DiscoVLA: Discrepancy Reduction in Vision, Language, and Alignment for Parameter-Efficient Video-Text Retrieval

Supplementary Material

A. Additional Experimental Settings

Datasets. Following common practice, we perform evaluations on four widely used benchmarks for video-text retrieval: (1) **MSRVTT** [53] includes 10,000 YouTube videos, each with 20 text descriptions. Following previous methods [13, 39], we train on the 'train+val' set with 9,000 video-text pairs and evaluate on the '1K-A' test set with 1,000 video-text pairs. (2) **LSMDC** [43] consists of 118,081 movie clips, each paired with a single description. We use 101,079 for training, 7,408 for validation, and 1,000 for testing, reporting results based on the test set. (3) **ActivityNet** [29] consists of 20,000 YouTube videos. Our evaluation utilizes the 'val1' split, which includes 10,009 videos for training and 4,917 for testing. Following previous methods [13, 59], we concatenate all sentence descriptions of a video into a single paragraph. (4) **DiDeMo** [1] comprises 10,000 videos with a total of 40,000 text descriptions. The training set contains 8,395 videos, while the test set contains 1,004 videos. Following previous methods [3, 31], we combine all descriptions of a video into a single query.

Evaluation Metrics. We evaluate the performance using common retrieval metrics such as Recall at K (R@K and K = 1, 5, 10), the sum of these recalls (R@sum), and Mean Rank (MnR). R@K measures the proportion of relevant items retrieved in the top K results for a given query. MnR calculates the mean rank of correct items. Note that for R@K, a higher score means better performance. Conversely, for MnR, a lower score indicates better results.

Implementation Details. Following previous parameter-efficient research [5, 22, 25, 58], we utilize the pre-trained CLIP model as our backbone. we implement the AdamW optimizer [36] with a batch size of 128. For all datasets, the initial learning rate is set to 6e-4, employing a cosine learning rate schedule [15] over 5 epochs. For MSRVTT and LSMDC, the max frame and caption length are set to 12 and 32. For ActivityNet and DiDeMo, the max frame and caption length are set to 32 and 64. In all experiments, the LoRA dimension and the adapter dimension r are set to 8. In Eq. 14, α and β are set to 0.3 and 1.0, respectively. For the number of IVFusion layers, we set $H_V = 4$ for the vision encoder and $H_L = 2$ for the text encoder.

B. Additional Experimental Results

Ablation study on α and β in Eq. (14). Figures 6 and 7 present ablation studies on hyperparameters α and β, respectively. These parameters control the trade-off among the video-level alignment loss $\mathcal{L}_A(\text{Sim}^{\text{vid}})$, the image-level alignment loss $\mathcal{L}_A(\text{Sim}^{\text{img}})$, and the distillation loss $\mathcal{L}_{KL}$. The parameter α represents the weight of $\mathcal{L}_A(\text{Sim}^{\text{img}})$. Increasing α improves both R@1 and R@sum significantly. Our DiscoVLA achieves optimal performance at $\alpha = 0.3$, beyond which it demonstrates parameter insensitivity. The parameter β represents the weight of $\mathcal{L}_{KL}$. While $\mathcal{L}_{KL}$ does not affect R@1, setting $\beta = 1.0$ yields a marked improvement in R@sum. Consequently, we set $\alpha = 0.3$ and $\beta = 1.0$ in our final implementation.

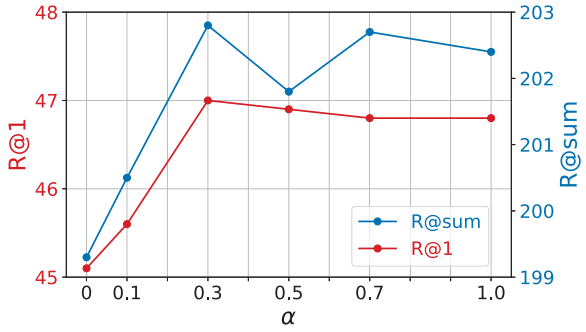

Figure 6. Ablation study on α in Eq. (14) for text-to-video results on MSRVTT using CLIP (ViT-B/32). α represents the weight of the image-level alignment loss $\mathcal{L}_A(\text{Sim}^{\text{img}})$. All other hyperparameters are kept constant.

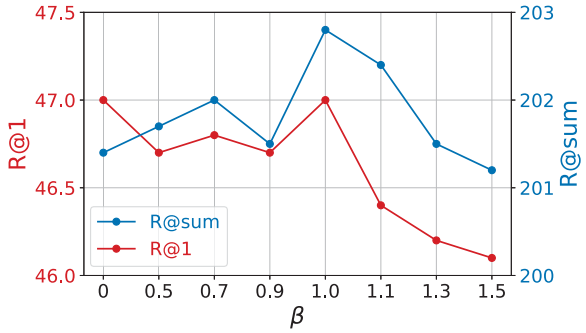

Figure 7. Ablation study on β in Eq. (14) for text-to-video results on MSRVTT using CLIP (ViT-B/32). β represents the weight of the distillation loss $\mathcal{L}_{KL}$. All other hyperparameters are kept constant.

Ablation study on the number of IVFusion layers H_V and H_L. The number of IVFusion layers in the vision and text encoders is denoted by H_V and H_L, respectively. IVFusion is applied to the upper layers of CLIP, as these layers extract high-level semantic information essential for cross-frame learning. This learning process requires an ad-

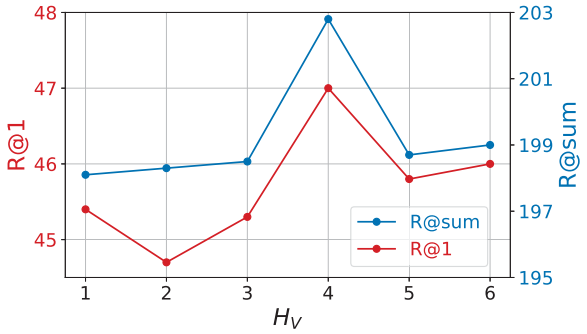

Figure 8. Ablation study on H_V for text-to-video results on MSRVTT using CLIP (ViT-B/32). H_V represents the number of IVFusion layers in the vision encoder. All other hyperparameters are kept constant.

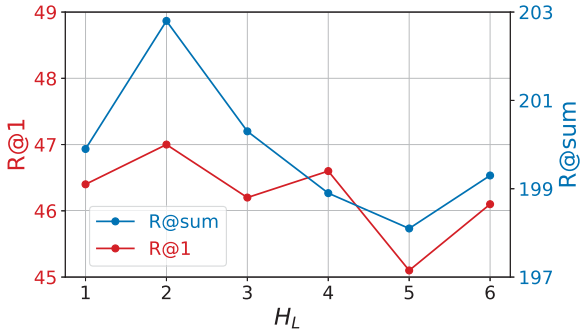

Figure 9. Ablation study on H_L for text-to-video results on MSRVTT using CLIP (ViT-B/32). H_L represents the number of IVFusion layers in the text encoder. All other hyperparameters are kept constant.

Post-processing		Text-to-Video			
		R@1	R@5	R@10	R@sum
MSRVTT	DiscoVLA	47.0	73.0	82.8	202.8
	+QB-Norm [4]	47.5	73.6	82.9	204.0
	+DSL [8]	**51.3**	**77.1**	**85.5**	**213.9**
ActivityNet	DiscoVLA	41.2	72.4	83.6	197.2
	+QB-Norm [4]	45.1	74.9	85.2	205.2
	+DSL [8]	**49.9**	**78.8**	**88.1**	**216.8**

Table 9. Effect of post-processing on MSRVTT and ActivityNet using CLIP (ViT-B/32).

equate number of IVFusion layers. From Figures 8 and 9, we observe that the best performance is achieved when $H_V = 4$ and $H_L = 2$. This difference in optimal values may be attributed to the inherent distinctions between the vision and text modalities. This configuration $H_V = 4$ and $H_L = 2$ demonstrate superior and robust performance across all benchmarks (see Tables 1-4).

Effect of post-processing. As shown in Table 9, we evaluate post-processing methods Q-Norm [4] and DSL [8] on top of our proposed DiscoVLA. While Q-Norm leads to a notable **8.0**% R@sum increase on ActivityNet, it has minimal effect on MSTVTT. In contrast, DSL provides consistent performance gains across both datasets, improving performance by **11.1**% R@sum on MSRVTT and **19.6**% on ActivityNet.