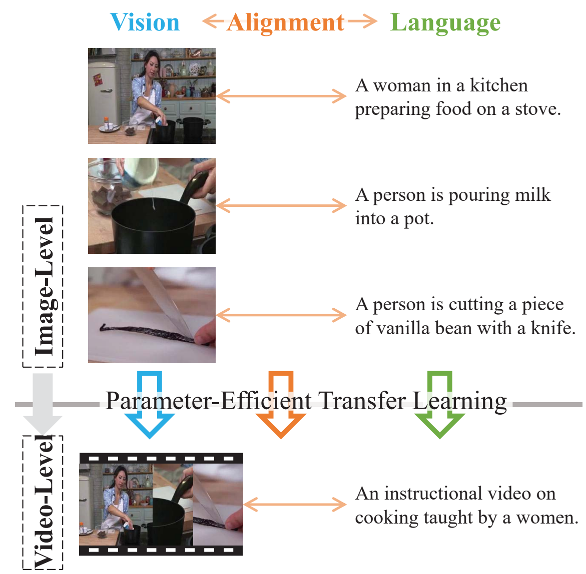
DiscoVLA: Discrepancy Reduction in Vision, Language, and Alignment for Parameter-Efficient Video-Text Retrieval

Leqi Shen[1,2,4*] **Guoqiang Gong**[5†] **Tianxiang Hao**[1,2] **Tao He**[6,7] **Yifeng Zhang**[5]
Pengzhang Liu[5] **Sicheng Zhao**[2‡] **Jungong Han**[3] **Guiguang Ding**[1,2‡]
[1] School of Software [2] BNRist [3] Department of Automation, Tsinghua University
[4] Hangzhou Zhuoxi Institute of Brain and Intelligence
[5] JD.com [6] GRG Banking Equipment Co., Ltd. [7] South China University of Technology

Abstract

*The parameter-efficient adaptation of the image-text pretraining model CLIP for video-text retrieval is a prominent area of research. While CLIP is focused on image-level vision-language matching, video-text retrieval demands comprehensive understanding at the video level. Three key discrepancies emerge in the transfer from image-level to video-level: vision, language, and alignment. However, existing methods mainly focus on vision while neglecting language and alignment. In this paper, we propose **Disc**repancy Reduction in **V**ision, **L**anguage, and **A**lignment (**DiscoVLA**), which simultaneously mitigates all three discrepancies. Specifically, we introduce Image-Video Features Fusion to integrate image-level and video-level features, effectively tackling both vision and language discrepancies. Additionally, we generate pseudo image captions to learn fine-grained image-level alignment. To mitigate alignment discrepancies, we propose Image-to-Video Alignment Distillation, which leverages image-level alignment knowledge to enhance video-level alignment. Extensive experiments demonstrate the superiority of our DiscoVLA. In particular, on MSRVTT with CLIP (ViT-B/16), DiscoVLA outperforms previous methods by **2.2%** R@1 and **7.5%** R@sum. The code is available at* `https://github.com/LunarShen/DsicoVLA`.



Method	Image-Level $\xrightarrow{\text{PETL}}$ Video-Level		
	Vision	Language	Alignment
Previous Methods	✓	✗	✗
Our DiscoVLA	✓	✓	✓

Figure 1. When parameter-efficient transferring image-level CLIP to video-text retrieval, we identify three key discrepancies: vision, alignment, and language. Unlike previous methods, our DiscoVLA aims to address all these discrepancies. More visualization examples are provided in Figure 4.

1. Introduction

The rapid expansion of online video content has generated a growing demand for effective video-text retrieval, focusing on matching video content with relevant textual descriptions. Inspired by the success of image-text pretraining approaches such as CLIP [42], recent efforts [9, 14, 24, 37, 48, 49, 51] have increasingly focused on extending these powerful capabilities to video-text retrieval tasks. However, fully fine-tuning these pretrained models requires updating numerous parameters for every dataset, resulting in large storage overhead [22, 58]. Therefore, we focus on parameter-efficient transfer learning methods, targeting strong performance with minimal trainable parameters.

Although both image-text pretraining and video-text retrieval involve vision-language (VL) matching, they exhibit

*lunarshen@gmail.com. Work done during an internship at JD.com.
†Project leader. ‡ Corresponding authors.

substantial discrepancies that hinder straightforward knowledge transfer. To be specific, image-text pretraining focuses on matching images with image captions, which represents image-level VL matching. In contrast, video-text retrieval focuses on matching videos with video captions, which represents video-level VL matching. The challenges of video-level VL matching, which demands understanding motion and time, create significant obstacles for transferring knowledge from image-text to video-text tasks.

In this work, we explore the discrepancies between image-level and video-level VL matching across three key aspects: vision, language, and alignment. Videos introduce a temporal dimension, which is absent in images. Similarly, the distinction between image captions and video captions arises from the varying levels of granularity, such as isolated image descriptions versus continuous video narratives. Thus, video-level alignment is inherently more complex than image-level alignment, as it must account for intricate spatio-temporal relationships. In Figure 1, the three images correspond to "preparing food", "pouring", and "cutting", while the entire video demonstrates an instructional cooking process. Image-level VL matching focuses on isolated images, while video-level VL matching emphasizes the overall understanding of continuous segments.

However, current methods [5, 22, 25, 58] struggle to fully address the discrepancies mentioned above. Some [5, 22, 25, 58] focus on the vision aspect, modeling temporal relationships to extract video-level features, while others [25, 58] attempt to reduce the modality gap between vision and language through shared parameter mechanisms. Despite these advances, these methods remain limited in both language and alignment discrepancies.

To tackle these challenges, we introduce **Disc**repancy Reducti**o**n in **V**ision, **L**anguage, and **A**lignment, termed **DiscoVLA**, a novel parameter-efficient method for video-text retrieval. Unlike previous methods that target a single gap, our method aims to simultaneously reduce all significant discrepancies, as illustrated in the table of Figure 1.

Specifically, we propose an Image-Video Features Fusion module (IVFusion), which addresses both vision and language discrepancies through a unified feature extraction approach. IVFusion effectively combines both image- and video-level features by utilizing a lightweight adapter. In addition to video-level alignment, we propose Pseudo Image-level Alignment (PImgAlign) which generates pseudo image captions and learns fine-grained image-level alignment. Furthermore, to reduce alignment discrepancy, we propose Image-to-Video Alignment Distillation (AlignDistill), transferring image-level alignment knowledge to improve video-level alignment. All proposed modules prioritize parameter efficiency: IVFusion requires minimal trainable parameters, while PImgAlign and AlignDistill operate without increasing inference parameters.

In summary, the main contributions are as follows:

- We reveal, for the first time, the necessity of addressing all three image-to-video discrepancies—vision, language, and alignment—to achieve parameter-efficient VTR.
- We introduce IVFusion to fuse image- and video-level features for both vision and language gaps. PImgAlign is introduced for fine-grained image-level alignment. We introduce AlignDistill to minimize alignment gaps.
- Our DiscoVLA achieves state-of-the-art results on MSRVTT, LSMDC, ActivityNet, and DiDeMo. Notably, on MSRVTT with CLIP (ViT-B/16), DiscoVLA reaches **50.5**% R@1, surpassing previous methods by **2.2**%.

2. Related Work

Video-Text Retrieval. Advances in vision-language pretraining [7, 12, 32–34, 42, 50, 52, 55] have achieved significant success across various downstream tasks [2, 10, 11, 18, 26, 45, 47, 54, 56, 57]. Leveraging powerful CLIP [42], recent video-text retrieval (VTR) methods [9, 14, 24, 37, 48, 49, 51] have achieved impressive performance enhancements. CLIP4Clip [37] is a pioneering CLIP-based method that aggregates image features via mean-pooling or transformers to obtain video features, inspiring subsequent studies. Additionally, Cap4Video [51] utilizes ZeroCap [46] and GPT-2 [41] to generate video-level captions as supplementary video content during both training and inference. In contrast, our DiscoVLA addresses the image-level and video-level gap by generating image-level captions used only in training to improve video-level alignment. These full fine-tuning methods impose substantial overhead from large trainable parameters.

Parameter-Efficient Transfer Learning. Parameter-efficient methods aim to adapt pre-trained models to new tasks with minimal fine-tuning. The mainstream methods in natural language processing and image tasks include techniques such as Prompt [23, 27, 28, 61, 62], Adapter [6, 17, 20, 60], and LoRA [21, 60]. Prompt [27] modifies the input space by adding learnable tokens. Adapter [20] introduces a lightweight bottleneck neural network consisting of up- and down-projection layers with a non-linear activation function. LoRA [21] aims to update up- and down-projection layers by learning low-rank decomposition matrices.

Recently, several studies have explored parameter-efficient VTR methods [5, 22, 25, 58] to incorporate temporal modeling into the frozen image-level CLIP backbone. VoP [22] employs trainable BiLSTMs [16] to produce prompt tokens in the vision encoder. DGL [58] introduces weight-sharing layers to generate prompts for both text and video encoders. MV-Adapter [25] introduces shared bottleneck structures in both video and text branches. However, these methods overlook language and alignment discrepancies. We address these limitations with DiscoVLA, a framework that facilitates robust video-level understanding.

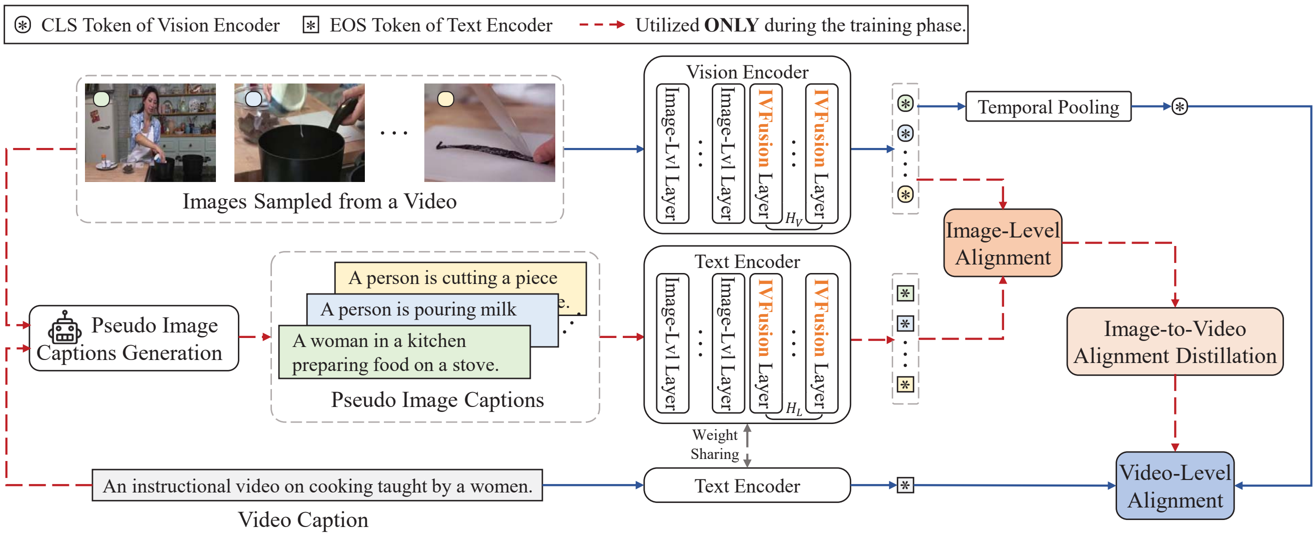


Figure 2. The overall framework of DiscoVLA. Initially, we generate pseudo image captions for each sampled image. In both vision and text encoders, we utilize image-level layers to acquire pretrained image-level knowledge and employ IVFusion layers to enhance spatio-temporal information. The single video caption is encoded through the text encoder, utilizing IVFusion layer as image-level (image-lvl) layer. Finally, AlignDistill is applied to distill image-level alignment to video-level alignment. For fair comparisons with previous methods, we do not generate pseudo image captions during the inference phase.

3. Methodology

3.1. Preliminary

Video-text retrieval is a task that involves retrieving relevant videos based on a given text query, or conversely, retrieving appropriate text descriptions for a video. Consistent with previous approaches [5, 22, 25, 58], we utilize the CLIP model, pre-trained on image-text data, as our backbone.

To process a given video, we first sample a sequence of frame images uniformly, which are then input into the CLIP vision encoder to extract image features v^{img}. These features are then aggregated through temporal average pooling to get video features v^{vid}. For a given video caption, it is passed through the CLIP text encoder to extract video caption features t^{vid}. Then, we compute the cross-modal contrastive loss [40], which functions as a video-level alignment term, maximizing the video-level similarity Sim^{vid} between matched pairs:

$$\text{Sim}_{ij}^{\text{vid}} = t_i^{\text{vid}^T} v_j^{\text{vid}}, \tag{1}$$

$$\mathcal{L}_{t2v}(\text{Sim}^{\text{vid}}) = \frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\text{Sim}_{ii}^{\text{vid}})/\tau)}{\sum_{j=1}^{B} \exp(\text{Sim}_{ij}^{\text{vid}})/\tau)}, \tag{2}$$

$$\mathcal{L}_{v2t}(\text{Sim}^{\text{vid}}) = \frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\text{Sim}_{ii}^{\text{vid}})/\tau)}{\sum_{j=1}^{B} \exp(\text{Sim}_{ji}^{\text{vid}})/\tau)}, \tag{3}$$

$$\mathcal{L}_{A}(\text{Sim}^{\text{vid}}) = -\frac{1}{2}[\mathcal{L}_{t2v}(\text{Sim}^{\text{vid}}) + \mathcal{L}_{v2t}(\text{Sim}^{\text{vid}})], \tag{4}$$

where B is the batch size, τ is the temperature parameter.

3.2. DiscoVLA

As shown in Figure 2, we propose DiscoVLA to address the challenges of vision, language, and alignment when transferring from image-level to video-level VL matching. First, we introduce Image-Video Features Fusion (IVFusion) in both vision and text encoders to enhance feature extraction. Next, we introduce Pseudo Image-Level Alignment (PImgAlign) for fine-grained image-level alignment, consisting of Pseudo Image Language Generation and Image-Level Alignment. Finally, Image-to-Video Alignment Distillation (AlignDistill) is proposed to leverage image-level alignment knowledge to enhance video-level alignment.

Importantly, PImgAlign is only employed during training. For a fair comparison with previous methods and due to its generation inefficiency, PImgAlign is omitted at inference. The final similarity used for retrieval is based on video-level similarity in Eq. (1).

We focus on parameter-efficient transfer learning for video-text retrieval, where the large backbone parameters remain fixed, and only a small number of additional parameters are trained. The trainable parameters are embedded in the parameter-efficient attention of each encoder layer, as illustrated in Figure 3.

3.3. Image-Video Features Fusion

Image-level and video-level variations exist in both vision and language modalities. Our proposed IVFusion fuses these image-level and video-level features in both the vision and language encoders. Below, we illustrate IVFusion's application within the vision encoder.

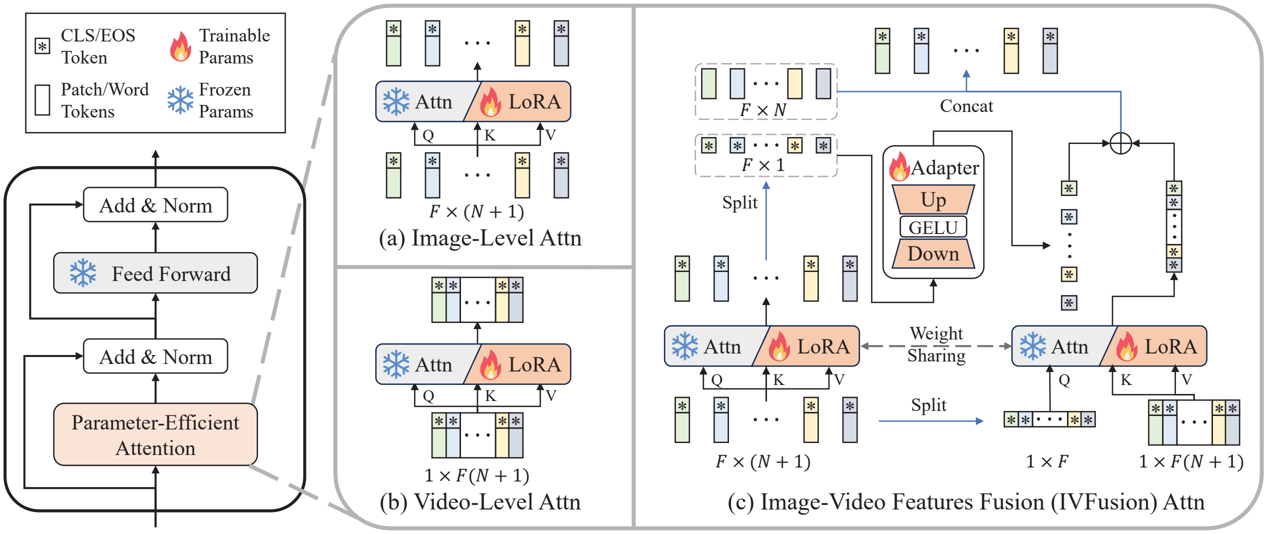


Figure 3. Illustration of the encoder layers for vision and text encoders. (a) Image-Level Attn operates on each of the F images or image captions individually. (b) Video-Level Attn concatenates tokens across all sampled images or image captions. (c) IVFusion Attn employs a lightweight adapter to integrate the efficiency of Image-Level Attn with the effectiveness of Video-Level Attn. Here, we illustrate the application of IVFusion within the vision encoder. The text encoder adopts the same approach for processing pseudo image captions.

Image-Level Attn. To achieve parameter-efficient transfer learning, we apply LoRA in the attention module of each encoder layer. As illustrated in Figure 3a, $F \times (N+1)$ tokens are processed via Image-level Attn, where F denotes the number of sampled images, and $N+1$ represents the combination of the global CLS token and patch tokens. Image-Level Attn processes each of the F images individually, using $N+1$ tokens per image. However, Image-Level Attn does not encode temporal information.

Video-Level Attn. A straightforward solution is Video-Level Attn, which concatenates tokens across all sampled images, enabling the attention module to learn spatio-temporal information. As shown in Figure 3b, $F(N+1)$ tokens are simultaneously processed during the attention module. Due to the quadratic complexity of the attention mechanism, Image-Level Attn has a complexity of $O(F(N)^2)$, while Video-Level Attn has a higher complexity of $O((FN)^2)$.

IVFusion Attn. To combine Image-Level Attn's efficiency with Video-Level Attn's effectiveness, we introduce IVFusion Attn in Figure 3c. We extend Image-Level Attn by incorporating a branch that uses all CLS tokens to extract spatio-temporal information, reducing complexity to $O(F(N)^2 + (F)^2 N)$:

$$[c_i^{(1)}, \hat{E}_i] = \mathrm{ATTN}(QKV = [c_i^{(0)}, E_i^{(0)}]), \tag{5}$$

$$[c_{1:F}^{(2)}] = \mathrm{ATTN}(Q = [c_{1:F}^{(0)}], KV = [c_{1:F}^{(0)}, E_{1:F}^{(0)}]), \tag{6}$$

$$\hat{c}_i = c_i^{(2)} + \sigma(c_i^{(1)} \mathbf{W}_{\mathrm{down}}) \mathbf{W}_{\mathrm{up}}, \tag{7}$$

where $c_i \in \mathbb{R}^{1 \times D}$ is CLS token of the i^{th} image and $c_{1:F} \in \mathbb{R}^{F \times D}$ are CLS tokens of all sampled images. Similarly, $E_i \in \mathbb{R}^{N \times D}$ are patch tokens of the i^{th} image and $E_{1:F} \in \mathbb{R}^{FN \times D}$ are patch tokens of all sampled images. In Eq. (7), we employ a trainable adapter to merge image-level features $c_i^{(1)}$ and video-level features $c_i^{(2)}$, where $\mathbf{W}_{down} \in \mathbb{R}^{D \times r}$ and $\mathbf{W}_{up} \in \mathbb{R}^{r \times D}$ represent the up- and down-projection layers, and σ denotes the nonlinear GELU activation [19]. Image-level features preserve pretrained spatial details, while video-level features capture temporal information. This fusion mechanism strengthens feature extraction.

Our parameter-efficient vision encoder comprises both image-level and IVFusion layers. The shallow image-level layers employ Image-Level Attn to capture spatial information, while the top H_v IVFusion layers leverage IVFusion Attn to focus on spatio-temporal information. Consequently, our vision encoder outputs enhanced image features $\left\{ v_i^{\text{img}} | 1 \leq i \leq F \right\}$.

3.4. Pseudo Image-Level Alignment

Unlike previous work focused on video-level alignment, we introduce PImgAlign to learn fine-grained image-level alignment via the decomposition of video content at the image level. To accomplish this, we first generate pseudo image captions in VTR datasets. In contrast to the straightforward sampling of images from a video, sampling image captions is impractical, as video captions generally lack the specific details needed for individual images. As a solution, we utilize the multimodal large language model (MLLM) LLaVA-NeXT [35] to generate pseudo image captions. For each sampled image from a video, we prompt the model using the relevant video caption for guidance:

> The provided image is a frame sampled from the video, which describes {*video caption*}. Based on the video's content, provide a caption for the provided image.

Unlike EA-VTR [38], which uses an image captioner, our method employs a strong MLLM. By incorporating video caption guidance, it generates higher-quality image captions that align with the overall video content. Similarly to visual feature extraction, we adopt our proposed IVFusion to extract pseudo image caption features $\left\{t_i^{\text{img}}|1 \leq i \leq F\right\}$.

Given image features and pseudo image caption features from a paired video and video caption, we introduce a fine-grained image-level similarity. Since image captions are typically more concise than the images they describe, a single caption may correspond to multiple images, as shown in Figure 4d. Therefore, we calculate the maximum similarity for each image across all image captions and vice versa. Specifically, we computed a similarity matrix between image and image caption features, selecting the maximum value from each row and column to obtain the final image-level similarity:

$$\text{Sim}_{ij-t2v}^{\text{img}} = \frac{1}{F}\sum_{n=1}^{F}\max_{1\leq m\leq F} {t_{i|n}^{\text{img}}}^T v_{j|m}^{\text{img}}, \tag{8}$$

$$\text{Sim}_{ij-v2t}^{\text{img}} = \frac{1}{F}\sum_{m=1}^{F}\max_{1\leq n\leq F} {t_{i|n}^{\text{img}}}^T v_{j|m}^{\text{img}}, \tag{9}$$

$$\text{Sim}_{ij}^{\text{img}} = \frac{1}{2}[\text{Sim}_{ij-t2v}^{\text{img}} + \text{Sim}_{ij-v2t}^{\text{img}}], \tag{10}$$

where $\text{Sim}_{ij}^{\text{img}}$ denotes the image-level similarity between the i^{th} video caption and the j^{th} video. $t_{i|n}^{\text{img}}$ denotes the n^{th} pseudo image caption features of the i^{th} video caption and $v_{j|m}^{\text{img}}$ denotes the m^{th} image features of the j^{th} video. PImgAlign achieves video-level alignment from the image level, effectively emphasizing spatial details.

However, in the inference stage, the lack of ground truth for video captions prevents pseudo image captions generation. Additionally, MLLM introduces a computational overhead. For fair comparisons with other methods, we exclude PImgAlign during inference.

3.5. Image-to-Video Alignment Distillation

Pretrained CLIP demonstrates a strong capability for image-level alignment. In contrast, video-text retrieval relies on comprehensive spatio-temporal alignment. Therefore, we propose AlignDistill to distill image-level alignment knowledge into video-level alignment, thus mitigating alignment discrepancies. Then, we optimize the Kullback-Leibler divergence [30] between image-level similarity and video-level similarity:

$$\mathcal{S}_{t2v}^{\text{img}} = [\text{s}_{i1}^{\text{img}}, \cdots, \text{s}_{iN_v}^{\text{img}}], \mathcal{S}_{v2t}^{\text{img}} = [\text{s}_{1i}^{\text{img}}, \cdots, \text{s}_{N_ti}^{\text{img}}], \tag{11}$$

$$\mathcal{S}_{t2v}^{\text{vid}} = [\text{s}_{i1}^{\text{vid}}, \cdots, \text{s}_{iN_v}^{\text{vid}}], \mathcal{S}_{v2t}^{\text{vid}} = [\text{s}_{1i}^{\text{vid}}, \cdots, \text{s}_{N_ti}^{\text{vid}}], \tag{12}$$

$$\mathcal{L}_{KL} = \frac{1}{2}(\mathbf{KL}(\mathcal{S}_{t2v}^{\text{img}} \parallel \mathcal{S}_{t2v}^{\text{vid}}) + \mathbf{KL}(\mathcal{S}_{v2t}^{\text{img}} \parallel \mathcal{S}_{v2t}^{\text{vid}})). \tag{13}$$

where $\text{s}_{ij}^{\text{img}}$ is $\text{Sim}_{ij}^{\text{img}}$ in Eq. (10) and $\text{s}_{ij}^{\text{vid}}$ is $\text{Sim}_{ij}^{\text{vid}}$ in Eq. (1). $\mathcal{S}_{t2v}^{\text{img}}$ and $\mathcal{S}_{t2v}^{\text{vid}}$ denote the probability distributions for the text-to-video task, while $\mathcal{S}_{v2t}^{\text{img}}$ and $\mathcal{S}_{v2t}^{\text{vid}}$ denote those for the video-to-text task. During the inference phase, We employ enhanced video-level similarity for retrieval tasks.

Finally, the overall training objective, which includes alignment loss (Eq. (4)) and distillation loss (Eq. (13)), is formulated as follows:

$$\mathcal{L} = \mathcal{L}_A(\text{Sim}^{\text{vid}}) + \alpha\mathcal{L}_A(\text{Sim}^{\text{img}}) + \beta\mathcal{L}_{KL}, \tag{14}$$

where α and β are used to balance the loss. $\mathcal{L}_A(\text{Sim}^{\text{img}})$ is the image-level alignment loss (see Eq. (4) and (10)).

4. Experiments

4.1. Experimental Settings

Following common practice, we perform evaluations on four widely used benchmarks: MSRVTT [53], LSMDC [43], ActivityNet [29], and DiDeMo [1]. 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). 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. Please see Appendix A for further details.

4.2. Comparisons with State-of-the-Art Methods

We compare our proposed DiscoVLA with state-of-the-art methods on popular benchmarks such as MSRVTT, LSMDC, ActivityNet, and DiDeMo.

Full fine-tuning refers to CLIP4Clip-meanP [37], which averages image features along the temporal dimension. All parameter-efficient methods follow this temporal average pooling strategy to obtain the final video features. Although Prompt [27], Adapter [20], and LoRA [21] are widely adopted for parameter-efficient image and NLP tasks, they lack effective temporal modeling for video-level understanding. In contrast, methods like RAP [5], VoP [22], DGL [58], and MV-Adapter [25] focus on temporal modeling in the vision modality.

As shown in Table 1, on MSRVTT, our DiscoVLA with CLIP (ViT-B/32) as the backbone achieves **47.0**% R@1 in

Method	# Params (M)	Text-to-Video					Video-to-Text				
		R@1↑	R@5↑	R@10↑	R@sum↑	MnR↓	R@1↑	R@5↑	R@10↑	R@sum↑	MnR↓
CLIP (ViT-B/32)											
Full fine-tuning	123.54	43.1	70.4	80.8	194.3	16.2	43.1	70.5	81.2	194.8	12.4
Prompt [27]	0.08	40.4	66.3	77.3	184.0	16.7	42.2	69.7	79.2	191.1	12.4
Adapter [20]	0.26	41.9	69.9	78.7	190.2	14.9	43.6	69.9	80.1	193.6	11.5
LoRA [21]	0.49	43.7	68.9	80.4	193.0	16.0	43.0	70.2	82.2	195.4	12.0
RAP [5]	1.06	44.8	71.4	81.5	197.7	14.4	44.0	71.9	82.4	198.3	10.1
$\text{VoP}^{\text{F+P}}$ [22]	0.4	43.5	69.3	79.3	192.1	14.8	43.6	71.2	81.2	196.0	11.0
VoP F+C [22]	14.10	44.7	70.5	79.2	194.4	16.2	42.1	70.0	80.6	192.7	13.4
DGL [58]	0.83	44.6	69.9	80.3	194.8	16.3	44.5	70.7	80.6	195.8	11.5
DiscoVLA	0.56	**47.0**	**73.0**	**82.8**	**202.8**	**14.1**	**47.7**	**73.6**	**83.6**	**204.9**	**10.0**
CLIP (ViT-B/16)											
MV-Adapter [25]	3.6	46.0	72.0	82.1	200.1	-	45.6	74.0	83.8	203.4	-
RAP [5]	1.06	46.5	73.9	82.0	202.4	12.1	45.3	**76.4**	84.8	206.5	9.1
$\text{VoP}^{\text{F+P}}$ [22]	0.4	47.1	72.4	81.8	201.3	12.9	-	-	-	-	-
$\text{VoP}^{\text{F+C}}$ [22]	14.10	47.7	72.4	82.2	202.3	12.0	-	-	-	-	-
DGL [58]	0.83	48.3	71.8	80.6	200.7	13.4	45.7	74.0	82.9	202.6	10.9
TempMe [44]	0.50	49.0	74.4	83.3	206.7	**11.9**	47.6	75.3	**85.4**	208.3	9.0
DiscoVLA	0.56	**50.5**	**75.6**	**83.8**	**209.9**	12.1	**49.2**	76.0	84.7	**209.9**	**8.6**

Table 1. Comparisons with state-of-the-art methods on MSRVTT. # Params denotes the number of trainable parameters. ↑ denotes higher values represent better performance and ↓ denotes lower values represent better performance. The best value for each metric is highlighted in bold. R@sum is defined as the total of R@1, R@5, and R@10. The gray row represents the fully fine-tuned CLIP4Clip-meanP [37].

Method	# Params (M)	Text-to-Video					Video-to-Text				
		R@1↑	R@5↑	R@10↑	R@sum↑	MnR↓	R@1↑	R@5↑	R@10↑	R@sum↑	MnR↓
Full fine-tuning	123.54	20.7	38.9	47.2	106.8	65.3	20.6	39.4	47.5	107.5	56.7
$\text{VoP}^{\text{F+P}}$ [22]	0.4	20.7	40.7	49.7	111.1	59.1	21.5	40.6	50.7	112.8	50.8
$\text{VoP}^{\text{F+C}}$ [22]	14.10	21.1	40.9	49.6	111.6	60.1	22.3	40.3	50.7	113.3	51.1
DGL [58]	0.83	21.4	39.4	48.4	109.2	64.3	-	-	-	-	-
DiscoVLA	0.56	**23.6**	**42.0**	**52.3**	**117.9**	**52.0**	**22.8**	**41.9**	**51.2**	**115.9**	**46.6**

Table 2. Comparisons with state-of-the-art methods on LSMDC using CLIP (ViT-B/32).

Method	# Params (M)	Text-to-Video					Video-to-Text				
		R@1↑	R@5↑	R@10↑	R@sum↑	MnR↓	R@1↑	R@5↑	R@10↑	R@sum↑	MnR↓
Full fine-tuning	123.54	40.5	72.4	-	-	7.4	42.5	74.1	85.8	202.4	6.6
RAP [5]	1.06	40.8	71.0	82.2	194.0	8.3	-	-	-	-	-
$\text{VoP}^{\text{F+P}}$ [22]	0.4	36.1	65.5	78.5	180.1	10.9	36.3	65.9	79.2	181.4	10.1
$\text{VoP}^{\text{F+C}}$ [22]	14.10	35.1	63.7	77.6	176.4	11.4	35.6	65.9	77.8	179.3	10.4
DGL [58]	0.83	38.6	69.2	81.6	189.4	9.0	-	-	-	-	-
DiscoVLA	0.56	**41.2**	**72.4**	**83.6**	**197.2**	**7.8**	**41.8**	**72.8**	**84.7**	**199.3**	**6.9**

Table 3. Comparisons with state-of-the-art methods on ActivityNet using CLIP (ViT-B/32).

the text-to-video task (t2v) and **47.7**% R@1 in the video-to-text task (v2t), significantly outperforming previous methods. Furthermore, when using CLIP (ViT-B/16), DiscoVLA achieves improvements of **2.2**% R@1 and **7.5**% R@sum. The comparison results on LSMDC, ActivityNet, DiDeMo are shown in Table 2-4. Our method outperforms all existing approaches across these datasets. Existing approaches overlook the critical image-to-video language and alignment discrepancies. In contrast, our proposed DiscoVLA addresses these discrepancies across all three areas: vision, language, and alignment.

4.3. Ablation Study

Ablation study on individual components. We evaluate the proposed components in Table 5. LoRA applies Image-Level Attn (see Figure 3a) in both the vision and text encoders. In B1, IVFusion significantly enhances performance by **8.7**% R@sum in t2v and **6.7**% R@sum in v2t, effectively integrating image-level and video-level features. In B2 and B3, we find that using PImgAlign or AlignDistill independently yields limited improvements. This is largely due to the notable gap between image-level and video-level (see Figure 1). Image-level PImgAlign struggles to impact video-level tasks, while AlignDistll lacks sufficient image-specific learning. When combined, however, they complement each other. PImgAlign captures fine-grained image-level alignments, and AlignDistill effectively transfers this knowledge to the video-level alignment. This improves **1.1**% R@sum in t2v and **2.8**% R@sum in v2t over B1. Finally, our full DiscoVLA improves **9.8**% R@sum in t2v and **9.5**% R@sum in v2t over LoRA.

Method	# Params (M)	Text-to-Video					Video-to-Text				
		R@1↑	R@5↑	R@10↑	R@sum↑	MnR↓	R@1↑	R@5↑	R@10↑	R@sum↑	MnR↓
Full fine-tuning	123.54	43.4	70.2	80.6	194.2	17.5	42.5	70.6	80.2	193.3	11.6
RAP [5]	1.06	42.6	70.4	79.6	192.6	18.0	-	-	-	-	-
$\text{VoP}^{\text{F+P}}$ [22]	0.4	45.3	72.3	80.4	198.0	13.8	44.7	71.2	81.1	197.0	9.9
$\text{VoP}^{\text{F+C}}$ [22]	14.10	46.4	71.9	81.5	199.8	**13.6**	44.4	71.8	81.8	198.0	9.5
DiscoVLA	0.56	**48.4**	**74.5**	**82.7**	**205.6**	14.0	**47.7**	**74.4**	**83.8**	**205.9**	**9.3**

Table 4. Comparisons with state-of-the-art methods on DiDeMo using CLIP (ViT-B/32).

Methods	Components			# Params	Text-to-Video				Video-to-Text			
	IVFusion	PImgAlign	AlignDistill	(M)	R@1↑	R@5↑	R@10↑	R@sum↑	R@1↑	R@5↑	R@10↑	R@sum↑
Pretraind CLIP	✗	✗	✗	0	30.8	53.8	63.3	147.9	26.6	50.1	62.0	138.7
LoRA	✗	✗	✗	0.49	43.7	68.9	80.4	193.0	43.0	70.2	82.2	195.4
B1	✓	✗	✗	0.54	46.5	**73.2**	82.0	201.7	46.6	72.6	82.9	202.1
B2	✓	✓	✗	0.56	**47.0**	73.1	81.3	201.4	45.7	72.9	82.4	201.0
B3	✓	✗	✓	0.54	46.6	71.6	81.0	199.2	45.9	71.9	82.2	200.0
DiscoVLA	✓	✓	✓	0.56	**47.0**	73.0	**82.8**	**202.8**	**47.7**	**73.6**	**83.6**	**204.9**

Table 5. Ablation study on the contribution of each proposed component on MSRVTT using CLIP (ViT-B/32). Pretrained CLIP denotes the zero-shot performance of CLIP without any additional training. Our proposed methods are built upon LoRA. B1, B2, and B3 denote various combinations of our proposed components.

Parameter-Efficient Attention	Text-to-Video			
	R@1↑	R@5↑	R@10↑	R@sum↑
Image-Level Attn	43.7	68.9	80.4	193.0
Video-Level Attn	45.7	70.9	80.8	197.4
IVFusion Attn *w/o* Adapter	46.0	71.0	80.4	197.4
IVFusion Attn	**46.5**	**73.2**	**82.0**	**201.7**

Table 6. Ablation study on IVFusion on MSRVTT using CLIP (ViT-B/32). We compare several implementations against our proposed IVFusion for Parameter-Efficient Attention (see Figure 3). Image-Level Attn represents the LoRA baseline.

Similarity in PImgAlign	Text-to-Video			
	R@1↑	R@5↑	R@10↑	R@sum↑
None	46.6	71.6	81.0	199.2
Video-Level	46.3	72.3	81.2	199.8
Paired Image-Level	46.1	**73.2**	82.4	201.7
Fine-grained Image-Level	**47.0**	73.0	**82.8**	**202.8**

Table 7. Ablation study on similarity in PImgAlign on MSRVTT using CLIP (ViT-B/32). None represents DiscoVLA without image-level alignment optimization, shown as B3 in Table 5. Paired Image-Level is based on matched image-caption pairs. Fine-grained Image-Level represents the full DiscoVLA.

Analysis of trainable parameters. The trainable parameters of our DiscoVLA consist of the LoRA and Adapter parts. As shown in Table 5, LoRA in both the vision and text encoders contain ∼0.49M parameters. In B1, an additional ∼0.05M parameters are introduced by the Adapter in the $H_V = 4$ IVFusion layers of the vision encoder. In B2, an additional ∼0.02M parameters are introduced by the Adapter in the $H_L = 2$ IVFusion layers of the text encoder, which are utilized for pseudo-image caption features in PImgAlign. Overall, the total number of trainable parameters of DiscoVLA is ∼0.56M.

Ablation study on IVFusion. In Table 6, we compare different strategies for parameter-efficient attention in Figure 3. Video-Level Attn concatenates all tokens across sampled images to capture spatio-temporal information,

Image Caption Generation	Text-to-Video			
	R@1↑	R@5↑	R@10↑	R@sum↑
DiscoVLA	47.0	73.0	82.8	202.8
W/O VidCap	46.1 (-0.9)	73.0	82.1 (-0.7)	201.2 (-1.6)

Table 8. Ablation study on image caption generation on MSRVTT using CLIP (ViT-B/32). W/O VidCap denotes generating pseudo image captions without video caption guidance.

achieving a notable **4.4**% R@sum improvement. To reduce the computational complexity, IVFusion applies spatio-temporal attention only to global CLS tokens and merges image-level features with video-level features. Without an adapter, averaging image- and video-level features results in a modest performance increase. In contrast, with a learnable adapter, IVFusion merges the features more effectively, achieving **201.7**% R@sum.

Ablation study on similarity in PImgAlign. Table 7 demonstrates the effectiveness of our proposed fine-grained image-level similarity in PImgAlign. The similarity is computed between sampled images and pseudo image captions from a paired video and video caption. The Video-Level baseline averages the image features and pseudo-image caption features along the temporal dimension to compute the video-level similarity, replacing fine-grained image-level similarity. However, this modification leads to a notable decrease in performance, emphasizing the importance of image-level similarity in PImgAlign. In contrast, we introduce the Paired Image-Level baseline, which computes similarity based on matched image-caption pairs. Compared to Video-Level, Paired Image-Level yields a **1.9**% gain in R@sum. Moreover, we observe in Figure 4d that a single image caption may describe multiple images. Consequently, our fine-grained image-level similarity, as defined in Eq. 10, selects the maximum similarity from all available pairs, further enhancing video-level alignment.

Figure 4. Visualization of discrepancies between video-level and image-level data. Each example consists of a paired video and video caption, with four sampled images from the video. Video captions are highlighted with orange solid lines, and pseudo image captions generated by LLaVA-NeXT [35] are indicated with green dashed lines. While video captions convey the general context of the video, image captions focus on the detailed context of each individual frame.

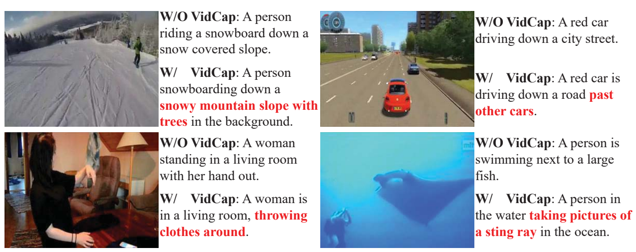


Figure 5. Comparison of image caption generation with (W/) and without (W/O) video caption guidance (VidCap) in PImgAlign.

Ablation study on image caption generation in PImgAlign. Figure 5 and Table 8 show qualitative and quantitative evaluations of video caption guidance. In Figure 5, video captions serve as the global context that enriches the understanding of individual frames (marked in red). In Table 8, without this guidance, image captions focus only on salient regions while missing globally relevant details, leading to performance drops. In PImgAlign, we employ MLLM to process both frame images and video descriptions, generating image captions that align with the overall video content. These high-quality captions further enhance DiscoVLA's ability to learn robust image-level alignment.

More Ablation Analysis. In Appendix B, we conduct additional ablation studies. Specifically, we evaluate the effects of hyperparameters α, β, H_V and H_V. For fairness, all results in the main paper are reported without post-processing. Complementary evaluations with Q-Norm [4] and DSL [8] as post-processing methods are provided in the appendix.

4.4. Qualitative Analysis

To compare image-level and video-level data, we conduct a visualization experiment in Figure 4. In Figure 4a and 4b, image-level data alone fails to capture the complete content of the video. Figure 4c and 4d demonstrate a strong correlation between the video-level and image-level content, but differences across images lead to varying image captions. Visually, a video is composed of a sequence of images, incorporating temporal information. Textually, image captions are specific to individual images, while video captions describe the overall video content. These discrepancies hinder the effective transfer of image-level CLIP. Our approach addresses this issue by utilizing image-level alignment to enhance video-level alignment.

5. Conclusion

In this paper, we identify three major discrepancies in the parameter-efficient transfer of image-level CLIP to video-text retrieval: vision, language, and alignment. To address these challenges, we propose DiscoVLA. Specifically, IV-Fusion fuses image-level and video-level features to tackle the vision and language discrepancies. For alignment discrepancies, we propose PImgAlign and AlignDistill, which learn and distill image-level alignment to enhance video-level alignment. Our DiscoVLA significantly outperforms existing methods, achieving state-of-the-art performance.

Acknowledgment

This work was supported by Beijing Natural Science Foundation (No. L223023), National Natural Science Foundation of China (Nos. 62441235, 62021002, 62441614), the Key R & D Program of Xinjiang, China (2022B01006), Zhejiang Provincial Natural Science Foundation of China under Grant (No. LDT23F01013F01), CCF-DiDi GAIA Collaborative Research Funds, China Postdoctoral Science Foundation (2024M750565), Guangdong S & T Program (2024B0101040008), and the Key Realm Research and Development Program of Guangzhou (No. 2024B01W0007).

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the International Conference on Computer Vision*, 2017. 5, 12

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the International Conference on Computer Vision*, 2015. 2

[3] 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 International Conference on Computer Vision*, 2021. 12

[4] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 8, 13

[5] Meng Cao, Haoran Tang, Jinfa Huang, Peng Jin, Can Zhang, Ruyang Liu, Long Chen, Xiaodan Liang, Li Yuan, and Ge Li. RAP: Efficient text-video retrieval with sparse-and-correlated adapter. In *Association for Computational Linguistics*, 2024. 2, 3, 5, 6, 7, 12

[6] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 2022. 2

[7] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 2023. 2

[8] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021. 8, 13

[9] Chaorui Deng, Qi Chen, Pengda Qin, Da Chen, and Qi Wu. Prompt switch: Efficient clip adaptation for text-video retrieval. In *Proceedings of the International Conference on Computer Vision*, 2023. 1, 2

[10] Xiaohan Ding, Yuchen Guo, Guiguang Ding, and Jungong Han. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In *Proceedings of the International Conference on Computer Vision*, 2019. 2

[11] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[12] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 2024. 2

[13] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Proceedings of the European Conference on Computer Vision*, 2020. 12

[14] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2

[15] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 12

[16] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 2005. 2

[17] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *Proceedings of the International Conference on Learning Representations*, 2021. 2

[18] Tao He, Leqi Shen, Yuchen Guo, Guiguang Ding, and Zhenhua Guo. Secret: Self-consistent pseudo label refinement for unsupervised domain adaptive person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 2

[19] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[20] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *Proceedings of the International Conference on Machine Learning*, 2019. 2, 5, 6

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 5, 6

[22] Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, and Donglin Wang. Vop: Text-video cooperative prompt tuning for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3, 5, 6, 7, 12

[23] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Proceedings of the European Conference on Computer Vision*, 2022. 2

[24] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2

[25] Xiaojie Jin, Bowen Zhang, Weibo Gong, Kai Xu, Xueqing Deng, Peng Wang, Zhao Zhang, Xiaohui Shen, and Jiashi Feng. Mv-adapter: Multimodal video transfer learning for video text retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3, 5, 6, 12

[26] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[27] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2, 5, 6

[28] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the International Conference on Computer Vision*, 2023. 2

[29] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the International Conference on Computer Vision*, 2017. 5, 12

[30] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997. 5

[31] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 12

[32] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 2021. 2

[33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*, 2022.

[34] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the International Conference on Computer Vision*, 2023. 2

[35] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 4, 8

[36] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *Proceedings of the International Conference on Learning Representations*, 2016. 12

[37] 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*, 2022. 1, 2, 5, 6

[38] Zongyang Ma, Ziqi Zhang, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Yingmin Luo, Xu Li, Xiaojuan Qi, Ying Shan, et al. Ea-vtr: Event-aware video-text retrieval. In *Proceedings of the European Conference on Computer Vision*, 2024. 5

[39] 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 *Proceedings of the International Conference on Computer Vision*, 2019. 12

[40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[41] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019. 2

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021. 1, 2

[43] Anna Rohrbach, Marcus Rohrbach, and Bernt Schiele. The long-short story of movie description. In *Pattern Recognition*, 2015. 5, 12

[44] Leqi Shen, Tianxiang Hao, Sicheng Zhao, Yifeng Zhang, Pengzhang Liu, Yongjun Bao, and Guiguang Ding. Tempme: Video temporal token merging for efficient text-video retrieval. *arXiv preprint arXiv:2409.01156*, 2024. 6

[45] Leqi Shen, Tao He, Sicheng Zhao, Zhelun Shen, Yuchen Guo, Tianshi Xu, and Guiguang Ding. X-reid: Cross-instance transformer for identity-level person reidentification. In *IEEE International Conference on Multimedia and Expo*, 2024. 2

[46] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[47] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 2024. 2

[48] Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. Text is mass: Modeling as stochastic embedding for text-video retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2

[49] Qiang Wang, Yanhao Zhang, Yun Zheng, Pan Pan, and Xiansheng Hua. Disentangled representation learning for text-video retrieval. *arXiv preprint arXiv:2203.07111*, 2022. 1, 2

[50] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 2

[51] Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4video: What can auxiliary captions do for text-video retrieval? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2

[52] Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *Proceedings of the International Conference on Learning Representations*, 2024. 2

[53] 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*, 2016. 5, 12

[54] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*, 2015. 2

[55] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language alignment. In *Proceedings of the International Conference on Learning Representations*, 2022. 2

[56] Fan Yang, Sicheng Zhao, Yanhao Zhang, Haoxiang Chen, Hui Chen, Wenbo Tang, Haonan Lu, Pengfei Xu, Zhenyu Yang, Jungong Han, et al. Llmi3d: Empowering llm with 3d perception from a single 2d image. *arXiv preprint arXiv:2408.07422*, 2024. 2

[57] Fan Yang, Ru Zhen, Jianing Wang, Yanhao Zhang, Haoxiang Chen, Haonan Lu, Sicheng Zhao, and Guiguang Ding. Heie: Mllm-based hierarchical explainable aigc image implausibility evaluator. *arXiv preprint arXiv:2411.17261*, 2024. 2

[58] Xiangpeng Yang, Linchao Zhu, Xiaohan Wang, and Yi Yang. Dgl: Dynamic global-local prompt tuning for text-video retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 1, 2, 3, 5, 6, 12

[59] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *Proceedings of the European Conference on Computer Vision*, 2018. 12

[60] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. *arXiv preprint arXiv:2206.04673*, 2022. 2

[61] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[62] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2022. 2