– Supplemental Document –

Now You See Me: Context-Aware Automatic Audio Description

Seon-Ho Lee* Korea University

Jue Wang Amazon AGI

David Fan† Meta FAIR

Zhikang Zhang Amazon AGI

Linda Liu Amazon Prime Video

Xiang Hao Amazon Prime Video

Vimal Bhat Amazon Prime Video

Xinyu Li Amazon AGI

1. More Implementation Detail

1.1. Detector Architecture

Context-aware Feature Enhancement: The architecture of the context-aware feature enhancement module h_{S4} is detailed in Figure S-1. For each S4 layer, the numbers within the brackets denote the output channels and dropout rate, respectively. For each dropout layer, the number within the brackets indicates the dropout rate. Also, for each linear layer, the number within brackets denotes the output channels. The feature enhancement module follows the Eq (3) in the main paper, which leverages the temporal capacity of the S4 model to capture longer temporal context into the short clip. Different from ViS4mer [4] and its variant [6] that reduce the sequence length by pooling or additional selection module, we enhance the short clip features (length of N) with longer context (length of N') and use enhanced feature for the following tasks, which further improve the effectiveness.

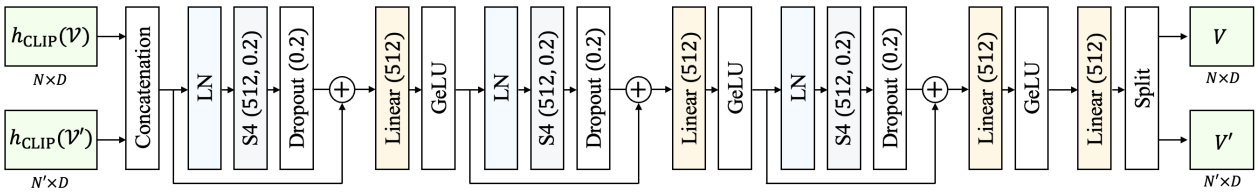


Figure S-1. An architecture of the feature enhancement module h_{S4}. LN indicates the layer normalization [1].

Detection-Specific Encoder: Figure S-2 illustrates the detailed architecture of the detection-specific encoder h_{det}. We adopt the same architecture for h^{w}_{det} in the language guided detection scheme.

MLP: Figure S-3 illustrates the architecture of prediction head in the detector, which is implemented by MLP.

1.2. Generator Architecture

Generation-Specific Encoder: Figure S-4 illustrates the detailed architecture of the generation-specific encoder f_V. Note that the generation-specific encoder takes the suppressed feature vector V_{sup} as its input and yields the prompt vector Z, which guides GPT2 to generate the desirable scripts. We construct the encoder with S4 layers and the perceiver [**?**] with the learnable latent vectors V_{lat}. We use the same architecture for $f_{V'}$ as well.

*Work done during an internship at Amazon Prime Video.
†Work done while at Amazon Prime Video.

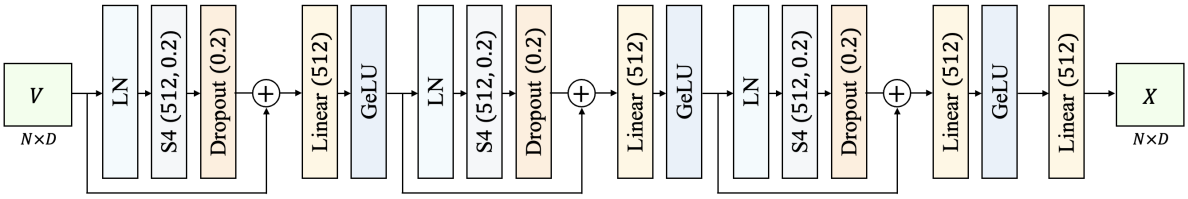


Figure S-2. An architecture of the detection-specific encoder h_{det}.

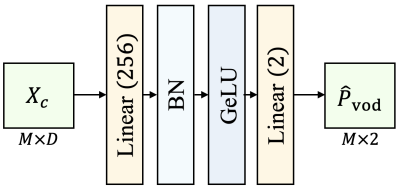


Figure S-3. An architecture of the MLP prediction head in the proposed detector. BN denotes batch normalization [**?**].

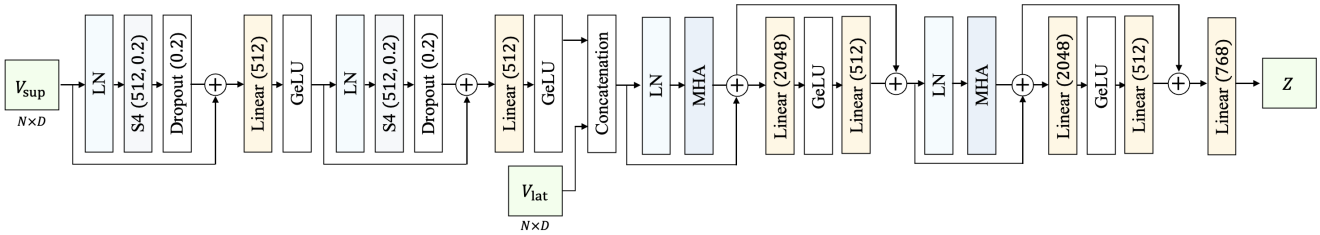


Figure S-4. An architecture of the generator f_{V}. MHA denotes the multi-head attention.

1.3. Pretraining

We pretrain the GPT2 on the AudioVault dataset [3]. Note that the AudioVault dataset contains the AD scripts only. Hence, for pretraining, we sample two consecutive AD scripts as in [3]. Then, the parameters of GPT2 are optimized so that it can predict the later AD script from the former AD script reliably. We employ AdamW optimizer [5] with a batch size of 128. The cosine learning rate scheduler is initialized at 0.0001 and the network is trained for 5 epochs with a linear warm-up period of 30,000 steps.

1.4. Training

During the training of $\mathrm{CA}^3\mathrm{D}$, we freeze h_{CLIP} and f_{GPT2}. Note that, to process the entire movie, we sample the movie clips using sliding window approach and then perform the AD event detection and AD script generation for each movie clip. However, some movie clips, such as the scenes with dialogues, include no AD event. To handle these movie clips properly, the detector should be trained on the movie clips with no AD event during training. Hence, for training, we also use the movie clips with no AD event, which are sampled from the movies in the training dataset. In such cases, the ground-truth probabilities P_{vod} and P_{lgd} for ℓ_{focal} are all zeros.

1.5. Inference

To process the entire movie $\mathcal{M}$, we sample the movie clips $\{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_T\}$ using a sliding window approach with step size S. If $S < N$, some frames are overlapped between adjacent sampled movie clips. We define the overlap ratio as

$$R = \frac{N - S}{N}. \tag{1}$$

At overlap ratio R, a frame can belong to $\lceil \frac{1}{R} \rceil$ different movie clips at most. In other words, we may have multiple detection results for some frames. In such cases, we select the detection result with the highest probability as our prediction. We set $R = 0.75$ by default.

2. More Experiments

2.1. Analysis

Analysis on N: Table S-1 compares the performances of AD event detection and AD script generation at different N's on the MADv2 dataset. Note that N is the length of the movie clip $\mathcal{V}$. In all tests, the best results are achieved at $N = 32$. Therefore, we use $N = 32$ as the default option.

Table S-1. Comparison of AD generation and detection performances at different N's.

	Detection			Generation	
N	Precision	Recall	F1	Rouge-L	CIDEr
24	47.6	81.6	58.3	11.2	9.1
32	54.4	82.2	65.4	11.3	9.4
40	52.8	81.6	64.1	11.2	9.0
48	51.9	62.3	56.6	10.7	8.5

Analysis on N': Table S-2 compares the performances of AD event detection and AD script generation with different N' on the MADv2 dataset. At both $N' = 32$ and $N' = 96$, the detection and generation performances degrade. Especially, CIDEr score is dropped by 0.4 at $N' = 96$, indicating that too much context information makes the training difficult. We set $N' = 64$ by default.

Table S-2. Comparison of AD generation and detection performances according to N'.

	Detection			Generation	
N'	Precision	Recall	F1	Rouge-L	CIDEr
32	51.1	83.2	62.9	10.8	9.3
64	54.4	82.2	65.4	11.3	9.4
96	54.2	77.7	63.9	10.8	8.9

Analysis on k: For feature suppression, we use k candidates with the highest probabilities among $\hat{P}_{\text{vod}}$ in (10). Table S-3 compares the results at different k's. In all tests, the best scores are obtained at $k = 1$. Therefore, we use $k = 1$ as the default option. Also, at $k = 50$, the script generation performances degrade severely. It is because the unreliable detection candidates affect the feature suppression.

Table S-3. Comparison of AD generation and detection performances at different k's.

	Detection			Generation	
k	Precision	Recall	F1	Rouge-L	CIDEr
1	54.4	82.2	65.4	11.3	9.4
10	53.8	79.6	64.1	11.3	9.3
50	52.8	78.7	63.2	11.0	8.9

Complexity: Table S-4 compares the throughput of the proposed algorithm and conventional algorithms during inference. CA^{3}D shows similar FPS with the others, although it performs two tasks. It is because most computation comes from GPT2.

Table S-4. Comparison of throughput.

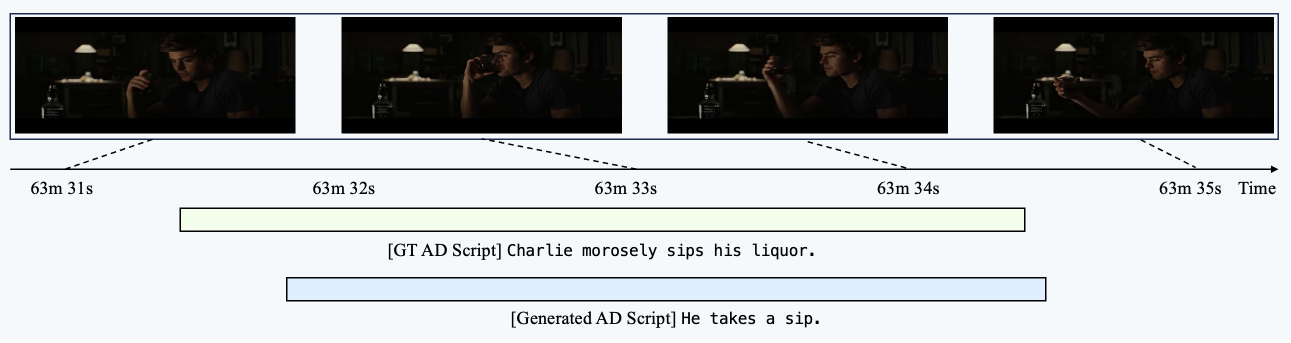
Method	AutoAD [3]	AutoAD II [2]	CA^{3}D
FPS	8.55	4.61	6.23

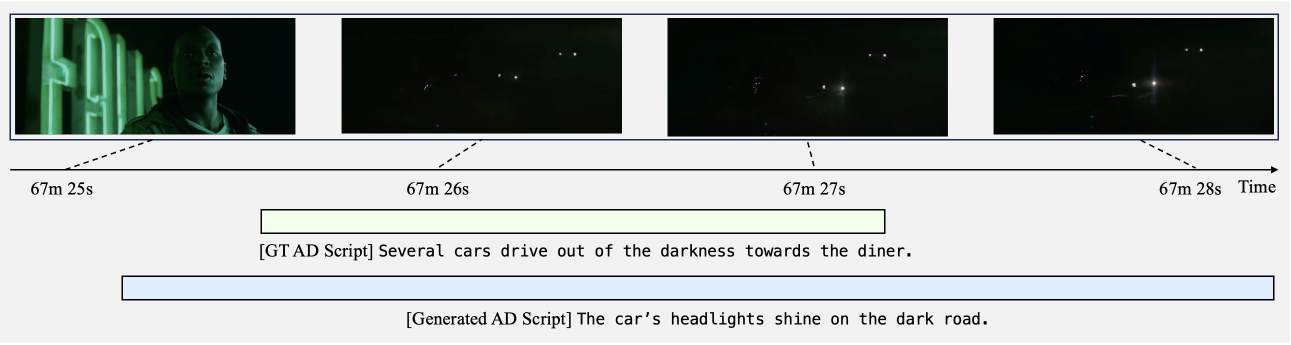
3. Limitation and Future Work

In Figure 3, the self-refinement module improves the results meaningfully at the iteration 1. However, at the iteration 2, the performances saturate because the refined results may not strong enough to make further changes. Hence, from the iteration 2, the performance gains are marginal especially compared to additional computation costs. Also, in AD script generation, it is important to understand the long-term movie context. However, $\mathrm{CA}^3\mathrm{D}$ divides the entire movie into the clips and processes them one by one. Thus, the range of movie context which the proposed algorithm can recognize is limited. To alleviate this, $\mathrm{CA}^3\mathrm{D}$ exploits visual context by using the feature enhancement module, but it is still required to extend the recognizable range of movie context. We leave this for future work.

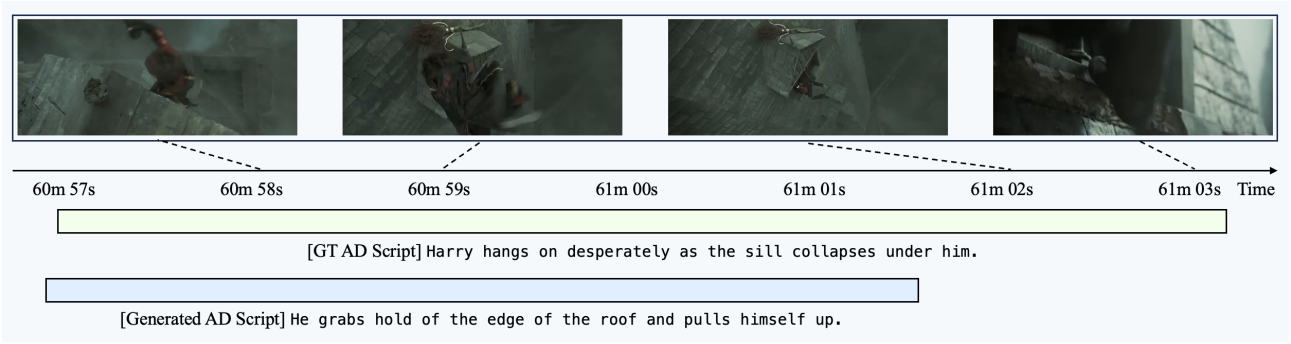
4. More Visualizations

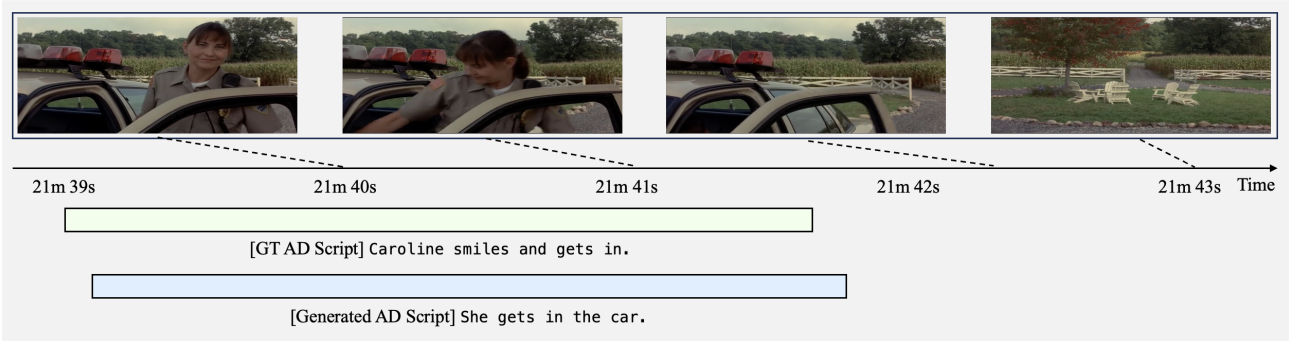
Figure S-5 shows more examples of AD event detection and AD script generation results on the MADv2 dataset.





(a)





(b)

Figure S-5. Examples of AD detection and generation results on the MADv2 test set.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 1

[2] 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*, 2023. 3

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

[4] Md Mohaiminul Islam and Gedas Bertasius. Long movie clip classification with state-space video models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[6] Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, and Raffay Hamid. Selective structured state-spaces for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6387–6397, 2023. 1