DistinctAD: Distinctive Audio Description Generation in Contexts

Bo Fang[1], Wenhao Wu[2,3], Qiangqiang Wu[1], Yuxin Song[2], Antoni B. Chan[1✉]

[1] Department of Computer Science, City University of Hong Kong
[2] Baidu Inc.   [3] The University of Sydney

{bofang6-c,qiangqwu2-c}@my.cityu.edu.hk, wenhao.wu@sydney.edu.au

songyuxin02@baidu.com, abchan@cityu.edu.hk

Abstract

*Audio Descriptions (ADs) aim to provide a narration of a movie in text form, describing non-dialogue-related narratives, such as characters, actions, or scene establishment. Automatic generation of ADs remains challenging due to: i) the domain gap between movie-AD data and existing data used to train vision-language models, and ii) the issue of contextual redundancy arising from highly similar neighboring visual clips in a long movie. In this work, we propose **DistinctAD**, a novel two-stage framework for generating ADs that emphasize distinctiveness to produce better narratives. To address the domain gap, we introduce a CLIP-AD adaptation strategy that does not require additional AD corpora, enabling more effective alignment between movie and AD modalities at both global and fine-grained levels. In Stage-II, DistinctAD incorporates two key innovations: (i) a Contextual Expectation-Maximization Attention (EMA) module that reduces redundancy by extracting common bases from consecutive video clips, and (ii) an explicit distinctive word prediction loss that filters out repeated words in the context, ensuring the prediction of unique terms specific to the current AD. Comprehensive evaluations on MAD-Eval, CMD-AD, and TV-AD benchmarks demonstrate the superiority of DistinctAD, with the model consistently outperforming baselines, particularly in Recall@k/N, highlighting its effectiveness in producing high-quality, distinctive ADs.*

1. Introduction

Audio description (AD) [20, 67] is a crucial accessibility service that provides verbal narration of visual elements in media content for individuals who are blind or have low vision. By offering succinct and vivid descriptions, ADs enable visually impaired audiences to fully comprehend and engage with non-dialogue-related narratives, *e.g.*, characters, facial expressions, non-verbal actions, or scene establishment. Recent studies also show ADs' value for
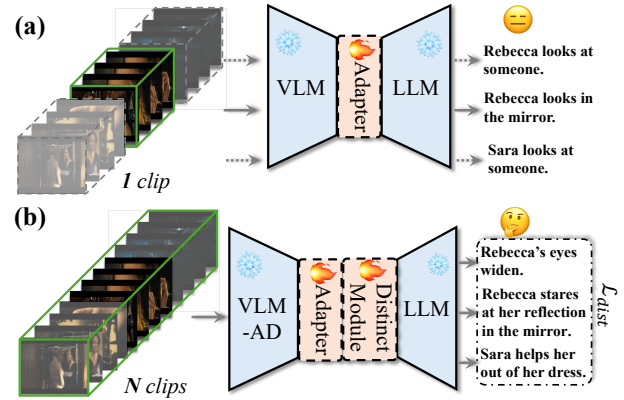

Figure 1. **(a)** Previous methods approach the AD task similar to video captioning, using only a single video clip as input, which leads to repetitive ADs due to highly similar neighboring clips. **(b)** Our DistinctAD method generates distinctive ADs across N consecutive clips, with three key innovations: VLM-AD adaptation, the Distinct Module, and explicit distinctive words prediction.

sighted viewers in supporting eye-free activities and facilitating child language development [31, 55], reinforcing its pivotal role in fostering inclusivity by bridging the perceptual gap between visual and non-visual elements. Crafting ADs requires careful attention to timing, language, and context to integrate smoothly with dialogue [22]. However, despite the availability of advanced AD platforms [9, 54], human-annotated methods are costly and difficult to scale, highlighting the need for automated generation systems, especially with the rise of user-generated content.

Advancements in Vision-Language Models (VLMs) and Large-Language Models (LLMs) have led to growing interest in automatic AD generation for media. Current approaches fall into two categories: (i) using powerful proprietary models like GPT-4 [3] or GPT-4V [53] in a training-free manner [14, 39, 82, 86], and (ii) fine-tuning open-source VLM components, such as visual-text adapters [5, 34, 42], for AD tasks [21–23, 40, 56, 74]. Both approaches have limitations: (i) Training-free methods often perform poorly and suffer from hallucinations due to the unique na-

ture of AD (*e.g.*, character names and narrative coherence), which differs from the common text data LLMs are trained on. (ii) Fine-tuning methods generally perform better but are still limited by insufficient data to fully adapt to the movie-AD domain and face the *context-repetition* issue.

Unlike video captioning [38, 62], ADs are generated on consecutive intervals (visual clips) throughout long videos [68], *e.g.*, movies. The *context-repetition* issue arises when models produce repetitive or similar descriptions for consecutive visual clips, especially when using prior ADs as prompts [21, 74]. This occurs because sequential clips often comprise redundant scenes or characters (and therein redundant visual features), leading models that only use the current visual clip to repeat the same information from the past, as shown in Fig. 1. However, audiences are more interested in the unique and distinct events of the current clip, rather than the common elements from the previous one.

In this paper, we propose **DistinctAD**, a two-stage framework for generating distinctive ADs within contexts. Given the domain gap between the movie-AD and VLM training data, we first bridge this gap in Stage-I by adapting VLMs, such as CLIP [58], to the movie-AD domain. Our adaptation strategy is inspired by a key observation (see Appendix §A): AD sentences encoded by the CLIP text encoder can be effectively reconstructed using simple LLMs like GPT-2 [57] with minimal fine-tuning, whereas AD reconstructions using CLIP visual features from the corresponding clips are often of poor quality. *This suggests that while CLIP's multi-modal embedding space is rich enough to represent AD information, its visual encoder is insufficient for extracting it.* To mitigate this domain gap, we adapt the CLIP vision encoder to better align with the frozen CLIP text encoder using existing paired video-AD data. The alignment involves global matching at video-sentence level, similar to CLIP pre-training. A challenge arises because video clips are labeled with whole ADs, and words may not appear in every frame but must be aggregated over frames. Therefore, we propose fine-grained matching at frame-word level for this *multiple-instance setting*.

For Stage-II, we propose a novel distinctive AD narrating pipeline based on the Expectation-Maximization Attention (EMA) [17] algorithm, which has demonstrated its efficacy in tasks such as semantic segmentation [35], video object segmentation [41], and text-video retrieval [27]. Differently, we apply EMA to contextual clips from long videos, which often exhibit high redundancy due to recurring scenes or characters. By extracting common bases from contextual information, DistinctAD reduces redundancy and generates compact, discriminative representations that enable the LLM decoder to produce more distinctive ADs. To further emphasize distinctiveness explicitly, we introduce a distinctive word prediction loss that filters out words that repeatedly appear in contexts, ensuring that the LLM decoder

focuses on predicting unique words specific to the current AD. With these two designs, DistinctAD produces contextually distinctive and engaging ADs that can provide better narratives for the audience.

In summary, our contributions are three-fold:
- We propose a CLIP-AD adaptation strategy tailored to movie-AD data, addressing the misalignment issue caused by the domain gap. Our adapted vision encoder can be seamlessly integrated into existing CLIP-based AD methods and stands to benefit from future improvements as more AD data becomes available.
- We introduce DistinctAD, which incorporates a Contextual EMA module and a distinctive word prediction loss, significantly enhancing the generation of distinctive ADs from consecutive visual clips with similar contexts.
- Comprehensive evaluations on MAD-Eval [21], CMD-AD [23], and TV-AD [82] highlight DistinctAD's superiority. Our outstanding performance in Recall@k/N demonstrates its effectiveness in generating high-quality ADs with both distinctiveness and technical excellence.

2. Related Work

Dense video captioning. A task closely related to AD is *dense* video captioning [29], which extends traditional video captioning [38, 45, 62, 63] by both generating a single caption for trimmed video segments as well as detecting and describing multiple events with grounded timestamps. Initial dense video captioning utilize a 2-stage pipeline [25, 26, 75, 79] by firstly performing localization and then describing events. Recent works [13, 18, 36, 51, 59, 64, 65, 75, 80, 83, 89] focus on training localization and captioning modules in an end-to-end manner to enhance inter-event associations. In contrast to these works, AD generation specifically aims to narrate a coherent story, maintain character-awareness, and complement the audio track without interfering with existing dialogue.

AD generation. Early AD systems relied heavily on specialized authoring tools [9] and skilled human contributors. Platforms like Rescribe [54] and LiveDescribe [9] have facilitated faster and more accurate AD creation; however, these methods are costly and do not scale efficiently for large volumes of visual content. Recent efforts have developed audio segmentation and transcription systems [8, 10, 11] to create high-quality video datasets with temporally aligned ADs [60, 61, 68, 69], advancing automatic AD research.

In general, current automatic AD generation systems can be categorized into two approaches: training-free and partial-fine-tuning. *Training-free* methods [39] generate ADs by leveraging proprietary models like GPT-4 [3] and GPT-4V [53]. MM-Narrator [86] enhances AD performance by multi-model in-context learning with memories. LLM-AD [14] and AutoAD-Zero [82] use prompts com-

prising visual frames with textual character names and colorful circles [66], enabling character-centric AD generation. However, training-free AD methods often suffer from high evaluation costs at scale and relatively poor performance due to domain-specific challenges and LLM hallucinations. *Partial-fine-tuning* methods [21–23, 40, 74], as well as our DistinctAD, only fine-tune a lightweight adapter [5, 34] between the pre-trained vision and text encoders. A representative example is the AutoAD series [21–23], which builds automatic AD systems and enriches them with character-aware prompts within different vision-language frameworks. However, previous studies tend to focus on constructing more accurate external character banks, whereas treating AD generation similarly to video captioning, overlooks AD's unique sequential structure of video clips. In contrast, our method emphasizes understanding the visual content within its temporal context, leading to more distinctive AD generation.

Distinctive captioning in images aims to articulate unique details that can help distinguishing targets from others. An intuitive way to promoting distinctiveness is through contrastive learning [16, 43, 47, 73], where generated captions are encouraged to align more closely with target images rather than distractors. In [12, 76–78], group-based distinctive attention is introduced to capture distinctiveness by comparing sets of similar images and re-weighting specific caption words. A recent closely related field is difference captioning [32, 33, 84], which aims to describe differences between a single pair of images. VisDiff [19] scales difference captioning to sets containing thousands of images with natural language. Our work differs from these distinctive captioning works in that we are the first to explore distinctiveness across dense, consecutive clips within hours-long movies, thereby generating ADs with better narrative.

3. Method

3.1. Stage-I: CLIP-AD Adaptation

AD and the visual content it describes exhibit a significant domain gap compared to typical large-scale web data. This gap often causes misalignment in current partial-fine-tuning techniques. Previous studies [21, 22, 74] try alleviating this problem by pre-training LLMs on text-only AD corpus, *e.g.* AudioVault [1]. However, misalignment persists at the initial stage of vision encoding, which is often neglected.

Inspired by our findings that AD sentences encoded by the CLIP text encoder can be effectively recovered using the GPT-2 language model (see §1 and Appendix §A), we identify that the primary issue of misalignment is caused by the CLIP vision encoder, *i.e.*, the discrepancy between visual embeddings and AD embeddings within the joint CLIP feature space. To address this, we propose adapting the CLIP vision encoder to the specific AD domain. However, due to
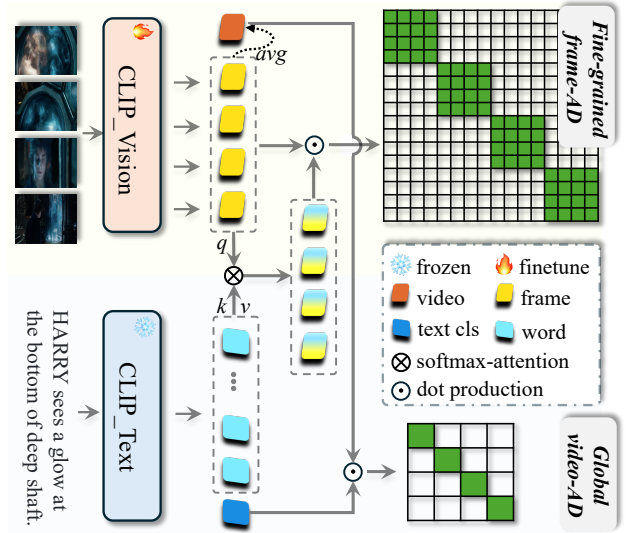

Figure 2. Illustration of Stage-I: CLIP-AD Adaptation. This process involves adapting the CLIP vision encoder to specific movie-AD data through global-level video-AD matching (bottom right) and fine-grained frame-AD matching (top right).

the unique multiple-instance learning setting for video-AD pairs (see §1), we consider both global matching and fine-grained frame-word matching in our adaptation method.

Global video-AD matching. A straightforward strategy involves adopting classical CLIP-style fine-tuning with video-AD pairs in large batches [46]. Formally, let video clip $\mathbf{V}_i = [\mathbf{f}_i^1; \cdots; \mathbf{f}_i^n] \in \mathbb{R}^{n \times C}$ be a collection of n frame embeddings, and corresponding AD $\mathbf{T}_i = [\mathbf{w}_i^0; \mathbf{w}_i^1; \cdots; \mathbf{w}_i^m] \in \mathbb{R}^{(m+1) \times C}$ be a collection of m word embeddings ($\mathbf{w}_i^j$) and the [CLS] token (denoted as $\mathbf{w}_i^0$), where C is the number of channels in the embedding space. We obtain the *global* video-level representation by averaging all frame embeddings in $\mathbf{V}_i$ using mean pooling: $\mathbf{v}_i = \frac{1}{n} \sum_{j=1}^n \mathbf{f}_i^j$. Following the standard CLIP, we use the [CLS] token as the *global* textual AD representation $\mathbf{t}_i = \mathbf{w}_i^0$. The **global video→AD matching** is performed by maximizing the sum of the main diagonal of a $B \times B$ similarity matrix, using the contrastive loss:

$$\mathcal{L}_{v \to AD} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(sim(\mathbf{v}_i, \mathbf{t}_i))}{\sum_{j=1}^B \exp(sim(\mathbf{v}_i, \mathbf{t}_j))}, \quad (1)$$

where B is the batch size, and the similarity function $sim(\cdot)$ is the vector inner product. This process is illustrated in the bottom right of Fig. 2. Similarly, we drive the AD→video loss $\mathcal{L}_{AD \to v}$ by maximizing the sum of the secondary diagonal (i.e., swapping the i and j indices in (1)). The final global-level contrastive loss is then the sum of the losses in both directions $\mathcal{L}_g = \mathcal{L}_{v \to AD} + \mathcal{L}_{AD \to v}$.

Fine-grained frame-AD matching. Matching global video to AD sentence [CLS] (and vice versa) aids in joint feature space alignment. However, this alignment is insufficient for effective adaptation due to the specific *multiple-instance*
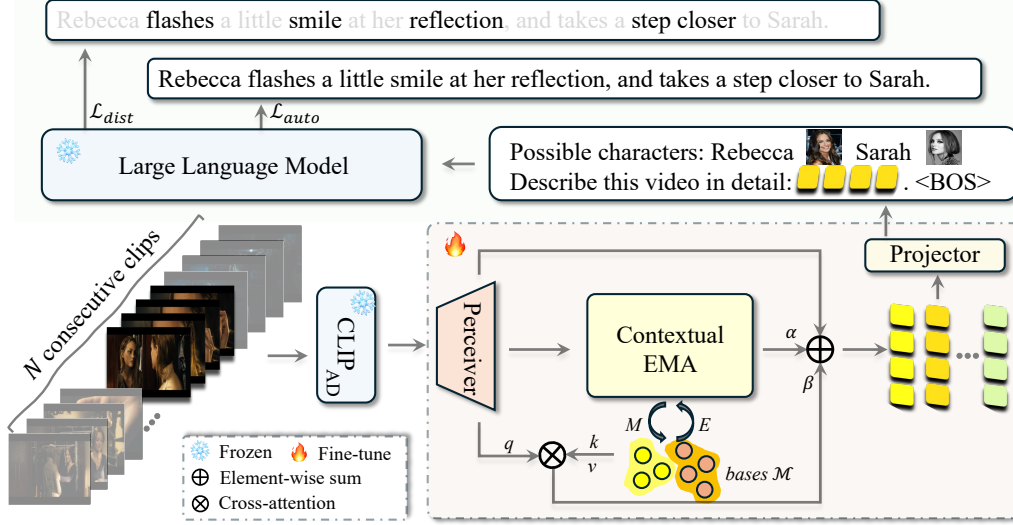
Figure 3. Pipeline of Stage-II: Distinctive AD Narration. Stage-II processes N consecutive video clips using the CLIP_{AD} vision encoder from Stage-I. We generate contextual-distinctive ADs by two key innovations: *i)* a Contextual EMA module to learn compact and discriminative visual representations for improved prompting of LLMs; *ii)* an extra distinctive word loss for predicting AD-specific terms.

setting of ADs, where only some words may correspond to a particular frame, but all words will have correspondence in aggregate. Thus, we propose a fine-grained matching loss at the frame-level to address this issue.

Formally, given the frame embeddings $\mathbf{V}_i$ and the word embeddings $\mathbf{T}'_i = [\mathbf{w}_i^1; \cdots; \mathbf{w}_i^m] \in \mathbb{R}^{m \times C}$, we calculate the weights of all words attending to each frame via softmax attention, taking $\mathbf{V}_i$ as the query and $\mathbf{T}'_i$ as the key. By then multiplying these attention weights by the value $\mathbf{T}'_i$, we obtain a frame-aware AD representation $\tilde{\mathbf{T}}_i \in \mathbb{R}^{n \times C}$:

$$\tilde{\mathbf{T}}_i = \text{Softmax}(\mathbf{V}_i \mathbf{T}'^T_i / \tau) \mathbf{T}'_i = [\tilde{\mathbf{t}}_i^1; \cdots; \tilde{\mathbf{t}}_i^n], \quad (2)$$

where for each frame, the words embeddings that are most similar to the frame-level visual feature have been aggregated (via softmax attention). The temperature parameter τ controls the aggregation process, where smaller τ incorporates more textual information.

The goal of the fine-grained matching is to pull a frame visual feature $\mathbf{f} \in \mathbf{V}_i$ closer to the frame-aware AD representations $\tilde{\mathbf{t}} \in \tilde{\mathbf{T}}_i$ in (2), corresponding to the positive set. To achieve this, we define the negative set $\tilde{\mathbf{T}}_{neg}$ as the frame-aware AD embeddings generated from other video clips (in the batch), and then use a Multi-Instance Loss [49],

$$\mathcal{L}_f = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\sum_{\tilde{\mathbf{t}} \in \tilde{\mathbf{T}}_i} \exp(sim(\mathbf{f}, \tilde{\mathbf{t}}))}{\sum_{\tilde{\mathbf{t}}_* \in \{\tilde{\mathbf{T}}_i \cup \tilde{\mathbf{T}}_{neg}\}} \exp(sim(\mathbf{f}, \tilde{\mathbf{t}}_*))}, \quad (3)$$

where $\mathbf{f} \in \mathbf{V}_i$ is a sampled frame from $\mathbf{V}_i$. This process is illustrated in the top right of Fig. 2.

Summary for Stage-I. The final objective for Stage-I is to minimize the sum of **global** and **fine-grained** aligning losses, balanced by a trade-off coefficient, $\mathcal{L}_\text{I} = \gamma \mathcal{L}_g + (1 -$

$\gamma)\mathcal{L}_f$. Note that during this adaptation process, the CLIP-Text encoder model remains frozen, and only the CLIP-Vision encoder is fine-tuned. Our fine-grained frame-AD matching is entirely parameter-free, as only the vision encoder will be utilized in the subsequent stage.

3.2. Stage-II: Distinctive AD Narration

The motivation for generating *distinctive* ADs stems from the observation that LLM often produce repetitive descriptions for adjacent clips [56, 82, 86]. Despite improved character recognition, the visual representation itself is not discriminative among neighboring (contextual) clips, leading to uninteresting ADs. Our goal is to create contextual-distinctive ADs that highlight current differences. We hypothesize, as verified in Appendix §B , that sequential clips from a long video often share redundant scenes or characters, leading similar visual features in contexts. Thus, we propose Stage-II: distinctive AD narration.

As shown in Fig. 3, we prepare N consecutive video clips (to be AD-described) $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$, each containing T uniformly sampled frames $\{\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_T\}$. Following [21, 22, 74], we employ a learnable Perceiver adapter [5] to resample T' prompt vectors for the T frame embeddings encoded by our Stage-I vision encoder, CLIP_{AD}. This process is formulated as:

$$\mathbf{h}_{\mathbf{x}_i} = \text{Perceiver}(\{\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_T\}) \in \mathbb{R}^{T' \times C}, \quad (4)$$
$$\mathbf{f}_i = \text{CLIP}_{\text{AD}}(\mathcal{F}_i). \quad (5)$$

We then introduce the Contextual EMA to capture compact, discriminative visual features for distinctive AD generation.

Contextual EMA. Expectation-Maximization Attention (EMA) [35] integrates the attention mechanism [81] into the classical EM [17] algorithm, which comprises three steps to

estimate a more compact set of bases: *Responsibility Estimation* (RE), *Likelihood Maximization* (LM), and *Data Reestimation* (DR). Inspired by this, we propose **Contextual EMA** to perform EMA on frames from N contextual clips, aiming to eliminate redundancy, learn compact representations, and explore distinctiveness.

Let $\mathcal{H} = \{\mathbf{h}_{\mathbf{x}_i}\}_{i=1}^N \in \mathbb{R}^{N \times T' \times C}$ represent N clip vectors from the Perceiver, and $\mathcal{M} = \{\mu_k\}_{k=1}^K \in \mathbb{R}^{K \times C}$ denote the randomly initialized base features, where C, K indicate the number of channel and bases. The RE step estimates the hidden variable $\mathcal{Z} = \{z_{nk}\}_{n=1,k=1}^{N \times T', K}$, where the responsibility z_{nk} represents the probability of the n-th *frame* belonging to the k-th base:

$$z_{nk} = \frac{\exp(\mathbf{h}_n \mu_k^T / \tau)}{\sum_{j=1}^K \exp(\mathbf{h}_n \mu_j^T / \tau)}, \qquad (6)$$

where τ determines the shape (peakiness) of distribution $\mathcal{Z}$. Then, the LM step updates the bases $\mathcal{M}$ by applying the weighted average on input $\mathcal{H}$, formulating the k-th base as:

$$\mu_k = \frac{\sum_{n=1}^{N \times T'} z_{nk} \mathbf{h}_n}{\sum_{n=1}^{N \times T'} z_{nk}}. \qquad (7)$$

The RE (E-step) and LM (M-step) are iteratively performed R times until convergence. Notably, since bases number K is much smaller than the embedding number $N \times T'$, we employ DR to reconstruct a compact version of $\mathcal{H}$ through:

$$\widehat{\mathcal{H}} \approx \mathcal{Z}\mathcal{M}. \qquad (8)$$

Here, $\widehat{\mathcal{H}} \in \mathbb{R}^{N \times T' \times C}$ retains the same shape as $\mathcal{H}$. We combine $\mathcal{H}$ and $\widehat{\mathcal{H}}$ element-wise with a hyperparameter α.

To enhance representation distinctiveness, we introduce an additional branch using cross-attention between raw $\mathcal{H}$ (query) and bases $\mathcal{M}$ (key and value), formulated as:

$$\widetilde{\mathcal{H}} = \mathrm{CrossAttention}(\mathcal{H}, \mathcal{M}), \qquad (9)$$

where $\widetilde{\mathcal{H}} \in \mathbb{R}^{N \times T' \times C}$. Linear layers projecting queries, keys, and values are omitted in Fig. 3 (see Appendix §C for details). Through (9), we process the distributions of $\mathcal{H}$ to attend on specific and informative bases, with improved linear separability (see Fig. 6). We combine $\mathcal{H}, \widehat{\mathcal{H}}, \widetilde{\mathcal{H}}$ element-wise around Contextual EMA to construct the final refined visual features. These features are then projected into the LLM embedding space using a single-layer projector:

$$\mathcal{H}_{sum} = \mathrm{Proj}(\mathcal{H} + \alpha\widehat{\mathcal{H}} + \beta\widetilde{\mathcal{H}}). \qquad (10)$$

Interleaved prompt as LLM's input. Following previous studies [22, 74, 82], we build our interleaved `prompt` enriched with character information, (see Fig. 3). To answer the "who is who" question when more than two characters are present, the corresponding actors' portrait images are projected as face tokens for reasoning. The `<BOS>` tag appended at the end indicates the beginning of AD generation.

Distinctive words highlighting. Our goal is to query a frozen LLM for AD generation using a vision-conditioned prompt. The typical supervision employs the commonly used auto-regressive loss function:

$$\mathcal{L}_{auto} = -\sum_n \log P_\theta(w_n | \mathrm{prompt}; w_{<n}), \qquad (11)$$

where w_n is the n-th token from the target AD. However, $\mathcal{L}_{auto}$ does not emphasize the distinctiveness specific to the current AD, which is our focus. To address this, we propose a distinctive word set w_d, created by filtering out duplicates, such as character names, prepositions, and pronouns, from the N context ADs of the target AD. During training, we explicitly encourage the LLM to predict the distinctive words in w_d by optimizing the distinctive loss $\mathcal{L}_{dist}$:

$$\mathcal{L}_{dist} = -\sum_{n=1}^N \sum_{i=1}^u \log P_\theta(w_n = w_d^i | \mathrm{prompt}, w_{<n}), \qquad (12)$$

where w_d^i denotes the i-th distinctive word in w_d and u is the size of the set. The final complete loss function for Stage-II is: $\mathcal{L}_{\mathrm{II}} = \mathcal{L}_{auto} + \mathcal{L}_{dist}$.

4. Experiments

4.1. Experiment Setup

Datasets. We follow the AD generation benchmark established in AutoAD [21], conducting experiments on the denoised **MAD-v2-Named** [68] and evaluating on **MAD-Eval-Named** split. Specifically, MAD-v2-Named includes ∼330k ADs from 488 movies for training and MAD-Eval has 6,520 ADs crawled from 10 movies for evaluation. We also evaluate on two recently introduced datasets. **CMD-AD** [23] (where "CMD" stands for Condensed Movie Dataset [7]) is a movie AD dataset that contains 101k ADs for more than 1432 movies, with 100 movies split for CMD-AD evaluation. **TV-AD** [82] is a recently proposed AD dataset based on TVQA [30], which contains ∼31k ADs for training and ∼3k ADs for evaluation.

Evaluation Metrics. Classic captioning metrics including **ROUGE-L** [37], **CIDEr** [72] and **SPICE** [6] are reported to evaluate the quality of generated ADs versus the ground-truth. Besides, we also report Recall@k within N Neighbours [22] (**R@k/N**), which calculates the average value of Recall@k for each AD with its N temporally adjacent GT texts, where BertScore [88] is used for text similarity matching. The R@k/N metric is based on retrieving the most similar ground-truth AD among N neighbors, and thus highlights the *distinctiveness* of generated ADs directly. **LLM-AD-eval** [23] employs LLMs to assess the quality of generated ADs, assigning scores from 1 (lowest) to 5 (highest). We utilize the LLM prompt from the original study [82] and apply open-source models LLaMA2-7B-Chat [70] and LLaMA3-8B-Instruct [4] for this evaluation.

Method	Pub.	VLM	LLM	ROUGE-L	CIDEr	SPICE	R@5/16
Training-free							
VLog [2]	-	-	GPT-4	7.5	1.3	2.1	42.3
MM-Vid [39]	ArXiv'23	GPT-4V	-	9.8	6.1	3.8	46.1
MM-Narrator [86]	CVPR'24	CLIP-L14	GPT-4	13.4	13.9	5.2	49.0
LLM-AD [14]	ArXiv'24	GPT-4V	-	13.5	20.5	-	-
AutoAD-Zero [82]	ACCV'24	VideoLLaMA2-7B	LLaMA3-8B	-	22.4	-	-
Partial-fine-tuning							
ClipCap [50]	ArXiv'21	CLIP-B32	GPT-2	8.5	4.4	1.1	-
CapDec [52]	ArXiv'22	-	-	8.2	6.7	1.4	-
AutoAD-I [21]	CVPR'23	CLIP-B32	GPT-2	11.9	14.3	4.4	42.1
AutoAD-II [22]	ICCV'23	CLIP-B32	GPT-2	13.4	19.5	-	50.8
AutoAD-III [23]	CVPR'24	EVA-CLIP	OPT-2.7B	-	22.8	-	52.0
AutoAD-III [23]	CVPR'24	EVA-CLIP	LLaMA2-7B	-	24.0	-	52.8
MovieSeq [40]	ECCV'24	CLIP-B16	LLaMA2-7B*	15.5	24.4	7.0	51.6
DistinctAD (Ours)		CLIP-B32	GPT-2	15.4	24.5	6.7	49.8
DistinctAD (Ours)		CLIP$_{AD}$-B32	GPT-2	16.4	25.5	7.4	51.7
DistinctAD (Ours)		CLIP$_{AD}$-B16	LLaMA2-7B	17.2	27.0	8.2	55.6
DistinctAD (Ours)		CLIP$_{AD}$-B16	LLaMA3-8B	**17.6**	**27.3**	**8.3**	**56.0**

Table 1. Comparisons of AD performance on the MAD-Eval benchmark. * indicates fine-tuning LLaMA2-7B model with LoRA [24]. CLIP$_{AD}$ is our CLIP vision encoder adapted using our Stage-I strategy.

Method	CIDEr	R@1/5	LLM-AD-eval
Video-BLIP2 [85]	4.8	22.0	1.89 | -
Video-LLaMA2 [87]	5.2	23.6	1.91 | -
AutoAD-II [22]	13.5	26.1	2.08 | -
AutoAD-III [23]	21.7	30.0	2.85 | -
AutoAD-Zero [82]	17.7	-	2.83 | 1.96
DistinctAD (LLaMA2)	22.3	32.9	**2.89** | 2.00
DistinctAD (LLaMA3)	**22.7**	**33.0**	2.88 | **2.03**
AutoAD-III† [23]	25.0	31.2	2.89 | 2.01

Table 2. Comparisons on CMD-AD. The LLM-AD-eval scores are evaluated with LLaMA2-7B (left) and LLaMA3-8B (right). † indicates pre-training on 3.4M HowTo-AD dataset [23, 48].

Method	CIDEr	R@1/5	LLM-AD-eval
AutoAD-III [23]	26.1	-	2.78 | 1.99
AutoAD-Zero [82]	22.6	30.6	**2.94** | 2.00
DistinctAD (LLaMA2)	27.2	31.8	2.86 | **2.01**
DistinctAD (LLaMA3)	**27.4**	**32.1**	2.89 | 2.00

Table 3. Comparisons on TV-AD. The LLM-AD-eval scores are evaluated using LLaMA2-7B (left) and LLaMA3-8B (right).

Implementation Details. To facilitate CLIP-AD adaptation in Stage-I, we collect the original raw movies from MAD from platforms like Amazon Prime Video. See Appendix F for details. We fine-tune the CLIP Vision encoder for 5 epochs with a fixed learning rate 5e-5 using the Adam optimizer [28] in Stage-I, with a batch size of 512. In Stage-II, we use a batch of 8 sequences, each containing 16 consecutive video AD-pairs from a movie. For each video clip, 8 frames are uniformly sampled. We use the AdamW [44] optimizer to train our model for 10 epochs, with a cosine-decayed learning rate and linear warm-up. The learning rate is set to 10^{-4} for both GPT-2 and LLaMA models. For external *character* information, we directly use the inference results from AutoAD-Zero [82] as it gives current best face

recognition performance. We use CLIP-B32 and GPT-2 on MAD-Eval-Named for ablations unless specified otherwise.

4.2. Comparisons with previous methods

We conduct comprehensive comparisons using the widely-adopted MAD-Eval benchmark [21] and two recently introduced AD datasets, CMD-AD [23] and TV-AD [82].

Comparisons on MAD-Eval are shown in Tab. 1. We primarily categorize previous studies into *Training-free* and *Partial-fine-tuning* approaches, as described in §2. Our method is a *Partial-fine-tuning* method. When using the same CLIP-B32 and GPT-2, our proposed DistinctAD achieves a CIDEr score of 24.5, surpassing previous AutoAD-I [21] (CIDEr 14.3) and AutoAD-II [22] (CIDEr 19.5). With our Stage-I adapted CLIP vision encoders (denoted as CLIP$_{AD}$), we observe stable improvements across all metrics, *e.g.* 25.5 *vs.* 24.5 on CIDEr and 51.7 *vs.* 49.8 on recall, validating the effectiveness of our Stage-I strategy. Notably, DistinctAD with CLIP$_{AD}$-B16 and LLaMA3-8B [4] achieves state-of-the-arts with CIDEr 27.3 and Recall@5/16 56.0. Our outstanding performance on R@k/N demonstrates DistinctAD's ability to generate distinctive ADs, well matching the uniqueness of the clip's contents.

Looking at the *training-free* methods, despite the capabilities of advanced proprietary VLMs, *e.g.* GPT-4V [53], and LLMs, *e.g.* GPT-4 [3], the performance of training-free methods remains inferior to those employing partial-fine-tuning. This discrepancy likely arises from the unique characteristics of AD and movie data, which exhibit a significant domain gap from common vision language training data. As such, these data types were not encountered during the pre-training of proprietary large-scale models.

Setting	CIDEr	R@5/16
None	6.7	34.0
Global $\mathcal{L}_g$	8.2	36.6
Fine-grained $\mathcal{L}_f$	7.7	35.2
$\gamma\mathcal{L}_g + (1-\gamma)\mathcal{L}_f$	**8.6**	**36.9**

(a) Stage-I components.

Coefficient γ	CIDEr
0.1	8.0
0.3	8.5
0.5	**8.6**
0.7	7.7

(b) Impact of coefficient γ.

Prompt	Stage-I	CIDEr	R@5/16
Contextual ADs [21]	✗	12.6 (17.8)	39.8 (43.1)
	✓	**14.1** (**19.0**)	**39.9** (**44.2**)
Character [22]	✗	22.0	45.6
	✓	**23.1**	**46.2**

(c) Impact of Stage-I w/ different prompts.

Table 4. Ablation studies for Stage-I. (a) Evaluation of global video-AD loss $\mathcal{L}_f$ and fine-grained frame-AD loss $\mathcal{L}_f$ on AD performance. (b) Analysis of the the impact of the coefficient γ. Both (a) and (b) are conducted with pure visual prompts. (c) Impact of Stage-I when combined with different prompts for the LLM decoder, including contextual ADs and character names. Performance in parentheses indicates results with ground-truth contextual ADs as prompts.
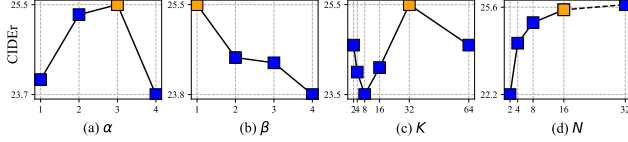

Figure 4. Ablation studies for hyperparameter in Stage-II, with final settings highlighted in orange. (a) Impact of α on the weight of compact representation $\widehat{\mathcal{H}}$. (b) Influence of β on cross-attended feature $\widetilde{\mathcal{H}}$. (c) Impact of K, which denotes the number of clusters in bases $\mathcal{M}$. (d) Effect of sampling N consecutive video clips. We switch to larger memory GPUs when N exceeds 16.

Ex#	$\alpha\widehat{\mathcal{H}}$	$\beta\widetilde{\mathcal{H}}$	$\mathcal{L}_{dist}$	CIDEr	R@5/16
A0	✗	✗	✗	23.1 (27.4)	46.2
B1	✓	✗	✗	23.7 (29.3)	46.6
B2	✗	✓	✗	23.4 (29.1)	46.1
B3	✓	✓	✗	23.3 (28.1)	48.0
C0	✗	✗	✓	24.7 (29.4)	49.5
C1	✓	✗	✓	24.3 (29.4)	50.7
C2	✗	✓	✓	25.3 (**30.4**)	51.5
C3	✓	✓	✓	**25.5** (29.8)	**51.7**

Table 5. Ablation studies for components in Stage-II. The CIDEr column shows scores with AutoAD-Zero's character [82] as prompt by default. CIDEr in parentheses represent performance with ground-truth character names.

Comparisons on CMD-AD and TV-AD. We further verify the generalizabilty of DistinctAD on the recently proposed CMD-AD and TV-AD benchmarks, with results presented in Tables 2 and 3. Both evaluations use adapted $CLIP_{AD}$-B16 model. DistinctAD outperforms AutoAD-Zero, AutoAD-II and AutoAD-III in terms of CIDEr and R@1/5 on both benchmarks Meanwhile, DistinctAD exhibits a lower CIDEr compared to AutoAD-III† on CMD-AD, which we conjecture is primarily due to AutoAD-III† pre-training on a very large-scale 3.4M transformed HowTo-AD dataset [23, 48], *which is currently publicly unavailable.* Despite this, DistinctAD achieves superior R@1/5, underscoring its exceptional ability to generate distinctive and high-quality ADs. This is further corroborated by its leading performance on the LLM-AD-eval metric.

4.3. Ablation studies

Effect of CLIP-AD Adaptation (I). Tab. 4a demonstrates the benefit of our Stage-I strategy, *i.e.* adapting CLIP the vi-

Consecutive N?	CIDEr	R@5/16
✓	**25.5**	51.7
✗	$23.8_{\downarrow1.7}$	$\textbf{52.5}_{\uparrow0.8}$

Table 6. Impact of whether sampling N consecutive clips or not.

sion encoder to the movie-AD domain via global video-AD matching $\mathcal{L}_g$ and fine-grained frame-AD matching $\mathcal{L}_f$. In Tab. 4b, the balancing coefficient γ performs best at 0.5. In Tab. 4c, our Stage-I strategy consistently enhances performance when combined with different prompts in the LLM decoder, such as contextual ADs in AutoAD-I [21] or character names in AutoAD-II [22]. This indicates that our AD-adapted CLIP vision encoder can integrate seamlessly into previous methods, including those training-free models that utilize CLIP-based visual extractors.

Effect of Distinctive AD narration (II). We evaluate the effectiveness of Stage-II components in Tab. 5, based on the default $\mathcal{H}$ (Perceiver's output) and full AD auto-regressive loss $\mathcal{L}_{auto}$. The baseline (A0) outperforms AutoAD-II in CIDEr (23.1 vs. 19.5), primarily due to more accurate character prompts from AutoAD-Zero and the adapted CLIP vision encoder from Stage-I. A0 with AutoAD-II's characters achieves a CIDEr score of 20.6, close to AutoAD-II's performance. Incorporating reconstructed feature $\widehat{\mathcal{H}}$ brings stable improvements on both CIDEr and recall (B1 & C1), highlighting the importance of compact representations in understanding visual semantics. Cross-attended feature $\widetilde{\mathcal{H}}$ works better together with distinctive word prediction loss $\mathcal{L}_{dist}$ (C2 *vs.* B2, C3 *vs.* B3). We conjecture this is because $\mathcal{L}_{dist}$ provides more definite supervision on re-weighting distinctive words, which guides $\widetilde{\mathcal{H}}$ to attend on concept-related bases. Overall, applying the full Stage-II pipeline brings significant and robust performance (C3).

Effect of Hyper-parameters. Fig. 4 summarizes the ablation studies on 4 hyper-parameters that potentially influence the results in Stage-II. Coefficient weights α and β yield optimal results when set to 3 and 1, respectively. This suggests the need to refine our final representations to be more compact for generating ADs. Fig. 4(c) shows setting bases number K to 32 yields best. A smaller K, *e.g.* 2, can still achieve notable CIDEr, as the Contextual EMA module does not significantly alter the final output $\mathcal{H}_{sum}$. However,
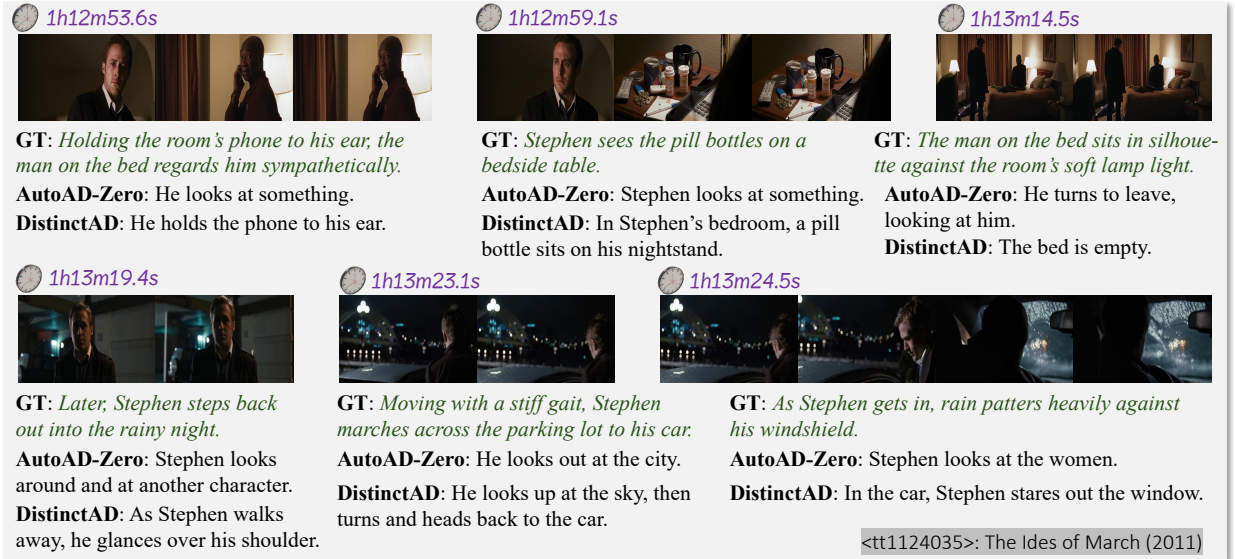
Figure 5. Qualitative results. We present ground-truth (GT) ADs, publicly released AutoAD-Zero outputs, and our DistinctAD predictions for several temporally consecutive movie clips. Movie frames are taken from The Ides of March (2011) [15]. Zoom in for details.

unsuitable values of K, *e.g.* 8 or 64, can negatively impact performance. Fig. 4(d) reveals that increasing the number of consecutive clips N (with K set to 32) enhances the CIDEr score, though this effect saturates when N exceeds 16. This demonstrates that more bases should be created to effectively summarize components with additional clips.

Do the N clips to be consecutive? Tab. 6 presents the results of sampling *non-consecutive* N clips during training. When using non-continuous clips, we observe a decrease in the CIDEr metric by 1.7 (25.5 *vs.* 23.8) because the Contextual EMA module struggles with unrelated contexts. However, the R@5/16 improves by 0.8, which indicates enhancement of the distinctiveness (uniqueness) of the generated AD when using more diverse visual contents.

Visualizations. To better understand what Contextual EMA learns, we show the t-SNE [71] visualizations of $\mathcal{H}, \widehat{\mathcal{H}}$ and $\widetilde{\mathcal{H}}$ (from §3.2) in Fig. 6, using the same perplexity value across all visualizations. With Contextual EMA, $\widehat{\mathcal{H}}$ exhibits more compact features compared to raw $\mathcal{H}$, Fig. 6(b). Interestingly in Fig. 6(c), cross-attention between $\mathcal{H}$ and bases $\mathcal{M}$ produces strip-like feature distributions pointing to specific base centers, enhancing contextual distinctiveness with improved linear separability and interpretability.

4.4. Qualitative results

Fig. 5 presents qualitative examples of our model. We compare the predictions of DistinctAD (using LLaMA3-8B) with ground-truth captions (GT) and publicly available AutoAD-Zero [82] outputs. Note that clips are sampled *consecutively* in time. Previous studies often struggle with similar contextual clips, such as those featuring closely-related scenes and characters, by repeating correct yet in-
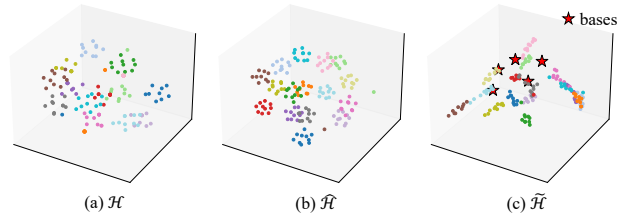

Figure 6. Visualizations of Contextual EMA. (a) A set of randomly generated 3D data $\mathcal{H}$, sampled from N types of samples. (b) Compact features $\widehat{\mathcal{H}}$ obtained via Data Re-estimation (DR). (c) Cross-attention outputs $\widetilde{\mathcal{H}}$ between $\mathcal{H}$ and bases $\mathcal{M}$.

significant action words, *e.g.* "look". In contrast, our DistinctAD effectively generates more engaging ADs by identifying distinctive objects in adjacent clips, *e.g.* "phone", "pill bottle", and "car", along with corresponding more specific behaviors. More examples can be found in Appendix §E.

5. Conclusion

This paper proposes DistinctAD, a two-stage framework for generating distinctive audio descriptions for better narrative. By addressing the domain gap between movie-AD data with a CLIP-AD adaptation strategy, and introducing a Contextual EMA module and a distinctive word prediction loss, our approach significantly improves the quality of AD generation. The effectiveness of DistinctAD is demonstrated through comprehensive evaluations on multiple benchmark datasets and ablations studies. Despite these promising results, DistinctAD is still limited by requiring numbers of parameters and the quality of the generated ADs still falls short of human annotations. Overall, automatic AD generation still remains a challenging task, and there is considerable scope for future advancements in this field.

Acknowledgement

References

[1] https://audiovault.net, 2019. 3

[2] Vlog: Video as a long document. https://github.com/showlab/VLog, 2023. GitHub repository. 6

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2, 6

[4] AI@Meta. Llama 3 model card. 2024. 5, 6, 2

[5] 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. *NeurIPS*, 35: 23716–23736, 2022. 1, 3, 4

[6] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398. Springer, 2016. 5

[7] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *ACCV*, 2020. 5

[8] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*, 2023. 2

[9] Carmen J Branje and Deborah I Fels. Livedescribe: can amateur describers create high-quality audio description? *Journal of Visual Impairment & Blindness*, 106(3):154–165, 2012. 1, 2

[10] Hervé Bredin and Antoine Laurent. End-to-end speaker segmentation for overlap-aware resegmentation. In *Interspeech*, 2021. 2

[11] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP*, pages 7124–7128. IEEE, 2020. 2

[12] Fuhai Chen, Rongrong Ji, Xiaoshuai Sun, Yongjian Wu, and Jinsong Su. Groupcap: Group-based image captioning with structured relevance and diversity constraints. In *CVPR*, pages 1345–1353, 2018. 3

[13] Shaoxiang Chen and Yu-Gang Jiang. Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In *CVPR*, pages 8425–8435, 2021. 2

[14] Peng Chu, Jiang Wang, and Andre Abrantes. Llm-ad: Large language model based audio description system. *arXiv preprint arXiv:2405.00983*, 2024. 1, 2, 6

[15] George Clooney. The ides of march. *Columbia Pictures*, 2011. 8

[16] Bo Dai and Dahua Lin. Contrastive learning for image captioning. *NeurIPS*, 30, 2017. 3

[17] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977. 2, 4

[18] Chaorui Deng, Shizhe Chen, Da Chen, Yuan He, and Qi Wu. Sketch, ground, and refine: Top-down dense video captioning. In *CVPR*, pages 234–243, 2021. 2

[19] Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. In *CVPR*, pages 24199–24208, 2024. 3

[20] Louise Fryer. An introduction to audio description: A practical guide, 2016. 1

[21] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. Autoad: Movie description in context. In *CVPR*, pages 18930–18940, 2023. 1, 2, 3, 4, 5, 6, 7

[22] Tengda Han, Max Bain, Arsha Nagrani, Gul Varol, Weidi Xie, and Andrew Zisserman. Autoad ii: The sequel-who, when, and what in movie audio description. In *ICCV*, pages 13645–13655, 2023. 1, 3, 4, 5, 6, 7

[23] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. Autoad iii: The prequel-back to the pixels. In *CVPR*, pages 18164–18174, 2024. 1, 2, 3, 5, 6, 7

[24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 6

[25] Vladimir Iashin and Esa Rahtu. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In *BMVC*, 2020. 2

[26] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *CVPRW*, pages 958–959, 2020. 2

[27] Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David Clifton, and Jie Chen. Expectation-maximization contrastive learning for compact video-and-language representations. *NeurIPS*, 35:30291–30306, 2022. 2

[28] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[29] 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

[30] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 5

[31] Elisa Lewis. Deep dive: How audio description benefits everyone, 2021. *Accessed on*, pages 11–13, 2023. 1

[32] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 3

[33] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: a multi-modal model with in-context instruction tuning. corr abs/2305.03726 (2023), 2023. 3

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 3

[35] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *ICCV*, pages 9167–9176, 2019. 2, 4

[36] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *CVPR*, pages 7492–7500, 2018. 2

[37] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5

[38] 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*, pages 17949–17958, 2022. 2

[39] 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. 1, 2, 6

[40] Kevin Qinghong Lin, Pengchuan Zhang, Difei Gao, Xide Xia, Joya Chen, Ziteng Gao, Jinheng Xie, Xuhong Xiao, and Mike Zheng Shou. Learning video context as interleaved multimodal sequences. *arXiv preprint arXiv:2407.21757*, 2024. 1, 3, 6

[41] Zhihui Lin, Tianyu Yang, Maomao Li, Ziyu Wang, Chun Yuan, Wenhao Jiang, and Wei Liu. Swem: Towards real-time video object segmentation with sequential weighted expectation-maximization. In *CVPR*, pages 1362–1372, 2022. 2

[42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024. 1

[43] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *ECCV*, pages 338–354, 2018. 3

[44] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[45] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 2

[46] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 3

[47] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In *CVPR*, pages 6964–6974, 2018. 3

[48] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640, 2019. 6, 7

[49] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, pages 9879–9889, 2020. 4

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

[51] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. Streamlined dense video captioning. In *CVPR*, pages 6588–6597, 2019. 2

[52] David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected clip. *arXiv preprint arXiv:2211.00575*, 2022. 6

[53] OpenAI. Gpt-4v(ision) system card. 2023. 1, 2, 6

[54] Amy Pavel, Gabriel Reyes, and Jeffrey P Bigham. Rescribe: Authoring and automatically editing audio descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 747–759, 2020. 1, 2

[55] Elisa Perego. Gains and losses of watching audio described films for sighted viewers. *Target*, 28(3):424–444, 2016. 1

[56] Haran Raajesh, Naveen Reddy Desanur, Zeeshan Khan, and Makarand Tapaswi. Micap: A unified model for identity-aware movie descriptions. In *CVPR*, pages 14011–14021, 2024. 1, 4

[57] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[58] 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. 2

[59] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *ICCV*, pages 8908–8917, 2019. 2

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

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

[62] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *CVPR*, pages 17959–17968, 2022. 2

[63] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 2

[64] Zhiqiang Shen, Jianguo Li, Zhou Su, Minjun Li, Yurong Chen, Yu-Gang Jiang, and Xiangyang Xue. Weakly supervised dense video captioning. In *CVPR*, pages 1916–1924, 2017. 2

[65] Botian Shi, Lei Ji, Yaobo Liang, Nan Duan, Peng Chen, Zhendong Niu, and Ming Zhou. Dense procedure captioning

in narrated instructional videos. In *ACL*, pages 6382–6391, 2019. 2

[66] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *ICCV*, pages 11987–11997, 2023. 3

[67] Joel Snyder. Audio description: The visual made verbal. In *International congress series*, pages 935–939. Elsevier, 2005. 1

[68] 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, 5, 1, 4

[69] 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

[70] 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. 5

[71] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008. 8

[72] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015. 5

[73] Gilad Vered, Gal Oren, Yuval Atzmon, and Gal Chechik. Joint optimization for cooperative image captioning. In *ICCV*, pages 8898–8907, 2019. 3

[74] Hanlin Wang, Zhan Tong, Kecheng Zheng, Yujun Shen, and Limin Wang. Contextual ad narration with interleaved multimodal sequence. *arXiv preprint arXiv:2403.12922*, 2024. 1, 2, 3, 4, 5

[75] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *CVPR*, pages 7190–7198, 2018. 2

[76] Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. Compare and reweight: Distinctive image captioning using similar images sets. In *ECCV*, pages 370–386. Springer, 2020. 3

[77] Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. Group-based distinctive image captioning with memory attention. In *ACMMM*, pages 5020–5028, 2021.

[78] Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. On distinctive image captioning via comparing and reweighting. *TPAMI*, 45(2):2088–2103, 2022. 3

[79] Teng Wang, Huicheng Zheng, Mingjing Yu, Qian Tian, and Haifeng Hu. Event-centric hierarchical representation for dense video captioning. *TCSVT*, 31(5):1890–1900, 2020. 2

[80] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *ICCV*, pages 6847–6857, 2021. 2

[81] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 4

[82] Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. Autoad-zero: A training-free framework for zero-shot audio description. *arXiv preprint arXiv:2407.15850*, 2024. 1, 2, 4, 5, 6, 7, 8

[83] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*, pages 10714–10726, 2023. 2

[84] Linli Yao, Weiying Wang, and Qin Jin. Image difference captioning with pre-training and contrastive learning. In *AAAI*, pages 3108–3116, 2022. 3

[85] Keunwoo Peter Yu. Videoblip, 2023. 6

[86] Chaoyi Zhang, Kevin Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. Mm-narrator: Narrating long-form videos with multimodal in-context learning. In *CVPR*, pages 13647–13657, 2024. 1, 2, 4, 6

[87] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 6

[88] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 5

[89] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, pages 8739–8748, 2018. 2