TemAPT: Efficient Temporal Adaptation of Vision-Language Models for Video Recognition

Mengyu Yang[1] , Ye Tian[1*], Jianwei Li[2], Lanshan Zhang[1], JiaKai Wu[1], YaYa Wei[3], Ziyan Zhong[3], Wendong Wang[1]
[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]Baidu Inc., Beijing, China
[3]Omni-Channel Operation Center China Telecommunications Corporation, Beijing, China

***Abstract*—Video recognition aims to interpret dynamic visual content, a task increasingly addressed by transferring knowledge from large-scale pre-trained Vision-Language Models (VLMs). However, existing parameter-efficient fine-tuning (PEFT) methods often face a trade-off: they either ignore complex temporal dynamics to maintain efficiency or incur high computational costs through heavy temporal modules. To bridge this gap, we propose TemAPT, a novel framework that enables robust temporal modeling of vision-language models with minimal overhead. Unlike methods relying on expensive attention mechanisms, TemAPT introduces a token-level temporal shift strategy via zero-cost channel re-indexing. This allows features to exchange information across frames and progressively enlarges the effective temporal receptive field throughout the encoder layers. Furthermore, to improve cross-modal alignment, we devise a semantic-conditioned adaptive prompt generator that dynamically aligns video representations with class-specific textual semantics. Extensive experiments on three widely used benchmarks demonstrate that TemAPT achieves a superior balance between accuracy and efficiency, outperforming state-of-the-art adaptation methods.**

***Index Terms*—Temporal Shift, Prompt Generation, Adaptor Tuning**

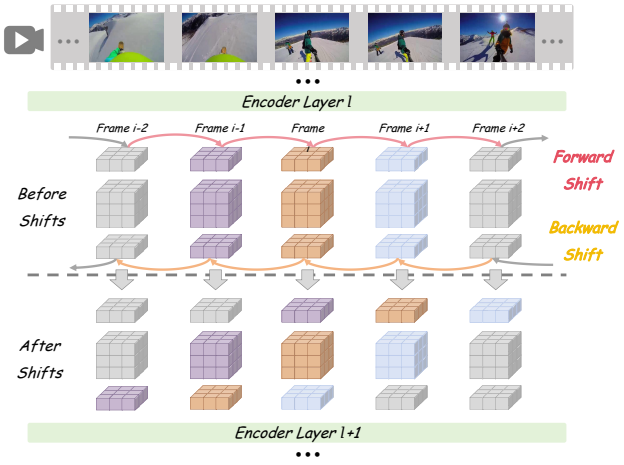

Fig. 1. **Token-level Temporal Shift Mechanism.** TemAPT achieves efficient temporal modeling via a zero-cost token-level temporal shift.

I. Introduction

The ability to perceive dynamic scenes and interpret complex temporal patterns is fundamental to human interaction with the visual world. To endow machines with similar capabilities, video understanding has become a pivotal research area in computer vision [1], [2]. Unlike static image analysis, video action recognition requires not only identifying spatial semantics within individual frames but also modeling the temporal dependencies across frames to capture motion patterns and event transitions. However, achieving robust spatiotemporal reasoning efficiently remains a formidable challenge, particularly for real-time applications where both accuracy and latency are critical constraints.

Recent advances in large-scale pre-training [3]–[5] have revolutionized visual representation learning. Vision-Language Models (VLMs), such as CLIP [3], have demonstrated remarkable generalization capabilities by aligning visual and textual semantics. However, directly adapting large pre-trained backbones to videos faces two practical obstacles. First, full fine-tuning is expensive and may degrade the transferable knowledge learned during pre-training [6]–[8], which is undesirable for data-limited video benchmarks and real-time deployment. Second, many temporal modeling strategies developed for video backbones become prohibitively costly when applied to large VLMs [9]–[11], especially those relying on heavy cross-frame attention or complex temporal modules.

To reduce training cost, parameter-efficient fine-tuning (PEFT) has emerged as a promising paradigm [1], [2], [6], [8]–[10], [12]–[14], which aims to adapt frozen image-based models to video domains by introducing a minimal number of learnable parameters. Early attempts often treated videos as unordered bags of frames or relied on simple temporal pooling, failing to capture complex motion dynamics. To address this, some methods introduced heavy MLP or 3D convolution adaptors to explicitly model motion information [10], [12], which significantly increase the computational overhead, rendering them ill-suited for resource-constrained or real-time applications. Other approaches adopt vision prompts or sparse spatiotemporal attention mechanisms [1], [2], [6], [14], to model temporal dynamics in a lightweight fashion, avoiding extra computation dedicated to temporal modeling. While relatively lightweight, these methods are inevitably prone to failure when the recognition relies heavily on frame-

* Corresponding author.
This work was supported in part by the National Natural Science Foundation of China under Grant 62502014, and in part by BUPT Excellent Ph. D. Students Foundation under Grant CX20242004.

to-frame interactions, as they tend to focus on spatial semantics while struggling to exchange information across the temporal dimension effectively.

To bridge the gap between high efficiency and effective temporal modeling, we propose TemAPT, a lightweight temporal adaptation method that adapts image-pretrained vision-language models for efficient video action recognition. As shown in Figure 1, TemAPT introduces token-level temporal shift with lightweight adaptors in a frozen encoder, enabling cross-frame interaction via zero-cost channel re-indexing while progressively enlarging the temporal receptive field through layered local exchanges. As the layer goes deeper, tokens propagate and aggregate cues over longer time spans through repeated local exchanges, yielding a progressively enlarged effective temporal receptive field. It naturally supports long-range temporal modeling, while maintaining near-constant computational overhead per layer. Then, a class-conditioned multi-frame aggregator guided by the global semantic summary further dynamically weights frame-level embeddings and forms a robust video-level representation. What's more, different from existing static learnable prompts, we further introduce an adaptive prompt generation mechanism that is conditioned on class semantics. During training, we freeze the pre-trained parameters, and only optimize the newly introduced visual adaptors and the adaptive prompt generator.

Our main contributions are as follows:

- We propose TemAPT, a simple yet effective framework that adapts image-pretrained models to video recognition, which effectively balances high computational efficiency with robust temporal modeling.
- We introduce a zero-cost token-level temporal shift mechanism that injects explicit cross-frame interaction without additional FLOPs.
- Extensive experiments on widely used benchmarks demonstrate the competitive performance of our TemAPT over the state-of-the-art methods.

II. Related Work

A. Large-scale Pre-trained Vision-Language Models

Large-scale vision-language models (VLMs) pre-trained on massive image–text pairs have become a cornerstone for transferable visual representation learning [3]–[5]. Early multimodal models [15], [16] typically relied on region features from object detectors and modeled cross-modal interactions. While they demonstrated initial success, their dependence on region extraction pipelines and limited pretraining scale constrained generalization. More recent contrastive pretraining frameworks, represented by CLIP [3] and ALIGN [4], learn aligned visual and textual embeddings in an end-to-end manner, significantly improving zero-shot transfer and semantic robustness. In parallel, VLMs [17], [18] have expanded capability toward instruction following and open-ended reasoning, but their original designs are largely image-centric and do not explicitly model temporal dynamics. Therefore, how to efficiently transfer strong image-based VLM representations to video understanding, without incurring prohibitive training/inference cost, remains a central challenge.

B. Parameter-efficient Video Adaptation for VLMs

As pre-trained models continue to scale up [3], [4], full fine-tuning becomes increasingly costly and can risk overwriting transferable knowledge, especially when downstream video datasets are limited [6], [19]. Parameter-efficient fine-tuning (PEFT) offers an attractive alternative by updating only a small number of newly introduced parameters while keeping the backbone largely frozen [7], [8], [20]. Recent works extend PEFT to video by adapting frozen VLMs [2], [9], [10], [12], [14], introducing modules like ST-Adapter [12] or video-specific prompting strategies [1], [2], [14]. However, a key limitation of existing PEFT-based video adaptations is their weak temporal inductive bias, relying on frame-wise adaptation with simple pooling or shallow temporal modeling, leading to a trade-off between temporal capacity and efficiency. These limitations motivate our work to inject explicit temporal interaction into the pre-trained visual encoder while maintaining efficiency.

III. Method

A. Overview

Figure 2 presents an overview of TemAPT, a parameter-efficient framework that adapts an image-pretrained vision-language model to video recognition with strong temporal modeling capability and minimal overhead. Note that the techniques introduced in this paper are versatile and can also be applied to other prompt methods. In this paper, we leverage CLIP [3] as a representative framework for its simplicity and generality. TemAPT consists of three key components: (i) a temporally shifted vision encoder, (ii) a semantic-conditioned adaptive prompt generator, and (iii) a class-conditioned multi-frame aggregator.

Given an input video $V = \{I_t\}_{t=1}^{T}$ and class names $\mathcal{C} = \{c_k\}_{k=1}^{K}$, the vision encoder E_v extracts frame representations $\{\mathbf{v}_t\}$ with the token-level temporal shift adaptation. Temporal context is progressively propagated into each $\mathbf{v}_t$ without expensive cross-frame attention. In parallel, the class-conditioned prompt generator estimates the condition semantics C_s and produces adaptive prompts p_k for each class. Then the frozen text encoder E_t outputs the corresponding text embedding $\mathbf{t}_k$. Guided by C_s, we aggregate $\{\mathbf{v}_t\}$ into a video-level representation and compute cosine similarity with $\mathbf{t}_k$ to obtain classification scores. During fine-tuning, all pre-trained parameters are frozen, and only the added adaptation components are updated, enabling efficient transfer.

B. Token-level Temporal Shift Mechanism

Since the pre-trained image encoder is originally designed for image understanding, it inherently possesses strong spatial modeling capabilities, but lacks explicit temporal modeling capacity. To remedy this, we introduce a token-level temporal shift mechanism within the pre-trained image encoder, to effectively endow it with the ability to process video.

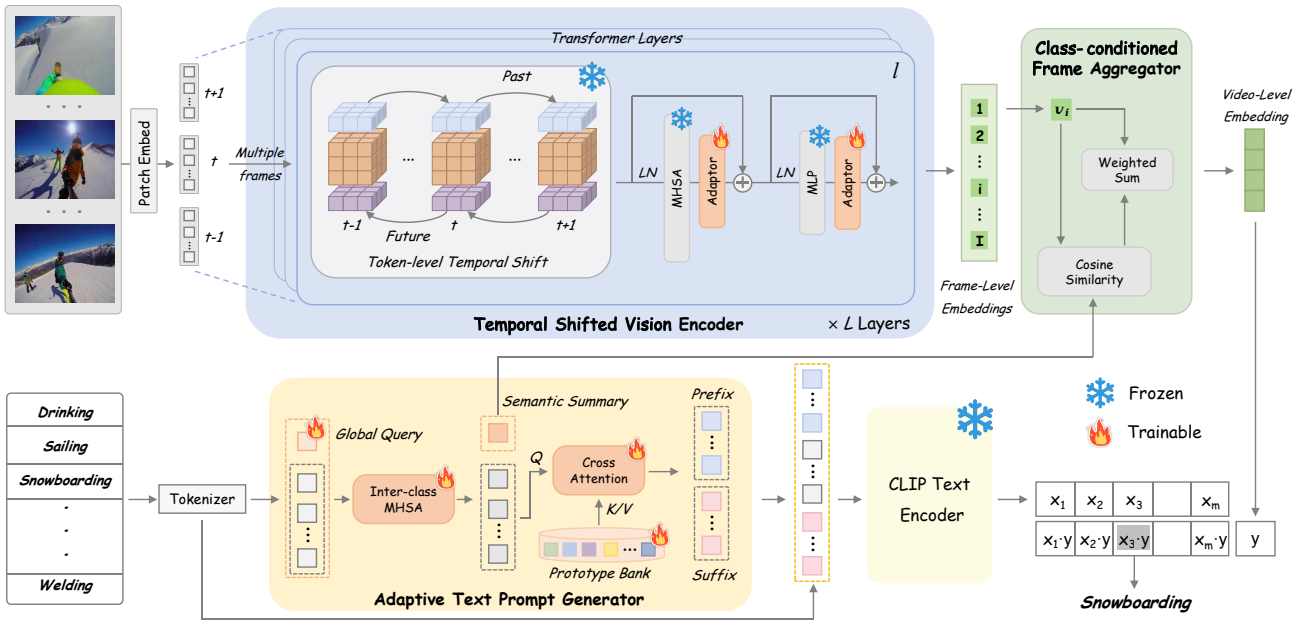

Fig. 2. **Overview of our proposed TemAPT**, which mainly consists of three components, namely the temporal shifted vision encoder, class-conditioned frame aggregator, and adaptive text prompt generator. During training, the pre-trained parameters are frozen, and only the newly introduced lightweight adaptation modules are optimized. Color version recommended for better visualization.

As shown in Figure 2, the token-level temporal shift mechanism injects temporal context through two key components: (i) parameter-free token-level temporal shift units that exchange partial feature channels across frames, and (ii) lightweight residual *adaptors* that selectively refine the resulting spatio-temporal mixed tokens. By repeating temporal exchanges and refining across multiple layers, tokens progressively propagate information over longer time spans as shown in the Figure 3, yielding an enlarged effective temporal receptive field as depth increases.

Given an input video $V = \{I_t\}_{t=1}^{T}$, each frame I_t is first tokenized into a sequence of $(N+1)$ tokens (with a CLS token) with feature dimension D. The token sequences from multiple frames are then jointly processed layer by layer through the encoder. At the l-th layer of the encoder, the input activations $\mathbf{O}^{(l-1)}$ can be denoted as $\{\mathbf{O}_t^{(l-1)}\}_{t=1}^{T} \in \mathbb{R}^{T\times(N+1)\times D}$. At first, the temporal shift operator $S(\cdot)$ re-indexes a subset of channels of $\mathbf{O}^{(l-1)}$ along the temporal axis t to get the input embedding $\mathbf{X}^{(l)}$ of the l-th layer as:

$$\mathbf{X}^{(l)} = \left[\mathbf{X}_1^{(l)},\ \mathbf{X}_2^{(l)}, ..., \mathbf{X}_T^{(l)},\right] = S(\{\mathbf{O}_t^{(l-1)}\}_{t=1}^{T}). \quad (1)$$

Concretely, we split the activations into three groups with shifted scale factor γ, which are $D\cdot\gamma$ forward-shift channels, $D\cdot\gamma$ backward-shift channels, and $D-2D\cdot\gamma$ remaining-unshifted channels as:

$$\mathbf{X}_t^{(l)} = \left[(\mathbf{O}_{t-1}^{(l-1)})_{\gamma\cdot D}^{\rightarrow},\ (\mathbf{O}_t^{(l-1)})_{D-2\cdot\gamma\cdot D}^{=},\ (\mathbf{O}_{t+1}^{(l-1)})_{\gamma\cdot D}^{\leftarrow},\right], \quad (2)$$

where $(\cdot)_{\cdot}^{\rightarrow}$ and $(\cdot)_{\cdot}^{\leftarrow}$ denote the shifted channel subsets, and $(\cdot)_{\cdot}^{=}$ denotes the remaining unshifted channels. $(\mathbf{O}_{t-1}^{(l-1)})_{\gamma\cdot D}^{\rightarrow}$ comprises the first $\gamma\cdot D$ channels providing past context, and $(\mathbf{O}_{t+1}^{(l-1)})_{D\cdot\gamma}^{\leftarrow}$ comprises last $\gamma\cdot D$ channels providing future context. The remaining $2\cdot\gamma\cdot D$ channels remain unchanged. For the temporal boundary frames where $t+1$ or $t-1$ do not exist, we apply zero-padding to the shifted channels. This operation introduces temporal context through local exchanges without constructing explicit pairwise attention between frames.

What's more, to effectively capture the temporal dynamics of video, we further introduce lightweight adaptors, which guide the pre-trained model to adapt to the temporally mixed sequences to explicitly model motion cues. As shown in Figure 2, we incorporate sequential lightweight adaptors within all layers of the pre-trained encoder to generate the temporal embedding $\mathbf{O}_t^{(l)}$ of frame t. Specifically, each block contains two lightweight adaptors. One is placed after the multi-head self-attention (MHSA), and the other after the MLP. For frame t, the adaptation block of l-th layer can be expressed as follows:

$$\begin{aligned}\hat{\mathbf{X}}_t^{(l)} &= \mathbf{X}_t^{(l)} + \text{adaptor}(\text{MHSA}(\text{LN}(\mathbf{X}_t^{(l)}))),\\ \mathbf{O}_t^{(l)} &= \hat{\mathbf{X}}_t^{(l)} + \text{adaptor}(\text{MLP}(\text{LN}(\hat{\mathbf{X}}_t^{(l)}))),\end{aligned} \quad (3)$$

where $\mathbf{X}_t^{(l)}$ represents the input activation of frame t for the l-th layer.

In all layers, the adaptors adopt a bottleneck structure to minimize the extra parameter overhead, consisting of two linear transformations for dimension up and down from D to r. To further enhance the adaptation of the temporal token shifting, we incorporate a non-linear activation $\sigma(\cdot)$ and a depth-wise 1D convolution layer DWConv1d$(\cdot)$ within the bottleneck. Formally, at the l-th layer, given the activation

$\mathbf{X}^{(l)} \in \mathbb{R}^{T\cdot(N+1)\cdot D}$, the adaptor computes:

$$\tilde{\mathbf{X}^{(l)}} = \sigma(\text{DWConv1d}(\text{LN}(\mathbf{X}^{(l)}) \cdot \mathbf{W}_{\text{down}})) \cdot \mathbf{W}_{\text{up}}, \quad (4)$$

where $\mathbf{W}_{\text{down}} \in \mathbb{R}^{D\times r}$ and $\mathbf{W}_{\text{up}} \in \mathbb{R}^{r\times D}$ are learnable projectors. $\text{LN}(\cdot)$ denotes layer normalization. After all L encoder layers, we take the temporally enhanced CLS token of each frame as the frame representation:

$$\mathbf{v}_t = \mathbf{O}_t^{(L)}[0] \in \mathbb{R}^D, \quad (5)$$

yielding a set of temporally-aware frame embeddings $\{\mathbf{v}_t\}_{t=1}^T$.

C. Adaptive Text Prompt Generation

Existing prompt-based adaptation methods typically learn static prompt tokens shared across categories, which can be inadequate for fine-grained recognition with subtle, diverse class semantics. To address this, we propose a global-aware adaptive prompt generator that models inter-class relationships and derives a holistic semantic condition to guide both prompt generation and video aggregation.

To provide conditions for the subsequent prompt generation and frame aggregation, we first evaluate the global semantics among all classes. Specifically, given the embeddings of all K classes $\{\mathbf{e}_k\}_{k=1}^K$, we introduce a learnable global query token $\mathbf{q}_g$ and concat it with the classes sequence as:

$$\mathbf{H}^{(0)} = \{\mathbf{q}_g; \mathbf{e}_1; ...; \mathbf{e}_K\}. \quad (6)$$

We first perform multi-head self-attention over the class set to explicitly model inter-class relationships as:

$$\mathbf{H}^{(1)} = \text{MHSA}(\text{LN}(\mathbf{H}^{(0)})). \quad (7)$$

The updated global token encodes holistic semantics across classes as:

$$\mathbf{H}_g = \mathbf{H}^{(1)}[0], \quad (8)$$

while each class token $\mathbf{H}_{1:K}^{(1)}$ integrates contextual information from other classes. Intuitively, $\mathbf{H}_g^{(1)}$ serves as a compact global semantic summary, which provides a principled basis for subsequent multi-frame aggregation on the visual side.

Subsequently, we refine the class-specific conditions via a cross-attention mechanism utilizing a learnable semantic prototype bank. We maintain a set of learnable vectors $\mathbf{M} = \{\mathbf{m}_j\}_{j=1}^J$, which serves as a compact semantic memory. To retrieve complementary semantic factors, the contextualized class tokens interact with this bank. During generation, each contextualized class token queries this bank to retrieve complementary semantic factors. Specifically, we compute the updated feature representation $\mathbf{H}^{(2)}$ with cross-attention between the input class tokens $\mathbf{H}_{1:K}^{(1)}$ and the prototype memory $\mathbf{M}$:

$$\mathbf{H}^{(2)} = \text{softmax}\left(\frac{(\mathbf{H}_{1:K}^{(1)}\mathbf{W}_Q)(\mathbf{M}\mathbf{W}_K)^\top}{\sqrt{d}}\right)(\mathbf{M}\mathbf{W}_V), \quad (9)$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ denote the learnable linear projections, and $\sqrt{d}$ is the scaling factor. The output $\mathbf{H}^{(2)}$ yields the final class-conditioned semantics $\mathbf{c}_k \in \mathbb{R}^D$ for each category:

$$\mathbf{c}_k = \mathbf{H}^{(2)}[k], \quad k = 1, \ldots, K. \quad (10)$$

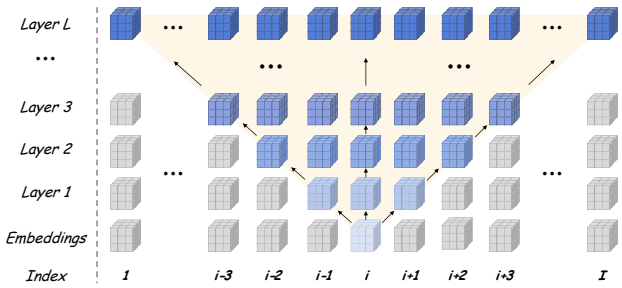

Fig. 3. **Receptive Field Growth Across Layers.** Best viewed in color.

Finally, we generate the adaptive prompt tokens associated with class k through a lightweight MLP:

$$\mathbf{p}_k = \text{MLP}(\mathbf{c}_k), \quad (11)$$

The adaptive prompts $\mathbf{p}_k$ are then inserted in front of the token sequence of class k, which is then fed into the frozen text encoder E_t to produce the final textual prototype $\mathbf{t}_k$. Compared with static learnable prompts shared across classes, our adaptive prompt generator produces class-aware prompts that better capture subtle semantic differences, thus improving discrimination in fine-grained settings.

D. Class-conditioned Multi-frame Aggregation

Standard video-language models typically utilize mean pooling to aggregate frame-level features into a video-level representation. However, this approach treats all frames equally, neglecting the fact that discriminative visual cues are often sparse and embedded in specific temporal segments. To address this, we propose a class-conditioned aggregation strategy that leverages the global semantic summary $\mathbf{H}_g$ (derived in Section III-C) to dynamically weigh the importance of each frame.

Formally, given the sequence of frame embeddings $\{\mathbf{v}_t\}_{t=1}^T$ and the global semantic token $\mathbf{H}_g$, we employ a cross-attention mechanism to compute the relevance of each frame to the target semantic concepts. We first project the global semantic token into a query vector, and the frame embeddings into key vectors:

$$q = \mathbf{H}_g\mathbf{W}_q^a, \quad k_t = \mathbf{v}_t\mathbf{W}_k^a, \quad (12)$$

where $\mathbf{W}_q^a$ and $\mathbf{W}_k^a$ are learnable projection matrices. The temporal attention weights $\boldsymbol{\alpha} = \{\alpha_t\}_{t=1}^T$ are computed via a softmax operation over the scaled dot-product similarities:

$$\alpha_t = \frac{\exp(q \cdot k_t^\top/\sqrt{D})}{\sum_{j=1}^T \exp(q \cdot k_t^\top/\sqrt{D})}. \quad (13)$$

The attention weights α_t explicitly reflect the semantic importance of the t-th frame relative to the global class context. Then the final video-level representation $\mathbf{v}_{video}$ is obtained by the weighted summation of frame embeddings:

$$\mathbf{v}_{video} = \sum_{t=1}^T \alpha_t\mathbf{v}_t + \mathbf{v}_{\text{mean}}, \quad (14)$$

where $\mathbf{v}_{\text{mean}} = \frac{1}{T}\sum \mathbf{v}_t$ is a residual connection to stabilize training.

TABLE I
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON KINETICS-400 [21].

Method & Arch.	Params.	Views	Top-1	Top-5	GFLOPs
ImageNet/JFT-300M/Kinetics Initialization					
MViT-B/16 [22]	37M	$64 \times 3 \times 3$	81.2	95.1	4095
UniFormer-B/16 [11]	50M	$32 \times 4 \times 3$	83.0	95.4	3108
TimeSformer-L/14 [23]	121M	$64 \times 1 \times 3$	80.7	94.7	7140
ViViT-L [24]	311M	$32 \times 1 \times 1$	80.6	92.7	3980
VideoSwin-B/16 [25]	-	$32 \times 4 \times 3$	82.7	95.5	3384
MViTv2-B/16 [26]	51M	$32 \times 1 \times 5$	82.9	95.7	1125
CLIP-400M Initialization (Fully Fine-tuning)					
ActionCLIP-B/16 [6]	142M	$32 \times 10 \times 3$	83.8	97.1	16890
X-CLIP-B/16 [9]	156M	$8 \times 4 \times 3$	83.8	96.7	3444
CLIP-400M Initialization (Parameter-efficient Fine-tuning)					
EVL-B/16 [10]	29M	$8 \times 1 \times 3$	82.9	-	444
VideoPrompt-B/16 [8]	10M	$16 \times 1 \times 5$	76.9	93.5	1435
ST-Adapter-B/16 [12]	11M	$32 \times 3 \times 1$	82.7	96.2	1821
AIM-B/16 [19]	11M	$8 \times 1 \times 3$	83.9	96.3	624
Vita-CLIP-B/16 [27]	39M	$16 \times 4 \times 3$	82.9	96.3	2280
m^2-CLIP-B/16 [14]	16M	$8 \times 4 \times 3$	83.4	96.3	2568
TemAPT ViT-B/16	6.86M	$8 \times 1 \times 3$	83.8	96.3	549
TemAPT ViT-B/16	6.86M	$16 \times 1 \times 3$	84.7	96.9	842

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON SOMETHING-SOMETHING V2 [28].

Method & Arch.	Params.	Views	Top-1	Top-5	GFLOPs
ImageNet/Kinetics Initialization					
TimeSformer-L/14 [23]	121M	$64 \times 1 \times 3$	62.4	-	7140
ViViT-L [24]	311M	$16 \times 4 \times 3$	65.4	89.8	11892
MViT-B/16 [22]	37M	$32 \times 1 \times 3$	67.1	90.8	510
MViTv2-B/16 [26]	51M	$32 \times 1 \times 3$	70.5	92.7	675
UniFormer-B/16 [11]	50M	$32 \times 1 \times 3$	71.2	92.8	777
CLIP-400M Initialization (Fully Fine-tuning)					
ViT-B/16 [3]	86M	$8 \times 1 \times 3$	44.0	77.0	419
ViT-L/14 [3]	303M	$8 \times 1 \times 3$	48.7	77.5	1941
CLIP-400M Initialization (Parameter-efficient Fine-tuning)					
VPT-B/16 [7]	6M	$8 \times 1 \times 3$	36.2	61.1	537
AdaptFormer-B/16 [13]	8M	$8 \times 1 \times 3$	51.3	70.6	544
Pro-tuning-B/16 [30]	9M	$8 \times 1 \times 3$	50.8	69.9	538
EVL-B/16 [10]	29M	$8 \times 1 \times 3$	61.0	-	512
AIM-B/16 [19]	14M	$8 \times 1 \times 3$	66.4	90.5	624
m^2-CLIP-B/16 [14]	16M	$8 \times 3 \times 1$	66.9	90.1	2568
ST-Adapter-B/16 [12]	11M	$32 \times 3 \times 1$	67.1	91.2	1467
TemAPT ViT-B/16	6.86M	$8 \times 1 \times 3$	67.4	91.0	537
TemAPT ViT-B/16	6.86M	$16 \times 1 \times 3$	68.5	91.9	835

IV. EXPERIMENTS

A. Experimental Setup

Datasets. We evaluate TemAPT on three standard video recognition benchmarks: Kinetics-400 (K400) [21], Something-Something V2 (SSV2) [28], and HMDB51 [29]. K400 provides large-scale and diverse data for general video recognition, SSV2 focuses on fine-grained temporal reasoning and subtle motion patterns, and HMDB51, with limited data and high intra-class variance, evaluates data efficiency and transferability.

Implementation Details: We instantiate TemAPT on CLIP ViT-B/16 [3]. During adaptation, the pre-trained image and text encoders are frozen. We train for 200 epochs using AdamW with a cosine learning rate schedule, a peak learning rate of 5×10^{-3}, and weight decay of 3×10^{-4}. We uniformly sample T frames from each video using Decord 0.6.0, resize frames to 256×256, center-crop to 224×224, and apply ImageNet normalization. All experiments are conducted with PyTorch 2.3.1 on four NVIDIA L20 GPUs.

We report Top-1/Top-5 accuracy on Kinetics-400 and Something-Something V2, and Top-1 accuracy on HMDB51, following prior works [6], [9], [10], [12]. Moreover, we report the number of learnable parameters and GFLOPs to measure the efficiency. The computation of class text embeddings (i.e., the text encoder forward for class prompts) is excluded since they are precomputed and cached, ensuring fair efficiency comparison across PEFT methods.

B. Comparisons with the State-of-the-arts

1) Results on Kinetics-400: Table I compares TemAPT with representative fully fine-tuned CLIP-based methods and parameter-efficient adaptations. With only 6.86M learnable parameters, TemAPT achieves 83.8% Top-1 accuracy under the $8 \times 1 \times 3$ setting, matching or surpassing competitive PEFT baselines while maintaining a low computational footprint. Notably, compared with methods that rely on heavier temporal modules or larger adaptation capacity (e.g., EVL and ST-Adapter), TemAPT attains a more favorable accuracy–efficiency balance, validating that our proposed TemAPT can serve as an effective temporal adaptation method for efficient image-pretrained VLMs transfer.

2) Results on Something-Something V2: Something-Something V2 requires fine-grained temporal reasoning where simple frame pooling often fails. As shown in Table II, TemAPT achieves 67.4% Top-1 with 8 frames and further reaches 68.5% when using 16 frames. It outperforms prior PEFT approaches such as VPT [7], AdaptFormer [13], and Pro-tuning [30], while remaining computationally efficient. These results suggest that explicit cross-frame feature exchange introduced by the temporal shift is crucial for motion-centric datasets.

3) Results on HMDB51: We further evaluate generalization on the smaller HMDB51 dataset. Table III shows that TemAPT achieves 74.2% Top-1 accuracy, outperforming CLIP-based

TABLE III
COMPARISON WITH STATE-OF-THE-ARTS ON HMDB51 [29].

Method	Pre-training	Top-1
ImageNet/Kinetics Initialization		
STC [31]	K400	72.6
R(2+1)D-34 [32]	K400	74.5
I3D [33]	ImageNet+K400	74.8
FASTER32 [34]	K400	75.7
SlowOnly-R101 [35]	OmniSource	79.0
CLIP-400M Initialization		
VideoPrompt-B/16 [8]	CLIP-400M	66.4
ST-Adapter-B/16 [12]	CLIP-400M	69.8
EVL-L/14 [10]	CLIP-400M	71.4
TemAPT ViT-B/16	CLIP-400M	74.2

TABLE IV
COMPARISON OF PRACTICAL EFFICIENCY. LATENCY AND THROUGHPUT ARE MEASURED WITH A BATCH SIZE OF 1 ON A SINGLE L20 GPU.

Method	GFLOPS	Latency (ms)	Throughput (V/s)
UniFormer-B/16 [11]	777	296.7	3.37
EVL-B/16 [10]	512	107.3	9.31
TemAPT ViT-B/16	429	66.9	14.94

PEFT baselines such as VideoPrompt [8] and ST-Adapter [12]. This indicates that our TemAPT is effective under limited training data, benefiting from freezing the strong CLIP backbone while learning a compact yet expressive temporal adaptation.

C. Practical Efficiency

Beyond GFLOPs, real deployment performance depends on end-to-end latency and throughput. We therefore measure practical efficiency on a single NVIDIA L20 GPU with batch size 1, reporting average latency per video and throughput (videos/s).

As summarized in Table IV, TemAPT achieves 66.9 ms latency and 14.94 videos/s, outperforming EVL [10] of 107.3 ms, 9.31 videos/s, and significantly faster than heavier video backbones, e.g., UniFormer [11]. This gain primarily stems from the temporal shift, which is parameter-free and introduces no extra attention computation. These results demonstrate that TemAPT is well-suited for real-time and resource-constrained video recognition.

TABLE V
ABLATION STUDY ON MAIN COMPONENTS OF TEMAPT.

Method	Params.	Kinetics-400	GFLOPs
CLIP with Temporal Pooling	0.0M	42.2	152.4
+ Temporal Shift	5.12M	78.6	173.3
+ Text Prompt	6.85M	80.6	181.5
+ Multi-frame Aggregation	6.86M	81.2	183.1

D. Ablation Studies

We conduct ablations on Kinetics-400 using 8 frames and a single spatial view to isolate the effect of each component.

1) Effectiveness of Main Components: Table V shows the incremental contributions of each module. A frozen CLIP with naive temporal pooling achieves only 42.2%, indicating limited video understanding. Token-level temporal shift boosts performance to 78.6%, demonstrating the importance of cross-frame interaction. The adaptive text prompt generator further improves accuracy to 80.6%, enhancing vision–language alignment, while class-conditioned multi-frame aggregation reaches 81.2%, benefiting from semantic-guided frame weighting. Overall, all components contribute consistently with modest overhead.

2) Analysis of the Temporal Shift Mechanism: We analyze the shift direction and the shift scale factor γ in Table VI. First, bidirectional shift performs best at 81.2% on K400 and 65.3% on SSV2, outperforming forward-only and backward-only variants. This suggests that leveraging both past and future context yields more balanced temporal cues. Second, varying γ reveals a clear trade-off: too small a shift limits temporal exchange, while too large a shift harms spatial representation capacity by overmixing channels. Empirically, γ=$1/8$ achieves the best overall performance, and we adopt it as the default.

TABLE VI
ABLATION STUDY ON TEMPORAL SHIFT MECHANISM.

Method	Kinetics-400	SSV2
Effectiveness of Shift Direction		
Forward-shift only	72.6	62.4
Backward-shift only	77.4	64.1
Bidirectional shift	81.2	65.3
Shift Scale Factor γ		
$\gamma = 1/16$	80.3	64.6
$\gamma = 1/8$	81.2	65.3
$\gamma = 1/4$	80.9	64.1
$\gamma = 1/2$	79.6	62.9

TABLE VII
ABLATION STUDY ON THE TEXT PROMPT GENERATION.

Method	Kinetics-400	GFLOPs
Class Modeling Method		
AvgPool	80.3	175.6
MLP	80.7	179.5
1D-Conv	80.9	179.9
Self-attention	81.2	183.1
Effectiveness of Semantic Bank		
w/o Bank	80.7	181.8
w Bank	81.2	183.1

3) Ablation on Adaptive Prompt Generation: Table VII investigates design choices of class modeling and semantic memory. Among different class modeling methods, self-attention achieves the strongest performance of 81.2%, indicating that explicitly modeling inter-class relationships is beneficial for producing discriminative prompts. Moreover, adding the semantic prototype bank provides a consistent gain (80.7% $\rightarrow$ 81.2%), suggesting that retrieving complementary semantic factors helps enrich class-conditioned representations beyond a single deterministic mapping.

TABLE VIII
ABLATION STUDY ON THE MULTI-FRAME AGGREGATION.

Method	Kinetics-400	GFLOPs
Learnable Weights	80.7	181.5
Self-attention	80.9	181.7
Learnable Query	80.7	182.9
Class-condition	81.2	183.1

4) Ablation on Multi-frame Aggregation: We compare multiple aggregation strategies in Table VIII. Simple learnable

weights and generic self-attention offer limited improvements. In contrast, our class-conditioned aggregation achieves the best performance of 81.2% with negligible extra compute. This validates our hypothesis that semantic-guided frame weighting can better focus on temporally sparse cues, leading to a more robust video representation.

V. Conclusion and Future Work

In this paper, we propose TemAPT, a parameter-efficient temporal adaptation framework for zero-cost temporal inductive bias injection into frozen VLMs, suggesting that lightweight structural re-indexing can substitute heavy temporal attention. Experiments on three benchmarks show that TemAPT achieves a strong accuracy–efficiency trade-off with favorable practical latency. A limitation is that shift-based local exchanges may be suboptimal for very long-range temporal reasoning, and our current evaluation focuses on trimmed action classification. Future work includes improving long-range token mixing, extending TemAPT to long-form videos and dense video understanding tasks (e.g., localization), and exploring broader video-language applications.

References

[1] X. Gao, Z. Chang, D. Kong, H. Zhou, and Y. Lu, "Mip-clip: Multimodal independent prompt clip for action recognition," *IEEE Transactions on Multimedia*, 2025.

[2] J. Wang, R. Zhao, R. Xiong, X. Wang, X. Fan, and T. Huang, "Sample: Semantic alignment through temporal-adaptive multimodal prompt learning for event-based open-vocabulary action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 14 409–14 419.

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

[4] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.

[5] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.

[6] M. Wang, J. Xing, J. Mei, Y. Liu, and Y. Jiang, "Actionclip: Adapting language-image pretrained models for video action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[7] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European conference on computer vision*. Springer, 2022, pp. 709–727.

[8] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, "Prompting visual-language models for efficient video understanding," in *European Conference on Computer Vision*. Springer, 2022, pp. 105–124.

[9] B. Ni, H. Peng, M. Chen, S. Zhang, G. Meng, J. Fu, S. Xiang, and H. Ling, "Expanding language-image pretrained models for general video recognition," in *European conference on computer vision*. Springer, 2022, pp. 1–18.

[10] Z. Lin, S. Geng, R. Zhang, P. Gao, G. De Melo, X. Wang, J. Dai, Y. Qiao, and H. Li, "Frozen clip models are efficient video learners," in *European Conference on Computer Vision*. Springer, 2022, pp. 388–404.

[11] K. Li, Y. Wang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "Uniformer: Unified transformer for efficient spatiotemporal representation learning," *arXiv preprint arXiv:2201.04676*, 2022.

[12] J. Pan, Z. Lin, X. Zhu, J. Shao, and H. Li, "St-adapter: Parameter-efficient image-to-video transfer learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 26 462–26 477, 2022.

[13] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "Adaptformer: Adapting vision transformers for scalable visual recognition," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 664–16 678, 2022.

[14] M. Wang, J. Xing, B. Jiang, J. Chen, J. Mei, X. Zuo, G. Dai, J. Wang, and Y. Liu, "A multimodal, multi-task adapting framework for video action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5517–5525.

[15] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[16] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[17] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[18] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

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

[20] H. Bahng, A. Jahanian, S. Sankaranarayanan, and P. Isola, "Exploring visual prompts for adapting large-scale models," *arXiv preprint arXiv:2203.17274*, 2022.

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

[22] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *ICCV*, 2021, pp. 6824–6835.

[23] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *ICML*, vol. 2, no. 3, 2021, p. 4.

[24] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *ICCV*, 2021, pp. 6836–6846.

[25] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *CVPR*, 2022, pp. 3202–3211.

[26] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "Mvitv2: Improved multiscale vision transformers for classification and detection," in *CVPR*, 2022, pp. 4804–4814.

[27] S. T. Wasim, M. Naseer, S. Khan, F. S. Khan, and M. Shah, "Vita-clip: Video and text adaptive clip via multimodal prompting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 034–23 044.

[28] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag *et al.*, "The" something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5842–5850.

[29] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563.

[30] X. Nie, B. Ni, J. Chang, G. Meng, C. Huo, S. Xiang, and Q. Tian, "Pro-tuning: Unified prompt tuning for vision tasks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 6, pp. 4653–4667, 2023.

[31] A. Diba, M. Fayyaz, V. Sharma, M. M. Arzani, R. Yousefzadeh, J. Gall, and L. Van Gool, "Spatio-temporal channel correlation networks for action classification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 284–299.

[32] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *CVPR*, 2018, pp. 6450–6459.

[33] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[34] L. Zhu, D. Tran, L. Sevilla-Lara, Y. Yang, M. Feiszli, and H. Wang, "Faster recurrent networks for efficient video classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 098–13 105.

[35] H. Duan, Y. Zhao, Y. Xiong, W. Liu, and D. Lin, "Omni-sourced webly-supervised learning for video recognition," in *European conference on computer vision*. Springer, 2020, pp. 670–688.