DANTE-AD: Dual-Vision Attention Network for Long-Term Audio Description

Adrienne Deganutti, Simon Hadfield, Andrew Gilbert
University of Surrey, UK
{a.deganutti, s.hadfield, a.gilbert}@surrey.ac.uk

Abstract

Audio Description is a narrated commentary that transforms visual content into accessible storytelling for vision-impaired audiences. Existing methods primarily rely on frame-level embeddings, limiting their ability to capture coherent long-term narratives. We introduce DANTE-AD, a novel video description model that uses a dual-vision Transformer-based architecture that enhances automated audio description generation. Our approach introduces a state-of-the-art sequential cross-attention mechanism that integrates frame- and scene-level embeddings to improve contextual awareness across video segments, enabling richer and more temporally coherent narration. Evaluated on a broad range of key scenes from well-known movies, DANTE-AD outperforms existing methods across both traditional NLP metrics and LLM-based evaluations, demonstrating its potential to create high-quality content generation for media accessibility.

1. Introduction

The booming television, streaming, and digital media industries provide entertainment to billions, yet they remain inaccessible to an estimated 338 million people with moderate to total blindness [3]. Beyond missing visuals, these audiences lose crucial storytelling elements like cinematography and symbolism that convey emotions and themes.

Audio Description (AD) addresses this by providing spoken narration of on-screen elements [29]. Unlike captions, AD enhances meaning by incorporating visual context and emotion. Increasing legal requirements have expanded AD adoption, yet its manual creation is costly and time-intensive, leaving many videos without coverage. Automating AD content generation is a pressing challenge.

Effective AD generation requires (i) recognising and describing video content and (ii) producing coherent, contextually relevant narratives. While video captioning tackles the first task for short clips, it lacks the depth and long-term coherence needed for AD [16], largely because most methods rely solely on frame-level information. As video duration increases and multiple scenes emerge, captioning struggles to account for concurrent events, resulting in a gap in generating detailed, context-aware descriptions for long-term content. We propose **DANTE-AD** (**D**ual-Vision **A**ttention **N**etwork for Long-**Te**rm **A**udio **D**escription), a multimodal framework that integrates frame-level information for fine-grained details with scene-level context to ensure continuity and coherence across scenes.

DANTE-AD uniquely integrate frame- and scene-level embeddings through sequential fusion, allowing scene-level context to guide frame-level attention. To our knowledge, our method is the first to employ multi-type visual embeddings for automated AD generation. In summary, our key contributions are the following.

1. Leveraging frame- and scene-level embeddings for complementary spatial and temporal modeling.
2. A Dual-Visual Attention Network, using sequential fusion to align global context with frame-level details.
3. Extensive evaluation on real long-term film clips with real-world metrics.

2. Related Works

Video captioning generates textual descriptions of video content by modelling relationships between visual and linguistic information. Early approaches used sequence-to-sequence models with attention mechanisms [31, 39], while recent methods integrate transformers [16, 27], reinforcement learning [22], and retrieval-based techniques [19, 35]. However, conventional video captioning struggles with capturing narrative depth and long-term dependencies [8], making it insufficient for audio description. This limitation motivates the exploration of techniques such as visual storytelling, multimodal integration, and long-term modelling, as discussed in the following sections.

2.1. Visual Storytelling

Visual storytelling enables models to convey spatial, narrative, and emotional context across domains such as Vision-Language Navigation [32], game generation [17], and social media analysis [26]. Unlike video captioning, AD requires

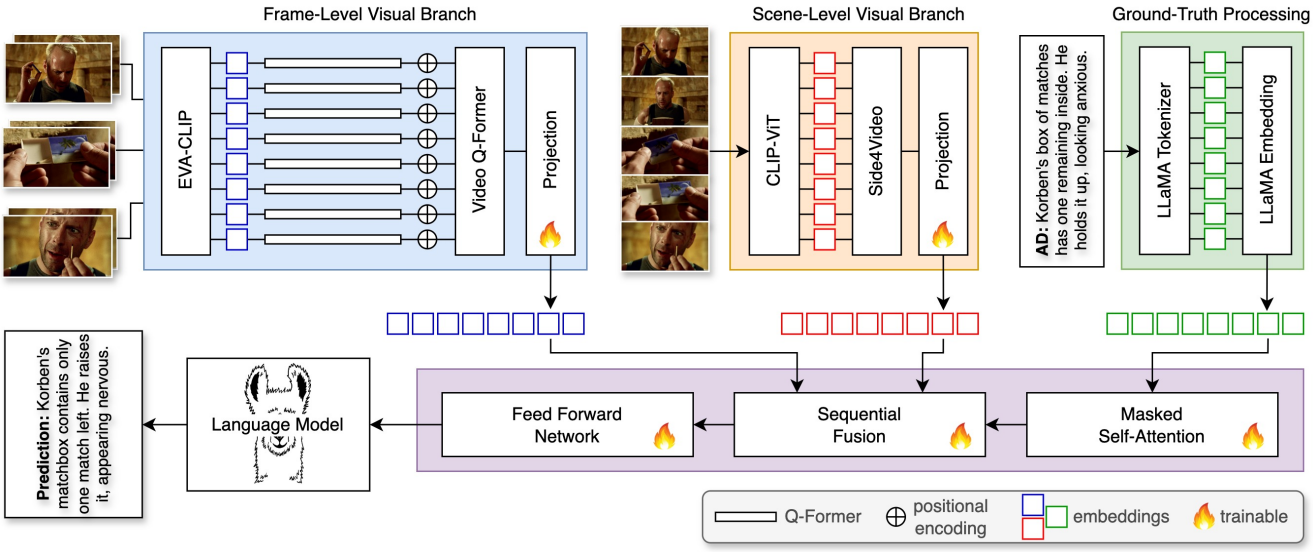

Figure 1. Our audio description pipeline features two visual branches (frame-level in blue, scene-level in red), which are fused in the Dual-Vision Attention Network (purple). The fused embeddings are decoded into natural language AD predictions by the LLaMA model.

coherent narration across frames, ensuring continuity and engagement. [7, 38] reduce redundant descriptions to improve flow, while others [9, 10, 20] enhance perceptual clarity through character identification. Our work extends this by leveraging dual visual embeddings for a more nuanced scene understanding and improved narrative comprehension.

2.2. Multimodal Video Captioning

Integrating additional modalities, such as audio [11, 25], subtitles [14], and motion cues [5], enhances video captioning. Moment localisation [14] and book-to-movie alignment [28, 44] further refine text-video correspondence. Our method introduces a second visual modality, combining frame-based representations with scene-level context, improving both scene understanding and narrative depth.

2.3. Long-Term Modelling

Unlike image captioning [37], video captioning requires capturing temporal dependencies. Sliding window approaches [4] address motion, but most methods [16, 25] focus on short-form datasets [25, 33, 34, 42], limiting long-term coherence. AD generation, however, depends on retaining long-term dependencies to prevent repetition. Recent work explores autoregressive decoding [23], memory clustering [43], and hierarchical captioning [12] for long-form descriptions. Our method enhances scene understanding through dual visual embeddings, improving long-term modelling as video duration increases.

3. Model Architecture

Our model shown in Fig. 1 features two parallel branches for extracting frame-level and scene-level embeddings. A dual-vision Transformer fuses spatial frame details with long-term scene context, and a frozen LLM decodes the fused logits into natural language.

3.1. Frame-Level Visual Branch

Given a film sequence of N frames, we uniformly subsample 8 frames to balance computational efficiency and temporal information. Frame-level dense embeddings are extracted using EVA-CLIP (ViT-G/14) from BLIP-2 [15], refined by a Q-Former through cross-attention with learnable query tokens for enhanced spatial awareness. Positional embeddings encode temporal structure, and the resulting features are flattened into a global video representation processed by the Video Q-Former from Video-LLaMA [41], and pre-trained on HowTo-AD. A linear projection aligns video embeddings with LLaMA2's context window, initialised with Movie-Llama2 [10] weights.

3.2. Scene-Level Visual Branch

To enhance video scene understanding, we extract global sequence representations using Side4Video (S4V) [36], a memory-efficient Transformer model. Specifically, we leverage its action recognition module, which operates a lightweight side network pre-trained on Kinetics-400 [13] alongside CLIP-ViT-B/16 [24]. Subsampled video frames are processed by S4V, where each frame is split into non-overlapping patches and projected into the S4V embedding space. The S4V block integrates temporal convolution,

[CLS] token shift, self-attention, and an MLP layer, capturing rich spatial-temporal information. The final scene-level representation is obtained via Global Average Pooling (GAP) and projected to the LLaMA2 embedding dimension $S \in \mathbb{R}^{B\times 1\times D_L}$.

3.3. Dual-Vision Transformer

Our dual-vision Transformer network integrates long-term AD context within individual video frames by cross-attending to frame and scene-level embeddings via a sequential fusion strategy [21]. It takes as input the frame embeddings, scene embeddings, and ground-truth AD segments. During training, ground-truth AD segments supervise alignment between video embeddings and text descriptions, enabling the model to predict the next AD token conditioned on visual embeddings and preceding tokens. As shown in Fig. 2, the Transformer consists of three layers, each with self-attention, cross-attention, and a Feed-Forward Network (FFN). Ground-truth captions are tokenized and encoded using the pre-trained LLaMA2-7B model, generating an embedded AD sequence autoregressively while conditioning on both visual embeddings and preceding word embeddings. Sinusoidal positional embeddings [30] are added to word embeddings to encode word order, and a causal attention mask ensures autoregressive training.

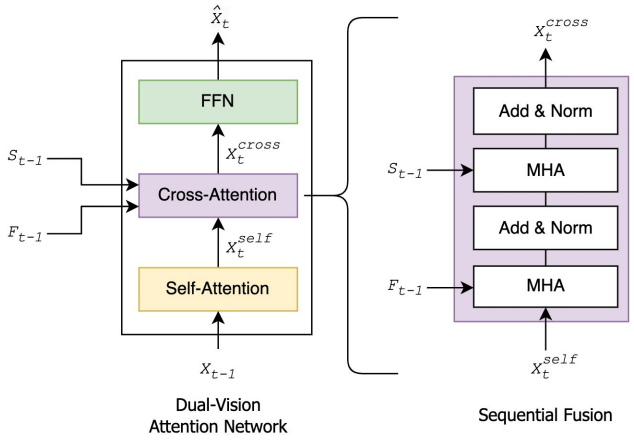

Figure 2. We propose a sequential fusion method within the Dual-Vision Attention Network to integrate frame- and scene-level embeddings. Ground-truth word embeddings are processed using a causal self-attention mask.

Sequential Cross-Attention Fusion. Cross-attention is applied independently to each visual embedding type using two stacked multi-head attention layers, allowing the model to first extract frame-level actions and events before integrating them into a broader scene narrative. The first attention layer processes frame-level embeddings F to capture fine-grained temporal details, while the second layer refines this representation by attending to scene-level embeddings S, incorporating higher-level contextual cues before generating the final description. Within the multi-head attention mechanism, frame or scene-level embeddings act as keys and values, while word embeddings ω serve as queries. During inference, the model omits ground-truth text, initiating autoregressive generation with an embedded `[BOS]` token and recursively feeding its own predictions into the decoder until the sequence is complete.

3.4. Language Model

We use the open-source LLaMA2-7B model for natural language decoding, keeping it frozen during training to preserve generalisation and reduce computational overhead. Autoregressive text generation is facilitated by prepending a `[BOS]` token to mark sequence start and appending a `[EOS]` token for termination, ensuring coherent and bounded outputs.

3.5. Training Details

The Q-Former and Video Q-Former are initialised with 32 queries as in [10, 41]. During training, only the frame- and scene-level projection layers and the dual-vision attention network are updated, while all other components remain frozen with pre-trained weights. Since AD generation is extensively pre-trained on HowTo-AD [10], we fine-tune for just 2 epochs to prevent overfitting. We use AdamW [18] with a learning rate of 3×10^{-5} and a cosine decay schedule for stable convergence. To optimize efficiency, frame- and scene-level embeddings are pre-computed and loaded offline, enabling training on a single RTX-4090 GPU (24GB).

4. Experiments

4.1. Datasets

We train and evaluate our method on the CMD-AD dataset [10], a version of the Condensed Movie Dataset (CMD) [1] adapted for audio description. CMD-AD contains short film scenes (about 2 minutes) annotated with human-generated audio descriptions transcribed using WhisperX [2]. Due to various encoding issues with the raw videos, our dataset is reduced from 101k to 96k AD segments. We will be releasing the audio visual feature embeddings to allow further research on this dataset.

4.2. Evaluation Metrics

We evaluate DANTE-AD using CIDEr, Recall@k/N, and the LLM-based LLM-AD-Eval. CIDEr measures word-level precision, while Recall@k/N uses BertScore to assess the model's top-k predictions. For better handling of long descriptions, LLM-AD-Eval uses LLaMA2-7B-chat and GPT-3.5-turbo to score similarity between predicted and ground-truth descriptions on a 0-5 scale, that we convert to percentages. Despite better alignment with human judgment, the subjective 0-5 scores introduce ambiguity. The prompts for LLM-AD-Eval are provided in Sec. 7.

Method	VLM	LLM	Cr ↑	R@1/5↑	LLM-AD-Eval (%)↑	
					LLaMA	GPT-3.5
Video-BLIP2 [40]	EVA-CLIP	OPT-2.7B	4.8	22.0	31.50	23.33
Video-Llama2 [6]	EVA-CLIP	LLaMA2-7B	5.2	23.6	31.83	23.83
AutoAD-II [9]	CLIP-B32	GPT-2	13.5	26.1	34.66	25.50
AutoAD-III [10]	EVA-CLIP	OPT-2.7B	22.3	29.8	46.33	37.50
AutoAD-III [10]	EVA-CLIP	LLaMA2-7B	25.0	31.2	48.67	**38.17**
DistinctAD [7]	CLIP_{AD}-B16	LLaMA3-8B	22.7	**33.0**	48.00	-
DANTE-AD	EVA-CLIP + S4V	LLaMA2-7B	**28.89**	28.01	**48.83**	34.50

Table 1. Comparisons of AD performance on the CMD-AD dataset. LLM-AD-Eval [10] is evaluated with LLaMA2-7B-chat (left) and GPT-3.5-turbo (right). We report the results for our method DANTE-AD using the sequential fusion of our visual embeddings.

4.3. Quantitative Results

We evaluate our model on the CMD-AD dataset, comparing it with prior AD methods and two non-AD video captioning benchmarks: Video-BLIP2 [40] and Video-LLaMA2 [6]. Trained on CMD-AD with pre-trained weights from HowTo-AD [10], DANTE-AD outperforms all prior methods on CIDEr and LLM-AD-Eval (using LLaMA2-7B), showing superior word-level precision and relative similarity.

4.4. Ablation

Effect of visual embedding ordering. We explore different cross-attention configurations to assess the causal relationship between scene and frame processing. Specifically, we evaluate whether prioritising frame-level embeddings ($F \implies S$) or scene-level context ($S \implies F$) improves performance. Results in Tab. 2 show that processing frame-level information first enhances contextual precision, while prioritising scene-level context generates more detailed descriptions, as seen in Fig. 3(b).

Method	Cr ↑	R@ ↑	LLM-AD-Eval (%) ↑	
			LL	G3.5
$S \implies F$	**35.71**	27.78	44.17	33.33
$F \implies S$	28.89	**28.01**	**48.83**	**34.50**

Table 2. Results of sequential cross-attention in our dual-vision model, comparing the impact of the order of frame- (F) and scene-level (S) embeddings on AD performance. CIDEr and LLM-AD-Eval reported using LLaMA2-7B and GPT-3.5.

4.5. Qualitative Results

Alongside our quantitative results, Fig. 3 compares our method with the ground truth AD. Fig. 3(a) highlights the challenge of using LLM-based similarity metrics, as they yield varying scores despite good alignment with the ground truth. Fig. 3(b) demonstrates the impact of sequential fusion ordering, with our $F \implies S$ approach generating more reasoned and detailed descriptions. Additional examples can be found in section Sec. 6.

Ground Truth: A fly lands on his eyelashes.
DANTE-AD: The camera zooms in on his eyeball.
LLM-AD-Eval: ***GPT-3.5***: 3 / 5 ***LLaMA***: 4 / 5

(a) Limitations of LLM-based evaluation metrics

Ground Truth: The Ferrari drives in the city.
DANTE-AD: (F -> S) He turns the corner and sees a police car parked on the side of the road.
DANTE-AD: (S -> F) They speed down the street, dodging other cars.

(b) Influence of embedding ordering

Figure 3. Qualitative results on CMD-AD-Eval with (a) comparison of quantitative results and (b) the effect of embedding ordering.

5. Conclusion

We introduced **DANTE-AD**, a dual-vision Transformer model that integrates frame- and scene-level representations for generating richer audio description in long-form video. Its multi-stage attention mechanism enhances spatial and temporal understanding, grounding visual details within narrative context. Results show DANTE-AD outperforms existing methods and offers emotive descriptions. We hope our contributions enable more effective automated AD generation solutions, improving accessible media content for the vision-impaired.

References

[1] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 3

[2] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*, 2023. 3

[3] Rupert Bourne, Jaimie D Steinmetz, Seth Flaxman, Paul Svitil Briant, Hugh R Taylor, Serge Resnikoff, Robert James Casson, Amir Abdoli, Eman Abu-Gharbieh, Ashkan Afshin, et al. Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the global burden of disease study. *The Lancet global health*, 9(2):e130–e143, 2021. 1

[4] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2025. 2

[5] Shaoxiang Chen, Wenhao Jiang, Wei Liu, and Yu-Gang Jiang. Learning modality interaction for temporal sentence localization and event captioning in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 333–351. Springer, 2020. 2

[6] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 4

[7] Bo Fang, Wenhao Wu, Qiangqiang Wu, Yuxin Song, and Antoni B Chan. Distinctad: Distinctive audio description generation in contexts. *arXiv preprint arXiv:2411.18180*, 2024. 2, 4

[8] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. Autoad: Movie description in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18930–18940, 2023. 1

[9] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13645–13655, 2023. 2, 4

[10] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. Autoad iii: The prequel-back to the pixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18164–18174, 2024. 2, 3, 4

[11] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 958–959, 2020. 2

[12] Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video recap: Recursive captioning of hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18198–18208, 2024. 2

[13] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[14] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer, 2020. 2

[15] 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 *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[16] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17949–17958, 2022. 1, 2

[17] Jialin Liu, Sam Snodgrass, Ahmed Khalifa, Sebastian Risi, Georgios N Yannakakis, and Julian Togelius. Deep learning for procedural content generation. *Neural Computing and Applications*, 33(1):19–37, 2021. 1

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[19] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9879–9889, 2020. 1

[20] A Nagrani and A Zisserman. From benedict cumberbatch to sherlock holmes: character identification in tv series without a script. In *British Machine Vision Conference, 2017*. British Machine Vision Association and Society for Pattern Recognition, 2017. 2

[21] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Grit: Faster and better image captioning transformer using dual visual features. In *European Conference on Computer Vision*, pages 167–184. Springer, 2022. 3

[22] Ramakanth Pasunuru and Mohit Bansal. Reinforced video captioning with entailment rewards. *arXiv preprint arXiv:1708.02300*, 2017. 1

[23] AJ Piergiovanni, Dahun Kim, Michael S Ryoo, Isaac Noble, and Anelia Angelova. Whats in a video: Factorized autoregressive decoding for online dense video captioning. *arXiv preprint arXiv:2411.14688*, 2024. 2

[24] 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 *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[25] Xuyang Shen, Dong Li, Jinxing Zhou, Zhen Qin, Bowen He, Xiaodong Han, Aixuan Li, Yuchao Dai, Lingpeng Kong, Meng Wang, et al. Fine-grained audible video description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10585–10596, 2023. 2

[26] Donghyuk Shin, Shu He, Gene Moo Lee, Andrew B Whinston, Suleyman Cetintas, and Kuang-Chih Lee. *Enhancing social media analysis with visual data analytics: A deep learning approach*. SSRN Amsterdam, The Netherlands, 2020. 1

[27] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019. 1

[28] Makarand Tapaswi, Martin Bauml, and Rainer Stiefelhagen. Book2movie: Aligning video scenes with book chapters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1827–1835, 2015. 2

[29] The Authority for Television On Demand Limited (ATVoD). Video on demand access services: Best practice guidelines for service providers, 2012. 1

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 3

[31] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015. 1

[32] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6629–6638, 2019. 1

[33] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4581–4591, 2019. 2

[34] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2

[35] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13525–13536, 2024. 1

[36] Huanjin Yao, Wenhao Wu, and Zhiheng Li. Side4video: Spatial-temporal side network for memory-efficient image-to-video transfer learning. *arXiv preprint arXiv:2311.15769*, 2023. 2

[37] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 2

[38] Zeng You, Zhiquan Wen, Yaofo Chen, Xin Li, Runhao Zeng, Yaowei Wang, and Mingkui Tan. Towards long video understanding via fine-detailed video story generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2

[39] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593, 2016. 1

[40] Keunwoo Peter Yu. Videoblip, 2023. 4

[41] 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. 2, 3

[42] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[43] Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18243–18252, 2024. 2

[44] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015. 2

DANTE-AD: Dual-Vision Attention Network for Long-Term Audio Description

Supplementary Material

6. Qualitative Examples

To showcase our results on DANTE-AD, we provide additional qualitative examples in Fig. 4.

Ground Truth: Chucky and Tiffany gaze at each other.
DANTE-AD: They stare at each other in the dim light.

Ground Truth: Richmond fans cheer on their feet, and Charles stuffs another running ball.
DANTE-AD: As the ball flies through the air, the kids cheer and jump up and down.

Ground Truth: Thing one grabs thing two by the feet and pounds them against the couch.
DANTE-AD: He grabs her by the ankles and tosses her in the air.

Ground Truth: His face contorts with emotion, and he swallows as tears well in his eyes.
DANTE-AD: His eyes fill with tears as he stares down at the floor.

Ground Truth: Chased by a Doberman Pinscher through a patio, Alfalfa dives into an Olympic-sized swimming pool.
DANTE-AD: He dives into the water and swims to the edge of the pool.

Ground Truth: O'Reilly fires his gun and an explosion rips through the house sending Moses flying.
DANTE-AD: The explosion blows the roof off the house and sends debris flying.

Figure 4. Qualitative results of our DANTE-AD method on CMD-AD-Eval.

7. LLM-AD-Eval Prompts

For accurate comparison with previous methods, we use the same prompts for LLM-AD-Eval as provided in [10]. For reference, the prompt for LLM-AD-Eval using LLaMA2-7B-chat is given in Fig. 5 and for GPT-3.5-turbo in Fig. 6.

Please evaluate the following movie audio description pair:
- Correct Audio Description: {*ground-truth AD segment*}
- Predicted Audio Description: {*predicted AD segment*}

Provide your evaluation only as a matching score where the matching score is an integer value between 0 and 5, with 5 indicating the highest level of match.

Please generate the response in the form of a Python dictionary string with keys 'score', where its value is the matching score in INTEGER, not STRING.

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: {'score': }.

Figure 5. LLM-AD-Eval prompt for evaluations using LLaMA2-7B-chat from [10].

System:

You are an intelligent chatbot designed for evaluating the quality of generative outputs for movie audio descriptions. Your task is to compare the predicted audio descriptions and determine its level of match, considering mainly the visual elements like actions, objects and interactions. Here's how you can accomplish the task:

Instructions:

- Check if the predicted audio description covers the main visual elements from the movie, especially focusing in the verbs and nouns.
- Evaluate whether the predicted audio description includes specific details rather than just generic points. It should provide comprehensive information that is tied to specific elements of the video.
- Consider synonyms or paraphrases as valid matches. Consider pronouns like 'he' or 'she' as valid matches with character names. Consider different character names as valid matches.
- Provide a single evaluation score that reflects the level of match of the prediction, considering the visual elements like actions, objects and interactions.

User:

Please evaluate the following movie description pair:
- Correct Audio Description: {*ground-truth AD segment*}
- Predicted Audio Description: {*predicted AD segment*}

Provide your evaluation only as a matching score where the matching score is an integer value between 0 and 5, with 5 indicating the highest level of match. Please generate the response in the form of a Python dictionary string with keys 'score', where its value is the matching score in INTEGER, not STRING.

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: {'score': }.

Figure 6. LLM-AD-Eval prompt for evaluations using GPT-3.5-turbo from [10].