# Appendices

## A. Video-to-Text Retrieval Results

| Methods | R@1 ↑ | R@5 ↑ | R@10 ↑ | MdR ↓ | MnR ↓ |
|---|---|---|---|---|---|
| CE [6] | 20.6 | 50.3 | 64.0 | 5.3 | 25.1 |
| MMT [3] | 27.0 | 57.5 | 69.7 | 3.7 | 21.3 |
| Straight-CLIP [10] | 27.2 | 51.7 | 62.6 | 5.0 | - |
| Support Set [9] | 28.5 | 58.6 | 71.6 | 3.0 | - |
| TeachText-CE+ [2] | 32.1 | 62.7 | 75.0 | 3.0 | - |
| CLIP4Clip-meanP [7] | 43.1 | 70.5 | 81.2 | **2.0** | 12.4 |
| CLIP4Clip-seqTransf [7] | 42.7 | 70.9 | 80.6 | **2.0** | 11.6 |
| X-Pool (ours) | **44.4** | **73.3** | **84.0** | **2.0** | **9.0** |

Table A1. $v2t$ results on the MSR-VTT-9K dataset.

| Methods | R@1 ↑ | R@5 ↑ | R@10 ↑ | MdR ↓ | MnR ↓ |
|---|---|---|---|---|---|
| Straight-CLIP [10] | 59.9 | 85.2 | 90.7 | **1.0** | - |
| TeachText-CE+ [2] | 27.1 | 55.3 | 67.1 | 4.0 | - |
| CLIP4Clip-meanP [7] | 56.6 | 79.7 | 84.3 | **1.0** | 7.6 |
| CLIP4Clip-seqTransf [7] | 62.0 | 87.3 | 92.6 | **1.0** | 4.3 |
| X-Pool (ours) | **66.4** | **90.0** | **94.2** | **1.0** | **3.3** |

Table A2. $v2t$ results on the MSVD dataset.

| Methods | R@1 ↑ | R@5 ↑ | R@10 ↑ | MdR ↓ | MnR ↓ |
|---|---|---|---|---|---|
| JSFusion [11] | 12.3 | 28.6 | 38.9 | 20.0 | - |
| Straight-CLIP [10] | 6.8 | 16.4 | 22.1 | 73.0 | - |
| TeachText-CE+ [2] | 17.5 | 36.0 | 45.0 | 14.3 | - |
| CLIP4Clip-meanP [7] | 20.6 | 39.4 | 47.5 | 13.0 | 56.7 |
| CLIP4Clip-seqTransf [7] | 20.8 | 39.0 | 48.6 | 12.0 | 54.2 |
| X-Pool (ours) | **22.7** | **42.6** | **51.2** | **10.0** | **47.4** |

Table A3. $v2t$ results on the LSMDC dataset.

## B. Number of Frames Experiment

Our experiments use 12 sampled frames by default following recent text-video retrieval literature [7], and we run additional experiments on the MSR-VTT-9K dataset by varying the number of sampled frames for both training and inference as shown in Figure B1. We observe worse performance for 6 frames likely due to important information being missing at this scale. As we increase the number of frames[1], we observe performance saturation which is consistent with findings in [7]. However, we note that the optimal number of sampled frames remains a dataset specific hyperparameter.

## C. Online Inference in a Large-Scale Production System

Since our model computes an aggregated video embedding conditioned on a given text, the embeddings from a

[1]"All" indicates inference with all frames at inference time after training on 12 sampled frames.

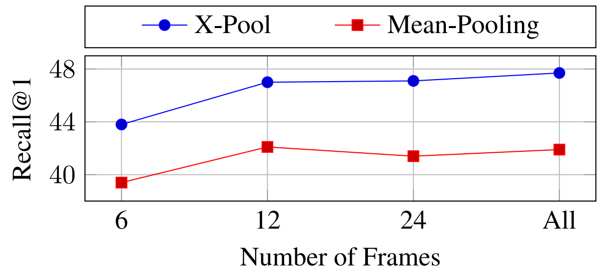


Figure B1. $t2v$ Recall@1 results on the MSR-VTT-9K dataset when varying the number of frames. "All" indicates inference with all frames.

video index set in $t2v$ cannot be entirely pre-computed because query texts are not a priori known during online inference. Instead, we can only pre-compute the frame embeddings of each index video, so fast nearest neighbour retrieval techniques [4, 5] cannot be readily applied. To address this in a production system with large-scale index sets, one commonly used approach is to use a high recall method to obtain a set of top retrieval candidates using using a nearest-neighbour search, and then use another method yielding high precision to re-rank the candidates [1, 8].

In our case, we can first mean-pool the pre-computed frame embeddings coming from X-Pool and then very efficiently obtain a set of $\mathcal{P}$ most similar candidates from the index set given a retrieval query. We can then run X-Pool's text-conditioned attention mechanism only on said candidates and then re-rank them for retrieval. That way, given $\mathcal{T}$ text queries and $\mathcal{V}$ index videos in $t2v$, instead of an $\mathcal{O}(\mathcal{T}\mathcal{V})$ complexity, we can achieve an $\mathcal{O}(\mathcal{T}\mathcal{P} + \mathcal{V})$ complexity where $\mathcal{P} << \mathcal{V}$ while maintaining good performance. In fact, we evaluated the performance of our model on the MSR-VTT dataset using the top-100 candidates from mean-pooling (i.e. $\mathcal{P} = 100$) and obtained the same performance in Recall@1, Recall@5 and Recall@10 as listed in our main results.

## References


[1] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198, 2016. 1

[2] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11583–11593, 2021. 1

[3] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 214–229. Springer, 2020. 1

[4] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 2019. 1

[5] Ting Liu, Andrew W Moore, Alexander G Gray, and Ke Yang. An investigation of practical approximate nearest neighbor algorithms. In *NIPS*, volume 12, page 2004. Citeseer, 2004. 1

[6] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019. 1

[7] 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. *arXiv preprint arXiv:2104.08860*, 2021. 1

[8] Junwei Ma, Satya Krishna Gorti, Maksims Volkovs, Ilya Stanevich, and Guangwei Yu. Cross-class relevance learning for temporal concept localization. *arXiv preprint arXiv:1911.08548*, 2019. 1

[9] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020. 1

[10] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using clip. In *Mexican Conference on Pattern Recognition*, pages 3–12. Springer, 2021. 1

[11] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018. 1