BiEfficient: Bidirectionally Prompting Vision-Language Models for Parameter-Efficient Video Recognition

Haichen He[1], Weibin Liu[1,*], Weiwei Xing[2]

[1] School of Computer Science and Technology, Beijing Jiaotong University, China
[2] School of Software Engineering, Beijing Jiaotong University, China
wbliu@bjtu.edu.cn

Abstract. Vision-language models (VLMs) pre-trained on large-scale image-text pairs have shown great success in various image tasks. However, how to efficiently transfer such powerful VLMs into video domain is still an open problem. Given that full finetuning VLMs for video tasks could be computationally expensive, recent studies turn their focus on parameter-efficient finetuning (PEFT). The great potential of VLMs lies in leveraging the bidirectional semantic connections between the two modalities of vision and language. Nevertheless, most current PEFT methods use the vision-only framework and usually ignore the semantic connections between vision and language. In this paper, we propose a novel method called BiEfficient, which use bidirectional prompting schemes to efficiently transfer the VLM to video recognition task with a small number of tunable parameters: 1) Vision-to-Language: we propose two prompt mechanisms, Pre-Prompt and Post-Prompt, which act before and after the text encoder respectively to generate discriminative video-level text representation for each input video. 2) Language-to-Vision: we propose Word-Guided Visual-Prompt, which enhances the temporal modeling of videos using textual knowledge in an almost parameter-free manner. Experiments on Kinetics-400, UCF-101, HMDB-51 demonstrate that the proposed method can achieve comparable or even better performance to the full finetuning methods with much fewer tunable parameters across closed-set and zero-shot video recognition benchmarks. Code is available here: https://github.com/WbLiuBJTULab/BiEfficient.

1 Introduction

Vision-language models (VLMs) are pre-trained on large-scale noisy image-text pairs from web (*e.g.*, CLIP [39], ALIGN [18], CoCa [61], Florence [62]), and have demonstrated remarkable success in various image tasks. However, extending such success to video tasks poses significant challenges. On the one hand, collecting a large-scale video-text pre-training dataset is much more difficult than collecting image-text pairs, on the other hand, training a VLM for video task

* Corresponding Author: Weibin Liu, Email: wbliu@bjtu.edu.cn

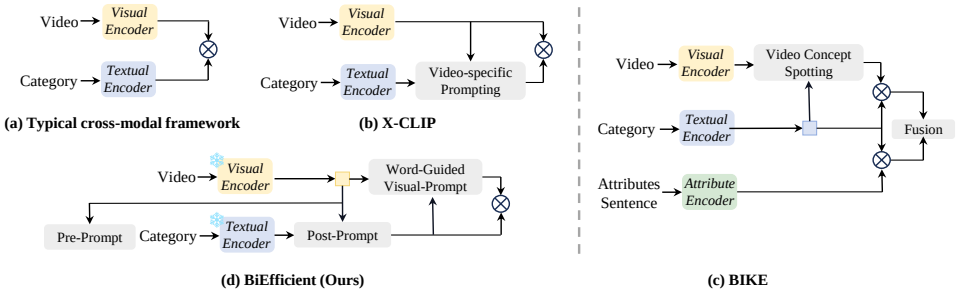

Fig. 1: Comparison with **(a)** the typical cross-modal framework and existing methods that focus on cross-modal communication. **(b)** X-CLIP [33]. **(c)** BIKE [56]. **(d)** Our proposed BiEfficient.

is highly resource-intensive. A prevalent method is to transfer knowledge from VLMs pre-trained on image-text pairs to the video domain. Nevertheless, as the model size increases, full finetuning is still too computationally expensive for video tasks, significantly restricting their deployment in real-world applications.

Recent studies turn their focus on parameter-efficient finetuning (PEFT). Originating from the field of natural language processing (NLP), PEFT methods only finetune a small number of parameters while keeping large pre-trained language models frozen [14, 16, 17, 22, 24, 36, 43, 63]. With the rise of large vision transformer (ViT) models [8], these techniques have been introduced to the computer vision community. PEFT methods can be broadly categorized into adapter-based methods and prompt-based methods. Adapter-based methods [12,32,34,60] insert lightweight trainable modules into the pre-trained models, allowing for efficient task-specific adaptations without significantly modifying the foundation models. In contrast, prompt-based methods [2,13,66,67] prepend a set of learnable tokens at the input point of the models for task finetuning, with minimal parameter updates.

However, existing PEFT methods usually fall short in leveraging the semantic connections between vision and language. PEFT methods AIM [60], ST-Adapter [34] and EVL [27] all use vision-only framework and discard language as supervision, leading to overfitting on the training dataset while losing capability of zero-shot on new data. Some existing works have explored the connections between vision and language, *e.g.*, X-CLIP [33](Fig.2(b)) and BIKE [56](Fig.2(c)). Nevertheless, the two methods both have drawbacks. X-CLIP only focus on vision-to-language direction, disregarding the language-to-vision direction, while BIKE adds preprocessing steps and extra branch. Moreover, the models are full finetuned for both methods, which is highly resource-intensive. X-CLIP and BIKE have demonstrated the effectiveness of leveraging the interaction of the two modalities, while their drawbacks also highlight the need for PEFT methods exploring the bidirectional semantic connections of the two modalities to deepen the understanding of multi-modal data.

In this paper, we propose a novel method named BiEfficient(Fig.2(d)), which leverages the Vision-to-Language and the Language-to-Vision knowledge by bidirectional prompting mechanisms.

For the first Vision-to-Language direction, we propose two prompt mechanisms, *i.e.*, Pre-Prompt and Post-Prompt. Pre-Prompt consists of two learnable components: a set of context tokens that effectively transfer pre-trained knowledge into downstream task and a learnable token (meta-token) that learned through a lightweight network (Meta-Net) acting as a discriminative context token for each video. Post-Prompt is builded on top of the text encoder, which allows text representation to extract the related visual context automatically to yield the video-level text representation. The two prompt mechanisms are inspired by that related visual context can make the text representation more discriminative for enhancing recognition. For example, if there is extra visual information about "grass", then the category is more like to be "kicking soccer ball" instead of "swimming".

For the second Language-to-Vision direction, we propose Word-Guided Visual-Prompt mechanism, which leverages textual knowledge to enhance the temporal modeling of videos in an almost parameter-free manner. We notice that only a few frames among the video clip are keyframes helpful for recognition, while most of the remaining frames are about background. So we use textual knowledge to highlight the keyframes which are semantically similar to the category for enhancing recognition. For example, if we highlight the keyframes "kicking the ball" and "running", then the actions "kicking soccer ball" and "running on the grass" that are both actions on the grass will be much easier to be distinguished.

We evaluate our model on three popular datasets (*i.e.*, Kinetics-400 [20], UCF-101 [42], HMDB-51 [21]) across two video recognition benchmarks, *i.e.*, closed-set and zero-shot video recognition. Comprehensive experiments demonstrate our proposed method is generally effective. Our contributions are summarized as follows:

- We propose a novel method for parameter-efficient finetuning of CLIP on video recognition tasks which use bidirectional prompting mechanisms to leverage the semantic connections of vision and language.
- We propose the Pre-Prompt and Post-Prompt mechanisms in Vision-to-Language direction to generate discriminative text representation using visual context, and the Word-Guided Visual-Prompt in Language-to-Vision direction to enhance temporal modeling of videos using textual knowledge.
- Extensive experiments on three datasets across two video recognition benchmarks demonstrate the effectiveness and efficiency of our proposed method.

2 Related Work

Vision-Language Models (VLMs) pre-trained on web-scale data have recently shown great success. By using natural language as supervision, VLMs can learn generalizable visual representation [18,39,61,62]. One of the most impressive works is CLIP [39]. The effectiveness of CLIP has inspired numerous

applications in various downstream tasks. For example, CoOp [67], CoCoOp [66], CoPL [13] leverage the strong generalization capability of CLIP to enhance the few-shot transfer on downstream recognition tasks. PointCLIP [64] applies CLIP to 3D recognition, leveraging the model's strong representation capabilities to handle 3D data effectively. For video understanding, some works [32,54,65] have explored using CLIP for video-text retrieval. All these works highlight the transformative impact of VLMs in vision tasks. In this work, we continue to explore transferring CLIP to video recognition tasks.

Video Recognition Convolutional neural networks (CNNs) have long been the standard backbone for video recognition tasks. Early works [6,10,41,49] learned spatial and temporal representation through parallel branches. Later works [37, 45,46,57] widely utilized 3D CNNs to jointly learn spatiotemporal representation. However, the computational demands of 3D CNNs are substantial, leading to the development of methods [25,26,29,30,48,53] which insert temporal modules into 2D CNNs for efficient temporal modeling. The rise of vision transformers (ViTs) [8] marked a significant shift in video recognition. Transformer-based methods [1, 3, 9, 28, 59] leverage the capabilities of self-attention mechanism to capture long-range temporal dependencies. Very recently, the application of VLMs *e.g.*, CLIP, in video recognition task has shown remarkable promise. By transferring knowledge from the pre-trained VLMs to video domain, these methods (*e.g.*, ActionCLIP [50], X-CLIP [33], BIKE [56], Text4Vis [55]) significantly enhancing video recognition performance, opening new avenues for research in this field. However, all these models are full finetuned on video data, which makes the training cost unaffordable to most of us. In this work, we focus on efficiently transferring CLIP to video recognition tasks with minimal training cost.

Parameter-Efficient Finetuning (PEFT) techniques [14,16,17,22,24,36,44, 63] originated from the field of natural language processing (NLP), and have been introduced to computer vision community. The goal of PEFT techniques is to reduce training cost by tuning only a small number of parameters while keeping the majority of the large pre-trained model frozen. Recently, researchers have explored applying PEFT to video recognition task. For example, methods like AIM [60] and ST-Adapter [34] insert lightweight adapter modules into the pre-trained model, allowing for task-specific adaptation without significantly modifying the foundation models. PromptCLIP [19] enhances video recognition by prepending a set of learnable tokens at the input point of the model. EVL [27] freezes the image encoder of CLIP and adding new trainable decoder branches specifically designed for temporal modeling. However, all these PEFT methods disregard the importance of cross-modal communication. In this work, we propose a PEFT method for video recognition, which leverages the semantic connections between vision and language to bidirectionally prompting CLIP for parameter-efficient video recognition, achieving improved performance across three datasets on two benchmarks.

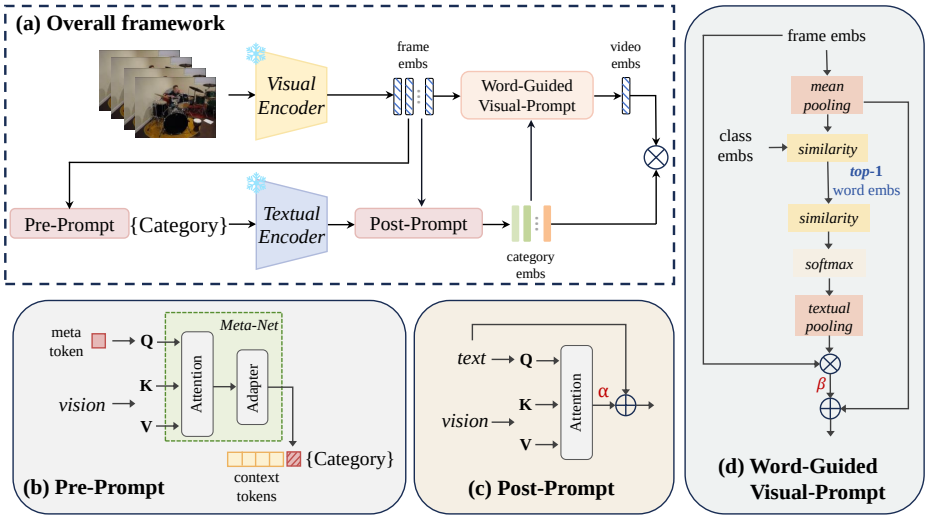

Fig. 2: Illustration of BiEfficient for video recognition. (a) BiEfficient prompts CLIP for video recognition bidirectionally: 1) *Vision-to-Language*: we propose (b) Pre-Prompt mechanism, which leveraging the vision context to efficiently generate video-level text prompt before the text encoder, and (c) Post-Prompt mechanism, which generates the final discriminative video-level category embeddings. 2) *Language-to-Vision*: we propose the (d) Word-Guided Visual-Prompt mechanism, which leverages the semantic similarity between video and categories to enhance temporal modeling for video.

3 Methodology

In this section, we first briefly present our framework in Sec. 3.1. Then, we describe the details of Pre-Prompt and Post-Prompt for the Vision-to-Language direction in Sec. 3.2, and the Word-Guided Visual-Prompt for the Language-to-Vision direction in Sec. 3.3. Finally, the training loss is introduced in Sec. 3.4.

3.1 Overview

An overview of our proposed BiEfficient is shown in Fig.2(a). Given a video clip V with T sampled frames, and a collection of categories $\mathbf{C} = \{C_1, C_2, \ldots, C_K\}$, where K is the number of categories. We feed the video clip V into the visual encoder $f(\cdot|\theta_v)$ to obtain the visual representation $\mathbf{v}$, where

$$\mathbf{v} = f(V|\theta_v) \tag{1}$$

Then, the Pre-Prompt generator $f(\cdot|\phi_{pre})$ is employed before the textual encoder to get a video-level textual prompt p. It takes $\mathbf{v}$ as input, formulated as:

$$p = f(\mathbf{v}|\phi_{pre}) \tag{2}$$

With the Pre-Prompt generator, we feed the category C into the textual encoder $f(\cdot|\phi_c)$ to get the text representation $\mathbf{c}$, and we employ the Post-Prompt generator $f(\cdot|\phi_{post})$ leveraging $\mathbf{v}$ to generate the discriminative video-level category embedding $\hat{\mathbf{c}}$, where

$$\mathbf{c} = f(p, C|\phi_c), \hat{\mathbf{c}} = f(\mathbf{c}, \mathbf{v}|\phi_{post}) \tag{3}$$

Then, we use the Word-Guided Visual-Prompt module $f(\cdot|\theta_t)$ to obtain the video embedding $\hat{\mathbf{v}}$, enhancing temporal modeling for video, where

$$\hat{\mathbf{v}} = f(\mathbf{v}, \hat{\mathbf{c}}|\theta_t) \tag{4}$$

Our goal is to ensure that $\hat{\mathbf{v}}$ and $\hat{\mathbf{c}}$ are similar if V and C are matched, and dissimilar otherwise. During training, the parameters θ_v and ϕ_c are initialized with weights from the pre-trained CLIP [39].

3.2 Vision-to-Language: two prompt mechanisms

In video recognition task, text refers to the category names, *i.e.*, a word or phrase, whose information is too insufficient. In order to get a good text description, current works usually use hand-crafted prompt(ActionCLIP [50], BIKE [56]) or learnable prompt(PromptCLIP [19]). In contrast, to learn robust and discriminative text representation, we propose two prompt mechanisms to improve performance for video recognition, leveraging visual context in the input video.

Pre-Prompt As depicted in Fig.2(b), given the i-th category $C_i \in \mathbf{C}$ and the visual representation $\mathbf{v}$. The category C_i is transformed into category embedding $\bar{\mathbf{c}}_i$ through CLIP *Tokenize* and *TokenEmbedding*, where

$$\bar{\mathbf{c}}_i = TokenEmbedding(Tokenize(C_i)) \tag{5}$$

First, we introduce M learnable context tokens $\{\mathbf{c}_1, \mathbf{c}_2, \dots \mathbf{c}_M\}$ to transform the pre-trained model into video recognition task efficiently. Then, on top of the M context tokens, we further learn a lightweight network (Meta-Net), denoted as $f(\cdot|\phi_m)$, which takes $\mathbf{v}$ and a learnable meta-token $\mathbf{c}_{init}$ as inputs. Meta-Net includes an Attention block allowing meta-token to extract helpful visual context automatically, and an Adapter block for visual-textual alignment. Its output $\mathbf{c}_v$ can act as a discriminative visual context token, where

$$\mathbf{c}_v = f(\mathbf{v}, \mathbf{c}_{init}|\phi_m) = \text{Adapter}(\text{Attention}(\mathbf{v}, \mathbf{c}_{init})) \tag{6}$$

Here, the Attention block is constructed by the standard multi-head self-attention and feed-forward network [47], which takes $\mathbf{c}_{init}$ as query, $\mathbf{v}$ as key and value. The Adapter block is built with a two-layer bottleneck structure(Linear-GELU [15]-Linear), with the hidden layer reducing the input dimension by $0.25\times$. The prompt for the i-th category C_i is thus enhanced by the visual context, *i.e.*, $\mathbf{t}_i = \{\mathbf{c}_1, \mathbf{c}_2, \dots \mathbf{c}_M, \mathbf{c}_v, \bar{\mathbf{c}}_i\}$. Then, we feed $\mathbf{t}_i$ into the textual encoder to obtain the category representation $\mathbf{c}_i$ of C_i. M is set to be 16 in this paper.

Post-Prompt In order to further improve the representation ability of categories, we continue to explore the impact of visual context on text information. To this end, we propose the Post-Prompt module, as depicted in Fig.2(c), which generates the final discriminative video-level category embeddings. For the i-th category $C_i \in \mathbf{C}$, the Post-Prompt module takes the category representation $\mathbf{c}_i$ and the visual representation $\mathbf{v}$ as inputs, and gets the video-level category embedding $\hat{\mathbf{c}}_i$ as follows:

$$\hat{\mathbf{c}}_i = \mathbf{c}_i + \alpha \times \mathrm{FFN}\left(\mathrm{MHSA}\left(\mathbf{v}, \mathbf{c}_i\right)\right) \tag{7}$$

Here the Post-Prompt module consists of a multi-head self-attention block and a feed-forward network [47], which takes $\mathbf{c}_i$ as query, $\mathbf{v}$ as key and value, allowing $\mathbf{c}_i$ to automatically extract the related visual information that can make itself more robust and discriminative. Finally, we combine $\mathbf{c}$ and $\mathbf{c}_i$ with a learnable parameter α to preserve the original information of $\mathbf{c}$ from being destroyed.

3.3 Language-to-Vision: Word-Guided Visual-Prompt

In Sec. 3.2, we have improved the representation ability of categories by leveraging visual context from the input video. In this section, we describe how to guide the temporal modeling by leveraging the semantic similarity between video and categories, as depicted in Fig.2(d).

Word-Guided Visual-Prompt Given a video clip V with T sampled frames, we can get the frame embeddings $\mathbf{v} \in \mathbb{R}^{T \times D}$, where D is the dimension size. We also have the category embeddings $\hat{\mathbf{c}} \in \mathbb{R}^{B \times N \times D}$, where B is the batch size, N is the number of words in the category name. According to CLIP [39], we can get two sets of embeddings through $\hat{\mathbf{c}}$: $\mathbf{z} \in \mathbb{R}^{B \times D}$ is a set of class embeddings, which can be regarded as category-level representation, and $\mathbf{w} \in \mathbb{R}^{B \times N \times D}$ is a set of word embeddings, which is fine-grained representation.

First, we perform meanpooling on $\mathbf{v} \in \mathbb{R}^{T \times D}$ to get a video-level embedding $\bar{\mathbf{v}} \in \mathbb{R}^{D}$. Then, we calculate the similarity of $\bar{\mathbf{v}}$ and $\left\{\mathbf{z}_i \in \mathbb{R}^{D} \,|i = 1, 2, \ldots, B\right\}$ to get $\hat{\mathbf{z}}$, which is semantically most similar to $\bar{\mathbf{v}}$. Also, we have $\hat{\mathbf{w}} \in \mathbb{R}^{N \times D}$ in the same category as $\hat{\mathbf{z}}$. To get the word-guided temporal weights for frames, we calculate the cosine similarities between each frame and each word as follows:

$$S\left(\mathbf{v}_i, \hat{\mathbf{w}}_j\right) = \frac{\langle \mathbf{v}_i, \hat{\mathbf{w}}_j \rangle}{\|\mathbf{v}_i\| \cdot \|\hat{\mathbf{w}}_j\|}, i = 1, 2, \ldots T, j = 1, 2, \ldots N \tag{8}$$

Next, we perform softmax in the temporal dimension to normalize the similarities for each frame, and perform meanpooling in the textual dimension to get the temporal weights $\mathbf{S} \in \mathbb{R}^{T}$. Finally, we can generate the text-enhanced video embedding $\hat{\mathbf{v}} \in \mathbb{R}^{D}$ as follows:

$$\hat{\mathbf{v}} = \bar{\mathbf{v}} + \beta \times \mathbf{v}^{\mathsf{T}}\mathbf{S} \tag{9}$$

where β is a learnable parameter to combine the meanpooling video embedding $\bar{\mathbf{v}}$ and the text-enhanced video embedding $\hat{\mathbf{v}}$.

3.4 Training Loss

BiEfficient consists of the visual encoder $f(\cdot|\theta_v)$, the textual encoder $f(\cdot|\phi_c)$, the Pre-Prompt module $f(\cdot|\phi_{pre})$, the Post-Prompt module $f(\cdot|\phi_{post})$, and the Word-Guided Visual Prompt module $f(\cdot|\theta_w)$. Model parameters θ_v and ϕ_c are initialized with the weights from pre-trained CLIP [39]. We freeze θ_v and ϕ_c, and jointly optimize ϕ_{pre}, ϕ_{post} and θ_w.

During training, given a video $V_i \in \{V_1, V_2, \dots V_B\}$ and a category $C_i \in \{C_1, C_2, \dots C_B\}$, where B is the batch size. Our objective is to maximize the similarity between $\hat{\mathbf{v}}_i$ and $\hat{\mathbf{c}}_i$ if V_i and C_i are matched, and minimize it otherwise. This is achieved with a symmetric cross entropy loss, following CLIP:

$$\mathcal{L} = -\frac{1}{2}\left(\log\frac{\exp\left(s\left(\hat{\mathbf{v}}_i, \hat{\mathbf{c}}_i\right)/\tau\right)}{\sum_{j=1}^{B}\exp\left(s\left(\hat{\mathbf{v}}_i, \hat{\mathbf{c}}_j\right)/\tau\right)} + \log\frac{\exp\left(s\left(\hat{\mathbf{c}}_i, \hat{\mathbf{v}}_i\right)/\tau\right)}{\sum_{j=1}^{B}\exp\left(s\left(\hat{\mathbf{c}}_i, \hat{\mathbf{v}}_j\right)/\tau\right)}\right) \tag{10}$$

where $s(\cdot,\cdot)$ is cosine similarity, τ refers to the temporature parameter for scaling.

4 Experiments

4.1 Experimental Settings

Datasets We evaluate our proposed BiEfficient on two popular benchmarks, *i.e.*, closed-set video recognition and zero-shot video recognition. For closed-set video recognition, we employ Kinetics-400 [20], which is a large scale action recognition dataset spanning 400 categories, including 240,000 training videos and 20,000 validation videos. For zero-shot video recognition, we employ UCF-101 [42] and HMDB-51 [21]. UCF-101 is an action recognition dataset that contains 13,320 videos from 101 categories, collected from YouTube. HMDB-51 is a collection of realistic videos from various sources, including 7,000 videos spanning 51 categories.

Implementation Details In this paper, the visual encoder and textual encoder are adopted from the pre-trained CLIP(ViT-B/16) [39], and are both frozen. We sample 8, 16 or 32 frames with a sparse sampling method. The spatial resolution of the input frames is 224×224. We use AdamW [31] as the optimizer with batch size of 32. The models are trained with 50 epochs and the weight decay is 0.2. The learning rate is 5×10^{-4}. It is warmed up for 5 epochs and decayed to zero following a cosine schedule for the rest of training. The temperature parameter is set to be 0.01. All models are trained with 4 NVIDIA RTX 3090 GPUs.

Baseline To establish a simple baseline, we employ the typical cross-modal framework, as shown in Fig.2(a). Technically, we add a 3-layer temporal transformer (9M parameters) on top of the visual encoder, commonly used in previous works [19, 50].

Table 1: Comparison to state-of-the-art on Kinetics-400. Views = #temporal clip × #spatial crop.

Method	Pre-train	Param (M)	Tunable Param(M)	Top-1	Top-5	Frames ×Views
Full finetuning						
Methods with random initialization						
MViT [38]	-	37	37	81.2	95.1	64×3×3
Methods with ImageNet pre-training						
NL I3D-101 [52]	IN-1K	61.8	61.8	77.7	93.3	128×10×3
UniFormer-B [23]	IN-1K	50	50	83.0	95.4	32×4×3
TimeSformer-L [3]	IN-21K	121	121	80.7	94.7	64×1×3
ViViT-L/16×2 [1]	IN-21K	311	311	80.6	92.7	32×1×1
VideoSwin-L [28]	IN-21K	197	197	83.1	95.9	32×4×3
Methods with web-scale image pre-training						
MTV-L [59]	JFT	876	876	84.3	96.3	32×4×3
TokenLearner-L/10 [40]	JFT	450	450	85.4	96.3	64×4×3
Methods with web-scale image-language pre-training						
ActionCLIP-B/16 [50]	CLIP	142	142	82.6	96.2	16×10×3
X-CLIP-B/16 [33]	CLIP	-	-	83.8	96.7	8×4×3
BIKE-B/16 [56]	CLIP	161	161	84.0	-	8×4×3
Text4Vis-B/16 [55]	CLIP	100	100	82.9	-	8×4×3
Frozen backbone						
PromptCLIP-B/16 A5 [19]	CLIP	-	6	76.6	93.3	-
ST-Adapter-B/16 [34]	CLIP	89	7	82.7	96.2	32×3×1
EVL-B/16 [27]	CLIP	175	59	84.2	-	32×3×1
AIM-B/16 [60]	CLIP	97	11	**84.7**	96.7	32×3×1
BiEfficient-B/16	CLIP	155	15	82.8	96.3	8×3×1
BiEfficient-B/16	CLIP	155	15	83.1	96.6	16×3×1
BiEfficient-B/16	CLIP	155	15	83.4	**96.7**	32×3×1

4.2 Main Results

Closed-set Video Recognition is the common scenario, where the model is trained and evaluated on videos from the same categories. We compare our method to the state-of-the-art on Kinetics-400 in Tab. 1. The state-of-the-art are under different pre-training, including random initialization, ImageNet-1k/21k [7] pre-training, web-scale image and image-language pre-training. We can observe that:

•*Comparison to the methods with full finetuning* Compared to the methods with random initialization and ImageNet pre-training, BiEfficient outperforms all these models with fewer tunable parameters and frames on lighter backbone. For example, BiEfficient-B/16 surpasses VideoSwin-L [28] by +0.3% with significantly fewer number of tunable parameters (15M vs 197M). Also, BiEfficient-B/16 (with 8 sampled frames) outperforms TimeSformer-L [3] (with 64 sampled

frames) by +2.1% with fewer tunable parameters (15M vs 121M), which greatly reduces the training cost.

Compared to the methods with web-scale image pre-training, our BiEfficient is also competitive. BiEfficient-B/16 (with 32 sampled frames) achieves comparable performance with MTV-L [59] (with 32 sampled frames) and TokenLearner-L/10 [40] (with 64 sampled frames) (83.4% vs 84.3%, 85.4%). Note that their backbone is larger than us, and we only need to tune less than 3% parameters of them.

Compared to the methods with web-scale image-language pre-training, our results are comparable to or even better than these full fintuning methods. With the same backbone ViT-B/16, BiEfficient (with 16 sampled frames) surpasses ActionCLIP-B/16 [50] (with 16 sampled frames) and Text4Vis [55] (with 8 sampled frames) by +0.5% and +0.2%, while tuning fewer parameters (15M vs 142M, 100M). When compared to ActionCLIP [50], X-CLIP [33] and BIKE [56], our BiEfficient achieves comparable performance (83.4% vs 83.8%, 83.8%, 84.0%) while significantly reduces the training cost by just tuning around 10% parameters of them.

• *Comparison to the methods frozen backbone* We compare our BiEfficient with previous efficient methods on the same backbone ViT-B/16. Compared to PromptCLIP A5 [19] (with 16 sampled frames), BiEfficient (with 8 sampled frames) can achieve much higher top-1 accuracy (82.8% vs 76.6%), while only introducing extra 9M parameters. With the same 32 sampled frames, BiEfficient also surpasses ST-Adapter [34] (83.4% vs 82.7%), while only introducing extra 8M parameters.

EVL [27] and AIM [60] achieves better results than us, we will analyze them. Our BiEfficient-B/16 achieves the same top-1 accuracy 82.8% with EVL-B/16 when the number of input frames is 8. EVL adds 12 layers of decoder blocks for strong temporal modeling, as a result, when the number of input frames increases, the temporal modeling capability of EVL is fully utilized to achieve higher performance. However, our BiEfficient is still comparable to EVL (83.4% vs 84.2%) while tunable parameters reducing 75%. AIM achieves 1.3% higher top-1 accuracy than us with 11M parameters. Because it reuse image pre-trained self-attention layers for temporal modeling thus significantly reduces tunable parameters, and its layer-by-layer jointly spatiotemporal modeling also helps a lot. Nevertheless, it is worthy to note that, EVL and AIM both use vision-only framework, which means that the categories will be mapped into a set of one-hot vectors, the classification head is only optimized for this set of categories while unable to generalize to new categories. Our method employs the cross-modal framework, and focus on exploiting the semantic relationship between vision and language, enhancing the zero-shot capability on new datasets.

Zero-shot Video Recognition is the novel scenario, where videos for training and evaluating are from different categories. We pre-train BiEfficient-B/16 on Kinetics-400 with 32 frames, and evaluate it on UCF-101 and HMDB-51. The results are shown in Tab. 2.

In contrast to the previous PEFT methods, our method is able to conduct zero-shot video recognition task due to the cross-modal framework. We evaluate BiEfficient-B/16 across three splits and report the mean top-1 accuracy. Our method demonstrates strong cross-dataset generalization capability and outperforms previous methods. For example, BiEfficient surpasses ER-ZSAR [5] by +6.5% on HMDB-51, and +8.5% on UCF-101. When compare to the CLIP-based full-finetuning methods, such as ActionCLIP-B/16 [50] and X-CLIP-B/16 [33], our BiEfficient also achieve comparable or even better performance. The results demonstrate that the proposed cross-modal framework, which leveraging the semantic relationship between vision and language, leads to improvement of generalization capability.

Table 2: Zero-shot results on HMDB-51 and UCF-101.

Method	HMDB-51	UCF-101
MTE [58]	19.7 ± 1.6	15.8 ± 1.3
ASR [51]	21.8 ± 0.9	24.4 ± 1.0
ZSECOC [35]	22.6 ± 1.2	15.1 ± 1.7
TS-GCN [11]	23.2 ± 3.0	34.2 ± 3.1
UR [68]	24.4 ± 1.6	17.5 ± 1.6
E2E [4]	32.7	48
ER-ZSAR [5]	35.3 ± 4.6	51.8 ± 2.9
ActionCLIP-B/16 [50]	40.8 ± 5.4	58.3 ±3.4
X-CLIP-B/16 [33]	**44.6 ± 5.2**	**72.0 ± 2.3**
BiEfficient	41.8 ± 5.2	60.3 ± 3.3

Training Cost We compare the training time (GPU hours) and tunable parameters (M) of our method and previous full finetuning methods in Tab. 3. As we can see, compared to UniFormer-B [23] (with 32 sampled frames), our method (with 16 sampled frames) achieve +0.2% performance improvement with around 33× training time reduction and 3× fewer tunable parameters. Compared to ActionCLIP [50], which is also pre-train ed with web-scale image-text pairs, our method still achieve +0.5% performance improvement with around 3× training time reduction and 9× fewer tunable parameters.

4.3 Ablation Analysis

We provide detailed ablation analysis on Kinetics-400 in this section. Unless specified otherwise, we use ViT-B/16 with 8 frames as backbone and 3 views for testing. The default settings in the paper is marked in gray in this section.

The Effect of Proposed Components. Tab. 4 demonstrates the effectiveness of our proposed components in Sec. 3. We can observe that through simply equipping a 3-layer temporal transformer on top of the frozen CLIP, we can establish a strong baseline (9M) achieving 80.4% top-1 accuracy. With the proposed Pre-Prompt module, we can obtain +1.6% performance improvement by introducing 3M tunable parameters. Then, appending the Post-Prompt module can further improve the accuracy by +0.4%, with extra 3M tunable parameters. It illustrates that the Pre-Prompt and Post-Prompt can generate discriminative textual representation for enhancing recognition by leveraging the visual

Table 3: Training cost comparison on Kinetics-400.

Method (#frames per view)	Pre-train	Top-1 (#views)	Training GPU Hours	Tunable Param(M)
VideoSwin-B [28] (32)	IN-21K	82.7 (4)	512 × V100	88
UniFormer-B [23] (32)	IN-1K	82.9 (4)	5000 × V100	50
ActionCLIP-B/16 [50] (16)	CLIP	82.6 (3)	480 × RTX 3090	142
BiEfficient-B/16 (16)	CLIP	**83.1** (3)	**152 × RTX 3090**	**15**

context. Finally, with the Word-Guided Visual-Prompt, BiEfficient can surpass the baseline by +2.4%. This shows that the Word-Guided Visual-Prompt can effectively use textual knowledge for enhancing temporal modeling.

Table 4: Component-wise analysis of BiEfficient. Baseline refers to the frozen CLIP followed by a 3-layer temporal transformer with no textual prompt.

Method	Tunable Param(M)	Top-1
Baseline	9	80.4
+ Pre-Prompt	12	82.0 (+1.6)
+ Post-Prompt	15	82.4 (+2.0)
+ Word-Guided Visual-Prompt	15	82.8 (+2.4)

Table 5: Comparison of different textual prompt.

Method	Top-1
w/o prompt	80.6
Hand-crafted prompt	80.9
Learnable prompt [19]	82.0
Pre-Prompt	82.4
Post-Prompt	82.5
BiEfficient	**82.8**

The Effect of Textual Prompt Tab. 5 demonstrates the effect of different textual prompt. As we can see, using no prompt leads to the worst result of 80.6% top-1 accuracy, while using hand-crafted prompt "This is a video about {}" and learnable prompt [19] (16 prompt vectors+X) results in a drop of 1.9% and 0.8%, respectively. Using the Pre-Prompt and Post-Prompt mechanism alone achieves 82.4% and 82.5% top-1 accuracy, respectively, demonstrating the effectiveness of the proposed textual mechanisms. Using the Pre-Prompt and Post-Prompt mechenisms together achieves the highest performance of 82.8%, demonstrating that the two textual prompt mechanisms can generate more discriminative text representation for enhancing recognition.

Optional Designs of Pre-Prompt We provide several optional designs of Pre-Prompt module in Tab. 6. A1 refers to the case where we directly concatenate the output of the visual encoder and the context tokens of length M. A2 refers to the case where we employ an Attention block constructed by the standard multi-head self-attention and feed-forward network [47], and take the context tokens as query, the output of visual encoder as key and value. The output of the Attention block is regarded as the text prompt. A3 refers to the case where we discard the Adapter block in the Pre-Prompt module. The results of A1 and A2 drop 1.0% and 0.8%, respectively. This demonstrates the proposed Pre-Prompt module can extracted the related visual context effectively. A3 leads to a

Table 6: Comparison of optional designs of Pre-Prompt.

Method	Param	Top-1.(%)	Top-5.(%)
A1	-	81.0	95.8
A2	3M	81.2	95.8
A3	3M	81.7	96.0
Pre-Prompt	3M	**82.0**	**96.1**

Table 7: Ablation study on the effect of frame meanpooling.

Method	Top-1.(%)	Top-5.(%)
w/o meanpooling	82.5	96.3
w/ meanpooling	**82.8**	**96.3**

Table 8: Ablation study on different source of top-1 word embeddings.

Top-1 Word Embs Source	Top-1.(%)	Top-5.(%)
Word Emb.	82.5	96.3
Class Emb.	**82.8**	**96.3**

Table 9: Comparison of varying values of α.

α	Top-1.(%)
1	82.1
0.1	**82.4**
0.01	82.3

Table 10: Comparison of varying values of β.

β	Top-1.(%)
1	81.8
0.1	**82.8**
0.01	82.5

drop of 0.3%, which illustrates that the Adapter block can align the meta-token carrying visual information with the text tokens.

The Effect of Meanpooling Video Embs. We discuss the effect of meanpooling video embeddings in this section. From Tab. 7, we can see that, in the case of without meanpooling, where the word-guided temporal modeled video embeddings are taken directly for classification without combined with the meanpooling video embeddings, the top-1 accuracy drops 0.3%. We conjecture this is because the video embeddings obtained from meanpooling contains more robust representation which can help recognition.

How to obtain the top-1 word embeddings? In order to demonstrate how to obtain the top-1 word embeddings in the Word-Guided Visual-Prompt, we disscuss two approaches. As we depict in Sec. 3.3, we can get two sets of embeddings through the category embedding: $\mathbf{z} \in \mathbb{R}^{B\times D}$ is a set of class embeddings, which can be regarded as category-level representation, and $\mathbf{w} \in \mathbb{R}^{B\times N\times D}$ is a set of word embeddings, which is fine-grained representation. Also, we can get the meanpooling video embedding $\bar{\mathbf{v}}$. For the first approach, we calculate the cosine similarity of each class embedding in $\mathbf{z} \in \mathbb{R}^{B\times D}$ and $\bar{\mathbf{v}}$ to obtain the top-1 word embeddings. For the second approach, we calculate the similarity of $\mathbf{w}_i$ and $\bar{\mathbf{v}}$ by $\sum_{n=1}^{N} s\left(\mathbf{w}_i^n, \bar{\mathbf{v}}\right)$, where $i = 1, 2, \ldots, B$, to get the top-1 word embeddings. We show the results for two approaches in Tab. 8. We can see that using the category-level class embeddings for similarity calculation works better.

The Parameter Analysis of $\boldsymbol{\alpha}$ α refers to the parameter that controls the combination of the original category embeddings and the visual-enhanced category embeddings in the Post-Prompt module. We provides the results of different initialization values of α in Tab. 9. As we can see, when α is initialized with 1

and 0.01, the top-1 accuracy drops 0.3% and 0.1%, respectively, demonstrating that too much or too little visual information both reduce recognition capability. Overall, the initialization value of 0.1 leads to better performance.

The Parameter Analysis of β β refers to the parameter that controls the combination of the original meanpooling video embeddings and the text-enhanced video embeddings in the Word-Guided Visual-Prompt module. We provide results of different initialization values of β in Tab. 10. We can observe that when the initialization value of β is 0.01, the result drops 0.3%. In addition, when β is initialized with 1, the result get a significantly drop of 1.0%. Overall, the initialization 0f 0.1 leads to a better combination of the meanpooling video embeddings and the text-enhanced video embeddings.

5 Conclusion

In this work, we propose a parameter-efficient finetuning method call BiEfficient for video recognition, which uses bidirectional prompting mechanisms to leverage the semantic connections of vision and language. In the Vision-to-Language direction, we propose the Pre-Prompt and the Post-Prompt mechanisms to generate discriminative text representation leveraging the visual context of each input video. In the Language-to-Vision direction, we propose the Word-Guided Visual-Prompt mechanism to enhance the temporal modeling for videos leveraging the textual knowledge. Extensive experiments on three datasets across two benchmarks has demonstrated the effectiveness and efficiency of our proposed method.

Acknowledgements This research is partially supported by Beijing Natural Science Foundation (No.L231005) and the National Key Research and Development Program of China (No.2021YFB2900704).

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021) 4, 9
2. Bahng, H., Jahanian, A., Sankaranarayanan, S., Isola, P.: Exploring visual prompts for adapting large-scale models. arXiv preprint arXiv:2203.17274 (2022) 2
3. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML. vol. 2, p. 4 (2021) 4, 9
4. Brattoli, B., Tighe, J., Zhdanov, F., Perona, P., Chalupka, K.: Rethinking zero-shot video classification: End-to-end training for realistic applications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4613–4623 (2020) 11
5. Chen, S., Huang, D.: Elaborative rehearsal for zero-shot action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13638–13647 (2021) 11

6. Christoph, R., Pinz, F.A.: Spatiotemporal residual networks for video action recognition. Advances in neural information processing systems **2**, 3468–3476 (2016) 4
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 9
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020) 2, 4
9. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6824–6835 (2021) 4
10. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019) 4
11. Gao, J., Zhang, T., Xu, C.: I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8303–8311 (2019) 11
12. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**(2), 581–595 (2024) 2
13. Goswami, K., Karanam, S., Udhayanan, P., Joseph, K., Srinivasan, B.V.: Copl: Contextual prompt learning for vision-language understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 18090–18098 (2024) 2, 4
14. He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., Neubig, G.: Towards a unified view of parameter-efficient transfer learning. arXiv preprint arXiv:2110.04366 (2021) 2, 4
15. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016) 6
16. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International conference on machine learning. pp. 2790–2799. PMLR (2019) 2, 4
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021) 2, 4
18. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021) 1, 3
19. Ju, C., Han, T., Zheng, K., Zhang, Y., Xie, W.: Prompting visual-language models for efficient video understanding. In: European Conference on Computer Vision. pp. 105–124. Springer (2022) 4, 6, 8, 9, 10, 12
20. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 3, 8
21. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision. pp. 2556–2563. IEEE (2011) 3, 8

22. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 (2021) 2, 4
23. Li, K., Wang, Y., Peng, G., Song, G., Liu, Y., Li, H., Qiao, Y.: Uniformer: Unified transformer for efficient spatial-temporal representation learning. In: International Conference on Learning Representations (2021) 9, 11, 12
24. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190 (2021) 2, 4
25. Li, Y., Ji, B., Shi, X., Zhang, J., Kang, B., Wang, L.: Tea: Temporal excitation and aggregation for action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 909–918 (2020) 4
26. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7083–7093 (2019) 4
27. Lin, Z., Geng, S., Zhang, R., Gao, P., De Melo, G., Wang, X., Dai, J., Qiao, Y., Li, H.: Frozen clip models are efficient video learners. In: European Conference on Computer Vision. pp. 388–404. Springer (2022) 2, 4, 9, 10
28. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022) 4, 9, 12
29. Liu, Z., Luo, D., Wang, Y., Wang, L., Tai, Y., Wang, C., Li, J., Huang, F., Lu, T.: Teinet: Towards an efficient architecture for video recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 11669–11676 (2020) 4
30. Liu, Z., Wang, L., Wu, W., Qian, C., Lu, T.: Tam: Temporal adaptive module for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13708–13718 (2021) 4
31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 8
32. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing **508**, 293–304 (2022) 2, 4
33. Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., Ling, H.: Expanding language-image pretrained models for general video recognition. In: European Conference on Computer Vision. pp. 1–18. Springer (2022) 2, 4, 9, 10, 11
34. Pan, J., Lin, Z., Zhu, X., Shao, J., Li, H.: St-adapter: Parameter-efficient image-to-video transfer learning. Advances in Neural Information Processing Systems **35**, 26462–26477 (2022) 2, 4, 9, 10
35. Qin, J., Liu, L., Shao, L., Shen, F., Ni, B., Chen, J., Wang, Y.: Zero-shot action recognition with error-correcting output codes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2833–2842 (2017) 11
36. Qing, Z., Zhang, S., Huang, Z., Wang, X., Wang, Y., Lv, Y., Gao, C., Sang, N.: Mar: Masked autoencoders for efficient action recognition. IEEE Transactions on Multimedia (2023) 2, 4
37. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3d residual networks. In: proceedings of the IEEE International Conference on Computer Vision. pp. 5533–5541 (2017) 4
38. Quanfu Fan, Richard Chen, R.P.: Can an image classifier suffice for action recognition? In: International Conference on Learning Representations (ICLR) (2022) 9

39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 1, 3, 6, 7, 8
40. Ryoo, M., Piergiovanni, A., Arnab, A., Dehghani, M., Angelova, A.: Tokenlearner: Adaptive space-time tokenization for videos. Advances in neural information processing systems **34**, 12786–12797 (2021) 9, 10
41. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems **27** (2014) 4
42. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012) 3, 8
43. Sung, Y.L., Cho, J., Bansal, M.: Lst: Ladder side-tuning for parameter and memory efficient transfer learning. Advances in Neural Information Processing Systems **35**, 12991–13005 (2022) 2
44. Sung, Y.L., Nair, V., Raffel, C.A.: Training neural networks with fixed sparse masks. Advances in Neural Information Processing Systems **34**, 24193–24205 (2021) 4
45. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015) 4
46. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018) 4
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 6, 7, 12
48. Wang, L., Tong, Z., Ji, B., Wu, G.: Tdn: Temporal difference networks for efficient action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1895–1904 (2021) 4
49. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016) 4
50. Wang, M., Xing, J., Liu, Y.: Actionclip: A new paradigm for video action recognition. arXiv preprint arXiv:2109.08472 (2021) 4, 6, 8, 9, 10, 11, 12
51. Wang, Q., Chen, K.: Alternative semantic representations for zero-shot human action recognition. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10. pp. 87–102. Springer (2017) 11
52. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018) 9
53. Wu, W., He, D., Lin, T., Li, F., Gan, C., Ding, E.: Mvfnet: Multi-view fusion network for efficient video recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 2943–2951 (2021) 4
54. Wu, W., Luo, H., Fang, B., Wang, J., Ouyang, W.: Cap4video: What can auxiliary captions do for text-video retrieval? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10704–10713 (2023) 4
55. Wu, W., Sun, Z., Ouyang, W.: Revisiting classifier: Transferring vision-language models for video recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2847–2855 (2023) 4, 9, 10

56. Wu, W., Wang, X., Luo, H., Wang, J., Yang, Y., Ouyang, W.: Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6620–6630 (2023) 2, 4, 6, 9, 10
57. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European conference on computer vision (ECCV). pp. 305–321 (2018) 4
58. Xu, X., Hospedales, T.M., Gong, S.: Multi-task zero-shot action recognition with prioritised data augmentation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 343–359. Springer (2016) 11
59. Yan, S., Xiong, X., Arnab, A., Lu, Z., Zhang, M., Sun, C., Schmid, C.: Multiview transformers for video recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3333–3343 (2022) 4, 9, 10
60. Yang, T., Zhu, Y., Xie, Y., Zhang, A., Chen, C., Li, M.: Aim: Adapting image models for efficient video understanding. In: International Conference on Learning Representations (2023) 2, 4, 9, 10
61. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022) 1, 3
62. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021) 1, 3
63. Zaken, E.B., Ravfogel, S., Goldberg, Y.: Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. arXiv preprint arXiv:2106.10199 (2021) 2, 4
64. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by clip. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8552–8562 (2022) 4
65. Zhao, S., Zhu, L., Wang, X., Yang, Y.: Centerclip: Token clustering for efficient text-video retrieval. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 970–981 (2022) 4
66. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022) 2, 4
67. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022) 2, 4
68. Zhu, Y., Long, Y., Guan, Y., Newsam, S., Shao, L.: Towards universal representation for unseen action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9436–9445 (2018) 11