DiffusionRet: Generative Text-Video Retrieval with Diffusion Model Supplementary Material

Peng Jin[1,3*] Hao Li[1,3*] Zesen Cheng[1,3] Kehan Li[1,3] Xiangyang Ji[4]
Chang Liu[4] Li Yuan[1,2,3†] Jie Chen[1,2,3†]

[1]School of Electronic and Computer Engineering, Peking University, Shenzhen, China [2]Peng Cheng Laboratory, Shenzhen, China
[3]AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, Shenzhen, China
[4]Department of Automation and BNRist, Tsinghua University, Beijing, China

{jp21, cyanlaser, kehanli}@stu.pku.edu.cn {lihao1984, yuanli-ece}@pku.edu.cn {liuchang2022, xyji}@tsinghua.edu.cn chenj@pcl.ac.cn

A. Appendix

This appendix provides the descriptions of datasets (Sec. A.1.1), implementation details (Sec. A.1.2), video-to-text retrieval performance on the LSMDC, MSVD, and ActivityNet Captions datasets (Sec. A.2.1), additional experiments for the out-domain retrieval (Sec. A.2.2), the reason for applying diffusion models (Sec. A.2.3), discussion of the limitations (Sec. A.2.4), the additional visualization of the diffusion process (Sec. A.3.1), the visualization of the text-frame attention map (Sec. A.3.2), and the visualization of the text-to-video retrieval examples (Sec. A.3.3).

A.1. Datasets and Implementation Details

A.1.1 Datasets

We compare the proposed DiffusionRet with other methods on five benchmark text-video retrieval datasets, including MSRVTT [37], LSMDC [33], MSVD [5], ActivityNet Captions [18], and DiDeMo [2].

MSRVTT. MSRVTT [37] contains 10,000 YouTube videos, each with 20 text descriptions. We follow the training protocol in [26, 10, 28] and evaluate on text-to-video and video-to-text search tasks on the 1K-A testing split with 1K video or text candidates defined by [39].

LSMDC. LSMDC [33] contains 118,081 video clips from 202 movies. The duration of videos in the LSMDC dataset is short. We follow the split of [10] with 1,000 videos for testing.

MSVD. MSVD [5] contains 1,970 videos. Each video has approximately 40 associated text description. Videos in the MSVD dataset are short in duration, lasting about 10 to 25 seconds. We follow the official split of 1,200 and 670 as the train and test set, respectively.

ActivityNet Captions. ActivityNet Captions [18] consists densely annotated temporal segments of 20K YouTube videos. Following [10, 29, 36], we concatenate descriptions of segments in a video to construct "video-paragraph" for retrieval. We report results on the "val1" split of 10,009 and 4,917 as the train and test set.

DiDeMo. DiDeMo [2] contains 10,464 videos annotated 40,543 text descriptions. We concatenate descriptions of segments in a video to construct "video-paragraph" for retrieval. We follow the training and evaluation protocol in [27].

A.1.2 Implementation Details.

Following previous works [27, 13, 14, 15], we utilize the CLIP (ViT-B/32) [31] as the pre-trained model. The dimension of the feature is 512. The temporal transformer [35, 23] is composed of 4-layer blocks, each including 8 heads and 512 hidden channels. The temporal position embedding [38] and parameters are initialized from the text encoder of the CLIP. We use the Adam optimizer [16] and set the batch size to 128. The initial learning rate is 1e-7 for the text encoder and video encoder and 1e-3 for other modules. We set the temperature $\hat{\tau}$ to 0.01 and $\tau^{'}$ to 1. For short video datasets, *i.e.*, MSRVTT, LSMDC, and MSVD, the word length is 32 and the frame length is 12. For long video datasets, *i.e.*, ActivityNet Captions and DiDeMo, the word length is 64 and the frame length is 64.

The training is divided into two stages. In the first stage, we train the feature extractor from the discrimination perspective. In the second stage, we optimize the generator from the generation perspective. For the MSRVTT and LSMDC datasets, the experiments are carried out on 2 NVIDIA Tesla V100 GPUs. For the MSVD, ActivityNet Captions, and DiDeMo datasets, the experiments are carried out on 8 NVIDIA Tesla V100 GPUs. In both of the

[1]Equal contribution.
[2]Corresponding author: Li Yuan, Jie Chen.

Method	LSMDC						MSVD					
	R@1↑	R@5↑	R@10↑	Rsum↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	Rsum↑	MdR↓	MnR↓
TT-CE [8] ICCV21	17.5	36.0	45.0	98.5	14.3	-	27.1	55.3	67.1	149.5	4.0	-
CLIP4Clip [27] Neurocomputing22	20.8	39.0	48.6	108.4	12.0	54.2	62.0	87.3	92.6	241.9	**1.0**	**4.3**
EMCL-Net [13] NeurIPS22	22.2	40.6	49.2	112.0	12.0	-	54.3	81.3	88.1	223.7	**1.0**	5.6
DiffusionRet (Ours)	**23.0**	**43.5**	51.5	**118.0**	**9.0**	40.2	**61.9**	**88.3**	**92.9**	**243.1**	**1.0**	4.5
+ QB-Norm [4]	22.8	43.2	**51.6**	117.6	**9.0**	**40.0**	60.3	86.4	92.0	238.7	**1.0**	4.5
Method	**ActivityNet Captions**						**DiDeMo**					
	R@1↑	R@5↑	R@10↑	Rsum↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	Rsum↑	MdR↓	MnR↓
TT-CE [8] ICCV21	23.0	56.1	-	-	4.0	-	21.1	47.3	61.1	129.5	6.3	-
CLIP4Clip [27] Neurocomputing22	41.4	73.7	85.3	200.4	**2.0**	6.7	41.4	68.2	79.1	188.7	2.0	12.4
EMCL-Net [13] NeurIPS22	42.7	74.0	-	-	**2.0**	-	45.7	74.3	82.7	202.7	2.0	10.9
HBI [14] CVPR23	42.4	73.0	86.0	201.4	**2.0**	6.5	46.2	73.0	82.7	201.9	2.0	**8.7**
DiffusionRet (Ours)	43.8	75.3	**86.7**	205.8	**2.0**	**6.3**	46.2	74.3	82.2	202.7	2.0	10.7
+ QB-Norm [4]	**47.4**	**76.3**	**86.7**	**210.4**	**2.0**	6.7	**50.3**	**75.1**	**82.9**	**208.3**	**1.0**	10.3

Table A: **Video-to-text retrieval performance on the LSMDC, MSVD, and ActivityNet Captions datasets.** "↑" denotes that higher is better. "↓" denotes that lower is better.

Method	MSRVTT					MSRVTT->ActivityNet Captions				
	R@1↑	R@5↑	R@10↑	Rsum↑	MdR↓	R@1↑	R@5↑	R@10↑	Rsum↑	MdR↓
CLIP4Clip [27]‡ Neurocomputing22	43.8	70.6	81.4	195.8	**2.0**	29.1	58.3	72.1	159.5	4.0
EMCL-Net [13]‡ NeurIPS22	47.0	72.3	82.6	201.9	**2.0**	28.7	56.8	70.6	156.1	4.0
DiffusionRet (Ours)	**49.0**	**75.2**	**82.7**	**206.9**	**2.0**	**31.5**	**60.0**	**73.8**	**165.3**	**3.0**

Table B: **Text-to-video retrieval performance in out-domain retrieval settings.** "*MSRVTT->ActivityNet Captions*" denotes that the generalization results on unseen ActivityNet Captions test setting using pre-trained models on the MSRVTT dataset."‡" denotes our own re-implementation of baselines. "↑" denotes that higher is better. "↓" denotes that lower is better.

tasks of text-to-video and video-to-text retrieval, we assume that only the candidate sets are known in advance. In the inference phase, we consider both the distance of video and text representations in the representation space and the joint probability of video and text. Code is available at https://github.com/jpthu17/DiffusionRet.

A.2. Additional Results and Discussions

A.2.1 Video-to-Text Retrieval

We compare the proposed DiffusionRet with other meth- ods on five benchmark. In addition to the text-to-video retrieval results in the main paper, we provide video-to-text retrieval results on the LSMDC, MSVD, ActivityNet Captions, and DiDeMo datasets in Tab. A. Extensive experiments on five datasets, including MSRVTT, LSMDC, MSVD, ActivityNet Captions, and DiDeMo, demonstrate that our method is capable of dealing with both short and long videos. DiffusionRet achieves consistent improvements across different datasets, which demonstrates the effectiveness of our method.

A.2.2 Out-domain Retrieval

Most text-video retrieval methods [27, 13, 14, 15] are evaluated using the same dataset, which may not reflect their ability to generalize to unseen data. To this end, we perform out-domain retrieval by pre-training a model on one dataset (referred to as the "source") and evaluating its performance on another dataset (referred to as the "target") that is not included in the training. In addition to the out-domain retrieval experiments in the main paper, we provide additional experiments in the out-domain retrieval setting (MSRVTT->ActivityNet Captions) in Tab. B. We find that discriminant approaches do not transfer well from in-domain to out-of-domain retrieval. For instance, EMCL-Net outperforms CLIP4Clip in in-domain retrieval, but its performance is slightly lower than CLIP4Clip in out-domain retrieval. In contrast, DiffusionRet achieves good performance in both in-domain and out-of-domain retrieval.

A.2.3 Why Diffusion Models

Diffusion models have demonstrated remarkable generative power in various fields. Besides the powerful generative

Figure A: **The visualization of the diffusion process of the probability distribution.** We highlight the ground truth in green, and show the process from randomly initialized noise input (x_{50}) to the final predicted distribution (x_0). The iterative refinement property and many-to-many nature of the diffusion model render it an effective approach for text-video retrieval.

power of diffusion models, we explain other advantages of applying the diffusion model rather than other generative approaches to cross-modal retrieval, mainly in two aspects. **First**, the coarse-to-fine nature of the diffusion model enables it to progressively uncover the correlation between text and video, rendering it a more effective approach for retrieval tasks than other generation training methods, such as generative adversarial network [11] and variational autoencoder [17]. **Second**, the many-to-many nature of the diffusion model makes it more suitable for generating joint probabilities than the auto-regressive networks [9, 32]. We recommend further investigation of the potential of the generative method for discriminant tasks in future research. In our future work, we will explore our algorithm in segmentation [22, 24] and visual question answering [21, 19, 20].

A.2.4 Limitations of our Work

Generative models have focused on generative tasks, *e.g.*, image generation [12, 34], natural language generation [3, 25], and audio generation [30]. Some other works have attempted to adapt the generative models for discriminant tasks, *e.g.*, image segmentation [1], visual grounding [7], and detection [6]. However, these precursor methods require additional discriminative training. To train on limited data, we optimize the proposed generation model from both generation and discrimination perspectives. Although such a hybrid training method can improve model performance with lim-

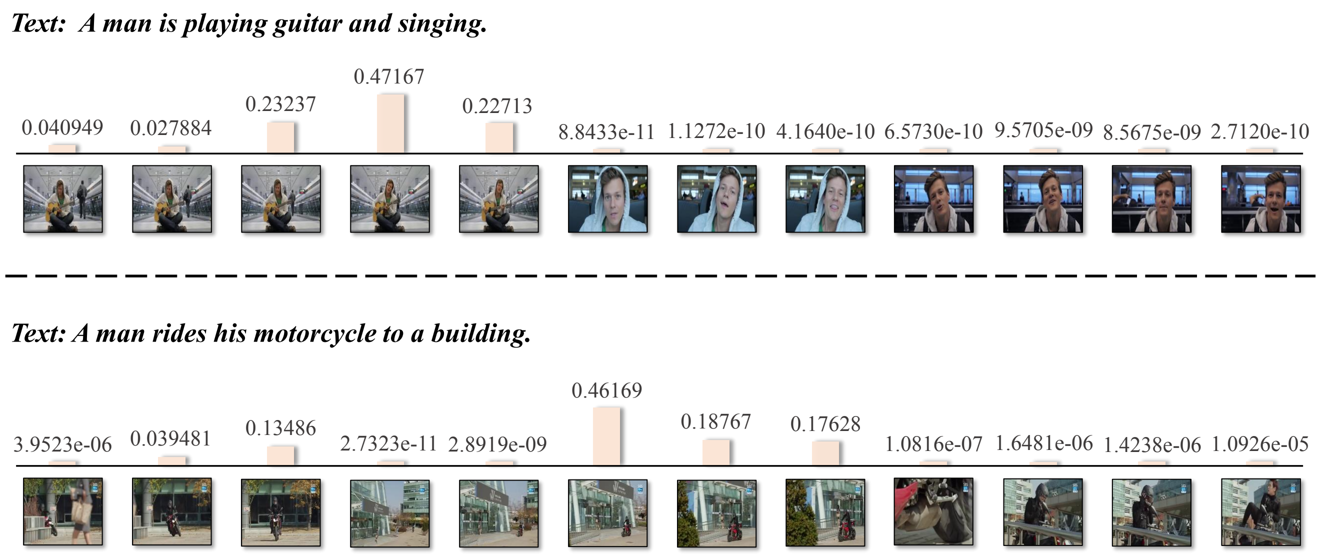

Figure B: **The visualization of the text-frame attention map.** These results demonstrate that our method can capture the correlation between text and frames.

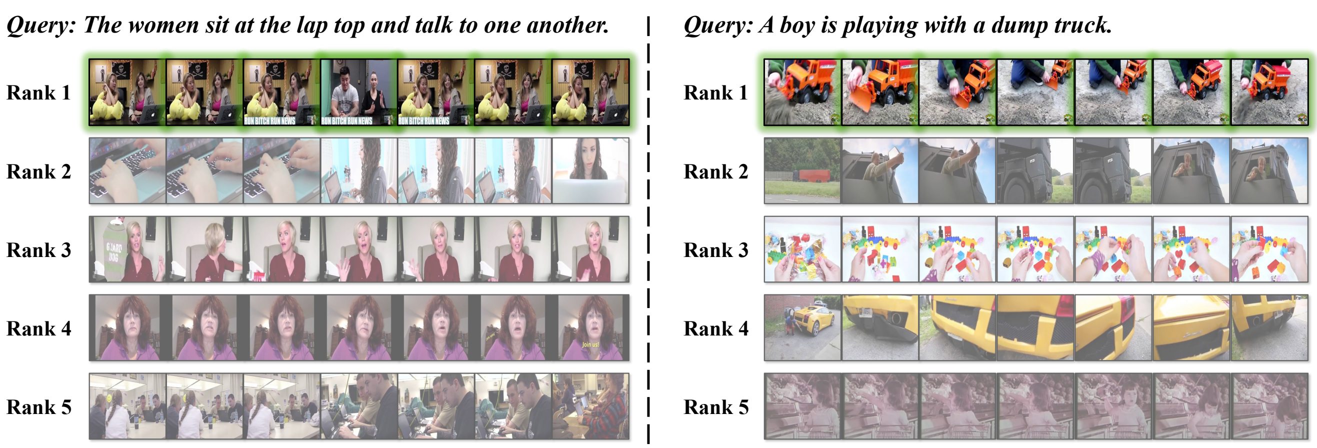

Figure C: **The visualization of the text-to-video results.** We highlight the ground truth in green. These results demonstrate that our method can mine the correlation between text and video effectively.

ited data, we believe that pure generative training is a more promising solution when the data is sufficient. We suggest exploring a pure generative training approach to the retrieval problem in the future.

A.3. Additional Visualizations

A.3.1 Diffusion Process

The coarse-to-fine nature of the diffusion model enables it to progressively uncover the correlation between text and video, rendering it an effective approach for cross-modal retrieval. To better understand the diffusion process, we show the additional visualization of the diffusion process in Fig. A. These results demonstrate that our method can progressively uncover the correlation between text and video.

A.3.2 Text-Frame Attention Map

To extract the joint encoding of text and video, we propose the text-frame attention encoder, which takes text representation as query and frame representation as key and value. To better understand the process of joint encoding of text and video, we show the visualization of the text-frame attention map in Fig. B. As shown in Fig. B, the text-frame attention encoder adaptively extracts the frames that are similar to the text so that fine-grained video features can be extracted. These results demonstrate that our method can capture the correlation between text and frames.

A.3.3 Text-to-Video Retrieval

We show two retrieval examples from the MSRVTT testing set for text-to-video retrieval in Fig. C. As shown in Fig. C, our method successfully retrieves the ground-truth video. These results demonstrate that our method can mine the correlation between text and video effectively.

References

[1] Tomer Amit, Eliya Nachmani, Tal Shaharbany, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021.

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017.

[3] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *NeurIPS*, pages 17981–17993, 2021.

[4] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *CVPR*, pages 5194–5205, 2022.

[5] David Chen and William Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, pages 190–200, 2011.

[6] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022.

[7] Zesen Cheng, Kehan Li, Peng Jin, Xiangyang Ji, Li Yuan, Chang Liu, and Jie Chen. Parallel vertex diffusion for unified visual grounding. *arXiv preprint arXiv:2303.07216*, 2023.

[8] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *ICCV*, pages 11583–11593, 2021.

[9] Brendan J Frey, Geoffrey E Hinton, and Peter Dayan. Does the wake-sleep algorithm produce good density estimators? In *NeurIPS*, 1995.

[10] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, pages 214–229, 2020.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020.

[13] Peng Jin, JinFa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David A. Clifton, and Jie Chen. Expectation-maximization contrastive learning for compact video-and-language representations. In *NeurIPS*, pages 30291–30306, 2022.

[14] Peng Jin, Jinfa Huang, Pengfei Xiong, Shangxuan Tian, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In *CVPR*, pages 2472–2482, 2023.

[15] Peng Jin, Hao Li, Zesen Cheng, Jinfa Huang, Zhennan Wang, Li Yuan, Chang Liu, and Jie Chen. Text-video retrieval with disentangled conceptualization and set-to-set alignment. *arXiv preprint arXiv:2305.12218*, 2023.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[18] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017.

[19] Hao Li, Jinfa Huang, Peng Jin, Guoli Song, Qi Wu, and Jie Chen. Weakly-supervised 3d spatial reasoning for text-based visual question answering. *TIP*, 32:3367–3382, 2023.

[20] Hao Li, Peng Jin, Zesen Cheng, Songyang Zhang, Kai Chen, Zhennan Wang, Chang Liu, and Jie Chen. Tg-vqa: Ternary game of video question answering. *arXiv preprint arXiv:2305.10049*, 2023.

[21] Hao Li, Xu Li, Belhal Karimi, Jie Chen, and Mingming Sun. Joint learning of object graph and relation graph for visual question answering. In *ICME*, pages 01–06, 2022.

[22] Kehan Li, Zhennan Wang, Zesen Cheng, Runyi Yu, Yian Zhao, Guoli Song, Li Yuan, and Jie Chen. Dynamic clustering network for unsupervised semantic segmentation. *arXiv preprint arXiv:2210.05944*, 2022.

[23] Kehan Li, Runyi Yu, Zhennan Wang, Li Yuan, Guoli Song, and Jie Chen. Locality guidance for improving vision transformers on tiny datasets. In *ECCV*, pages 110–127, 2022.

[24] Kehan Li, Yian Zhao, Zhennan Wang, Zesen Cheng, Peng Jin, Xiangyang Ji, Li Yuan, Chang Liu, and Jie Chen. Multi-granularity interaction simulation for unsupervised interactive segmentation. *arXiv preprint arXiv:2303.13399*, 2023.

[25] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *arXiv preprint arXiv:2205.14217*, 2022.

[26] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*, 2019.

[27] 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.

[28] 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.

[29] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F. Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *ICLR*, 2021.

[30] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *ICML*, pages 8599–8608, 2021.

[31] 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, 2021.

[32] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI*, 2018.

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

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[36] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: Global-local sequence alignment for text-video retrieval. In *CVPR*, pages 5079–5088, 2021.

[37] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016.

[38] Runyi Yu, Zhennan Wang, Yinhuai Wang, Kehan Li, Yian Zhao, Jian Zhang, Guoli Song, and Jie Chen. Position embedding needs an independent layer normalization. *arXiv preprint arXiv:2212.05262*, 2022.

[39] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, pages 471–487, 2018.