Rethinking Noisy Video-Text Retrieval via Relation-aware Alignment Supplementary Material

Huakai Lai[1] Guoxin Xiong[1] Huayu Mai[1] Xiang Liu[3] Tianzhu Zhang[1,2]*
[1]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition,
University of Science and Technology of China;
[2]National Key Laboratory of Deep Space Exploration, Deep Space Exploration Laboratory;
[3]Dongguan University of Technology
{tbhk, xgx, mai556}@mail.ustc.edu.cn, liuxiang@dgut.edu.cn, tzzhang@ustc.edu.cn

1. More Quanlitative Results

Results at higher noise ratios. We conduct experiments with over 50% noise rates shown in Tab. 1. Our method achieves R@1 of 38.7 under 75% noise, which is higher than RVTR [4] under 50% noise, further demonstrating the robustness of our method.

Results under different batch sizes. Different batch sizes can affect the agent construction. We experiment on batch size as shown in Tab. 2 and find that a small batch size affects agent selection, while an adequate size ensures its reliability.

Potential of more noisy data. We conduct experiments from two aspects to demonstrate the potential of our method in leveraging more noisy data, as shown in Tab. 3. *Clean only* refers to training with only 50% of the clean data in the training set. The 2-nd row indicates that the remaining 50% of noisy data is added to the 50% clean data for training. The 3-rd row denotes further training with an additional 200K noisy data pairs from the WebVid dataset [1]. The following two points can be observed: First, compared to using only clean data, our method shows a more significant improvement when 50% noisy pairs are added. Second, after incorporating 200K noisy WebVid data pairs scraped from the web, the result gain of our method becomes even more pronounced. The above results fully demonstrate the potential of our method to utilize noisy data.

Table 1. Comparison with higher noise ratio on MSR-VTT 1k-A.

	Text-to-Video			Video-to-Text		
Noise	R@1	R@5	R@10	R@1	R@5	R@10
RVTR [4] (50%)	36.2	61.6	73.7	36.0	61.7	73.0
Ours (50%)	44.6	71.1	81.2	44.3	71.9	82.4
Ours (75%)	38.7	65.9	76.4	37.6	65.0	75.9

Table 2. Comparison of text-to-video R@1 results with different batch sizes on MSR-VTT 1k-A.

Batch Size	R@1	Batch Size	R@1
16	40.7	128	44.6
32	41.5	256	45.0
64	44.2	512	44.9

Table 3. Comparison of results under different noisy data on MSR-VTT 1k-A, with each row having the same clean data.

	Text-to-Video			Video-to-Text		
Noise	R@1	R@5	R@10	R@1	R@5	R@10
Clean Only	40.8	69.2	79.8	41.8	70.1	80.1
Ours (50%)	44.6	71.1	81.2	44.3	71.9	82.4
Ours (50%) + Webvid	46.9	72.9	82.7	46.7	73.2	83.0

2. More Qualitative Results

We present several retrieval results for text-to-video and video-to-text from the MSR-VTT testing set [3] in Fig. 1 and Fig. 2, respectively, where the model is trained with 50% noise. For each instance, we present the retrieval results of CLIP4Clip [2] and our proposed method, with ground-truth highlighted in green and incorrect results in red. It can be observed that our method effectively mitigates the impact of noise correspondence, enabling the retrieval of accurate results, whereas CLIP4Clip [2] is significantly affected by noise correspondence, leading to the retrieval of semantically unrelated results. For example, for text-to-video results in the Fig. 1, the videos retrieved by CLIP4Clip in the top-right and bottom-left are completely unrelated to the query semantics, whereas our method accurately retrieves the correct results. In the bottom right example, the retrieval results of both methods are relatively similar, however, the result retrieved by CLIP4Clip corresponds to 'black dress', whereas our method accu-

*Corresponding author

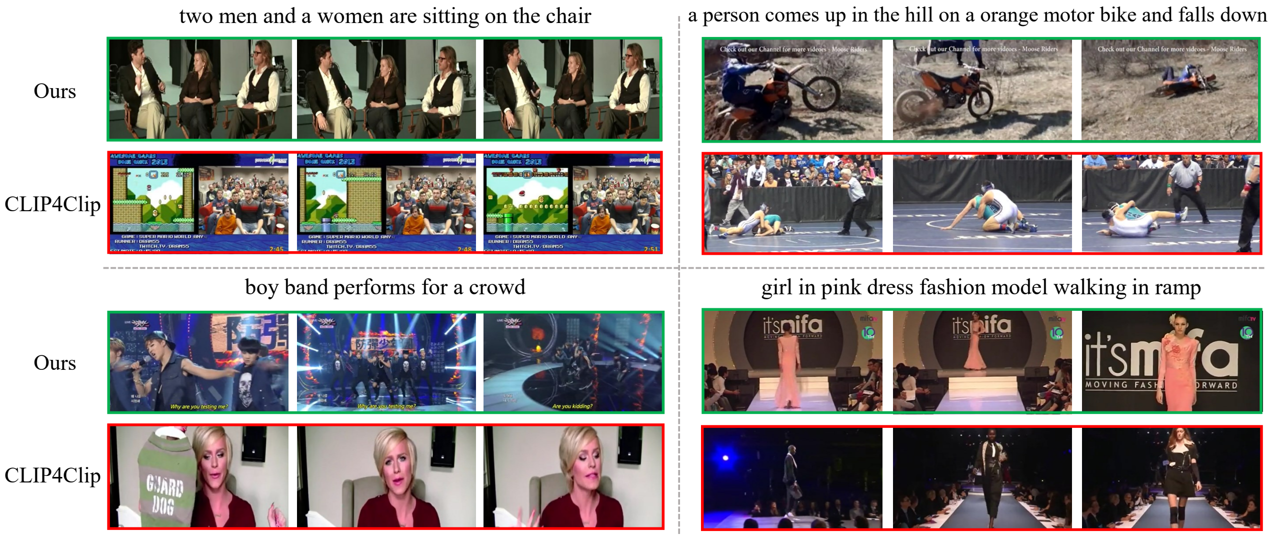
two men and a women are sitting on the chair
a person comes up in the hill on a orange motor bike and falls down
Ours
CLIP4Clip
boy band performs for a crowd
girl in pink dress fashion model walking in ramp
Ours
CLIP4Clip

Figure 1. Text-to-video retrieval results for CLIP4Clip and our method under 50% noise in MSR-VTT. For each query text, we present the top-1 retrieval video of each method, highlighting the ground truth in green and incorrect results in red.

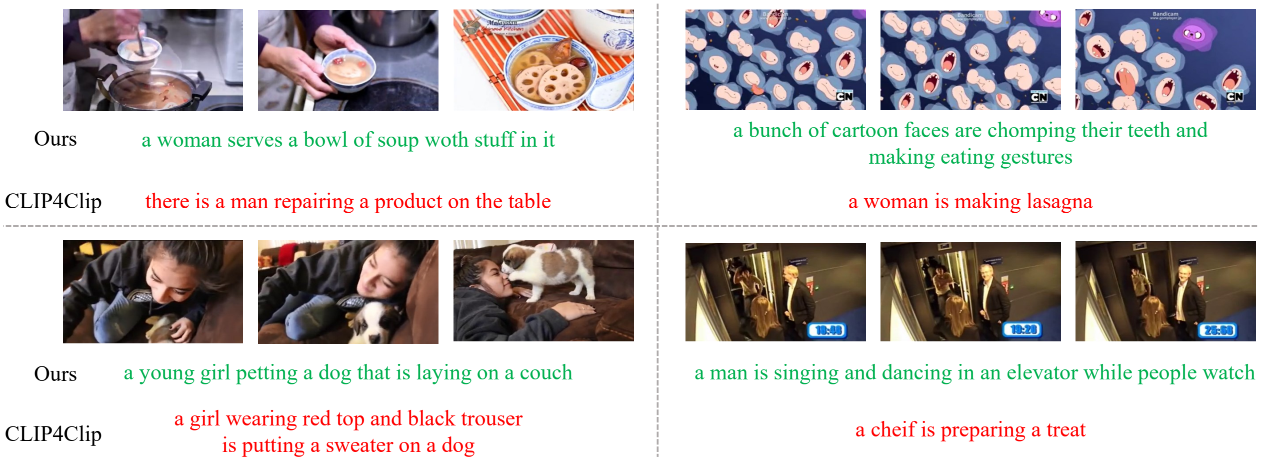
Ours
a woman serves a bowl of soup woth stuff in it
CLIP4Clip
there is a man repairing a product on the table
a bunch of cartoon faces are chomping their teeth and making eating gestures
a woman is making lasagna
Ours
a young girl petting a dog that is laying on a couch
CLIP4Clip
a girl wearing red top and black trouser is putting a sweater on a dog
a man is singing and dancing in an elevator while people watch
a cheif is preparing a treat

Figure 2. Video-to-text retrieval results for CLIP4Clip and our method under 50% noise in MSR-VTT. For each query video, we present the top-1 retrieval text of each method, highlighting the ground truth in green and incorrect results in red.

rately identifies ‘pink dress’. For video-to-text results in the Fig. 2, for the top-left, top-right, and bottom-right examples, the retrieval results of CLIP4Clip are all unrelated to the semantics of query video. In contrast, our method achieves accurate cross-modal semantic alignment under the influence of noise correspondence. For the bottom left example, both methods identify ‘a girl’ and ‘a dog’, while CLIP4Clip incorrectly identifies ‘wearing a red top and black trouser’. Overall, our method demonstrates high robustness under noise correspondence, achieving reliable and accurate alignment, which can be attributed to the proposed relation-aware purified consistency based on ranking distribution.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021. 1

[2] 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. 1

[3] 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. 1

[4] Huaiwen Zhang, Yang Yang, Fan Qi, Shengsheng Qian, and

Changsheng Xu. Robust video-text retrieval via noisy pair calibration. *IEEE Transactions on Multimedia*, 25:8632–8645, 2023. 1