Learning Audio-guided Video Representation with Gated Attention for Video-Text Retrieval
— *Supplementary Materials* —

Boseung Jeong[1] Jicheol Park[2] Sungyeon Kim[1] Suha Kwak[1,2]

[1]Dept. of CSE, POSTECH [2]Graduate School of AI, POSTECH

{boseung01, jicheol, sungyeon.kim, suha.kwak}@postech.ac.kr

http://cvlab.postech.ac.kr/research/AVIGATE

This supplementary material provides additional details of the audio resampler and experimental results, which we could not include in the main paper. We first describe the details of the audio resampler that reduces the number of audio embeddings to a fixed length by employing a query-based transformer [1, 3, 8] and the details of MLP in Eq. (4) of the main paper in Sec. A. We also provide implementation details of our method in Sec. B. We then present further quantitative results, including the effect of post-processing and entire video-text retrieval results on VATEX [15] and Charades [13] in Sec. C. Moreover, we conduct additional ablation studies on hyperparameters, such as the layer depth of the gated fusion transformer, the scaling factor, the maximum margin in Eq. (5) of the main paper and the type of the gate mechanism, and the effect of freezing modality encoders in Sec. D. Lastly, we provide more qualitative results, further illustrating the effectiveness of AVIGATE in Sec. E.

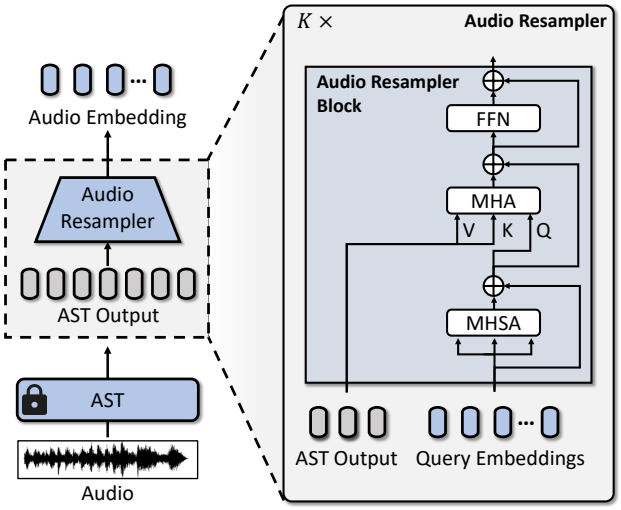

Figure A1. The overall architecture of audio resampler.

A. More Architectural Details

To efficiently fuse audio embeddings with frame embeddings while reducing computational overhead, we introduce an audio resampler using a query-based transformer framework [1, 3, 8] that utilizes a cross-attention mechanism with M learnable query embeddings. Specifically, the audio input is fed into Audio Spectrogram Transformer (AST) [6], and the output is then passed to the audio resampler to reduce the number of audio embeddings to a fixed length of M while preserving essential information. As shown in Figure A1, the audio resampler comprises K audio resampler blocks, each with multi-head self-attention (MHSA), multi-head cross-attention (MHA), and a feed-forward network (FFN). We set K as 4 for default.

MHSA first allows the learnable query embeddings to interact and capture contextual relationships among themselves, refining their initial representations. This is followed by MHA, where the query embeddings attend to the output of AST, extracting audio embeddings with a fixed length of M. The FFN then processes the audio embeddings to refine them. This sequence of operations enables the audio resampler to reduce the number of audio embeddings efficiently while preserving critical information, facilitating seamless fusion with the frame embeddings in subsequent stages.

The MLP in Eq. (4) consists of two layers with dimensions $\mathbb{R}^{2D\times D/2}$ and $\mathbb{R}^{D/2\times 1}$, using a QuickGELU as the non-linearity between them.

B. More Implementation Details

The details of the training configurations of our method across datasets are provided in Table A1. We follow [10, 12] for most configurations, such as the image encoder, training epochs, optimizer, batch size, max frames, max words, learning rate for CLIP encoders, and temperature τ.

Source dataset	MSR-VTT [17]	VATEX [15]	Charades [13]
Image encoder	2 CLIP-ViTs (B/32 and B/16)		
Total epochs	5		
Optimizer	Adam [9]		
Embedding dimension D	512		
Batch size	128	128	64
Max frames	12	12	32
Max words	32	32	64
Resampled audio length	12		
Depth of Gated Fusion Transformer L	4		
Learning rate for Non-CLIP parameters	1e−4	1e−4	5e−4
Learning rate for CLIP encoders	1e−7		
Temperature τ in Eq.(6)	Learnable (After training: 0.01)		
Maximum margin δ in Eq.(5)	0.1	0.05	0.1
Scaling factor λ in Eq.(5)	0.2		
Scaling factor α in Eq.(8)	50		

Table A1. Training configurations of various datasets.

Methods	Modality	Text-to-Video Retrieval			Video-to-Text Retrieval			RSum
		R@1	R@5	R@10	R@1	R@5	R@10	
CLIP ViT-B/32								
CAMoE [4]	V+T	44.6	72.6	81.8	45.1	72.4	83.1	399.6
+DSL	V+T	47.3 (+2.7)	74.2 (+1.6)	84.5 (+2.7)	49.1 (+4.0)	74.3 (+1.9)	84.3 (+1.2)	413.7 (+14.1)
TS2-Net [11]	V+T	47.0	74.5	83.8	45.3	74.1	83.7	408.4
+DSL	V+T	51.1 (+4.1)	76.9 (+2.4)	85.6 (+1.8)	-	-	-	-
UATVR [5]	V+T	47.5	73.9	83.5	46.9	73.8	83.8	409.4
+DSL	V+T	49.8 (+2.3)	76.1 (+2.2)	85.5 (+2.0)	51.1 (+4.2)	74.8 (+1.0)	85.1 (+1.3)	422.4 (+13.0)
UCoFiA [16]	V+T	48.2	73.3	82.3	-	-	-	-
+SK norm	V+T	49.4 (+1.2)	72.1 (-0.9)	83.5 (+1.2)	47.1	74.3	83.0	409.4
AVIGATE (Ours)	A+V+T	50.2	74.3	83.2	49.7	75.3	83.7	416.4
+DSL	A+V+T	**53.9** (+3.7)	**77.0** (+2.7)	**86.0** (+2.8)	**53.0** (+3.3)	**78.2** (+2.9)	**85.4** (+1.7)	**433.5** (+16.9)
CLIP ViT-B/16								
TS2-Net [11]	V+T	49.4	75.6	85.3	46.6	75.9	84.9	417.7
+DSL	V+T	54.0 (+4.6)	79.3 (+3.7)	87.4 (+2.1)	-	-	-	-
TEFAL [7]	A+V+T	49.9	76.2	84.4	-	-	-	-
+DSL+QB-Norm	A+V+T	52.0 (+2.1)	76.6 (+0.4)	86.1 (+1.7)	-	-	-	-
UATVR [5]	V+T	50.8	76.3	85.5	48.1	76.3	85.4	422.4
+DSL	V+T	53.5 (+2.7)	79.5 (+3.2)	88.1 (+2.7)	54.5 (+6.4)	79.1 (+2.8)	**87.9** (+2.5)	442.6 (+20.2)
AVIGATE (Ours)	A+V+T	52.1	76.4	85.2	51.2	77.9	86.2	429.0
+DSL	A+V+T	**56.3** (+4.2)	**80.8** (+4.4)	**88.1** (+2.9)	**57.4** (+6.2)	**80.2** (+2.3)	87.4 (+1.2)	**450.2** (+21.2)

Table A2. Text-to-video and video-to-text retrieval results on the MSR-VTT 9k split. The post-processing techniques such as DSL [4], QB-Norm [2], and SK norm are used for further performance boosting.

C. More Quantitative Results

Effect of Post-Processing: Post-processing techniques have been widely adopted in video-text retrieval to enhance performance by refining similarity scores. Previous methods [4, 5, 7, 11, 16] adopt the post-processing techniques, including Dual Softmax Loss (DSL) [4], Querybank Norm (QB-Norm) [2], and the Sinkhorn-Knopp algorithm (SK-Norm), for further improvements in retrieval accuracy. We also explore the effect of the post-processing technique by adopting DSL that applies inverted softmax [14] during inference. We report the retrieval performance of AVIGATE with and without post-processing in Table A2 compared with existing methods. Our model, AVIGATE, consistently achieves superior performance across all evaluation metrics for both text-to-video and video-to-text retrieval tasks, outperforming all previous methods by significant margins. Specifically, for the CLIP ViT-B/32 backbone, AVIGATE with post-processing achieves R@1 of 53.9% for text-to-video retrieval. Furthermore, in video-to-text retrieval, AVIGATE with DSL achieves R@1 of 53.0%. Similarly, for the CLIP ViT-B/16 backbone, AVIGATE achieves substantial gains over existing methods. When using post-processing, our method achieves R@1 of 56.3% for text-to-video retrieval, representing a considerable 2.3%p improvement over TS2-Net [11]. In video-to-text retrieval, AVIGATE also outperforms other methods with R@1 of 57.4%.

Entire Performance on VATEX [15] and Charades [13]: We present the complete video-text retrieval results on VA-

Methods	Modality	Text-to-Video Retrieval			Video-to-Text Retrieval			RSum
		R@1	R@5	R@10	R@1	R@5	R@10	
CLIP ViT-B/32								
AVIGATE (Ours)	A+V+T	63.1	90.7	95.5	76.6	97.3	98.8	522.0
+DSL	A+V+T	**70.7** (+7.6)	**93.4** (+2.7)	**95.5** (+1.4)	**85.3** (+8.7)	**99.1** (+1.8)	**99.8** (+1.0)	**545.2** (+23.2)
CLIP ViT-B/16								
AVIGATE (Ours)	A+V+T	67.5	93.2	96.7	80.7	97.8	99.5	535.4
+DSL	A+V+T	**74.6** (+7.1)	**95.3** (+2.1)	**97.8** (+1.1)	**88.7** (+8.0)	**99.3** (+1.5)	**99.9** (+0.3)	**555.6** (+20.2)

Table A3. Text-to-video and video-to-text retrieval results on VATEX. The post-processing technique, DSL [4], is used for further performance boosting.

Methods	Modality	Text-to-Video Retrieval			Video-to-Text Retrieval			RSum
		R@1	R@5	R@10	R@1	R@5	R@10	
CLIP ViT-B/32								
AVIGATE (Ours)	A+V+T	18.8	40.0	51.8	17.2	40.4	51.7	219.9
+DSL	A+V+T	**21.3** (+2.5)	**42.4** (+2.4)	**54.4** (+2.7)	**20.0** (+2.8)	**43.0** (+2.6)	**54.9** (+3.2)	**236.0** (+16.1)
CLIP ViT-B/16								
AVIGATE (Ours)	A+V+T	24.1	48.5	61.3	22.9	48.4	61.0	266.2
+DSL	A+V+T	**27.5** (+3.4)	**52.7** (+4.2)	**64.5** (+3.2)	**27.1** (+4.2)	**52.7** (+4.3)	**65.0** (+4.0)	**289.5** (+23.3)

Table A4. Text-to-video and video-to-text retrieval results on Charades. The post-processing technique, DSL [4], is used for further performance boosting.

TEX and Charades in Table A3, including both text-to-video and video-to-text retrieval. The results are reported using two variants of the CLIP ViT backbone, CLIP ViT-B/32 and CLIP ViT-B/16. Moreover, we assess the effect of the post-processing technique, DSL [4], for further performance boosts. On VATEX, with the CLIP ViT-B/32 backbone, AVIGATE achieves notable results in text-to-video retrieval, with R@1 of 63.1% and 76.6% for text-to-video retrieval and video-to-text retrieval, respectively. When applying DSL, we observe significant improvements across all metrics. Specifically, it improves AVIGATE by a large margin, 7.6%p and 8.7%p in R@1 for text-to-video retrieval and video-to-text retrieval, respectively. When using the larger backbone, CLIP ViT-B/16 backbone, AVIGATE demonstrates the scalability across different backbone sizes, achieving R@1 of 67.5% for text-to-video retrieval and 80.7% for video-to-text retrieval. Moreover, the use of DSL consistently boosts the retrieval accuracy overall, with 20.2%p improvements in RSum. On Charades, with the CLIP ViT-B/32 backbone, AVIGATE achieves R@1 of 18.8% in text-to-video retrieval and 17.2% in video-to-text retrieval, which modestly increase to 21.3% and 20.0% when DSL is applied. Employing the larger backbone, CLIP ViT-B/16, AVIGATE attains R@1 of 24.1% in text-to-video retrieval and 22.9% in video-to-text retrieval, with DSL boosting these figures to 27.5% and 27.1%.

D. More Ablation Studies

We further conduct ablation studies using varying hyperparameters in AVIGATE. Similar to the main paper, we report text-to-video retrieval results on the MSR-VTT dataset [17] with CLIP ViT-B/32. Table A5 presents the whole results of the ablation studies.

Layer Depth of Gated Fusion Transformer: We present the impact of the number of layers of the gated fusion transformer (L) in Table A5(a) and observe that the performance gradually improves up to L=4, where the best performance is achieved.

Hyperparameters λ and δ in Eq. (5): We investigate the impact of the scaling factor λ and the maximum margin δ in Eq. (5) of the manuscript. It is worth noting that the adaptive margin in Eq. (5) becomes 0 when λ or δ are set to 0, leading the loss in Eq. (6) to the conventional contrastive loss. As shown in Table A5(b), when λ is set to 0.2, the model yields the best performance. Meanwhile, setting λ to 0.1 results in a slight decrease in performance, indicating that a smaller scaling factor may not provide sufficient margin adjustment. However, increasing λ to 0.3 does not lead to further improvements. Similarly, Table A5(c) presents the effect of varying the maximum margin δ. We observe that the performance gradually improves up to $\delta = 0.1$. Increasing δ beyond 0.1 degrades performance due to excessively large margins pushing negative pairs too far apart.

Gate Mechanism Type: Our method employs a soft gate mechanism, which allows for continuous modulation of the contribution of audio during fusion. To evaluate the effectiveness of the soft gate mechanism, we compare the soft gate with a hard gate mechanism, which assigns a gating score of 1 if it exceeds a predefined threshold and 0 otherwise. As shown in Table A5(d), using the hard gate under-

performs our method. Unlike using the hard gate mechanism, our method facilitates the effective use of relevant audio cues while minimizing the impact of irrelevant or noisy audio signals; it enables the model to leverage informative audio more precisely, thereby improving retrieval accuracy.

Effect of freezing AST: We freeze AST to reduce training costs. Fine-tuning AST is impractical since it processes 1,214 tokens per input audio, far more than 50 tokens for each video frame in ViT-B/32. A solution is to largely reduce the batch size, which however degrades performance since the contrastive loss is highly dependent on the batch size. The results of freezing and fine-tuning AST with tiny input batches are reported in Table A5(e), while freezing AST outperforms fine-tuning it. The results are attributed to the characteristics of AST pre-trained on the audio classification dataset, allowing it to extract discriminative embeddings from audio inputs. Therefore, we decided to freeze the AST instead of fine-tuning that requires a burden of computational and memory costs.

Freezing both CLIP image and text encoders: As shown in Table A5(f), freezing the CLIP image and text encoders leads to noticeably lower performance, highlighting the importance of fine-tuning both encoders, as also demonstrated in prior work such as CLIP4Clip [12]. Fine-tuning is essential for capturing task-specific video and text information and improving the alignment between them.

Ablated Setting	Text-to-Video Retrieval R@1	R@5	R@10
(a) Layer depth of Gated Fusion Transformer: L			
L=1	49.0	74.0	82.6
L=2	49.8	74.0	83.0
L=4	50.2	74.3	83.2
L=6	49.5	74.2	82.6
(b) Scaling factor in Eq.(5): λ			
λ=0.0	48.0	75.1	83.4
λ=0.1	49.4	74.8	83.8
λ=0.2	50.2	74.3	83.2
λ=0.3	50.0	74.4	83.2
(c) Maximum margin in Eq.(5): δ			
δ=0.00	48.0	75.1	83.4
δ=0.05	49.4	75.1	83.6
δ=0.10	50.2	74.3	83.2
δ=0.15	49.3	74.8	83.8
δ=0.20	48.3	74.4	83.9
(d) Gate mechanism type			
Hard Gate	49.3	75.0	82.5
Soft Gate	50.2	74.3	83.2
(e) Effect of freezing AST (Batch size:32)			
Freezing	48.2	75.3	83.7
Fine-tuning	48.0	73.5	83.4
(f) Effect of freezing CLIP encoders			
Freezing	41.1	68.5	78.2
Fine-tuning	50.2	74.3	83.2

Table A5. Ablation studies on hyperparameters. gray corresponds to our default setting.

E. More Qualitative Results

We further present additional qualitative results that illustrate the effectiveness of AVIGATE in leveraging audio information for text-to-video retrieval. Figure A2 shows the Top-1 retrieved videos from our method, including the corresponding audio signals, to highlight how audio cues influence retrieval outcomes.

In Figure A2(a) and (b), we present a scenario where the audio provides valuable information that enables improving retrieval performance. AVIGATE, which incorporates audio through the gated fusion transformer, successfully retrieves the correct video corresponding to the text query. In contrast, the method without audio information (*i.e.*, w/o Audio) fails to retrieve the true matches. This comparison highlights the benefit of utilizing informative audio cues.

Conversely, Figure A2(c) and (d) present another scenario where the audio input contains irrelevant information, such as background noise. AVIGATE effectively filters out the uninformative audio signals through the gating mechanism. The gating function assigns low gating scores, allowing the model to focus only on the visual cues. As a result, AVIGATE successfully retrieves the correct videos. In contrast, the method without the gating function (*i.e.*, w/o Gate) is impacted by the noisy audio and fails to retrieve the true matches.

These qualitative results demonstrate that the gated fusion transformer successfully filters out irrelevant audio while leveraging valuable audio information when the audio contributes positively.

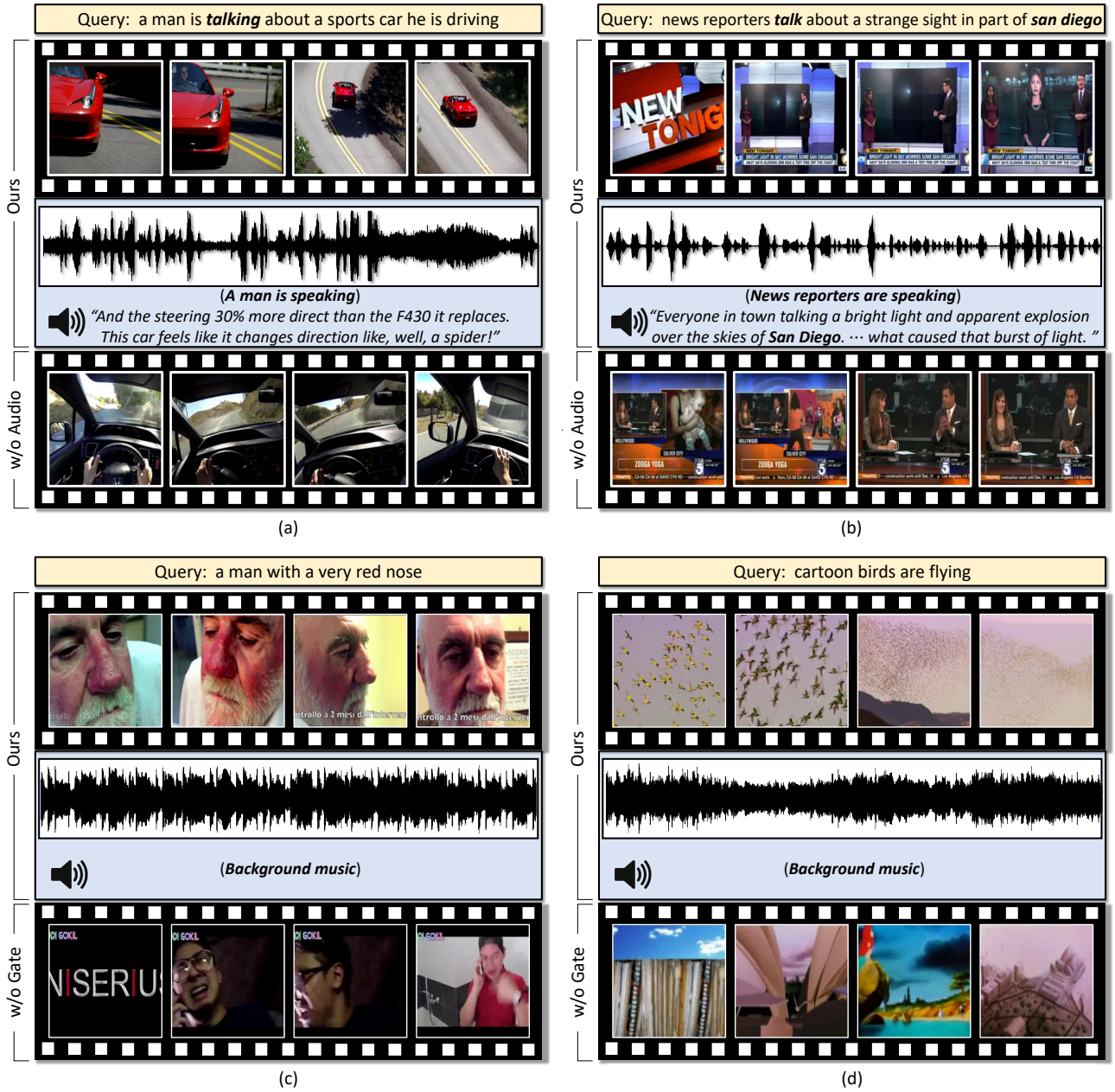

Figure A2. Top-1 text-to-video retrieval results of our method on MSR-VTT, where they are true matches. The audio provides informative cues for accurate retrieval, where "a man is talking" in the query text is not visible (a) and "talk· · · san diego" in the query text is not visible but audible (b). However, neglecting these informative audio signals (*i.e.*, w/o Audio) fails to retrieve true matches. Meanwhile, the irrelevant audio is filtered by the gated fusion transformer, leading to accurate retrieval results (c) and (d); without the gating mechanism (*i.e.*, w/o Gate), it leads to retrieving false matches due to the irrelevant audio.

References

[1] 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. *Proc. Neural Information Processing Systems (NeurIPS)*, 2022. 1

[2] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 1

[4] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021. 2, 3

[5] Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Yuxin Song, Weiping Wang, Xiangbo Shu, Xiangyang Ji, and Jingdong Wang. Uatvr: Uncertainty-adaptive text-video retrieval. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2023. 2

[6] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech*, 2021. 1

[7] Sarah Ibrahimi, Xiaohang Sun, Pichao Wang, Amanmeet Garg, Ashutosh Sanan, and Mohamed Omar. Audio-enhanced text-to-video retrieval using text-conditioned feature alignment. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2023. 2

[8] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proc. International Conference on Machine Learning (ICML)*, 2021. 1

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2015. 2

[10] Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. Eclipse: Efficient long-range video retrieval using sight and sound. In *Proc. European Conference on Computer Vision (ECCV)*, 2022. 1

[11] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *Proc. European Conference on Computer Vision (ECCV)*, 2022. 2

[12] 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, 2022. 1, 4

[13] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proc. European Conference on Computer Vision (ECCV)*, 2016. 1, 2

[14] Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proc. International Conference on Learning Representations (ICLR)*, 2017. 2

[15] 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 *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[16] Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Unified coarse-to-fine alignment for video-text retrieval. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2023. 2

[17] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3