Contextual AD Narration with Interleaved Multimodal Sequence

Supplementary Material

Appendix

A. Overview

The supplementary material provides more implementation details, ablation analyses and qualitative results to show deep insights into our method. Specifically, Appendix B describes the architecture details of our model and experiment settings. Appendix C provides more ablation studies on our Uni-AD. To further show the effectiveness of our model design, we show more qualitative results of Uni-AD in Appendix D.

B. Implementation Details

B.1. Architecture details

We show the architecture details of our visual mapping network for MAD-eval-Named benchmark and character-refinement module in Tab. S1. Note that for CMDAD and TVAD datasets, we use the training set of CMDAD to train our character-refinement module and apply InternVideo2.5 [15] to extract video frame features. The character-refinement performances are shown in Tab. S2.

Table S1. Architecture details of visual mapping network and character-refinement module in Uni-AD.

Visual mapping network (GPT-2-B32)	projection layer	512→768
	num latent	30
	num blocks	2
	channel	768
	num head	12
	ffn dimension	3072
Visual mapping network (LLaMA-L14)	projection layer	768→4096
	num latent	30
	num blocks	2
	channel	4096
	num head	32
	ffn dimension	16384
Character-refinement module (for MAD-eval-Named)	num blocks	3
	channel	768
	num head	12
	ffn dimension	3072
Character-refinement module (for CMDAD&TVAD)	projection layer	4096→4096
	num blocks	1
	channel	4096
	num head	32
	ffn dimension	16384

B.2. Datasets

MAD-Named [2, 12]: The MAD-Named benchmark consists of two parts: MAD-v2-Named for training and MAD-eval-Named for testing. Specifically, MAD-v2-Named contains 334,296 ADs and 628,613 subtitles collected from 488 movies, while MAD-eval-Named contains 6,520 ADs and 10,602 subtitles collected from 10 movies. Annotation includes the start and end time of each AD and the AD contents without any post-processing on character names. We notice there are many homophonic name mismatches in MAD-Named, such as an actor's name being 'Gray' in character information from IMDb but 'Grey' in the ad annotation. We thus processed these mismatched information to ensure that the same actor's name remains consistent.

Table S2. Architecture details of visual mapping network and character-refinement module in Uni-AD.

Dataset	Precision↑	Recall↑
MAD-eval-Named	0.41	0.77
CMDAD	0.26	0.94
TVAD	0.27	0.94

CMDAD [4]: CMDAD is a movie AD dataset that contains 101k ADs for more than 1432 movies, with 100 movies split for evaluation.

TVAD [16]: TVAD is a recently proposed TV-series AD dataset, which contains 31k ADs for training and 3k ADs for evaluation.

AudioVault-AD [2]: AudioVault-AD is a text-only dataset composed of 3.3 million AD utterances collected from 7,057 movies downloaded from the AudioVault website. Movies in AudioVault-AD are not included in the MAD dataset.

B.3. Baselines

In this paper, we compare our Uni-AD with the following AD generation methods:

ClipCap [7]. The ClipCap model converts the CLIP [11] feature of visual inputs into embeddings with a mapping network. Then the output embeddings will be used as prefix to prompt GPT-2 [10] to generate corresponding captions.

AutoAD-I [2]. AutoAD-I follows ClipCap and concatenate previous AD descriptions and subtitles in movie with visual embeddings to prompt the fronzen GPT-2 [10] for AD generation. This approach further apply partial-data pretrain to address the issue of insufficient AD data.

AutoAD-II [3]. This method applies a Flamingo-style [1] architecture for AD generation and introduces an external Character Bank to enable their model to label characters appearing in the movie. AutoAD-II also presents an AD temporal proposal module to determine whether AD should be inserted in the given pause in dialogue.

AutoAD-III [4]. AutoAD-III follows BLIP2 [5] to use

Q-former architecture to bridge the visual space with the language space. Then the model can generate textual outputs with a large language model. AutoAD-III also proposed a large-scale HowTo-AD dataset for pre-training.

AutoAD-Zero [16]. AutoAD-Zero designs a pipeline for character recognition with face detection methods and prompts LLM by circling character faces. A two-stage training-free method is proposed for AD generation, which consists of (i) VLM-Based Video Description and (ii) LLM-Based AD Summary.

MM-Narrater [17]. MM-Narrater employs specialized vision and audio expert models to extract multimodal information from the input video clip. The outputs, along with movie subtitles and previous AD descriptions are used to build prompt to query GPT-4 [8] or GPT-4V [9] for AD generation. Besides, MM-Narrater utilizes retrieval enhancement and in context learning to improve the quality of generated AD.

B.4. Metircs

In this paper, we use both classic captioning metrics and newly proposed metrics for evaluation. Classic captioning metrics include ROUGE-L [6] and CIDEr [14]. In this section, we mainly introduce newly proposed metrics: R@k/N, CRITIC and LLM-AD-eval.

R@k/N [3]: R@k/N is a retrieval metric that distinguishes the predicted text among a set of neighbours. The parameters k and N mean within a temporal window of N neighbouring reference ADs, whether the predicted AD can retrieve the corresponding reference AD at top-k position.

CRITIC [4]: CRITIC assesses the precision of character recognition in generated ADs. Specifically, a co-referencing model is utilized to substitute ambiguous pronouns in ADs with official names from the character banks. Subsequently, two sets of names from predicted and ground truth ADs are compared, and the IoU is computed to yield a CRITIC score.

LLM-AD-eval [4]: LLM-AD-eval utilises LLMs to judge the quality of generated ADs by scoring them between 1 (lowest) and 5 (highest). We use llama2-7b-chat [13] for the evaluation in our experiments.

C. More Ablation Studies

In this section, we explore more ablation studies on our Uni-AD, which are not displayed in the main paper due to space limitation. All experiments are conducted with the character-refinement module and no pre-training is applied.

C.1. Ablation on visual mapping network

As stated in Sec. 3.2 of the main paper, there are multiple reasons why we choose a multi-layer transformer encoder with a fixed number of learnable vectors as our mapping network. Here we compare our visual mapping network with two different visual mapping designs: MLP and multi-layer transformer encoder without learnable vectors. Results in Tab. S3 show that no interaction between video frames (MLP as visual mapping network) gets the worst performance. Allowing interaction between video frames(transformer encoder as visual mapping network) brings better results, but the length of visual embeddings is limited to be consistent with the number of frames(8 in our experiments). Our visual mapping network with 30 learnable vectors performs the best.

Table S3. **Ablation on the structure of visual mapping network.** *latent* denotes the number of learnable vectors in our visual mapping network. Experiments are conducted with Uni-AD(GPT-2-B32)

Visual mapping network	RL↑	C↑	R@5/16↑
MLP	14.0	20.2	45.5
Transformer encoder	15.2	23.4	49.2
Ours(latent=10)	15.4	22.4	49.0
Ours(latent=30)	**15.7**	**23.7**	**49.4**

Table S4. **Ablation on the impact of sharing visual mapping network**. *Share?* shows whether we encode both video and image with one single visual mapping network.

Methods	Share?	RL↑	C↑	R@5/16↑
Uni-AD (GPT-2-B32)	✗	**15.7**	**23.7**	**49.4**
	✓	15.5	22.9	48.8
Uni-AD (LLaMA-L14)	✗	**16.5**	**25.9**	52.5
	✓	16.3	25.8	**53.6**

C.2. Ablation on sharing visual mapping network

Since in Uni-AD, the structure of video mapping network is the same as image mapping network, we in this section study the impact of using a single visual mapping network to encode both video and image. The results are shown in Tab. S4. We can see that encoding video and image with two separate mapping networks is important to Uni-AD(GPT-2-B32), while not necessary for Uni-AD(LLaMA-L14). This reflects that when visual features and LLM are good enough, images and videos can be mixed together for training the mapping network.

C.3. Ablation on image-video interaction

In the main paper, we encode video and image into visual tokens and apply the frozen LLM for interaction between video and image. To study whether more interaction between image and video can benefit AD generation, we replace the input of visual mapping network as concatenation of image and video. Specifically, we replace the input with concatenation of current character's image and the video for image mapping network. For video mapping network, we replace the input with concatenation of all recognized

Table S5. **Ablation on more interaction between image and video**. *Inter-I.?* shows whether we concatenate character's image and the video clip as input of image mapping network. *Inter-V.?* shows whether we concatenate all character images and the video clip as input of video mapping network.

Methods	Inter-I.?	Inter-V.?	RL↑	C↑	R@5/16↑
Uni-AD (GPT-2-B32)	✗	✗	**15.7**	23.7	**49.4**
	✗	✓	15.7	23.1	49.2
	✓	✗	15.7	23.6	49.1
	✓	✓	15.5	**23.8**	49.1
Uni-AD (LLaMA-L14)	✗	✗	**16.5**	**25.9**	52.5
	✗	✓	16.5	25.4	**53.2**
	✓	✗	16.5	25.7	52.2
	✓	✓	16.3	25.0	52.5

characters and the video. In this way, we investigate whether the visual mapping network can extract better character and video representations by more interaction. The results are shown in Tab. S5, which indicates that more interaction between image and video for visual mapping can not further benefit our Uni-AD.

C.4. Ablation on the Threshold in Character-Refinement Module

We conduct ablation study on the impact of threshold in Character-Refinement Module to our Uni-AD and the results in Tab. S6 show that the threshold has a considerable impact on AD generation. High threshold may lead to excessive loss of character information thus gets poor results.

Table S6. **Ablation on the Threshold value in Character-Refinement Module**.

Threhold	RL↑	C↑	R@5/16↑	Threhold	RL↑	C↑	R@5/16↑
0.3	16.5	25.7	52.4	0.5	**16.8**	**27.3**	53.3
0.7	16.2	24.8	**55.4**	0.9	12.8	16.1	54.3

D. Additional Qualitative Analyses

Figure S1 and Figure S2 shows more qualitative results of Uni-AD on the MAD-eval dataset. Note that our character-refinement module can not only recognize AD-related characters, but also serve as a character information denoiser. For example, in sample (a) where characters *Graham* and *Merrill* do not appear in the video clip but are included in the character bank, our character-refinement module removes these noises and provides more precise character information. Though AD without character-refinement module also focuses on describing the female police officer, it can not figure out who she is since there are noises in the initial character bank, thus mistakes *Caroline* as *Merrill's* mom. In sample (c), we find that with more learnable vectors, our model can take the female character who appears at the beginning into AD generation. However,

Figure S1. Qualitative analysis on character-refinement module and contextual information. Movies are selected from (a): Signs (2002), (b): The Ides of March (2011).

Figure S2. Qualitative analysis on number of learnable vectors and comparison with other approaches. Movies are selected from (c): The Ides of March (2011), (d): Les Misérables (2012).

the female character is just ignored by our Uni-AD with fewer learnable vectors.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 1

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

[3] 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*, pages 13599–13609. IEEE, 2023. 1, 2

[4] 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. IEEE, 2024. 1, 2

[5] 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 *ICLR*. JMLR.org, 2023. 1

[6] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 2

[7] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: CLIP prefix for image captioning. *CoRR*, abs/2111.09734, 2021. 1

[8] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 2

[9] OpenAI. GPT-4V(ision) system card. 2023. 2

[10] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 1

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1

[12] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba Heilbron, Chen Zhao, Silvio Giancola, and Bernard Ghanem. MAD: A scalable dataset for language grounding in videos from movie audio descriptions. In *CVPR*, pages 5016–5025. IEEE, 2022. 1

[13] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. 2

[14] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575. IEEE Computer Society, 2015. 2

[15] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025. 1

[16] 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. *CoRR*, abs/2407.15850, 2024. 1, 2

[17] 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. *CoRR*, abs/2311.17435, 2023. 2