examples for tasks involving intricate multimodal fusion. To this end, we propose to query LLM 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 $\mathcal{Q}$. This process evolves our demonstration format from simple $(\mathcal{Q}, \mathcal{A})$ pairs to more comprehensive $(\mathcal{Q}, \mathcal{R}, \mathcal{A})$ tuples.

Instead of the conventional random sampling from the entire pool $\mathcal{P}$, we propose selecting the most straightforward examples, quantified by the shortest number of reasoning steps. These are compiled into a simpler subset pool $\mathcal{P}_{simple}$, from which we conduct our demonstration sampling. This method ensures the inclusion of more intuitive and concise examples in our MM-ICL process. We present detailed ablation study in §5.4, validating that complexity serves as a robust measure for selecting effective ICL examples for improving AD generation.

4. Segment-based GPT-4 Evaluator

The lack of standard AD annotation guidelines, varying cultural background and preferences of human annotators imply that AD is an inherently subjective recurrent text generation process, leading to notable inter-annotator disagreements [21] and challenges in evaluation using traditional reference-based captioning metrics. To this end, inspired by [32, 35], we propose a segment-based GPT-4 evaluator `SegEval` to measure the recurrent AD generation, in terms of multi-domain qualities.

Suppose L ADs form one segment `Seg`. For each `Seg`, the evaluator takes into consideration an oracle context window `Ctx` of length W, to measure its multi-aspect scores. Specifically, we gather W - 1 adjacent segments to form `Ctx`, which consists of $\frac{W-1}{2}$ past and $\frac{W-1}{2}$ future segments surrounding the targeted `Seg`. Given a pair of predicted (PD) and ground-truth (GT) AD segments, `SegEval` would treat them as outputs of two separate AD generation systems, and query GPT-4 to reason and mark their raw marks independently. The final score is calculated as the ratio r of these raw marks between predicted and human-annotated AD, via post-processing. If the re-scaled r is higher than 1.0, it indicates that GPT-4 might favour the predicted AD over human annotations under the specific aspect. Besides, this rescaling operation makes it comparable among different approaches, sharing human annotations as the marking standard. Noticeably, although GPT-4 is unaware of the segment source that which `Seg` is the GT or PD, we always form `Ctx` from GT annotations to set the oracle for investigating contextual influences.

Overall, as shown in Figure 3, `SegEval` can measure context-irrelevant, short-context and long-context scores by flexibly changing the value of W. For example, it could measure *text-level qualities* such as originality and consistency (when $W = 1$), while it could also mark *sequence-level qualities* such as coherence, diversity and specificity (when $W > 1$). The details of each marking criteria are provided in supplementary (§D).

5. Experiments

5.1. Evaluation Setup

Datasets. We conduct experiments on the AD generation benchmark established in AutoAD [22], where MAD-v2-`Named` and MAD-eval-`Named` are released as training and testing splits, respectively. **MAD-v2-`Named`** consists of 334,296 ADs and 628,613 subtitles from 488 movies, while **MAD-eval-`Named`** is compromised of 6,520 ADs and 10,602 subtitles from 10 movies.

Metrics. Following AutoAD [22], we report three traditional captioning metrics to measure the quality of ADs generated versus human-annotated ones, including ROUGE-L [29] (**R-L**), CIDEr [59] (**C**) and SPICE [9] (**S**). Besides, we follow AutoAD-II to benchmark the text sequence generation over their recall-based metric 'Recall@k within Neighbours' (**R@k/N**), where the text similarity is measured by BertScore [67]. We also report Bleu-1 [46] and METEOR [11] for ablation studies. To reduce experimental variability, each experiment of MM-Narrator is repeated *three* times in Tables 1 to 6, as well as Figure 5, with `mean` (and `std`) reported.

5.2. Comparison with State-of-the-Art Approaches

Fine-tuning-based SOTAs. We first compare our training-free framework against the fine-tuning-based SOTAs, including ClipCap [41], ClipDec [42] and AutoAD-I [22]. As shown in Table 1, our training-free approach outperforms its fine-tuning-based counterparts [22, 41, 42], in terms of ROUGE-L, SPICE and R@k/N, especially the AutoAD-I [22] (R-L 12.1 vs 11.9; S 4.5 vs 4.4; R@k/N 48.0 vs 42.1) which is proposed to conduct partial data pretraining over an extra large-scale text-only AV-AD dataset [2, 22] (consisting of 3.3M ADs from over 7k movies) to address the lack of paired training data for AD generation. Unlike [21, 22] who report to struggle with benefiting from character dialogues, our MM-Narrator could better integrate multimodal information and effectively identify characters from appropriate subtitle usage (shown as model D in §5.3).

Training-free LLM/LMM Baselines. We next compare our MM-Narrator with LLM and LMM baselines: (a) VLog [6] and (b) VideoChat-Text [27] are two LLM-based methods for multimodal video understanding. They convert multimodal perceptions into natural languages via several pretrained models [26, 49, 62, 64], and then utilizes a LLM to generate texts based on task-specific prompts. To make a fair comparison, we make them query GPT-4 with the same AD generation prompt as we use in MM-Narrator. (c) MM-Vid [30] is a LMM system which generates AD through incorporating external knowledge with clip-level video description generated by GPT-4V [45, 66].

As shown in Table 2, our MM-Narrator (*w/o MM-ICL*) would outperform VLog and VideoChat, which is mainly attributed to the proposed short-term memory queue and long-term visual memory to effectively leverage relevant contextual information recalled from past ADs. In addition, while MM-Narrator is based on GPT-4 (text-only), it also surpasses the GPT-4V(ision) based MM-Vid system in terms of R-L and SPICE. The results suggest that a memory-augmented LLM can be comparably valuable to the perception-enhanced ones. Furthermore, with our proposed MM-ICL, MM-Narrator outperforms these training-free LLM/LMM counterparts by a large margin. Finally, the bottom two rows of Table 2 further validate the effectiveness of the proposed MM-ICL design.

Method	Training-Free	R-L (↑)	C (↑)	S (↑)	R@5/16 (↑)
ClipCap [41]	✗	8.5	4.4	1.1	36.5
ClipDec [42]	✗	8.2	6.7	1.4	-
AutoAD-I [22]	✗	11.9	**14.3**	4.4	42.1
MM-Narrator	✓	**12.1**	11.6	**4.5**	**48.0**

Table 1. Comparisons with fine-tuning-based state-of-the-art methods on MAD-eval-`Named` benchmark. Note: the random guess will result in a R@5/16 of 31.3%.

Method	LLM/LMM	R-L (↑)	C (↑)	S (↑)	R@5/16 (↑)
VLog [6]	GPT-4	7.5	1.3	2.1	42.3
VideoChat [27]	GPT-4	7.9	2.4	1.8	42.5
MM-Vid [30]	GPT-4V	9.8	6.1	3.8	46.1
MM-Narrator					
w/o MM-ICL	GPT-4	10.3	4.9	3.8	47.1
w/ MM-ICL	GPT-4	**12.1**	**11.6**	**4.5**	**48.0**

Table 2. Comparisons with training-free LLM/LMM baselines on MAD-eval-`Named` benchmark.

Method	Training-Free	R-L (↑)	C (↑)	S (↑)	R@5/16 (↑)
AutoAD-II † [21]	✗	13.4	19.5	-	50.8
MM-Narrator †	✓	13.4	13.9	5.2	49.0

Table 3. Evaluation on MAD-eval-`Named` benchmark, with an external character bank annotated and utilized for improved character recognition (denoted as †).

Utilizing External Character Bank. Previously, all discussed methods share the same and only knowledge source to assist in character recognition. More specifically, they, like us humans, mostly identify characters and infer their names through hearing (i.e., auditory cues) *alone* when watching movies. Given this single source of gaining character information, our MM-Narrator would convey contextual information via retrieving visual and temporal memories. However, these methods suffer from an unavoidable limitation: The character identities would unfortunately remain mystery until their names are being first-time called in dialogues.

To alleviate that, following AutoAD-II [21] we also investigate how our method could benefit from incorporating an external character bank. To construct this character bank, [21] exploits actor portrait images (from an external movie database) to retrieve a few most similar frames for each main character in each movie. Unlike [21] who trains an auxiliary character recognizer from these retrieved frames, we maintain our training-free designs by simply concatenating these frames into short video clips to introduce each character (with ADs as their names). Next, we prepend these video clips to the long-form videos, such that they could work compatibly with our register-and-recall mechanism. As shown in Table 3, our MM-Narrator (*w/ ExtCharBank*) could further boost its performance and generate outcomes comparable to the fine-tuning-based AutoAD-II.

Qualitative Results. Qualitative comparisons over MAD-

Figure 4. Qualitative comparisons between ClipCap, MM-Vid, AutoAD-II, and our MM-Narrator, where the latter two approaches are equipped with the external character bank. The movies are from (a) *Signs (2002)*, (b) *Ides of March (2011)*, (c) *Charlie St. Cloud (2010)*, and (d) *Les Misérables (2012)*. Zoom in for details.

eval dataset are shown as Figure 4, while the qualitative demonstrations of applying our MM-Narrator on other long-form videos (external to the MAD-eval dataset) are shown in Figure 1. Additional qualitative results are included in supplementary (Figure 10 and 11).

5.3. Building MM-Narrator From Image Captioner

As shown in Figure 5, we quantitatively demonstrate how our training-free MM-Narrator are developed step by step. Starting from (A) an image captioner, we elaborate how multimodal perception benefits MM-Narrator to form an intricate multimodal understanding over video content. Specifically, it includes adding (B) multiple frames, (C) subtitles, and (D) a task-specific hint[1]. Noticeably, simply adding the dialogues (C) might not result in an immediate performance gain. However, with prompt engineering in (D), MM-Narrator pays more attention to effectively leverage multimodal clues for character-centric AD generation.

Next, we illustrate how we transform MM-Narrator into recurrent AD narrator to produce story-coherent AD, with incorporation of past memories and complexity-based MM-ICL. Specifically, MM-Narrator maintains (E) a short-term memory queue, learns from (F) multimodal demonstrations via MM-ICL, and retrieves (G) long-term visual memory for character re-identification, which could be further boosted with (H) an external character bank.

5.4. Ablations on Multimodal In-Context Learning

We investigated three groups of MM-ICL proposed to augment the baseline. Specifically, we built random R1 and similarity-based S1, by adapting classic ICL techniques [13, 33] from conventional NLP tasks into multimodal AD generation. Next, we presented our complexity-based design as C2.

[1] *"Hint: try to infer character names from subtitles for AD generation."*

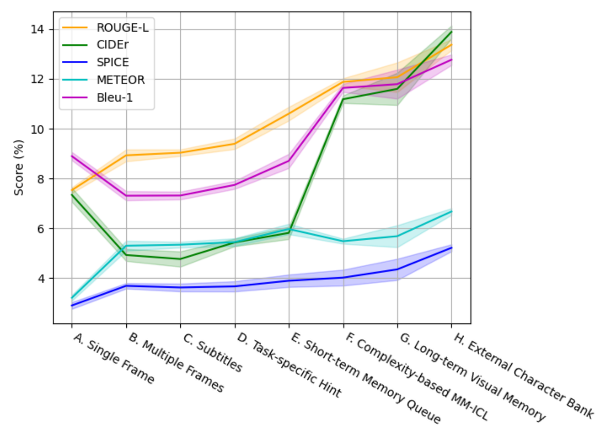

Figure 5. Ablations on each component for MM-Narrator.

Model	Pool Size	Demo. Format	CoT	R-L (↑)	C (↑)	B-1 (↑)
Baseline w/o MM-ICL						
B1	-	-	✗	$11.8_{\pm0.1}$	$8.6_{\pm0.1}$	$9.7_{\pm0.2}$
Random MM-ICL						
R1	100%	$(\mathcal{Q}, \mathcal{A})$	✗	$13.2_{\pm0.1}$	$12.9_{\pm0.2}$	$12.2_{\pm0.1}$
R2	100%	$(\mathcal{Q}, \mathcal{R}, \mathcal{A})$	✓	$13.4_{\pm0.1}$	$13.4_{\pm0.2}$	$12.7_{\pm0.1}$
R3	10% *random*	$(\mathcal{Q}, \mathcal{A})$	✗	$13.3_{\pm0.1}$	$13.0_{\pm0.1}$	$12.3_{\pm0.0}$
R4	10% *random*	$(\mathcal{Q}, \mathcal{R}, \mathcal{A})$	✓	$13.3_{\pm0.1}$	$13.4_{\pm0.1}$	$12.6_{\pm0.0}$
Similarity-based MM-ICL						
S1	100%	$(\mathcal{Q}, \mathcal{A})$	✗	$13.5_{\pm0.0}$	$13.1_{\pm0.0}$	$12.6_{\pm0.1}$
Complexity-based MM-ICL						
C1	10% *shortest*	$(\mathcal{Q}, \mathcal{A})$	✗	$13.2_{\pm0.1}$	$13.3_{\pm0.3}$	$12.3_{\pm0.1}$
C2	10% *shortest*	$(\mathcal{Q}, \mathcal{R}, \mathcal{A})$	✓	$13.4_{\pm0.0}$	$13.9_{\pm0.1}$	$12.8_{\pm0.0}$
C3	10% *longest*	$(\mathcal{Q}, \mathcal{R}, \mathcal{A})$	✓	$13.3_{\pm0.1}$	$12.7_{\pm0.2}$	$12.4_{\pm0.1}$

Table 4. Our different MM-ICL designs for MM-Narrator †. The baseline (B1) and
each representative MM-ICL implementation (R1, S1 and C2) are highlighted.

Method	LLM/LMM	Text-level Quality		Sequence-level Quality					
		Context-irrelevant Scores		Short-context Scores			Long-context Scores		
		`Orig.`	`Cons.`	`Cohe.`	`Dive.`	`Spec.`	`Cohe.`	`Dive.`	`Spec.`
		±0.02	±0.02	±0.01	±0.06	±0.04	±0.01	±0.01	±0.03
GT	-	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ClipCap [41]	GPT-2	0.43	0.42	0.26	0.35	0.35	0.26	0.42	0.33
VLog [6]	GPT-4	1.03	0.88	0.34	0.55	0.52	0.32	0.57	0.43
MM-Vid [30]	GPT-4V	0.85	0.78	0.51	0.81	0.66	0.53	0.84	0.62
MM-Narrator	GPT-4	$1.05_{\pm0.10}$	$1.03_{\pm0.05}$	$0.52_{\pm0.06}$	$0.70_{\pm0.06}$	$0.66_{\pm0.04}$	$0.57_{\pm0.05}$	$0.70_{\pm0.02}$	$0.61_{\pm0.05}$
MM-Narrator	GPT-4V	$\mathbf{1.49}_{\pm0.10}$	$1.45_{\pm0.05}$	$0.94_{\pm0.07}$	$1.01_{\pm0.04}$	$1.13_{\pm0.08}$	$0.87_{\pm0.04}$	$1.05_{\pm0.04}$	$\mathbf{1.14}_{\pm0.05}$
MM-Narrator †	GPT-4	$0.95_{\pm0.02}$	$1.06_{\pm0.01}$	$0.62_{\pm0.04}$	$0.75_{\pm0.01}$	$0.76_{\pm0.01}$	$0.62_{\pm0.04}$	$0.80_{\pm0.03}$	$0.71_{\pm0.03}$
MM-Narrator †	GPT-4V	$1.45_{\pm0.14}$	$\mathbf{1.46}_{\pm0.04}$	$\mathbf{0.98}_{\pm0.03}$	$\mathbf{1.06}_{\pm0.04}$	$\mathbf{1.24}_{\pm0.09}$	$\mathbf{0.94}_{\pm0.02}$	$\mathbf{1.09}_{\pm0.05}$	$1.12_{\pm0.03}$

Table 5. Evaluating AD generation with `SegEval` on MAD-eval-`Named` benchmark, with segment size L set to 5. The context window sizes W are set as 1 / 3 / 11 to compute context-irrelevant / short-context / long-context scores, respectively. `Orig.`, `Cons.`, `Cohe.`, `Dive.`, and `Spec.` stand for *originality*, *consistency*, *coherence*, *diversity*, and *specificity*, respectively. The scoring variances of these GPT-4 evaluators are denoted below for references, which are estimated by three repeated evaluations over the same inference outputs. These re-scaled scores measure the corresponding AD prediction (PD) qualities of each specific method, compared to the shared marking standards set by ground-truth (GT) ADs. For example, given a pair of PD and GT segments, without revealing to the evaluator which segment is GT or PD, if it reasons and marks the raw qualities (`R.Q.`) as 8 and 5 for PD and GT segments, respectively, we derive the re-scaled score r as $\frac{\texttt{R.Q.}_{\texttt{PD}}}{\texttt{R.Q.}_{\texttt{GT}}} = \frac{8}{5} = 1.6$. † indicates our incorporation with ExtCharBank.

Method	R-L (↑)	C (↑)	M (↑)	B-1 (↑)
MM-Narrator				
+ GPT-4	$12.1_{\pm0.4}$	$11.6_{\pm0.4}$	$5.7_{\pm0.2}$	$11.8_{\pm0.3}$
+ GPT-4V	$11.8_{\pm0.1}$	$7.0_{\pm0.2}$	$6.5_{\pm0.1}$	$9.3_{\pm0.1}$
MM-Narrator †				
+ GPT-4	$13.4_{\pm0.0}$	$13.9_{\pm0.1}$	$6.7_{\pm0.0}$	$12.8_{\pm0.0}$
+ GPT-4V	$12.8_{\pm0.0}$	$9.8_{\pm0.2}$	$7.1_{\pm0.0}$	$10.9_{\pm0.0}$

Table 6. Comparisons over classic reference-based captioning scores, when incorporating our MM-Narrator with GPT-4V.

The results as shown in Table 4, verify our hypothesis that complexity serves as an appropriate measure for selecting effective ICL demonstrations for improving AD generation. It also indicates that our proposed complexity-based design (C2) is more preferable than classic ones (R1, S1) for AD generation, especially the CIDEr score. In supplementary (§C), we further discuss three sub-questions to elaborate an in-depth analysis, including 1) *Does CoT help?* 2) *Are more intuitive examples helpful for AD Generation?* and 3) *Does complexity-based MM-ICL work effectively?*

5.5. Evaluating AD Generation with GPT-4

In Table 6, we observe a few performance drop on classic reference-based captioning scores when incorporating MM-Narrator with GPT-4V [45]. As shown in Figure 4, the decrease in performance can be primarily attributed to the more detailed and much richer ADs generated by our method, which diverge from the typically shorter human-annotated ADs in MAD-eval-`Named`. This suggests that taking human annotated AD as oracles to measure AD-level captioning scores might be unsuitable for advanced LMM approaches, which further motivates our proposal of evaluating recurrent text generation with GPT-4.

Adjusting W, our proposed `SegEval` could flexibly measure both *text-level* and *sequence-level* qualities. As shown in Table 5, the performance ranking order observed in `SegEval` aligns with our other experimental results, validating the reliability of `SegEval` as an evaluation tool, except for GPT-4V based MM-Vid where ours falls short on *diversity*. Furthermore, when employing GPT-4V as our vision expert, MM-Narrator not only outperforms others by a large margin, but also closely mirrors the quality of human annotated ADs in multiple aspects, gaining more favor from the source-agnostic GPT-4 evaluator.

Compared to classic reference-based captioning scores, `SegEval` could better reflect the recurrent text generation qualities with GPT-4. One human validation on `SegEval` is shown in Figure 3, and more examples can be found in supplementary (Figure 9). Moreover, `SegEval` could be easily extended to support more comprehensive evaluation perspectives by querying it with extra customized marking criteria.

6. Conclusion

MM-Narrator represents a significant leap in automatic audio description (AD) generation for long-form videos, leveraging the power of GPT-4 and innovative multimodal in-context learning (MM-ICL). This recurrent AD narrator excels in generating story-coherent and character-centric AD by combining immediate textual context with long-term visual memory. Its training-free design, coupled with our proposed complexity-based MM-ICL demonstration selection strategy, outperforms both existing fine-tuning-based and LLM-based approaches in most scenarios, as measured by traditional captioning metrics. Furthermore, we introduce a GPT-4 empowered evaluator for a more comprehensive measurement of recurrent text generation qualities. Its results suggest that MM-Narrator generates AD comparable to human annotations across several considered aspects.

References

[1] American council of the blind. https://adp.acb.org/. 2

[2] Audiovault. https://audiovault.net/. 6

[3] Google text-to-speech: a python library and cli tool to interface with google translate's text-to-speech api. https://gtts.readthedocs.io/en/latest/. 13

[4] Langchain. https://langchain.readthedocs.io/. 3

[5] Pyscenedetect: Video scene cut detection and analysis tool. https://www.scenedetect.com/. 13

[6] VLog: Video as a Long document. https://https://github.com/showlab/VLog, 2023. GitHub repository. 3, 6, 8

[7] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys*, 52(6):1–37, 2019. 2

[8] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 3

[9] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *ECCV*, 2016. 6

[10] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, 2023. 4, 13

[11] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop*, 2005. 6

[12] Aanisha Bhattacharya, Yaman K Singla, Balaji Krishnamurthy, Rajiv Ratn Shah, and Changyou Chen. A video is worth 4096 tokens: Verbalize story videos to understand them in zero shot. *EMNLP*, 2023. 3

[13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. 3, 7, 13

[14] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90% ChatGPT quality, March 2023. 3

[15] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 3

[16] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 3, 4

[17] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. PaLM-E: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 3

[18] Jenny Rose Finkel, Trond Grenager, and Christopher D Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, pages 363–370, 2005. 4

[19] Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*, 2022. 12

[20] Hila Gonen, Srini Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. Demystifying prompts in language models via perplexity estimation. *arXiv preprint arXiv:2212.04037*, 2022. 3

[21] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. AutoAD II: The Sequel - who, when, and what in movie audio description. In *ICCV*, 2023. 2, 3, 5, 6

[22] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. AutoAD: Movie description in context. In *CVPR*, 2023. 2, 3, 5, 6, 13

[23] Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. Understanding in-context learning via supportive pretraining data. *arXiv preprint arXiv:2306.15091*, 2023. 4

[24] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017. 2

[25] Elisa Lewis. Deep dive: How audio description benefits everyone, 2021. Accessed on: 2023-11-13. 2

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 6

[27] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-Centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 3, 6

[28] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, et al. VALUE: A multi-task benchmark for video-and-language understanding evaluation. In *NeurIPS Datasets and Benchmarks Track*, 2021. 2

[29] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. *ACL*, 2004. 6

[30] Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, et al. MM-Vid: Advancing video understanding with GPT-4V (ision). *arXiv preprint arXiv:2310.19773*, 2023. 2, 3, 6, 8

[31] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. SwinBert: End-to-end transformers with sparse attention for video captioning. In *CVPR*, 2022. 2

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023. 3, 5

[33] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? *arXiv preprint arXiv:2101.06804*, 2021. 3, 4, 7, 13

[34] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023. 3

[35] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. GPTEval: NLG evaluation using GPT-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023. 5

[36] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 35:2507–2521, 2022. 3

[37] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023. 3

[38] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 3

[39] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. MetaICL: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021. 3

[40] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022. 3

[41] Ron Mokady, Amir Hertz, and Amit H Bermano. ClipCap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 6, 8, 13

[42] David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected clip. *EMNLP Findings*, 2022. 6

[43] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022. 4

[44] OpenAI. GPT-4 technical report. 2023. 3, 14

[45] OpenAI. GPT-4V(ision) system card. 2023. 6, 8, 14

[46] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 6

[47] Elisa Perego. Gains and losses of watching audio described films for sighted viewers. *Target*, 28(3):424–444, 2016. 2

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 3, 4, 13

[49] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, 2023. 6

[50] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *CVPR*, pages 3202–3212, 2015. 2

[51] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *Int. J. Comput. Vis.*, 123:94–120, 2017. 2

[52] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*, 2021. 3

[53] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. MAD: A scalable dataset for language grounding in videos from movie audio descriptions. In *CVPR*, pages 5026–5035, 2022. 2

[54] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. MovieChat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2023. 3

[55] Taylor Sorensen, Joshua Robinson, Christopher Michael Rytting, Alexander Glenn Shaw, Kyle Jeffrey Rogers, Alexia Pauline Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. An information-theoretic approach to prompt engineering without ground truth labels. *arXiv preprint arXiv:2203.11364*, 2022. 3

[56] Eshaan Tanwar, Manish Borthakur, Subhabrata Dutta, and Tanmoy Chakraborty. Multilingual llms are better cross-lingual in-context learners with alignment. *arXiv preprint arXiv:2305.05940*, 2023. 3

[57] Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*, 2015. 2

[58] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3

[59] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015. 6

[60] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *CVPR*, pages 7464–7475, 2023. 3

[61] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, October 2019. 2

[62] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali

Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. 2022. 6

[63] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837, 2022. 2, 3

[64] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. GRiT: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022. 3, 6

[65] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016. 2

[66] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The Dawn of LMMs: Preliminary explorations with GPT-4V (ision). *arXiv preprint arXiv:2309.17421*, 2023. 6

[67] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 6

[68] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022. 3