The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)

DGL: Dynamic Global-Local Prompt Tuning for Text-Video Retrieval

Xiangpeng Yang[1], **Linchao Zhu**[2], **Xiaohan Wang**[2], **Yi Yang**[2] *

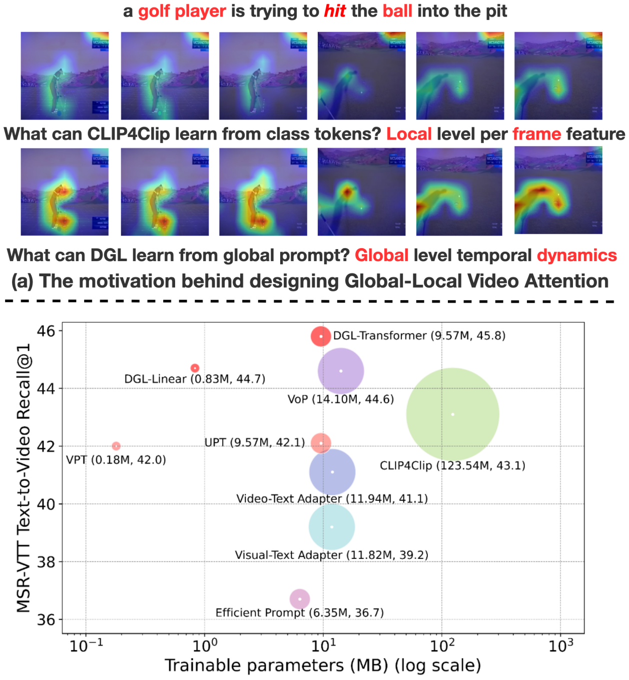
[1] ReLER, AAII, University of Technology Sydney
[2] CCAI, Zhejiang University
Xiangpeng.Yang@student.uts.edu.au, wxh1996111@gmail.com, {zhulinchao,yangyics}@zju.edu.cn

Abstract

Text-video retrieval is a critical multi-modal task to find the most relevant video for a text query. Although pretrained models like CLIP have demonstrated impressive potential in this area, the rising cost of fully finetuning these models due to increasing model size continues to pose a problem. To address this challenge, prompt tuning has emerged as an alternative. However, existing works still face two problems when adapting pretrained image-text models to downstream video-text tasks: (1) The visual encoder could only encode frame-level features and failed to extract global-level general video information. (2) Equipping the visual and text encoder with separated prompts failed to mitigate the visual-text modality gap. To this end, we propose ***DGL***, a cross-modal **D**ynamic prompt tuning method with **G**lobal-**L**ocal video attention. In contrast to previous prompt tuning methods, we employ the shared latent space to generate local-level text and frame prompts that encourage inter-modal interaction. Furthermore, we propose modeling video in a global-local attention mechanism to capture global video information from the perspective of prompt tuning. Extensive experiments reveal that when only **0.67%** parameters are tuned, our cross-modal prompt tuning strategy DGL outperforms or is comparable to fully finetuning methods on MSR-VTT, VATEX, LSMDC, and ActivityNet datasets. Code will be available at https://github.com/knightyxp/DGL

Introduction

With the recent advancement of large-scale contrastive image-text pretraining methods *i.e.,* CLIP (Radford et al. 2021), the field of TVR (Text-Video Retrieval) has experienced many works (Luo et al. 2022; Gorti et al. 2022; Zhao et al. 2022; Liu et al. 2022; Ma et al. 2022; Wang et al. 2022a) to adapt image-text pretrained models like CLIP to the video-text domain and already achieve the promising performance. These approaches incur a large storage burden in actual scenarios because they need to store distinct new models for different tasks. However, as the capacity of pretrained models is rapidly expanding nowadays, *i.e.,* BEIT-3 (Wang et al. 2022b) has 1.9B parameters, and BLIP-L/14 has 578M parameters. Fully finetuning the entire model for each downstream task requires maintaining separate model weights for every dataset, hindering the feasibility of deployment given the growing model capacities.



(a) The motivation behind designing Global-Local Video Attention

(b) Different Parameter-efficient methods comparison on MSRVTT

Figure 1: In Fig (a), we observed that frame attention methods, like CLIP4Clip, often emphasize non-semantic corners, missing the protagonist's action. This led us to design the global-local video attention for capturing global-level cross-frame dynamics. Fig (b) showcases a performance comparison on MSRVTT: DGL outperforms six PEFL methods and fine-tuned CLIP4Clip while updating minimal parameters.

To address this problem, inspired by the recent success of prompt tuning in both NLP (Lester, Al-Rfou, and Constant 2021; Liu et al. 2021) and common visual recognition tasks (Jia et al. 2022), we continue to introduce **prompt tuning** to the cross-modal domain. In this way, we only need to store the parameters of a few prompt vectors for various retrieval tasks and keep the pretrained model backbone frozen, thus

*Corresponding author

reducing the total parameter cost.

Efficient Prompt (Ju et al. 2022) is the first work that has attempted prompt tuning in this area, introducing learnable prompt vectors in the text input while viewing the video as separate frames. Despite incorporating an additional transformer for temporal encoding, the performance remains unsatisfactory. VoP (Huang et al. 2023), another recent prompt tuning approach, designs three kinds of video-specific prompts but optimizes the dual branches' prompts independently. We argue these current methods still fail to handle two key challenges when applying prompt tuning in text-video retrieval. (1) Cross-modal alignment: Existing schemes, like those in VoP and Efficient Prompt, optimize the two branches separately, making it challenging for the model to learn mutual cross-modal information effectively. (2) General video information extraction: Since CLIP is pretrained on image-text pairs, its primary focus is on local-level frame features rather than holistic, global-level video information. This inherent design leads to potential pitfalls when used directly for TVR tasks. In the top images of Fig 1 (a), the attention weights of CLIP4Clip's CLS tokens reveal this limitation. Specifically, the CLS tokens overlook the action of "hit the ball" and instead allocate more attention to the upper-right corner – a semantically void region.

To address the aforementioned issues, we propose dynamic global-local prompt tuning (coined as **DGL**) for text-video retrieval. Our approach generates dynamic local-level prompts (text and frame prompts) from a shared latent space. This allows for joint optimization and ensures the alignment of the two modalities. Moreover, we propose global-local video attention to model videos from both the global and local levels, capturing inter-frame temporal information with the global prompt and focusing on each frame's information with the local frame prompts.

From a qualitative standpoint, the bottom image of Fig 1(a) clearly shows the effectiveness of **DGL**. In contrast to CLIP4Clip, our DGL can focus on the "hit" action and the ball's trajectory into the pit. This demonstrates that our method can efficiently capture temporal dynamics. Besides, on the quantitative front, as shown in Fig 1(b), our method achieves the best trade-off between trainable parameters and performance. More specifically, with only 9.57M parameters updated, DGL achieves 45.8 R@1 on MSRVTT. These results demonstrate the importance of cross-modal interaction and a comprehensive understanding of video information. We undertake extensive experiments on four benchmarks, including MSR-VTT, VATEX, LSMDC, and ActivityNet. Our contributions can be summarized as follows:

- We propose to generate dynamic cross-modal prompts from the shared latent space to ensure the cross-modal interaction.
- We propose a global-local attention mechanism for a comprehensive understanding of input video, facilitating effective learning of cross-frame temporal dynamics.
- Compared to the fully finetuning CLIP4Clip and other prompt tuning methods, our DGL achieves superior or equivalent performance on R@1 across four public datasets while reducing 99.3% of trainable parameters.

Related Work

Text-Video Retrieval. Text-video retrieval is a prevalent task in multimodal learning. Previous works like (Zhu and Yang 2020; Wang, Zhu, and Yang 2021; Sun et al. 2019; Bain et al. 2021; Lei et al. 2021; Liang et al. 2023) utilize abundant video information for multimodal learning. With the pretrained models like CLIP (Radford et al. 2021) gaining traction, CLIP4Clip (Luo et al. 2022) proposed to fine-tune CLIP on text-video retrieval by adding extensive similarity calculation mechanisms, which shows good performance on several benchmarks. This inspired follow-up research (Gorti et al. 2022; Bogolin et al. 2022; Liu et al. 2022; Zhao et al. 2022; Fang et al. 2021) which delved deeper into cross-modal learning. Recent research (Wu et al. 2023; Jin et al. 2023a,b) have introduced external tools for enhanced retrieval but predominantly utilize features extracted from CLIP. Our approach continues to build upon the foundation set by CLIP4Clip, emphasizing efficient parameter learning within the encoder.

Parameter Efficient Methods. Fully fine-tuning is a common approach to adapting pretrained models into downstream tasks, but it can be inefficient due to large parameter sizes and time costs. To address this, parameter-efficient learning (PEFL) has been proposed, including adapter and prompt tuning methods. **Adapters** (Houlsby et al. 2019) offers a plug-and-play approach by adding modules to pretrained networks. VL-Adapter (Sung, Cho, and Bansal 2022) further extends the adapter to vision-and-language tasks. Recently, (Jiang et al. 2022) introduced a weight-share mechanism and adopted the query-scoring frame features reweighting method proposed in (Bain et al. 2022) to boost performance. (Zhang et al. 2023) proposed a temporal adaptation and cross-modal interaction modules. **Prompt tuning** (Lester, Al-Rfou, and Constant 2021) is another parameter-efficient choice by introducing additional learnable parameters at the model's input. (Liu et al. 2021) further applies prompts to each encoder layer for more knowledge probing. Adaptations to models like CLIP for specific tasks have been explored in (Zhou et al. 2022b,a), with further refinements in image and cross-modal domains (Jia et al. 2022; Zang et al. 2022; Khattak et al. 2022). In the text-video retrieval task, Efficient Prompt (Ju et al. 2022) tried incorporating additional prompts into text queries but overlooked the potential of visual prompts in this context. While (Huang et al. 2023) made advancements with video-specific prompts, they still hard to address cross-modal interactions in prompt tuning. In this paper, we delve deeper into cross-modal prompt tuning and seek effective ways to represent videos, considering their inherent complexity compared to text.

Methods

In this section, we will illustrate the details of our method. Firstly, we provide a comprehensive overview of the proposed DGL framework. Furthermore, we will introduce how to design global-local video attention to learn discriminative features from holistic video information.

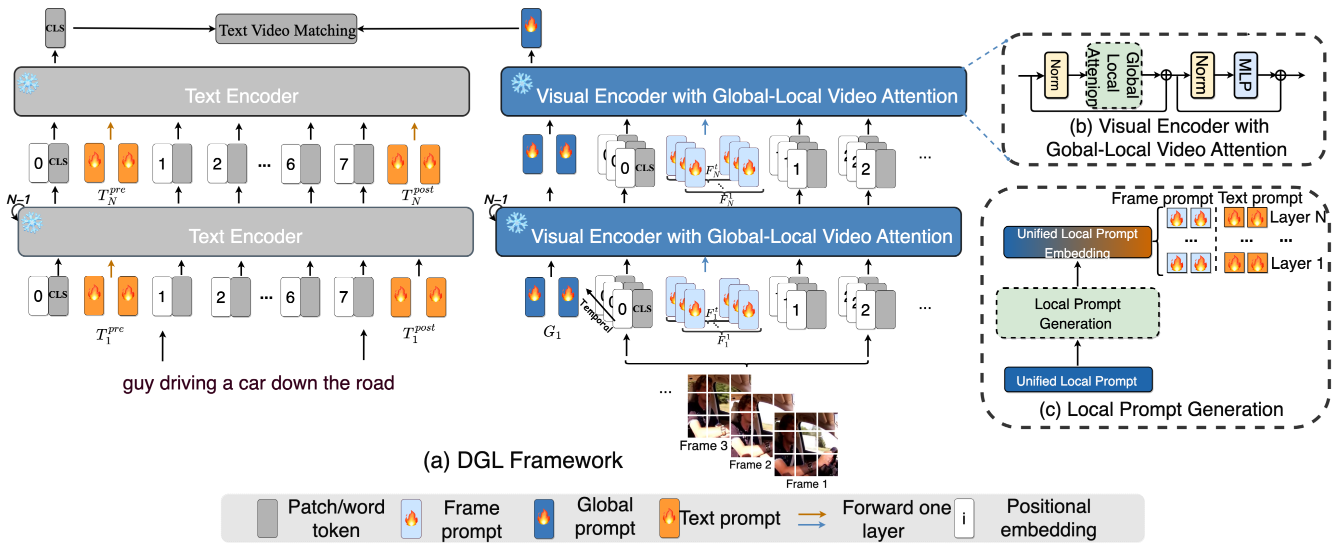


Figure 2: Overview of our Dynamic Global-Local prompt tuning Framework. DGL consists of the Local Prompt Generation, Text Encoder, and Visual Encoder. During downstream training, all the encoders are frozen, and only the parameter pictured with fire is trainable. The Local Prompt Generation ensures cross-modal interaction at the word-frame level, and the Global-Local Video Attention hints to the visual encoder to extract general video information from different perspectives.

Global-Local Prompt Tuning Framework

Current parameter-efficient methods in TVR, such as Efficient Prompt (Ju et al. 2022), and VoP (Huang et al. 2023) often neglect the crucial interaction between the visual and text modalities, as they focus only on inserting prompt vectors in the text input or prompting the dual branches separately. Additionally, deep prompts are necessary for more complex tasks like TVR, as evidenced by studies like VPT (Jia et al. 2022), which showed shallow prompts to be less effective for traditional visual tasks like classification and segmentation. Meanwhile, classic fully finetuning methods like CLIP4Clip (Luo et al. 2022), TS2Net, Cap4Video, and HBI (Liu et al. 2022; Wu et al. 2023; Jin et al. 2023a) still process videos as discrete frames, owing to CLIP's text-image pretraining. This complicates modeling inter-frame relationships and capturing temporal information.

To address these issues, we introduce DGL, a dynamic global-local prompt tuning method that facilitates global video-level information learning and ensures cross-modal alignment between frame-level visual features and word-level text features. Our DGL framework, as illustrated in Fig 2, consists of a local prompt generation module, a text encoder, and a visual encoder.

In the text branch, following (Ju et al. 2022), we design to learn a set of deep text prompts, including prefix text prompts $T_i^{pre} = \{T_i^{pre;j} \in \mathbb{R}^d | j \in \mathbb{N}, 1 \leq j \leq n_{pre}\}$ and postfix text prompts $T_i^{post} = \{T_i^{post;j} \in \mathbb{R}^d | j \in \mathbb{N}, 1 \leq j \leq n_{post}\}$ for each layer index i. The prefix text prompts are added to the input text query before the word embedding, while the postfix text prompts are placed afterward. Here, d is the dimension of the prompt vectors, n_{pre} and n_{post} denote the numbers of prefix and postfix prompts, respectively.

In the visual branch, to perform global-local video attention, we consider learning a single layer of n_g global prompts $G = \{G^j \in \mathbb{R}^d | j \in \mathbb{N}, 1 \leq j \leq n_g\}$ to capture the global information and a set of deep frame prompts $F_i^k = \{F_i^{k;j} \in \mathbb{R}^d | j \in \mathbb{N}, 1 \leq k \leq t, 1 \leq j \leq n_f\}$ to extract frame information. Here, k is frame index, t is the number of frames, j is the length index for frame prompts. n_f and n_g are the length of each frame prompts and global prompts.

The visual encoder input contains global video prompts, frame patch tokens, and frame prompts. The text encoder input consists of text prompts and word tokens. Given the different choices of prompt generation modules, we dub our method as DGL-Transformer and DGL-Linear, respectively.

Local Prompt Generation

We utilize two methods to generate local-level cross-modal prompts from the shared latent space and optimize them jointly. The first approach is the unified prompt transformer, and the second approach is the unified linear projection. We describe the details of these two methods as follows:

Unified Prompt Transformer. To exploit the cross-modal interaction at the fine-grained level, inspired by UPT (Zang et al. 2022), we propose to generate frame prompts and text prompts from a unified transformer (short as "trans"). For each layer, we merge text and visual prompts to form the unified prompt $U_i = [T_i^{pre}, T_i^{post}, F_i^k]$, processed in a unified prompt transformer for cross-modal interaction. We learn layer-wise unified prompts U_i^{trans} for both text and visual encoders. After transformation, U_i^{trans} splits into three parts $\{T_i^{pre}, T_i^{post}, F_i^k\}$, where the first two are sent into the text encoder and the last into the visual encoder. Notably, our unified prompt transformer only has a single layer. The hidden dimension matches the visual encoder's. Besides, we use an MLP Layer to adjust the text prompts' dimension.

Unified Linear Projection. To further reduce the parameter cost, we consider using two simple linear layers, U^{pre}_{linear} and U^{post}_{linear}, to map the frame prompts to the text prefix and text postfix prompts in each encoder layer respectively. This process can be formulated as follows:

$$T^{pre}_i = U^{pre}_{linear}(F^k_i) \qquad 1 \leq i \leq N \tag{1}$$

$$T^{post}_i = U^{post}_{linear}(F^k_i) \qquad 1 \leq i \leq N \tag{2}$$

Here i is the layer index, N is the total layers, and is 12 in our DGL, the same as the total encoder layers in CLIP. The length of each frame prompt and text prefix/postfix prompts are the same. All layers' multi-modal local prompts share the two projection layers. Therefore, the unified linear projection minimizes the parameter cost.

The two local prompt generation modules enable efficient interaction by mapping different modal prompts from the shared latent space (either the lightweight transformer or the linear layers), which ensures the cross-modal alignment between video frames and text words.

Text Encoder and Visual Encoder

Text Encoder. In the text branch, combined with the text prefix prompts T^{pre}_i and text postfix prompts T^{post}_i, we get the i_{th} layer's text embedding as follows:

$$[\_, W_i, \_] = L^t_i([T^{pre}_{i-1}, W_{i-1}, T^{post}_{i-1}]) \tag{3}$$

where $[\cdot, \cdot, \cdot]$ refers to the concatenation operation, L^t_i represents the i_{th} text encoder layer, W_i is the word embedding of i_{th} text encoder layer. The prefix text prompts and postfix prompts are updated by the local prompt generation module in each layer. We get the final text representation by projecting from the final layer's word embedding W_N.

Visual Encoder. For the i_{th} ViT layer, combined with the global prompts G_i and local frame prompts F^k_i, the prompt augmented ViT layer can be formulated as:

$$[G_i, C^k_i, \_, E^k_i] = L^v_i([G_{i-1}, C^k_{i-1}, F^k_{i-1}, E^k_{i-1}]) \tag{4}$$

where i is the layer index, and k is the frame index. L^v_i represents the i_{th} visual encoder layer. C^k_i, E^k_i represent each frame's [CLASS] embeddings and patch embeddings, respectively. Besides, since the global prompts G_i are only prepended in the first visual layer, therefore, $\{G_i | i \neq 1\}$ means the global prompt embedding for the i_{th} ViT layer.

Leveraging Global-Local Video Attention

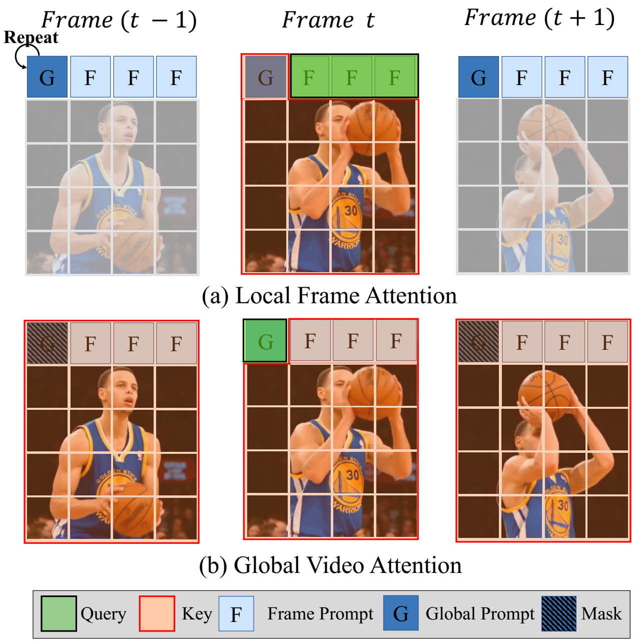
In the text-video matching task, some text caption is related to single or short-latest frames. Thus the local frames feature is a must and basic. While some text query is a summary of behaviors of a video, therefore global video feature is also significant. Following (Xue et al. 2022), we propose to devise local frame and global video attention in a share-parameter manner to extract frame and global video features based on the frame prompts and global prompts, respectively.

Local Frame Attention. Specifically, for the local frame attention, we want each frame prompt could perceive each local frame information. As shown in Fig 3 (a), in the i_{th} ViT layer, we concatenate the [CLASS] embeddings, frame prompt embeddings, and frame patch embeddings along the temporal dimension k, therefore, $[C^k_{i-1}, F^k_{i-1}, E^k_{i-1}]$ serve as the query Q^{loc}_{i-1}. To ensure each frame prompt could perceive global information, we repeat global prompt embeddings k times and concatenate them all. Therefore, we get $[G^k_{i-1}, C^k_{i-1}, F^k_{i-1}, E^k_{i-1}]$ as the key K^{loc}_{i-1}. Our Local frame attention can be formulated as follows:

$$Q^{loc}_{i-1} = [C^k_{i-1}, F^k_{i-1}, E^k_{i-1}] \tag{5}$$

$$K^{loc}_{i-1} = [G^k_{i-1}, C^k_{i-1}, F^k_{i-1}, E^k_{i-1}] \tag{6}$$

$$[C^k_i, F^k_i, E^k_i] = Att(Q^{loc}_{i-1}, K^{loc}_{i-1}, V^{loc}_{i-1}) \tag{7}$$



(a) Local Frame Attention

(b) Global Video Attention

Figure 3: Illustration of Global-Local Video Attention. The patches in the green mask serve as the queries in self-attention, and the patches in the orange mask are the key or value in self-attention. In the local frame attention, frame prompts serve as the query to investigate fine-grained local information in each frame; In global video attention, the global prompt acts as the query to excavate the global-level video information from all frames.

Global Video Attention. For global video attention, as shown in Fig 3 (b), global prompts G_i need to learn global discriminant information. Therefore global prompts are attended to all frames' patch embeddings and prompt embeddings. This process can be formulated as follows:

$$Q^{glo}_{i-1} = G_{i-1} \tag{8}$$

$$K^{glo}_{i-1} = [G_{i-1}, \cdots, X^k_{i-1}, V^k_{i-1}, E^k_{i-1}] \tag{9}$$

$$G_i = Att(Q^{glo}_{i-1}, K^{glo}_{i-1}, V^{glo}_{i-1}) \tag{10}$$

For each visual encoder layer, our local frame attention and global video attention are multi-head attention, using the

CLIP pretrained visual encoder corresponding layer's parameter. But the query, key, and value are not the same. We perform the two attentions in query mode so that the query in the proposed mechanisms can perceive local frame-level and global video-level representations. Additionally, the two attention mechanisms are implemented in the sharing-parameter manner, which has two advantages: *(1)* Sharing parameter could reduce half parameters cost in the visual encoder. *(2)* Sharing parameter could excavate the pretrained CLIP visual encoder's potential extremely, which is validated in Table 3 (b).

Similarity Calculation. Since the global prompts perceive each frame's information, we consider them a combination of fine-grained frames and global-discriminant video representations. Thus, we output the first global prompt, computing its similarity with the text representation.

Compared to the parameter-rich calculator, such as using four transformer layers to fuse temporal information in CLIP4Clip (Luo et al. 2022), our method is parameter-free in similarity computing. In addition, compared to the video feature re-weighting methods, like computing query-related frame features by cross-attention in (Gorti et al. 2022), through inner-product (Bain et al. 2022) or TopK (Liu et al. 2022), and then re-weighting the output video feature according to the frame-query similarity, our method is faster and query-agnostic. Especially in realistic applications, we could save much inference time because we only compute video and text features once. In contrast, the above re-weight methods need to be computed twice.

Objective Function. In the training process, following (Luo et al. 2022), we still adopt symmetric cross-entropy loss. During downstream training, both the text and visual encoders are frozen. For various text-video retrieval scenarios, only the parameters of text/visual prompts and the local prompt generation module need to be stored. We only need to reuse a copy of the pretrained model (*i.e.* CLIP), which reduces storage costs to the greatest extent.

Discussion of other PEFL methods. Adapters employ down-projection with nonlinear activation and up-projection mappings in each layer. However, adapters need a large intermediate compression dimension to maintain performance, undermining efficiency. Our test (Table 1) shows that adapter methods exceed the GPU memory of fully fine-tuned CLIP4Clip by over 50%, using above 30GB against CLIP4Clip's 20.8GB. Also, these methods alter the original model's structure, complicating deployment. Therefore, we mainly focus on prompt tuning in this study.

Experiments

Datasets and Evaluation Metrics. We conduct experiments on four datasets including MSR-VTT (Xu et al. 2016), LSMDC (Rohrbach et al. 2015), ActivityNet (Heilbron et al. 2015) and VATEX (Wang et al. 2019).

To measure the retrieval performance, we use standard metrics: recall at rank k (R@K, higher is better) and mean rank (MnR, lower is better). R@K computes the percentage of correct videos among the top K videos retrieved, we report the R@1, R@5, and R@10 results for each experiment. Mean rank computes the average rank of all correct answers.

Compared Baselines. We evaluate our approach against six strong baselines.

Efficient Prompt (Ju et al. 2022) introduces prompt tuning in TVR by adding text prompts to text encoder input and a two-layer transformer for temporal modeling.

VPT (Jia et al. 2022) is a visual recognition method using prompt tuning, with VPT-deep showing notable results.

UPT (Zang et al. 2022) generates both visual and text prompts from a unified transformer layer.

Visual-Text Adapter. Following (Houlsby et al. 2019), we add visual/text adapters after self-attention in each layer.

Video-Text Adapter. Based on Visual-Text Adapter, we replace the adapter in the visual encoder with ST-adapter (Pan et al. 2022) to enhance the capability of extracting temporal information, which is inserted before multi-head attention.

CLIP4CLip (Luo et al. 2022) is the fully finetuning baseline. We only compare the mean-pooling type for fairness, since our similarity calculator is also parameter-free.

Implementation Details. We use the CLIP (ViT-B/32) as the pre-trained model. During training, all the original parameters of CLIP are frozen unless explicitly mentioned. We apply a warm-up strategy followed by a cosine learning rate policy, using the AdamW optimizer with decoupled weight decay set to 0.2. The initial learning rate is 1e-2 for LSMDC and 5e-3 for the other three datasets. The max epochs are 10 for all datasets. Following CLIP4Clip, we uniformly sample 12 frames for MSRVTT, LSMDC, and VATEX and set the caption token length to 32. For ActivityNet, the frame length and caption length are set to 64. All the videos' short sides resize to 224, and the frame per second (fps) is 3. By default, the lengths of the frame prompts, text prefix/postfix prompts, and global prompts are all set to 4. Also, the depth of frame prompts and text prefix/postfix prompts is set to 12 by default. The inner dim of the adapter is set to 368. All experiments are done with mixed precision.

Results on Benchmarks

Results on MSR-VTT. Table 1 presents MSRVTT-9K results. With only 0.83MB parameters trainable, DGL-Linear (ViT-B/16) enhances R@1 by 2.7% over CLIP4Clip. For ViT-B/32, DGL-Transformer tops the list and exceeds Efficient Prompt and VoP in $T{\rightarrow}V$ R@1 by 9.1% and 1.2%. Across the board, DGL surpasses all adapters and prompt methods. Notably, DGL-Linear consumes just 18.75 GB of GPU memory, less than CLIP4Clip's 20.8 GB.

Results on the other three datasets. Table 2 displays the retrieval results on VATEX, LSMDC, and ActivityNet. For ViT-B/32, tuning only 0.83MB parameters, DGL surpasses CLIP4Clip by 0.3% and 0.7% in $T{\rightarrow}V$ R@1 on VATEX and LSMDC and comparable performance on ActivityNet. This underscores our method's efficiency. Notably, we outperform Efficient Prompt by 8.0% in LSMDC's $T{\rightarrow}V$ R@1 and achieve a 4.0% lead over VoP on ActivityNet.

Ablation Study

In this section, we thoroughly ablate DGL on MSR-VTT-9K using DGL-Transformer (ViT-B/32) unless specified.

Type	Methods	Trainable Params(MB) ↓	Memory Usage(GB) ↓	Text → Video R@1↑	R@5↑	R@10↑	MnR↓	Video → Text R@1↑	R@5↑	R@10↑	MnR↓
CLIP-ViT-B/32											
Finetune	CLIP4Clip	123.54	20.80	43.1	70.4	80.8	16.2	43.1	70.5	81.2	12.4
Adapter	Visual-Text Adapter	11.82	30.71	39.2	65.7	76.1	17.6	40.7	68.8	77.6	13.7
	Video-Text Adapter	11.94	31.59	41.1	67.0	77.1	17.4	42.6	68.4	78.4	13.8
Prompt	Efficient Prompt (Ju et al. 2022)	6.35	-	36.7	64.6	-	-	-	-	-	-
	VPT (Jia et al. 2022)	**0.18**	20.98	42.0	66.6	77.3	19.2	39.4	66.8	77.2	16.2
	UPT (Zang et al. 2022)	9.57	23.46	42.1	67.7	78.2	16.5	42.6	70.3	79.3	12.3
	VoP^{F+C} (Huang et al. 2023)	14.10	-	44.6	69.9	80.3	16.3	44.5	70.7	80.6	11.5
	DGL-Linear(Ours)	0.83	**18.75**	44.7	70.5	79.2	16.2	42.1	70.0	80.6	13.4
	DGL-Transformer(Ours)	9.57	20.69	45.8	69.3	79.4	16.3	43.5	70.5	80.7	13.1
	+ QB-Norm(Bogolin et al. 2022)	9.57	20.69	47.0	70.4	81.0	16.4	44.9	70.7	79.6	13.3
CLIP-ViT-B/16											
	CLIP4Clip	123.54	25.70	45.6	71.2	80.9	15.2	43.2	72.5	80.7	10.9
	VoP^{F+C}	14.10	-	47.7	72.4	82.2	12.0	-	-	-	-
	DGL-Linear(Ours)	0.83	22.86	48.3	71.8	80.6	13.4	45.7	74.0	82.9	10.9
	+ QB-Norm(Bogolin et al. 2022)	0.83	22.86	**49.7**	73.1	82.3	15.1	**47.8**	74.1	83.3	10.6

Table 1: Retrieval results on the MSR-VTT-9K dataset.

Type	Methods	VATEX R@1↑	R@5↑	R@10↑	MnR↓	LSMDC R@1↑	R@5↑	R@10↑	MnR↓	ActivityNet R@1↑	R@5↑	R@10↑	MnR↓
Finetune	CLIP4Clip	55.9	89.2	95.0	3.9	20.7	38.9	47.2	65.3	40.5	72.4	-	7.5
Adapter	Visual-Text Adapter	53.1	85.0	92.3	4.9	18.0	34.4	43.5	75.2	33.5	64.8	77.5	10.9
	Video-Text Adapter	53.5	85.0	92.4	4.7	18.3	35.5	44.0	74.8	36.4	66.1	79.6	10.0
Prompt	Efficient Prompt (Ju et al. 2022)	-	-	-	-	13.4	29.5	-	-	-	-	-	-
	VoP^{F+P} (Huang et al. 2023)	-	-	-	-	20.7	40.7	49.7	59.1	36.1	65.5	78.5	10.9
	DGL-Transformer(Ours)	54.3	85.5	92.3	4.9	21.2	37.8	48.8	66.5	40.1	69.5	80.9	9.1
	DGL-Linear(Ours)	56.2	87.1	93.5	4.1	21.4	39.4	48.4	64.3	38.6	69.2	81.6	9.0
	+ QB-Norm(Bogolin et al. 2022)	**57.3**	87.0	93.3	4.2	**21.6**	39.3	49.0	64.4	**43.1**	72.3	82.7	8.6

Table 2: Combined Retrieval Results for VATEX, LSMDC, and ActivityNet Datasets.

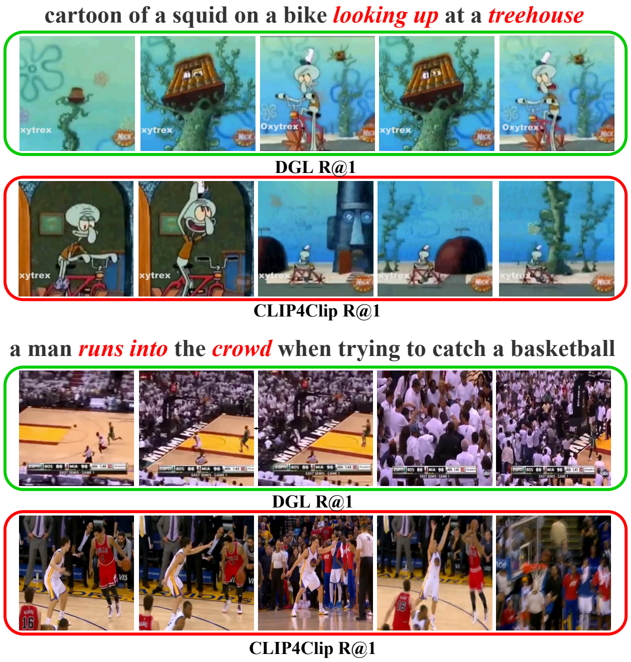

cartoon of a squid on a bike looking up at a treehouse
DGL R@1
CLIP4Clip R@1
a man runs into the crowd when trying to catch a basketball
DGL R@1
CLIP4Clip R@1

Figure 4: Visualization of text-video retrieval results. Frames in the green box are DGL R@1 results, while those in the red box are CLIP4clip R@1 results.

The output of the visual encoder. We compare four different types of visual encoder output as the final video representation as shown in Table 3 (a). We found that using the first global prompt performs best, this is because our attention is designed for the global prompt, which can perceive global information and local frame details.

Ablation of global-local video attention. We assess global-local video attention by testing global/local attention and shared parameters separately. Table 3 (b) shows that global-local video attention with sharing parameters outperforms the other methods, demonstrating that both global and local information is crucial for text-video retrieval.

Effect of the postfix text prompt. Following Efficient Prompt (Ju et al. 2022), which adds text prefix and postfix prompts only in the input layer, we extend this to all encoder layers. Table 3 (c) shows [4+X+4] deep text prompts outperforming [8+X] or [4+X], maximizing prompt potential.

Verify DGL on other baselines. We evaluated DGL on different structures and other CLIP-based methods. **(1)** Following Token Mix (Liu et al. 2023), we integrated DGL with BLIP (ViT-B/16) (Li et al. 2022), applying global-local video attention in the frozen visual encoder. Table 3 (d) top part shows that DGL surpasses the fully finetuning/PEFL method. **(2)** For CLIP-based method comparison, we focus on parameter-efficient designs and compare with X-CLIP (Ma et al. 2022) by freezing the CLIP backbone for fairness. The bottom part demonstrates DGL's effectiveness.

Why project Linear from visual to text? Visual features are more complex than textual, as videos typically contain more information. Projecting simpler text to complex visual features is challenging. Table 3 (e) shows Visual2Text pro-

Visual Output	R@1↑	R@5↑	R@10↑	MnR↓
	Text → Video			
First Global Prompt	**45.8**	69.3	79.4	16.3
Avg Global Features	43.5	69.7	79.7	16.9
Avg Local Features	41.5	68.3	77.2	16.0
Avg GL Features	42.5	68.8	77.6	15.6
	Video → Text			
First Global Prompt	43.5	70.5	80.7	13.1
Avg Global Features	43.1	70.0	79.9	12.5
Avg Local Features	41.9	70.9	79.3	12.4
Avg GL Features	44.3	69.6	79.9	12.4

(a) Comparison of different visual output. "GL" indicates Global-Local.

global	local	share	R@1↑	R@5↑	R@10↑	MnR↓
			Text → Video			
✓			40.4	67.8	77.3	18.3
	✓		42.0	68.7	78.2	16.9
✓	✓		43.5	70.1	79.0	17.1
✓	✓	✓	**45.8**	69.3	79.4	16.3
			Video → Text			
✓			41.3	67.8	76.8	14.3
	✓		42.2	69.7	78.6	12.6
✓	✓		43.1	69.4	80.0	13.6
✓	✓	✓	43.5	70.5	80.7	13.1

(b) Ablation of global-local video attention

Position	R@1↑	R@5↑	R@10↑	MnR↓
	Text → Video			
4+X	43.9	70.0	79.0	16.1
8+X	45.0	70.2	80.3	15.0
4+X+4	**45.8**	69.3	79.4	16.3
	Video → Text			
4+X	42.4	70.5	80.3	12.8
8+X	43.0	70.1	81.3	12.4
4+X+4	43.5	70.5	80.7	13.1

(c) Effect of text prompt position

		Text → Video		
Methods	UP(M)↓	R@1↑	R@5↑	R@10↑
BLIP(Li et al. 2022)				
Full(Liu et al. 2023)	226.51	47.6	73.4	81.8
Token Mix	7.07	47.1	70.8	80.5
DGL(ours)	**0.30**	**48.6**	71.4	79.7
X-CLIP(Ma et al. 2022)				
X-CLIP*	8.0	39.6	66.8	76.4
+DGL-Linear	**2.9**	**44.0**	69.9	79.6

(d) Verifying DGL on other baselines, "*" indicates freeze backbone.

Direction	R@1↑	R@5↑	R@10↑	MnR↓
	Text → Video			
T→V	43.4	69.6	79.7	16.2
V→T	**44.7**	70.5	79.2	16.2
	Video → Text			
T→V	43.0	70.0	79.5	12.6
V→T	42.1	70.0	80.6	13.4

(e) Abalation experiment of DGL-Linear projection direction.

Method	R@1↑	R@5↑	R@10↑	MnR↓
	Text → Video			
Baseline	43.8	68.7	80.2	16.2
DGL-Transformer	**45.8**	69.3	79.4	16.3
	Video → Text			
Baseline	43.9	69.4	80.1	12.2
DGL-Transformer	43.5	70.5	80.7	13.1

(f) Effect of generating cross-modal prompts from the shared latent space.

Table 3: Ablation studies on the MSRVTT-9K dataset

jection achieves higher R@1, validating our claim.

Generating from the shared latent space. DGL-Transformer enhances cross-modal interaction and local consistency by generating prompts from a shared latent space. Compared with the divided prompt baselines while maintaining global-local video attention, the 2% T→V R@1 improvement in Table 3 (f) demonstrates its effectiveness.

Retrieval results comparison. In Fig 4 above, our DGL model captures global details like"look up at a tree house", while CLIP4Clip sees local cues, such as "cartoon of a squid on a bike." In Fig 4 below, DGL identifies actions like "run into the crowd" and "catch a basketball," whereas CLIP4Clip only recognizes "catch a basketball." Thus, the results show that DGL perceives global video information.

What can global prompt learn? As shown in Fig 5, we visualize the attention weights of the global prompt on each frame, which is the output of the visual encoder. The top figure illustrates that our global prompt effectively focuses on the temporal dynamics of "wrestling." The bottom figure demonstrates that our global prompt can associate local information to extract global information. For example, it attends to "talks" in the first frame, "two generals" in the second and third frames, "war" in the fourth frame, and "king" in the fifth frame, and after summarizing and generalizing this information, successfully retrieves the relevant video clip. Our visualization results demonstrate that the global prompt in DGL can effectively capture temporal dynamics and global information.

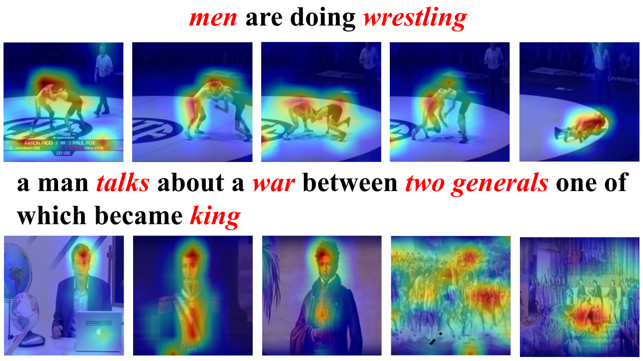

men are doing wrestling
a man talks about a war between two generals one of which became king

Figure 5: Visualization of global prompt, we plot the global prompt's attention weight on each frame. The red text in the query corresponds to the video's discriminative features.

Conclusion

In this work, we propose DGL, which generates local-level prompts for text and vision branches from a shared latent space, enhancing cross-modal interaction. Also, we propose a new attention mechanism for creating local and global prompts tailored to videos, which stands out in comparison to the existing literature where each frame is encoded separately by a fixed encoder. Extensive experiments show that, compared to the fully finetuning method or naive PEFL methods, our method only trains 0.83M parameters and outperforms them on four text-video retrieval datasets.

Acknowledgments

This work was supported in part by the Australian Research Council (ARC) under Grant DP200100938. We thank Yiyuan Yang and Shuai Zhao for their helpful discussions.

References

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. 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*, 1728–1738.

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2022. A CLIP-Hitchhiker's Guide to Long Video Retrieval. *arXiv preprint arXiv:2205.08508*.

Bogolin, S.-V.; Croitoru, I.; Jin, H.; Liu, Y.; and Albanie, S. 2022. Cross modal retrieval with querybank normalisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5194–5205.

Fang, H.; Xiong, P.; Xu, L.; and Chen, Y. 2021. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*.

Gorti, S. K.; Vouitsis, N.; Ma, J.; Golestan, K.; Volkovs, M.; Garg, A.; and Yu, G. 2022. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5006–5015.

Heilbron, F. C.; Escorcia, V.; Ghanem, B.; and Niebles, J. C. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE conference on computer vision and pattern recognition (CVPR)*, 961–970. IEEE.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2790–2799. PMLR.

Huang, S.; Gong, B.; Pan, Y.; Jiang, J.; Lv, Y.; Li, Y.; and Wang, D. 2023. VoP: Text-Video Co-Operative Prompt Tuning for Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6565–6574.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, 709–727. Springer.

Jiang, H.; Zhang, J.; Huang, R.; Ge, C.; Ni, Z.; Lu, J.; Zhou, J.; Song, S.; and Huang, G. 2022. Cross-modal adapter for text-video retrieval. *arXiv preprint arXiv:2211.09623*.

Jin, P.; Huang, J.; Xiong, P.; Tian, S.; Liu, C.; Ji, X.; Yuan, L.; and Chen, J. 2023a. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2472–2482.

Jin, P.; Li, H.; Cheng, Z.; Li, K.; Ji, X.; Liu, C.; Yuan, L.; and Chen, J. 2023b. Diffusionret: Generative text-video retrieval with diffusion model. *arXiv preprint arXiv:2303.09867*.

Ju, C.; Han, T.; Zheng, K.; Zhang, Y.; and Xie, W. 2022. Prompting visual-language models for efficient video understanding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, 105–124. Springer.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2022. Maple: Multi-modal prompt learning. *arXiv preprint arXiv:2210.03117*.

Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T. L.; Bansal, M.; and Liu, J. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7331–7341.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.

Liang, C.; Wang, W.; Zhou, T.; Miao, J.; Luo, Y.; and Yang, Y. 2023. Local-global context aware transformer for language-guided video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, X.; Ji, K.; Fu, Y.; Du, Z.; Yang, Z.; and Tang, J. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Liu, Y.; Xiong, P.; Xu, L.; Cao, S.; and Jin, Q. 2022. Ts2-net: Token shift and selection transformer for text-video retrieval. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*, 319–335. Springer.

Liu, Y.; Xu, L.; Xiong, P.; and Jin, Q. 2023. Token Mixing: Parameter-Efficient Transfer Learning from Image-Language to Video-Language. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2022. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304.

Ma, Y.; Xu, G.; Sun, X.; Yan, M.; Zhang, J.; and Ji, R. 2022. X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, 638–647.

Pan, J.; Lin, Z.; Zhu, X.; Shao, J.; and Li, H. 2022. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35: 26462–26477.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rohrbach, A.; Rohrbach, M.; Tandon, N.; and Schiele, B. 2015. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3202–3212.

Sun, C.; Myers, A.; Vondrick, C.; Murphy, K.; and Schmid, C. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7464–7473.

Sung, Y.-L.; Cho, J.; and Bansal, M. 2022. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5227–5237.

Wang, Q.; Zhang, Y.; Zheng, Y.; Pan, P.; and Hua, X.-S. 2022a. Disentangled representation learning for text-video retrieval. *arXiv preprint arXiv:2203.07111*.

Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O. K.; Singhal, S.; Som, S.; et al. 2022b. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.

Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.-F.; and Wang, W. Y. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4581–4591.

Wang, X.; Zhu, L.; and Yang, Y. 2021. T2vlad: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5079–5088.

Wu, W.; Luo, H.; Fang, B.; Wang, J.; and Ouyang, W. 2023. Cap4Video: What Can Auxiliary Captions Do for Text-Video Retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10704–10713.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5288–5296.

Xue, H.; Sun, Y.; Liu, B.; Fu, J.; Song, R.; Li, H.; and Luo, J. 2022. CLIP-ViP: Adapting Pre-trained Image-Text Model to Video-Language Representation Alignment. *arXiv preprint arXiv:2209.06430*.

Zang, Y.; Li, W.; Zhou, K.; Huang, C.; and Loy, C. C. 2022. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*.

Zhang, B.; Jin, X.; Gong, W.; Xu, K.; Zhang, Z.; Wang, P.; Shen, X.; and Feng, J. 2023. Multimodal video adapter for parameter efficient video text retrieval. *arXiv preprint arXiv:2301.07868*.

Zhao, S.; Zhu, L.; Wang, X.; and Yang, Y. 2022. Centerclip: Token clustering for efficient text-video retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 970–981.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhu, L.; and Yang, Y. 2020. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8746–8755.