MV-Adapter: Exploring Parameter Efficient Learning for Video Text Retrieval

Supplementary Material

Abstract

This supplementary material presents the following ablations, qualitative results and analysis: (1) Ablations conducted on the design and optimal settings of Cross Modality Tying (CMT). (2) Ablations of Temporal Adaptation (TA) and CMT on more datasets. (3) Additional qualitative results and case studies. (4) More detailed efficiency analysis.

7. More Ablations

Cross Modality Tying. We conduct extensive experiments to validate the design and settings of CMT.

- **Design.** As shown in Eq. (4), CMT calibrates the weights of `Downsample` by element-wise multiplication of $\beta_{\text{cal}} \in \mathbb{R}^d$ with each column of W_{down}. Here, the dimension of β_{cal} is equal to the in-channels of W_{down}, which we refer to as the "Down-In" design. We also examine the "Down-Out" design, where the dimension of β_{cal} is equal to the out-channels of W_{down}, i.e., d'. In this scenario, the dimension of the cross modality factor f_C remains unchanged, but the projection matrix M_S is reconfigured from $d^m \times d$ to $d^m \times d'$ to align the out-channel dimensions. Beyond comparing "In" and "Out", we also explore the calibration of `Upsample` weights with CMT, leading to a comparison of four variants. As indicated in Tab. 7, calibrating the in-channels of the downsampling matrix yields the best performance.
- **Optimal settings.** Further ablations are conducted to identify the optimal CMT settings, including the encoder layers where CMT is applied, and the dimension of the modality factor $f_C \in \mathbb{R}^{d_m}$. The results are shown in Tab. 8 and Tab. 9 respectively. Considering both R@1 and R@Sum, CMT achieves optimal results when applied to the final 2 layers with a factor dimension of 32. We hypothesize that applying CMT in higher layers is more effective, as feature spaces of the two modalities converge more closely, whereas earlier application might negatively impact the training process.

Results on More Datasets. In the main paper, we present ablations on MSR-VTT due to space limit. Tab. 10 shows ablation results on other datasets. These new results lead to similar conclusions: equipping the model with temporal adaptation (TA) consistently improves performance across all datasets, confirming the importance of TA in our tasks; Additionally, by facilitating modality alignment, CMT significantly enhances the performance of each model. Since CMT incurs negligible extra parameters (less than 0.1% of vanilla CLIP), it can be conveniently used to boost model capabilities.

Designs	T2V	V2T
Down-In	46.2/**202.1**	**47.2/205.9**
Down-Out	**46.8**/201.8	47.0/204.9
Up-In	46.5/200.7	46.3/204.5
Up-Out	46.0/200.5	46.5/203.7

Table 7. Ablations on the design of CMT on MSR-VTT [46] using the CLIP (ViT-B/16) backbone [38]. We set factor dimension to 32, and use CMT in the last 2 layers of encoders by default.

Layers	T2V	V2T
No CMT	45.9/200.9	46.6/204.4
Last 1	**46.6**/201.3	46.8/204.1
Last 2	46.2/**202.1**	**47.2/205.9**
Last 3	45.8/201.0	46.6/205.8
Last 6	45.9/201.0	46.8/204.9
Last 12	45.4/199.7	46.5/203.9

Table 8. Ablations on layers using CMT on MSR-VTT [46] using the CLIP (ViT-B/16) backbone [38] with factor dimension 32. "Last n" refers to using CMT in the last n layers of the visual/text encoders.

Dim	T2V	V2T
8	**46.7**/202.0	46.5/204.7
16	46.4/202.1	46.5/205.1
32	46.2/202.1	**47.2/205.9**
64	46.3/200.6	46.0/204.1

Table 9. Ablations on the factor dimension in CMT on MSR-VTT [46]. The backbone used is CLIP (ViT-B/16) [38].

8. Case Study

We summarize two failure modes of our method from comprehensive case studies:

- Since we use a fixed number of frames per video in training, our method may fail to capture fine movements in long videos as the differences between frames aggregate.
- Results may be incorrect when the caption is related to the audio contents.

Examples of these two modes are put into `long_video` and `audio` directories respectively. These directories have been zipped together with this document and uploaded to CMT as supplementary files. Each directory contains one retrieval result consisting of caption.txt (the query used to search), gt_*.mp4, and pred_*.mp4 (the ground truth and

Settings	MSVD								LSMDC							
	Text-to-Video				Video-to-Text				Text-to-Video				Video-to-Text			
	R@1	R@5	R@10	Sum	R@1	R@5	R@10	Sum	R@1	R@5	R@10	Sum	R@1	R@5	R@10	Sum
Ours	**49.4**	**78.3**	**87.0**	**214.8**	**71.8**	**93.0**	**96.4**	**261.2**	23.2	**43.9**	53.2	**120.3**	**24.0**	**42.8**	52.1	**118.8**
w/o TA	49.3	78.2	**87.0**	214.5	70.8	93.2	96.1	260.2	23.1	42.7	52.4	118.2	23.6	42.2	**52.4**	118.3
w/o CMT	49.0	**78.3**	86.9	214.2	71.0	92.3	96.2	259.4	**23.4**	43.2	**53.3**	119.9	23.0	41.6	51.8	116.5

Settings	Didemo								ActivityNet							
	Text-to-Video				Video-to-Text				Text-to-Video				Video-to-Text			
	R@1	R@5	R@10	Sum	R@1	R@5	R@10	Sum	R@1	R@5	R@10	Sum	R@1	R@5	R@10	Sum
Ours	**44.3**	**72.1**	**80.5**	**196.8**	42.7	73.0	**81.9**	**197.6**	42.7	74.2	**85.8**	202.7	**44.0**	74.4	86.0	204.4
w/o TA	43.5	71.9	80.4	195.8	**43.2**	72.2	81.2	196.6	42.1	72.9	84.5	199.6	42.9	73.4	85.5	201.8
w/o CMT	43.8	71.8	80.4	195.9	42.4	**73.2**	81.5	197.1	**42.9**	**74.5**	85.7	**203.1**	43.6	**75.0**	**86.5**	**205.2**

Table 10. Ablation of TA and CMT modules by removing one at a time from MV-Adapter. "Sum" represents the sum +of R@1/5/10 in Text-to-Video or Video-to-Text task. The backbone is CLIP (ViT-B/16) [38].

predicted video, where * represents the index number of videos in MSR-VTT [46].)

For the sake of clarity, we present a detailed analysis of each example below.

long_video_0. In this case, the clips featuring "people" are quite concentrated and "fade" quickly. Given the video's duration of 27 seconds, the long intervals between extracted frames result in key information being omitted, making it impossible to match the description. Therefore, another video, where the fading effect and the characters are clearer, is returned instead.

long_video_1. The groundtruth video is long and the shot that corresponds to the target in the query (walking down a short runway) is relatively short. Since the number of input frames is fixed, non-target information in videos tends to dominate the input, making the model fail to parse out the fine movement ("walking" and "short runway") from the groundtruth. Eventually, the model returns a similar video that contains "walking" but on a "long runway" (should be a short runway).

audio_0. Audio information is necessary in order to determine the topic of talking.

audio_1. Though the retrieved result is visually similar to groundtruth, the contents of the talk do not match that of the query text. With the help of audio content (like transcripts from ASR), the results can be corrected.

9. Efficiency Analysis

MV-Adapter has three types of newly added parameters: down-sampling, a lightweight transformer and up-sampling, the parameter complexity of which are $O(d \times d')$, $O(d' \times d)$, and $O((d')^2)$, respectively. Temporal calibration's parameter complexity is also $O((d')^2)$. CMT only introduces $O(d^m + d^m \times d)$ parameters. As a result, MV-Adapter is rather parameter-efficient as $d^m, d' \ll d$. The total increase in parameters compared with the CLIP backbone is about 2.4% when d is 768 and d' is 64. Due to its small number of tunable parameters, MV-Adapter is highly parameter-efficient in both deployment and training stages.