A Multimodal, Multi-Task Adapting Framework for Video Action Recognition

Mengmeng Wang[1], **Jiazheng Xing**[1], **Boyuan Jiang**[2], **Jun Chen**[1],
Jianbiao Mei[1], **Xingxing Zuo**[3]*, **Guang Dai**[4], **Jingdong Wang**[5], **Yong Liu**[1]*
[1]Zhejiang University
[2]Youtu Lab,Tencent
[3]Technical University of Munich
[4]SGIT AI Lab, State Grid Corporation of China
[5]Baidu Inc
{mengmengwang, jiazhengxing, junc, jianbiaomei}@zju.edu.cn, byronjiang@tencent.com
xingxing.zuo@tum.de, guang.gdai@gmail.com, wangjingdong@outlook.com, yongliu@iipc.zju.edu.cn

Abstract

Recently, the rise of large-scale vision-language pretrained models like CLIP, coupled with the technology of Parameter-Efficient FineTuning (PEFT), has captured substantial attraction in video action recognition. Nevertheless, prevailing approaches tend to prioritize strong supervised performance at the expense of compromising the models' generalization capabilities during transfer. In this paper, we introduce a novel Multimodal, Multi-task CLIP adapting framework named M^2-CLIP to address these challenges, preserving both high supervised performance and robust transferability. Firstly, to enhance the individual modality architectures, we introduce multimodal adapters to both the visual and text branches. Specifically, we design a novel visual TED-Adapter, that performs global Temporal Enhancement and local temporal Difference modeling to improve the temporal representation capabilities of the visual encoder. Moreover, we adopt text encoder adapters to strengthen the learning of semantic label information. Secondly, we design a multi-task decoder with a rich set of supervisory signals to adeptly satisfy the need for strong supervised performance and generalization within a multimodal framework. Experimental results validate the efficacy of our approach, demonstrating exceptional performance in supervised learning while maintaining strong generalization in zero-shot scenarios.

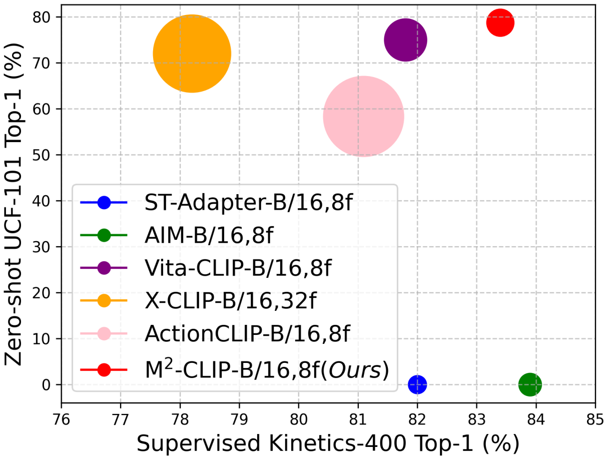

Figure 1: Performance Comparison: Zero-shot *vs* supervised accuracy. The circle area represents the number of tunable parameters, where models with better performance are positioned towards the right and upper side, with a small circle area. Our M^2-CLIP achieves the best zero-shot performance with very few tunable parameters.

Introduction

Over the past few years, there has been a remarkable surge of large-scale vision-language pre-trained models (VLM) like CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021), and Florence (Yuan et al. 2021). As a result, researchers have actively delved into methods to effectively adapt these large models to their specific domains. In this paper, we focus on transferring the influential CLIP model to the domain of video action recognition, emphasizing its crucial role in driving advancements in this field.

Undoubtedly, transferring knowledge from the powerful CLIP holds great promise due to its robust representation capability and impressive generalization performance. The most intuitive approach is to directly add temporal modeling to CLIP's image encoder and then finetune the entire network (Wang et al. 2023; Tu et al. 2023; Ni et al. 2022). However, finetuning comes with a high computational cost and may potentially impact the original generalization capabilities of CLIP. With the emergence of PEFT, researchers have begun to explore freezing the original CLIP parameters and introducing various adapters (Liu et al. 2023; Park et al. 2023) or prompts (Wasim et al. 2023; Ju et al. 2022), only training the newly added parameters. Notably, PEFT has motivated a reevaluation of the traditional unimodal video classification framework. By directly utilizing CLIP's visual branch in conjunction with added adapters, coupled with a Linear classification layer at the end, these approaches have demonstrated impressive results in supervised scenarios (Lin

*Corresponding authors: Yong Liu and Xingxing Zuo.

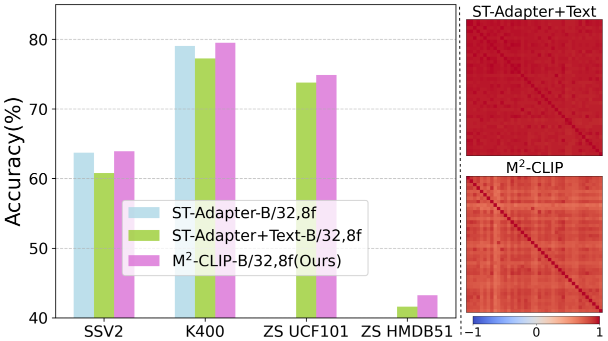

Figure 2: Analysis of transferring a unimodal framework into a multimodal one. (a) Performance comparison. Note that ST-Adapter is not able to zero-shot transferring, thus having no results in zero-shot (ZS) UCF101 and HMDB51. (b) Inter-class correlation maps of the top 40 correlated SSv2 label features of ST-Adapter+text *vs* the corresponding 40 label features of our method. The redder the color, the stronger the feature coupling. Our M^2-CLIP ultimately improved the performance on the four datasets and significantly reduced the correlation among the features of different labels.

et al. 2022b; Pan et al. 2022; Yang et al. 2023; Park et al. 2023; Zhao et al. 2023). However, it is worth noting that excluding the text branch in these approaches leads to the loss of CLIP's generalization capabilities, which are among the fundamental attractions of the CLIP itself.

PEFT can also be applied to multimodal CLIP transfer frameworks, directly affecting the visual branch (Liu et al. 2023) or the text branch (Ju et al. 2022), or even both simultaneously (Wasim et al. 2023). It significantly improves efficiency and reduces the number of learnable parameters. However, freezing the multimodal backbone causes a drop in supervised accuracy, leaving a gap compared to the performance of the unimodal frameworks, even when incorporating strong unimodal adapters. We experiment to validate this observation further, as shown in the left of Fig. 2. Using ST-Adapter (Pan et al. 2022) as a representative of the unimodal frameworks, we introduce CLIP's text branch to transform ST-Adapter into a multimodal framework. Just as anticipated, by freezing the CLIP parameters while learning the adapters, we have indeed observed a noticeable decrease in supervised performance. The reason is that the text branch of CLIP lacks sufficient discriminative features, particularly for action verbs, as shown in the right of Fig. 2. Then, the contrastive learning loss of CLIP itself makes it hard to learn discriminative features for videos when training with relatively small datasets compared with its original training set, especially when textual data is scarce.

To mitigate the performance degradation while ensuring generalization, we propose a new multimodal, multi-task CLIP transfer framework, dubbed as M^2-CLIP. Firstly, we focus on multimodal adapting to construct stronger architectures, adding adapters to both the text and visual branches. Specifically, to better represent the temporal information of videos, we design a novel TED-Adapter, capable of simultaneously integrating global temporal enhancement and local temporal difference modeling. In addition, we introduce a kind of naive adapter to the text branch to capture additional semantic information related to action labels, which significantly improves the first issue. Secondly, we devise a multi-task decoder for tapping into more substantial learning potential. The decoder consists of four components. (a) The first is the original contrastive learning head, which aims to align the pairwise video-text representations. (b) The second head is a cross-modal classification head, which can highlight the discriminative capabilities of cross-modal features. (c) Thirdly, we design a cross-modal masked language modeling head at the final layer of the text branch, promoting the focus of visual features on verbs for recognition. (d) Lastly, we incorporate a visual feature classifier at the end of the visual branch to facilitate the distinction of visual features across diverse categories.

In summary, our contributions are threefold: 1) We propose a novel multi-modal, multi-task adapting framework to transfer the powerful CLIP to video action recognition tasks. This method achieves strong supervised performance while ensuring state-of-the-art zero-shot transferability as shown in Fig.1. 2) We design a new visual TED-adapter that performs Temporal Enhancement and Difference modeling to enhance the representation capabilities of the video encoder. Simultaneously, we introduce the adapters for the text encoder to make the label representation learnable and adjustable. 3) We introduce a multi-task decoder to improve the learning capability of the whole framework, adeptly achieving a balance between supervised performance and generalization.

Related Works

Early action recognition algorithms (Bertasius, Wang, and Torresani 2021; Arnab et al. 2021; Jiang et al. 2019) mostly relied on end-to-end finetuning of models pretrained on ImageNet (Deng et al. 2009). Recently, the advent of large-scale image-language pretrained models like CLIP brought about significant changes (Wang et al. 2023; Ni et al. 2022). The PEFT technique initially emerged in the NLP field (Houlsby et al. 2019) to address the challenges of full finetuning for large-scale language models. In video action recognition, this technique has also become a research hotspot in recent years (Ju et al. 2022; Lin et al. 2022b; Yang et al. 2023; Xing et al. 2023; Liu et al. 2023). EVL (Lin et al. 2022b) first proposed to leverage frozen CLIP image features with a lightweight spatiotemporal Transformer decoder to enhance video recognition tasks. ST-Adapter (Pan et al. 2022) introduced a parameter-efficient spatiotemporal adapter, which effectively harnesses the power of CLIP's image models for video understanding. AIM (Yang et al. 2023) presented spatial, temporal, and joint adaptations to finetune pretrained image transformer models. These cost-effective methods have achieved SOTA performance, but they are all single-modal transfers, neglecting the text branch and thereby losing CLIP's generalization ability. Vita-CLIP (Wasim et al. 2023) attempted to address this issue by adding prompts to both branches for transfer. However, we observed that its

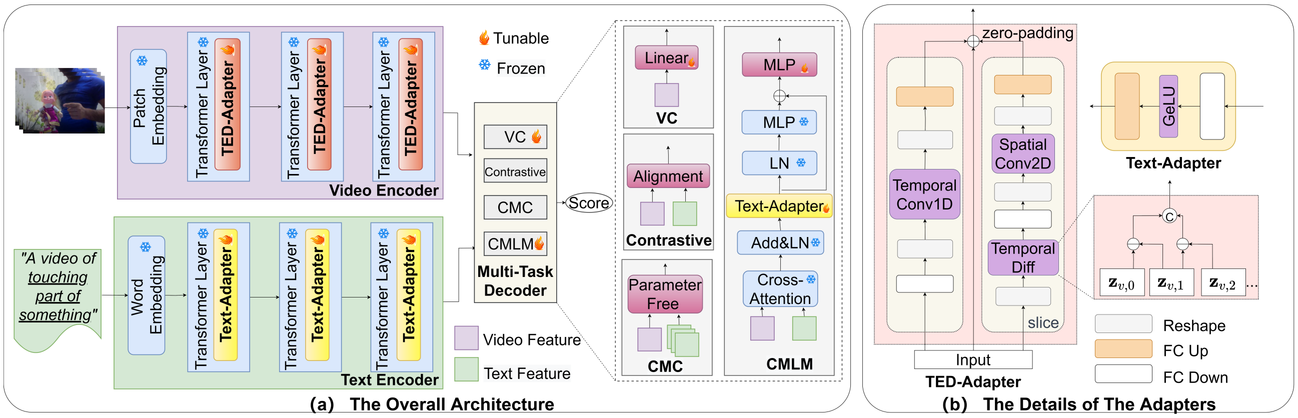

Figure 3: (a) Overview of M^2-CLIP: An example of integrating an adapter into each transformer layer is illustrated. M^2-CLIP consists of a video encoder, a text encoder and a multi-task decoder, where the backbones of the two encoders are frozen and assisted by the proposed trainable TED-Adapter and Text-Adapter. The multi-task decoder has four different heads that utilize multi-task constraints to improve the joint representation of the entire multimodal framework. (b) Detailed Structure of proposed adapters, where $L = 1 + M$ and $h \times w = M$.

performance on temporally strong-correlated datasets was suboptimal, and their additional summary attention layers introduced in its visual branch increased the number of learnable parameters. In this work, we aim to apply PEFT to multi-modal frameworks to ensure competitive supervised performance with minimal increase in learnable parameters while maintaining strong generalization capabilities.

Method

Architecture Overview

As illustrated in Fig. 3a, our framework has three key components: a video encoder, a text encoder, and a multi-task decoder. Here we will introduce the overview of the whole architecture and leave the details of the multimodal adapters and multi-task decoder in the following two sections.

Formally, the input to the framework is given as a video $\mathbf{V} \in \mathbb{R}^{T\times H\times W\times 3}$ of spatial size $H \times W$ with T sampled frames, and a text label $\mathbf{y}$ from a predefined label set $\mathcal{Y}$.

Video Encoder: $\mathbf{E}_v$ consists of L_v transformer layers $\{\mathcal{E}_v^{(i)}\}_{i=1}^{L_v}$ and the proposed corresponding visual TED-Adapters $\{\mathcal{A}_v^{(j)}\}_{j=1}^{L_{A_v}}$. The t-th frame of the input divided into non-overlapping patches $\{\mathbf{P}_{t,i}\}_{i=1}^{M} \in \mathbb{R}^{P^2\times 3}$, $M = HW/P^2$. Then they are then projected into patch embeddings $\mathbf{X}_{v,t} \in \mathbb{R}^{M\times d_v}$, prepended with a learnable class token $\mathbf{C}_t$ and added with a positional encoding $\mathbf{e}_v$. Mathematically, the frame-level input is constructed as:

$$[\mathbf{c}_t^{(0)}, \mathbf{z}_{v,t}^{(0)}] = [\mathbf{C}_t, \mathbf{X}_{v,t}] + \mathbf{e}_v. \qquad (1)$$

If we place the visual adapter before every transformer layer, the input will be sequentially processed as,

$$[\mathbf{c}_t^{(i)}, \mathbf{z}_{v,t}^{(i)}] = \mathcal{E}_v^{(i)}(\mathcal{A}_v^{(i)}([\mathbf{c}_t^{(i-1)}, \mathbf{z}_{v,t}^{(i-1)}])) \quad i = 1, 2, \cdots, L_v. \qquad (2)$$

To obtain the final video representation $\mathbf{v}$, $\mathrm{c}_t^{(L_v)}$ (the class tokens) of the last transformer layer is projected to a common video-language (VL) space by $\mathbf{v}_t = \mathbf{h}_v(\mathrm{c}_t^{(L_v)})$, and averaged along the temporal dimension,

$$\mathbf{v} = \mathtt{AvgPool}([\mathbf{v}_1, \cdots, \mathbf{v}_T]) \qquad \mathbf{v} \in \mathbb{R}^{d_{vl}}. \qquad (3)$$

Language Encoder: Similarly, $\mathbf{E}_l$ consists of L_l transformer layers $\{\mathcal{E}_l^{(i)}\}_{i=1}^{L_l}$ and its corresponding text adapter $\{\mathcal{A}_l^{(j)}\}_{j=1}^{L_{A_l}}$. The input words are tokenized and projected into word embeddings $\mathbf{X}_l \in \mathbb{R}^{N\times d_l}$, where N is the text length. The input to the encoder is constructed as:

$$\mathbf{z}_l^{(0)} = \mathbf{X}_l + \mathbf{e}_l. \qquad (4)$$

Taking the example of inserting the text adapter before each transformer layer, the feature of each layer is obtained as:

$$\mathbf{z}_l^{(i)} = \mathcal{E}_l^{(i)}(\mathcal{A}_l^{(i)}(\mathbf{z}_l^{(i-1)})) \quad i = 1, 2, \cdots, L_l. \qquad (5)$$

The final VL space text representation $\mathbf{w} \in \mathbb{R}^{d_{vl}}$ of the label $\mathbf{y}$ is obtained by $\mathbf{w} = \mathbf{h}_l(\mathbf{z}_{l,N}^{(L_l)})$, where $\mathbf{z}_{l,N}^{(L_l)}$ is the last token of $\mathbf{z}_l^{(L_l)}$ and $\mathbf{h}_l$ is a projection layer.

Decoder: Once the output features from the two encoders are obtained, they are fed into our specially designed multi-task decoder. In the training process, the role of the decoder is to impose constraints on the feature representations generated by the encoders, facilitating semantic alignment between the two modalities and enabling differentiation between features of different categories. Once a model completes its training, the decoder is versatile, capable of generating classification scores for supervised learning and conducting zero-shot classification. The detailed design of the decoder's structure will be elaborated in the following section.

Visual and Textual Adapters

To better transfer CLIP to this task and enhance the semantic representation of action verbs in the labels, we introduce adapters for both the visual and text branches to improve their respective representation capabilities.

Video TED-Adapter: Adapting CLIP's image branch to the video branch requires additional temporal modeling modules, which can be approached from two perspectives, global temporal enhancement and local temporal difference modeling. The former is the intuitive global temporal aggregation referred to as spatiotemporal features (Lin, Gan, and Han 2019; Feichtenhofer et al. 2019), where temporal attentions or temporal convolutions are applied to multiple frames' features to aggregate the similar action subject. It has been extensively explored in CLIP's transfer (Pan et al. 2022; Yang et al. 2023; Liu et al. 2023). The latter is short-term frame-wise feature difference learning, which seeks to capture the local motion patterns and dynamics between adjacent frames. This kind of feature has been mentioned in earlier computation-efficient convolutional algorithms (Jiang et al. 2019; Wang et al. 2022; Li et al. 2020) but remains unexplored in the context of CLIP's transfer. To explore both the two kinds of temporal modeling in a unified structure, we design a novel TED-Adapter, which learns the **T**emporal **E**nhancements and temporal **D**ifferences in the meanwhile.

As shown in Fig. 3b, we first adopt a 1D temporal convolution for temporal feature enhancement. For the input of a TED-Adapter layer including the class token and patch tokens $\mathbf{Z} = \{[\mathbf{c}_t, \mathbf{z}_{v,t}]\}_{t=1}^{T} \in \mathbb{R}^{T\times(1+M)\times d_v}$, we perform the following operations:

$$\mathbf{Z}_E = \mathtt{Conv1D}(\mathbf{Z}\mathbf{W}_{dn})\mathbf{W}_{up}, \tag{6}$$

where $\mathbf{W}_{dn}$ and $\mathbf{W}_{up}$ are the down-projection and up-projection weights. $\mathtt{Conv1D}$ represents the 1D-convolution for spatiotemporal modeling operating on the temporal dimension. Note that the reshape operations are omitted in this section for simplicity but are shown in Fig. 3b.

Next, for the temporal difference modeling, we subtract the previous frame's feature from the current frame and then employ a 2D spatial convolution to learn useful information from the adjacent feature differences automatically. Formally, when given the input patch tokens $\mathbf{z}_{v,t}$ the of the t-th frame,

$$\mathbf{z}_{D,t} = \mathtt{Conv2D}((\mathbf{z}_{v,t} - \mathbf{z}_{v,t-1})\mathbf{W}_{dn})\mathbf{W}_{up}, \tag{7}$$

where $\mathtt{Conv2D}$ represents the 2D spatial convolution. For the first frame, we set its feature differences to zeros.

Finally, the output of TED-Adapter can be obtained by fusing the two kinds of temporal features together. Moreover, a residual summation is applied to preserve the information in the input:

$$\mathcal{A}_v([\mathbf{c}_t, \mathbf{z}_{v,t}]) = \mathbf{Z}_E + \mathbf{Z}_D + \mathbf{Z}, \tag{8}$$

where $\mathbf{Z}_D = \{[\mathbf{O}, \mathbf{z}_{D,t}]\}_{t=1}^{T}$, and $\mathbf{O}$ is a zero matrix which has the same shape as $\mathbf{c}_t$.

The TED-Adapter is simply placed before the Multi-Head Self-Attention (MHSA) by default unless otherwise specified. By incorporating the temporal enhancement and temporal difference operations, the proposed TED-Adapter can capture spatiotemporal features and local finer motion patterns, which are both crucial for this task.

Text Adapter: In action recognition, the textual labels describing actions are often short and succinct, emphasizing the actions themselves, such as "unfolding something" and "hurdling". However, we observed that CLIP's text encoder alone might not effectively distinguish such label text features, as shown in Fig. 2. To address this, we introduce adapters to the text branch to learn better semantic representations for the action labels. We directly utilized the basic adapter (Houlsby et al. 2019) structure here as shown in Fig. 3b. Specifically, given the input text tokens of a text adapter layer $\mathbf{z}_l$, we perform the text adapter like:

$$\mathcal{A}_l(\mathbf{z}_l) = \mathbf{z}_l + \mathtt{Act}(\mathbf{z}_l\mathbf{W}_{dn})\mathbf{W}_{up}, \tag{9}$$

where $\mathtt{Act}$ means a non-linear activation function and we use GeLU here.

The text adapters are inserted before the Feed-Forward Networks (FFN) of the transformer layer by default. By incorporating the text adapter, the model can enhance its understanding of the action labels, capturing more discriminative semantic information. This allows for improved alignment between the textual and visual representations, resulting in more accurate and effective video action recognition.

Multi-Task Decoder

As previously described, we observed that when utilizing CLIP's multimodal framework, relying solely on contrastive learning did not perform as well as the equivalently configured unimodal framework. To address this, we propose a multi-task decoder equipped with four distinct learning tasks, each corresponding to a separate head, as shown in the right part of Fig. 3a. This approach aims to leverage multiple task constraints to improve the joint representation power of the multimodal framework.

Multimodal Contrastive Learning Head (Contrastive). This is the original training objective of CLIP. To pull the pairwise video representation $\mathbf{v}$ and label representation $\mathbf{w}$ close to each other, symmetric similarities are defined between the two modalities:

$$\begin{aligned} p_i^{\mathbf{v2y}}(\mathbf{V}) &= \frac{\exp(\cos(\mathbf{v}, \mathbf{w}_i)/\tau)}{\sum_{j=1}^{B}\exp(\cos(\mathbf{v}, \mathbf{w}_j)/\tau)}, \\ p_i^{\mathbf{y2V}}(\mathbf{y}) &= \frac{\exp(\cos(\mathbf{w}, \mathbf{v}_i)/\tau)}{\sum_{j=1}^{B}\exp(\cos(\mathbf{w}, \mathbf{v}_j)/\tau)}, \end{aligned} \tag{10}$$

where $\cos$ means cosine similarity, τ is a temperature parameter and B is the number of training pairs. The ground-truth is defined as 0 for negative pairs and 1 for positive pairs. We use Kullback-Leibler divergence as the video-text contrastive loss to optimize this head as ActionCLIP.

When a model is trained, it will be ready for zero-shot classification:

$$p(\hat{\mathbf{y}}|\mathbf{V}) = \frac{\exp(\cos(\mathbf{v}, \mathbf{w}_{\hat{y}})/\tau)}{\sum_{i=1}^{C}\exp(\cos(\mathbf{v}, \mathbf{w}_i))}. \tag{11}$$

In practice, the text input can be prompted like "$\mathtt{a\ video\ of}$ $< \hat{\mathbf{y}} >$", where $\hat{\mathbf{y}}$ is a category name of C classes. The process of predicting $\hat{\mathbf{y}}$ of a certain video $\mathbf{V}$ is to find the highest similarity score calculated by Eq. 11.

Cross-Modal Masked Language Modeling Head (CMLM). Unlike the original CLIP, which primarily deals with image-text paired data, our action labels predominantly focus on verbs. To enhance CLIP's text branch for better representation of action-related words and help the learning of the text adapters, we introduce an additional CMLM head, which urges the text branch to predict masked words from the other text and video tokens. Specifically, given the framewise video features $[\mathbf{v}_1, \cdots, \mathbf{v}_T]$ and the text features $\mathbf{z}_l^{(L_l)}$, we perform a cross-attention operation to obtain cross-modal features. Due to the limited amount of textual data, directly learning the parameters of this attention layer can be challenging. Our approach to addressing this is to initialize the parameters of this attention layer using the parameters of the final transformer layer in the text branch and then freeze these parameters and add a text adapter of Eq. (9) before the FFN of the transformer layer. Then we only learn the parameters of this text adapter. The process can be presented as:

$$\begin{aligned} \mathbf{w}^* &= \mathbf{z}_l^{(L_l)} + \mathtt{CA}(\mathtt{LN}(\mathbf{z}_l^{(L_l)}), \mathtt{LN}([\mathbf{v}_1, \cdots, \mathbf{v}_T])), \\ \hat{\mathbf{w}} &= \mathcal{A}_l(w^*), \;\; \mathbf{w}_m = \hat{\mathbf{w}} + \mathtt{MLP}(\mathtt{LN}(\hat{\mathbf{w}})), \end{aligned} \tag{12}$$

where CA, LN and MLP indicate the cross attention layer, layer norm and a MLP layer, respectively. Then, we attach a BERT MLM head (Devlin et al. 2018) to predict the masked words with a cross-entropy loss, as shown in Fig. 3a.

Cross-Modal Classification Head (CMC). Since the action labels are predefined within a given set (C classes), we can compute the complete label feature set for each iteration, enabling us to carry out cross-modal feature classification. In this work, we employed a straightforward parameter-free cross-modal fusion approach, which directly computes cosine similarity using the Eq. (11). Note that it differs from Eq. (10), which considers video-text matching within a training batch and can not cover all the action labels' representations. After obtaining these similarities, the goal is to ensure that a video's representation is similar to the textual representation of its corresponding label rather than the textual representations of other categories. To achieve this, we ingeniously transform the problem into a 1-in-C classification task and add a classification constraint to the cross-modal similarity scores with the cross-entropy loss.

Visual Classification Head (VC). Furthermore, we introduced a straightforward classification head to the video branch to enhance the distinction between different categories in video features. Given video features $\mathbf{v}$, we directly appended a Linear layer for classification, training with cross-entropy loss. Importantly, with the inclusion of this classification head, we can directly use its output for supervised classification tasks. For zero-shot experiments, we still employ Eq. (11). The addition of this classification head enables the model to learn to discriminate the video features between different action categories more effectively.

In summary, by introducing these four learning tasks, we tap into a richer set of supervisory signals, guiding the model to better align the visual and textual modalities while simultaneously capturing various aspects of semantic information. This multi-task approach not only mitigates the performance disparity of supervised learning but also preserves CLIP's remarkable generalization capabilities.

Experiments

Experimental Setup

We evaluate our M^2-CLIP for supervised learning in two primary datasets: Kinetics-400 (K400) (Kay et al. 2017) and Something-Something-V2 (SSv2) (Goyal et al. 2017). For the generalization evaluation, we test our model on UCF101 (Soomro, Zamir, and Shah 2012) and HMDB51 (Kuehne et al. 2011). We employ ViT-B/16 based CLIP as our backbone and use a sparse frame sampling strategy with 8, 16, or 32 frames during training and inference.

Fully-Supervised Experiments

We present our results of K400 and SSV2 in Tab. 1 and Tab. 2, respectively, comparing our approach with SOTAs trained under various transfer methods, including full finetuning, unimodal and multimodal PEFT from frozen CLIP.

On K400, our 8-frame M^2-CLIP-B/16 model surpasses models pretrained by ImageNet (Deng et al. 2009), achieving higher performance with fewer learnable parameters and computational requirements. Compared to end-to-end finetuned CLIP models with the same ViT-B/16 backbones, our approach demonstrates comparable results. With just 11% of the adjustable parameters, we exceed ActionCLIP (Wang et al. 2023)'s results. In addition, while our results fall slightly short when compared to the leader performances achieved by BIKE (Wu et al. 2023) and ILA (Tu et al. 2023), it's important to note that they employed a much larger network architecture (ViT-L) and had 14 times the number of tunable parameters as ours. Compared with the unimodal PEFT approaches, our method achieves comparable or even superior results. Our 8-frame M^2-CLIP-B/16 model outperforms 8-frame ST-Adapter-B/16 (Pan et al. 2022) by 1.4%. Note that while unimodal methods exhibit high performance in supervised settings, they lack support for zero-shot generalization. In contrast, our approach achieves competitive results with them and demonstrates strong generalizations. Lastly, compared with multimodal PEFT approaches, our method achieves superior results. Vita-CLIP (Wasim et al. 2023) is a multimodal Prompt-based method while we use adapters. It is evident that we achieve higher performance with only 41% trainable parameters.

As for SSv2, our approach achieves comparable performance and even surpasses several full-finetuned methods with similar configurations, such as Mformer-B (Patrick et al. 2021) and ILA-ViT-B (Tu et al. 2023), while utilizing fewer trainable parameters and computational resources. Compared with unimodal PEFT methods, our 8-frame M^2-CLIP model surpasses AIM-ViT-B/16 (Yang et al. 2023) and EVL-ViT-B/16 (Lin et al. 2022b) in Top-1 performance and maintains a competitive position compared to other methods. In the domain of multimodal PEFT approaches, our method outperforms the recent method Vita-CLIP (Wasim et al. 2023) by a large margin of over 18%. The results demonstrate that our proposed multimodal adapters and the

Method	Pre-train	Tunable Param	#Frames	Top-1(%)	Top-5(%)	GFLOPs	Zero-shot
Full Finetuning							
Swin-B (Liu et al. 2022)	IN-21k	88	32 × 4 × 3	82.7	95.5	282	✗
MViTv2-B (Li et al. 2022b)	✗	52	32 × 5 × 1	82.9	95.7	225	✗
Uniformer V2-B/16 (Li et al. 2022a)	CLIP-400M	115	8 × 3 × 4	85.6	97.0	154	
ActionCLIP-B/16 (Wang et al. 2023)	CLIP-400M	142	32 × 10 × 3	83.8	96.2	563	✓
X-CLIP-B/16 (Ni et al. 2022)	CLIP-400M	132	16 × 4 × 3	84.7	96.8	287	✓
BIKE-L/14 (Wu et al. 2023)	CLIP-400M	230	16 × 4 × 3	**88.1**	97.9	830	✓
S-ViT-B/16 (Zhao et al. 2023)	CLIP-400M	-	16 × 3 × 4	84.7	96.8	340	✗
ILA-ViT-L/14 (Tu et al. 2023)	CLIP-400M	-	8 × 4 × 3	88.0	**98.1**	673	✓
PEFT: unimodal visual framework (frozen CLIP)							
EVL-B/16 (Lin et al. 2022b)	CLIP-400M	86	8 × 1 × 3	82.9	-	444	✗
ST-Adapter-B/16 (Pan et al. 2022)	CLIP-400M	7	8 × 1 × 3	82.0	95.7	148	✗
ST-Adapter-B/16 (Pan et al. 2022)	CLIP-400M	7	32 × 1 × 3	82.7	96.2	607	✗
AIM-B/16 (Yang et al. 2023)	CLIP-400M	11	8 × 1 × 3	83.9	96.3	202	✗
AIM-B/16 (Yang et al. 2023)	CLIP-400M	11	32 × 1 × 3	84.7	96.7	809	✗
DUALPATH-B/16 (Park et al. 2023)	CLIP-400M	10	32 × 1 × 3	**85.4**	**97.1**	237	✗
PEFT: multimodal framework (frozen CLIP)							
STAN-conv-B/16 (Liu et al. 2023)	CLIP-400M	-	8 × 1 × 3	83.1	96.0	238	✓
Vita-CLIP B/16 (Wasim et al. 2023)	CLIP-400M	39	8 × 4 × 3	81.8	96.0	97	✓
Vita-CLIP B/16 (Wasim et al. 2023)	CLIP-400M	39	16 × 4 × 3	82.9	96.3	190	✓
M^2-CLIP-B/16	CLIP-400M	16	8 × 4 × 3	83.4	96.3	214	✓
M^2-CLIP-B/16	CLIP-400M	16	16 × 4 × 3	83.7	96.7	422	✓
M^2-CLIP-B/16	CLIP-400M	16	32 × 4 × 3	**84.1**	**96.8**	842	✓

Table 1: Performance comparison on K400. The per-view GFLOPs is reported. #Frame means frames×crops×clips.

Model	#Frames	Top-1(%)
Full Finetuning		
Mformer-B (Patrick et al. 2021)	16×1×3	66.5
MViTv2-B (Li et al. 2022b)	32×1×3	70.5
ILA-ViT-B/16 (Tu et al. 2023)	8×4×3	65.0
ILA-ViT-B/16 (Tu et al. 2023)	16×4×3	66.8
Uniformer V2-B/16 (Li et al. 2022a)	32×1×3	**70.7**
S-ViT-B/16 (Zhao et al. 2023)	16×2×3	69.3
PEFT: unimodal visual framework (frozen CLIP)		
ST-Adapter-B/16 (Pan et al. 2022)	8×1×3	67.1
ST-Adapter-B/16 (Pan et al. 2022)	32×1×3	69.5
EVL-ViT-B/16 (Lin et al. 2022b)	16×1×3	61.7
DUALPATH-B/16 (Park et al. 2023)	32×1×3	**70.3**
AIM-ViT-B/16 (Yang et al. 2023)	8×1×3	66.4
AIM-ViT-B/16 (Yang et al. 2023)	32×1×3	69.1
PEFT: multimodal framework (frozen CLIP)		
STAN-conv-B/16 (Liu et al. 2023)	8×1×3	65.2
Vita-CLIP-B/16 (Wasim et al. 2023)	16×-	48.7
M^2-CLIP-B/16	8×1×3	66.9
M^2-CLIP-B/16	32×1×3	**69.1**

Table 2: Performance comparison on SSv2.

multi-task decoder are helpful strategies for efficient multi-modal CLIP-based image-to-video knowledge transfer.

Zero-shot Experiments

In this experiment, we employ the M^2-CLIP-B/16 model pre-trained on K400, with 8 frames as input, for conducting generalization experiments on UCF101. We use the outputs of the contrastive learning head for classification. It's worth noting that the model used for testing is consistent with the one mentioned in the supervised experiments in Tab. 1, as Vita-CLIP(Wasim et al. 2023).

Our approach demonstrates impressive generalization capabilities as shown in Tab. 3. All methods in this table, except ResT (Lin et al. 2022a), are transferred from CLIP. In comparison to methods that require full finetuning, our results on both datasets outperform them by a significant margin. For instance, we surpass X-CLIP (Ni et al. 2022) by 6.7% on UCF101. Moreover, our method requires far fewer trainable parameters than these approaches. Among other methods that do not require full finetuning and are based on Prompts, our method also outperforms the majority. Compared to Vita-CLIP, our model uses only 41% of the trainable parameters of them and has better supervised results. Furthermore, our accuracy on UCF101 is 2.7% higher than theirs, making our approach the leader among all the methods on this dataset.

Ablation and Analysis

In this section, unless otherwise specified, we use ViT-B/16 as the backbone and 8 input frames on K400.

Effectiveness of Components. In Tab. 4, we first construct a CLIP frozen baseline and use its zero-shot performance as

Method	**UCF101(%)**	**FT**
ResT-101 (Lin et al. 2022a)	34.4	N/A
ActionCLIP (Wang et al. 2023)	58.3 ± 3.4	✓
X-CLIP-B/16 (Ni et al. 2022)	72.0 ± 2.3	✓
A5 (Ju et al. 2022)	69.3 ± 4.2	✗
CoOp (Zhou et al. 2022b)	66.6	✗
Co-CoOp (Zhou et al. 2022a)	68.2	✗
MaPLe (Khattak et al. 2023)	68.7	✗
Vita-CLIP-B/16 (Wasim et al. 2023)	75.0 ± 0.6	✗
M^2-CLIP-B/16	**78.7** ± 1.2	✗

Table 3: Comparison for zero-shot performances on HMDB51 and UCF101. FT means CLIP finetuning.

Components	**Top-1(%)**
Baseline(CLIP zero-shot)	56.5
+ TED-Adapter	80.5
+ Multimodal-Adapter	81.4
+ Multimodal-Adapter + Multi-Task Decoder	83.4

Table 4: Ablations for Components.

1-5	**6-12**	**TE**	**TD**	**Top-1(%)**
✓		✓	✓	80.4
	✓	✓	✓	82.9
✓	✓	✓		83.1
✓	✓		✓	81.8
✓	✓	✓	✓	**83.4**

Table 5: Ablations for TED-Adapter.

Number	**Supervised K400** (%)	**ZS UCF101 (%)**
0	83.1	75.6
1	**83.4**	**79.2**
6	83.3	78.1
12	83.2	76.4

Table 6: Ablations for Text-Adapter.

Contrastive	**CMC**	**MLM**	**VC**	**Top-1(%)**
✓				81.4
✓	✓			82.1
✓	✓	✓		82.4
✓	✓	✓	✓	**83.4**

Table 7: Ablations for Multi-task decoder.

our baseline. Then, we gradually introduced our contributions based on this baseline. We have observed a significant performance boost when we introduce the learnable temporal modeling module, TED-Adapter, demonstrating its effectiveness. Subsequently, we add the Text-Adapter, which formulates the multimodal adapters with TED-Adapter to CLIP and further improves the performance. Finally, by adding the Multi-task decoder, we form the final M^2-CLIP, incorporating multiple learning objectives and resulting in a substantial improvement. In summary, our proposed multimodal adapter and multi-task decoder are both highly effective and modular components that can be easily integrated into CLIP transfer frameworks as plug-and-play modules.

Ablations for Video TED-Adapter. We next conducted ablation experiments on Video TED-Adapter in Tab. 5. We use one TED-Adapter with bottleneck width 384 as ST-Adapter (Pan et al. 2022). First, we attempt to add the TED-Adapter to the front half and the back half of the network. We have observed that although more TED-Adapters generally lead to better results, the benefits gained from adding them to deeper layers outweigh those from shallow layers. Secondly, as our TED-Adapter includes temporal enhancement (TE) and temporal difference (TD), we conducted separate experiments, revealing that the improvement from TE is more pronounced, but the combination of both yields the best performance.

Ablations for Text-Adapter. We progressively added a varying number of Text-Adapters from deep to shallow. We show the performance changes in Tab. 6, along with the corresponding zero-shot transfer performance on UCF101. It can be observed that as the number of Text-Adapters increases, the model's performance first improves and then starts to decline slightly, with the best performance achieved when adding just one Text-Adapter. The zero-shot results exhibit a similar trend of variation. We believe the reason is that the text data contains only label information, and having too many Text-Adapters may lead the model to overfit on these labels, ultimately affecting overall performance and generalization. Therefore, our final model includes one Text-Adapter, balancing performance and generalization.

Ablations for Multi-Task Decoder. In Tab. 7, we evaluate the impact of the individual heads in the decoder. It is evident that each head in our model contributes positively to the results. By incorporating CMC on top of the original CLIP's contrastive learning, we achieve additional 0.7% improvements in performance, and notably, this enhancement is parameter-free. Furthermore, the inclusion of CMLM further boosted the results. Lastly, the addition of the VC head elevated the performance to 83.4%. This analysis validate multi-task decoder's effectiveness in enhancing the multimodal framework's overall learning.

Conclusion

In this paper, we introduced a novel multimodal, multi-task adapting approach that addresses the challenging task of transferring a large vision-language model, CLIP, to the domain of video action recognition. Our core innovation lies in integrating multimodal adapters and a multi-task decoder into the multimodal framework. Comprehensive experiments on various datasets showcase our method's remarkable zero-shot performance while maintaining promising supervised results with few tunable parameters.

Acknowledgments

This work is funded by Collective Intelligence & Collaboration Laboratory (Open Fund Project No.QXZ23012301).

References

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *ICCV*, 6836–6846.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? In *ICML*, 813–824.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. Ieee.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *ICCV*, 6202–6211.

Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. 2017. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 5842–5850.

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.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. *CoRR*, abs/2102.05918.

Jiang, B.; Wang, M.; Gan, W.; Wu, W.; and Yan, J. 2019. Stm: Spatiotemporal and motion encoding for action recognition. In *ICCV*, 2000–2009.

Ju, C.; Han, T.; Zheng, K.; Zhang, Y.; and Xie, W. 2022. Prompting Visual-Language Models for Efficient Video Understanding. In *ECCV*. Springer.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19113–19122.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. HMDB: a large video database for human motion recognition. In *ICCV*, 2556–2563.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Wang, L.; and Qiao, Y. 2022a. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*.

Li, Y.; Ji, B.; Shi, X.; Zhang, J.; Kang, B.; and Wang, L. 2020. Tea: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 909–918.

Li, Y.; Wu, C.-Y.; Fan, H.; Mangalam, K.; Xiong, B.; Malik, J.; and Feichtenhofer, C. 2022b. Improved multiscale vision transformers for classification and detection. In *CVPR*.

Lin, C.-C.; Lin, K.; Wang, L.; Liu, Z.; and Li, L. 2022a. Cross-modal representation learning for zero-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19978–19988.

Lin, J.; Gan, C.; and Han, S. 2019. Tsm: Temporal shift module for efficient video understanding. In *ICCV*.

Lin, Z.; Geng, S.; Zhang, R.; Gao, P.; de Melo, G.; Wang, X.; Dai, J.; Qiao, Y.; and Li, H. 2022b. Frozen CLIP Models are Efficient Video Learners. *arXiv preprint arXiv:2208.03550*.

Liu, R.; Huang, J.; Li, G.; Feng, J.; Wu, X.; and Li, T. H. 2023. Revisiting temporal modeling for clip-based image-to-video knowledge transferring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6555–6564.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022. Video swin transformer. In *CVPR*.

Ni, B.; Peng, H.; Chen, M.; Zhang, S.; Meng, G.; Fu, J.; Xiang, S.; and Ling, H. 2022. Expanding Language-Image Pretrained Models for General Video Recognition. In *ECCV*.

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.

Park, J.; Lee, J.; Sohn, K.; and xxx. 2023. Dual-path Adaptation from Image to Video Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2203–2213.

Patrick, M.; Campbell, D.; Asano, Y.; Misra, I.; Metze, F.; Feichtenhofer, C.; Vedaldi, A.; and Henriques, J. F. 2021. Keeping your eye on the ball: Trajectory attention in video transformers. In *NeurIPS*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *ICML*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Tu, S.; Dai, Q.; Wu, Z.; Cheng, Z.-Q.; Hu, H.; and Jiang, Y.-G. 2023. Implicit temporal modeling with learnable alignment for video recognition. *arXiv preprint arXiv:2304.10465*.

Wang, M.; Xing, J.; Mei, J.; Liu, Y.; and Jiang, Y. 2023. ActionCLIP: Adapting Language-Image Pretrained Models for Video Action Recognition. *IEEE Transactions on Neural Networks and Learning Systems*.

Wang, M.; Xing, J.; Su, J.; Chen, J.; and Liu, Y. 2022. Learning spatiotemporal and motion features in a unified 2d network for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3347–3362.

Wasim, S. T.; Naseer, M.; Khan, S.; Khan, F. S.; and Shah, M. 2023. Vita-CLIP: Video and text adaptive CLIP via Multimodal Prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23034–23044.

Wu, W.; Wang, X.; Luo, H.; Wang, J.; Yang, Y.; and Ouyang, W. 2023. 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*, 6620–6630.

Xing, J.; Wang, M.; Hou, X.; Dai, G.; Wang, J.; and Liu, Y. 2023. Multimodal Adaptation of CLIP for Few-Shot Action Recognition. *arXiv preprint arXiv:2308.01532*.

Yang, T.; Zhu, Y.; Xie, Y.; Zhang, A.; Chen, C.; and Li, M. 2023. Aim: Adapting image models for efficient video action recognition. *arXiv preprint arXiv:2302.03024*.

Yuan, L.; Chen, D.; Chen, Y.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; Liu, C.; Liu, M.; Liu, Z.; Lu, Y.; Shi, Y.; Wang, L.; Wang, J.; Xiao, B.; Xiao, Z.; Yang, J.; Zeng, M.; Zhou, L.; and Zhang, P. 2021. Florence: A New Foundation Model for Computer Vision. *CoRR*, abs/2111.11432.

Zhao, Y.; Luo, C.; Tang, C.; Chen, D.; Codella, N.; and Zha, Z.-J. 2023. Streaming Video Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14602–14612.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional Prompt Learning for Vision-Language Models. In *CVPR*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to Prompt for Vision-Language Models. *IJCV*.