Reliable and Diverse Hierarchical Adapter for Zero-shot Video Classification

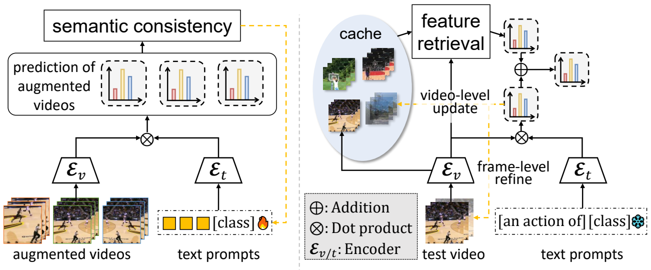
Wenxuan Ge, **Peng Huang**, **Rui Yan**, **Hongyu Qu**, **Guosen Xie** and **Xiangbo Shu**
Nanjing University of Science and Technology
{gwx, penghuang, ruiyan, quhongyu, shuxb}@njust.edu.cn, gsxiehm@gmail.com

Abstract

Adapting pre-trained vision-language models to downstream tasks has emerged as a novel paradigm for zero-shot learning. Existing test-time adaptation (TTA) methods such as TPT attempt to fine-tune visual or textual representations to accommodate downstream tasks but still require expensive optimization costs. To this end, Training-free Dynamic Adapter (TDA) maintains a cache containing visual features for each category in a parameter-free manner and measures sample confidence based on prediction entropy of test samples. Inspired by TDA, this work aims to develop the first training-free adapter for zero-shot video classification. Capturing the intrinsic temporal relationships within video data to construct and maintain the video cache is key to extending TDA to the video domain. In this work, we propose a reliable and diverse Hierarchical Adapter for zero-shot video classification, which consists of Frame-level Cache Refiner and Video-level Cache Updater. Before each video sample enters the corresponding cache, it needs to be refined at frame level based on prediction entropy and temporal probability difference. Due to the limited capacity of the cache, we update the cache during inference based on the principle of diversity. Experiments on four popular video classification benchmarks demonstrate the effectiveness of Hierarchical Adapter. The code is available at https://github.com/Gwxer/Hierarchical-Adapter.

1 Introduction

Large amounts of labeled data, such as K400 [Kay *et al.*, 2017] and K600 [Carreira *et al.*, 2018], are typically required in deep learning based video classification tasks for training. However, annotating sufficient samples is time-consuming and resource-intensive, which poses significant challenges for practical applications. In recent years, many works [Rasheed *et al.*, 2023; Wang *et al.*, 2023; Ju *et al.*, 2022] focus on leveraging zero-shot learning techniques to transfer the learned knowledge to predict novel categories. They have attempted to tune an off-the-shelf visual encoder, originally developed for the image domain, on large-scale video datasets for zero-shot video classification. Though these works have improved the model architecture to enhance motion semantic representations, they still suffer from the issue of data shift, which degrades the classification performance and is a challenge commonly encountered in real-world scenarios.



(a) Test-time Prompt Tuning (Training-required)

(b) Hierarchical Adapter (Training-free)

Figure 1: Motivation of this work. Existing training-required Test-time Prompt Tuning optimizes the learnable prompt via semantic consistency loss. To get rid of the computational burden of gradient descent, we construct and update a video cache in a training-free manner, and adjust video-text inter-modal similarity with intra-modal similarity between the test video and the cache.

Test-time adaptation (TTA) offers an alternative approach for zero-shot transfer, which can effectively address the domain shift problem. Existing TTA methods can be divided into two types, *i.e.*, training-required and training-free. Training-required TTA approaches [Feng *et al.*, 2023; Yan *et al.*, 2024; Yan *et al.*, 2025; Qu *et al.*, 2025a], represented by Test-time Prompt Tuning (TPT) [Shu *et al.*, 2022a], fine-tune visual or textual representations via an unsupervised semantic consistency loss. Specifically, TPT fine-tunes the context of the prompt by constraining the consistency of augmented samples, which helps precisely retrieve the knowledge of Vision-Language Models (VLMs). Recent works have attempted to improve TPT, such as enhancing the diversity of augmented samples through generative models [Feng *et al.*, 2023] and enriching text descriptions based on large language models [Yan *et al.*, 2024].

However, training-required approaches involve significant computational overhead during inference, which hinders their practical applications in computation-limited downstream tasks. Training-free TTA methods [Udandarao *et al.*, 2023;

Karmanov *et al.*, 2024; Zhang *et al.*, 2024b] typically use support set or cache to store prototypes for each category of the target domain, and then use these prototypes to adjust the predictions of the VLM, bridging the gap between the source domain and the target domain. The support set or the cache can be derived from generative models [Udandarao *et al.*, 2023], historical test samples [Karmanov *et al.*, 2024], or boosting samples [Zhang *et al.*, 2024b].

Although these methods have made progress in the image domain, directly applying them to zero-shot video classification would encounter the following issues. i) **How to represent a multi-frame video**: Simply averaging multi-frame representations tends to introduce substantial noise, thereby compromising representation quality. Extracting only a single key frame, on the other hand, leads to the attenuation or loss of crucial temporal cues. ii) **How to represent a complicated action**: Although entropy-based cache update algorithms can select high-confidence samples, these samples tend to exhibit high visual similarity, making it difficult to construct a semantically diverse action feature bank.

To this end, we propose a reliable and diverse Hierarchical Adapter, which is the first training-free adapter for zero-shot video classification. This framework consists of two core modules. **Frame-level Cache Refiner (FCR)**: To capture rich and effective motion features, we design a comprehensive frame selection strategy with two metrics: prediction entropy and temporal probability difference. Prediction entropy independently measures the model's confidence in the prediction results for each frame, while temporal probability difference assesses the model's sensitivity to temporal discriminative cues. We also propose a two-step top-k approach to effectively integrate them. **Video-level Cache Updater (VCU)**: To represent the underlying data pattern, we introduce a diversity criterion to improve the cache update algorithm.

To summarize, we make the following contributions:

- We propose a criterion for selecting reliable and diverse video data that effectively captures the underlying data manifold, which is conductive to enhancing cache representations.
- We design a hierarchical adapter, a novel adaptation strategy in test-time adaptation of VLMs for video classification, which improves training-free dynamic adapters by progressively filtering out unreliable and redundant data at frame-/video- level.
- Extensive experiments over four benchmarks demonstrate that the reliable and diverse hierarchical adapter achieves superior performance while maintaining competitive computational efficiency.

2 Related Work

2.1 Zero-shot Activity Recognition

Zero-shot action recognition [Liu *et al.*, 2011; Qian *et al.*, 2022; Rasheed *et al.*, 2023] refers to the process of identifying actions in videos without having seen any samples of those activities during model training, which is crucial for real-world applications with limited annotated data. Early studies mainly focus on designing semantic representation of actions. Many attempts have been made in this regard, such as using manually defined attributes to represent actions [Liu *et al.*, 2011; Gan *et al.*, 2016b], mining objects as attributes [Jain *et al.*, 2015; Gan *et al.*, 2016a; Gao *et al.*, 2019], and utilizing word embeddings of action names or action descriptions as semantic representations [Qian *et al.*, 2022; Mandal *et al.*, 2019; Qin *et al.*, 2017; Xu *et al.*, 2017]. Differently, recent studies generally delve into adapting large pre-trained VLMs (*e.g.*, CLIP [Radford *et al.*, 2021]) to zero-shot video recognition. For instance, ViFi-CLIP [Rasheed *et al.*, 2023] fully tunes CLIP on videos with minimal design changes. ActionCLIP [Wang *et al.*, 2023] introduces temporal encoder to strengthen the video representation. PromptCLIP [Ju *et al.*, 2022] also adopts a lightweight Transformer on the top of the CLIP image encoder for temporal modeling.

2.2 Prompt-based Learning for VLMs

Prompt learning, derived from natural language processing, has been studied extensively to leverage the existing knowledge of VLMs to boost their generalization. CoOp [Zhou *et al.*, 2022b], a typical example of prompt learning for VLMs, learns prompt context knowledge by inserting learnable vectors into the class embeddings and optimizing the tokens using supervised classification loss. CoCoOp [Zhou *et al.*, 2022a] extends CoOp by conditioning the text prompts on image embeddings to solve the issue of overfitting. Although these methods have demonstrated significant performance improvements, their reliance on a large amount of training data from the target domain hinders their practical applications in downstream tasks. To this end, Shu *et al.* propose a new paradigm, test-time tuning (TPT) [Shu *et al.*, 2022a], which optimizes prompts dynamically via an unsupervised semantic consistency loss during inference. TPT has attracted much attention and has been extensively explored in recent researches [Feng *et al.*, 2023; Yan *et al.*, 2024; Zhang *et al.*, 2024a; Abdul Samadh *et al.*, 2024]. For example, DiffTPT [Feng *et al.*, 2023] leverages generative models (*i.e.*, Stable Diffusion) to augment test images, making augmented views more diverse. PromptAlign [Abdul Samadh *et al.*, 2024] extends TPT with the token alignment strategy, which enforces to bridge the data shift in the test data. DTS-TPT [Yan *et al.*, 2024] transfers TPT to the video domain, considering the diversity of motion semantics. Nevertheless, these methods require gradient descent during inference, which is computationally expensive and time-consuming, thereby conflicting with the principles of test-time adaptation. This paper seeks to achieve efficient test-time adaption by leveraging test samples cache.

2.3 Memory-based Learning for VLMs

In recent years, it has become a trend to apply memory-based learning to various tasks in computer vision [Zhang *et al.*, 2022; Karmanov *et al.*, 2024; Zhang *et al.*, 2024b; Udandarao *et al.*, 2023; Qu *et al.*, 2024; Qu *et al.*, 2025b] and natural language processing [Grave *et al.*, 2017; Merity *et al.*, 2016]. As a parameter-free technique, memory-based learning enhances test-time adaption by providing efficient inference. Tip-Adapter [Zhang *et al.*, 2022] is the first to adapt

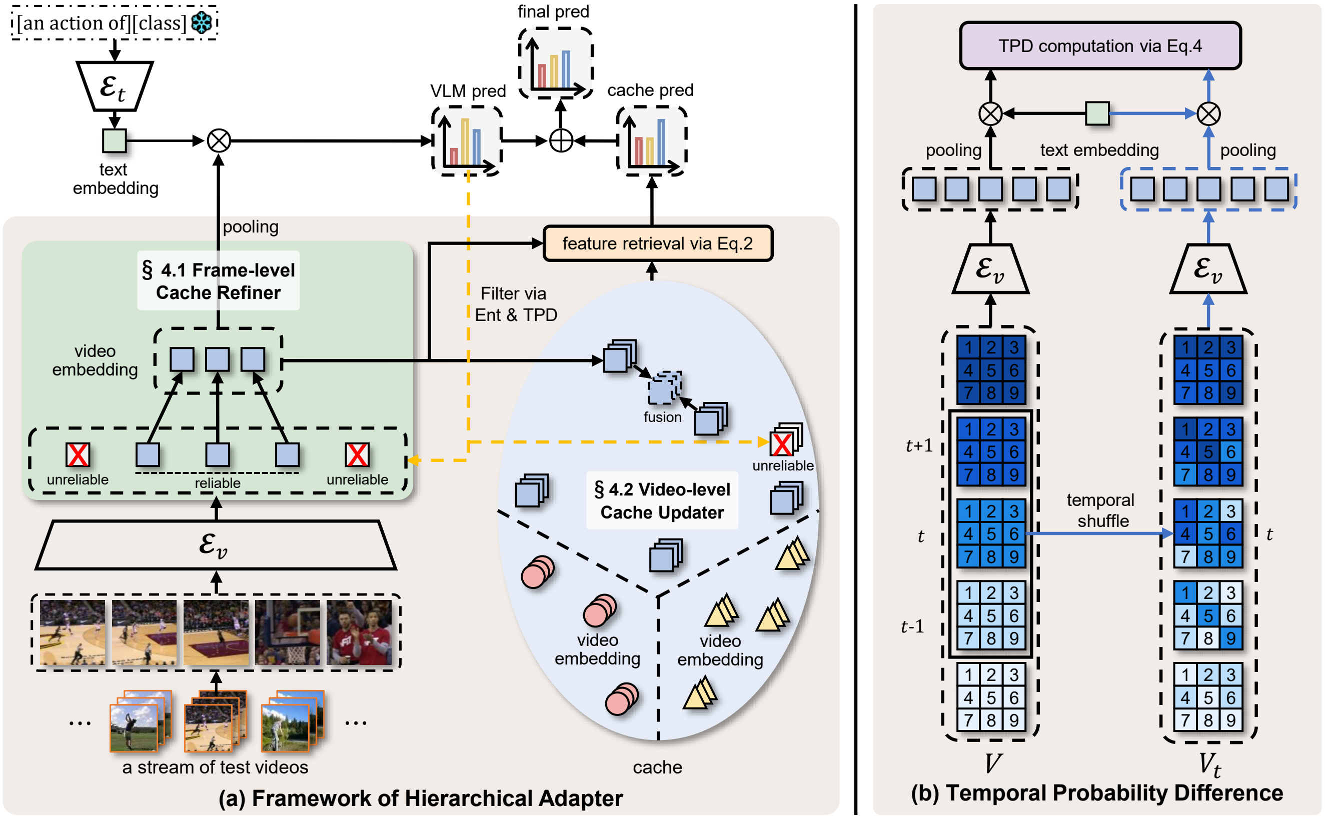


Figure 2: Overview of the proposed reliable and diverse **Hierarchical Adapter**. (a) Framework of Hierarchical Adapter, which first refines cache samples and then update the dynamic cache for zero-shot video classification. i) **Frame-level Cache Refiner**: It selects reliable frames from the test video based on prediction entropy and temporal probability difference. ii) **Video-level Cache Updater**: It dynamically updates the cache during inference based on diversity criterion. (b) Details of temporal probability difference. Each frame performs patch-level temporal shuffling with its neighboring K frames to compute the temporal probability difference.

VLMs via memory banks, in which a key-value cache model is constructed from the few-shot training set. SuS-X [Udandarao *et al.*, 2023] forms a support set with the help of Stable Diffusion, which generates salient and informative support images. To overcome the unavailability of source data during zero-shot inference and alleviate data shift caused by generative models, TDA [Karmanov *et al.*, 2024] designs a dual-cache model constructed from reliable images and their corresponding pseudo-labels during testing. BoostAdapter [Zhang *et al.*, 2024b] combines instance-agnostic historical cache in TDA with instance-aware boosting cache, achieving promising results. In this work, we extend memory-based learning to the video domain for zero-shot activity recognition.

3 Preliminary

3.1 Problem Statement

This work focuses on transductive test-time adaption for zero-shot video classification. Given a set of C class labels $Y = \{y_1, y_2, \cdots, y_C\}$ and a sequence of N testing videos $V = \{v_1, v_2, \cdots, v_N\}$, the task of zero-shot video classification is to predict the label of each testing video v as $\hat{y_c} \in Y$ while Y is unseen to the pre-trained model. Specifically, inference is performed under a transductive setting, prediction for i-th test sample v_i may therefore depend on the representations and predictions of the first $i-1$ samples.

3.2 Memory-based Adaption

Zero-shot Matching. Given an arbitrary pre-trained VLM (such as CLIP [Radford *et al.*, 2021] and ViFi-CLIP [Rasheed *et al.*, 2023]) consisting of a visual encoder $\mathcal{E}_v$ and a textual encoder $\mathcal{E}_t$, we first encode the test video $\boldsymbol{v}_i$ as $\boldsymbol{f}_{v_i} = \mathcal{E}_v(\boldsymbol{v_i})$ and get text features (*i.e.*, zero-shot classifier) as $\{\boldsymbol{f}_c|\boldsymbol{f}_c = \mathcal{E}_t(\boldsymbol{t}_c)\}_{c=1}^{C}$, where $\boldsymbol{t}_c$ is a manual-crafted text prompt corresponding to class label $\boldsymbol{y}_c$, *e.g.*, "*a video of a person doing* [class $\boldsymbol{y}_c$]." The zero-shot matching result of $\boldsymbol{v}_i$ and $\boldsymbol{y}_c$ can be denoted as

$$\boldsymbol{p}_c(\boldsymbol{v}_i) = \boldsymbol{f}_c^{\mathrm{T}} \boldsymbol{f}_{v_i}, \tag{1}$$

which represents the cosine similarity between multi-modal features after normalizing $\boldsymbol{f}_c$ and $\boldsymbol{f}_{v_i}$ respectively. Furthermore, the complete output logits of the test video $\boldsymbol{v}_i$ is $\boldsymbol{p}(v_i) = [\boldsymbol{p}_1(v_i), \boldsymbol{p}_2(v_i), \cdots, \boldsymbol{p}_C(\boldsymbol{v}_i)] \in R^C$.

Training-free Dynamic Adapter. To address the issue of data shift between source domain and target domain, a widely used approach is to construct a cache based on target domain, leveraging the knowledge of target domain data distribution to guide model inference [Zhang *et al.*, 2022; Karmanov *et al.*, 2024]. In few-shot action recognition, the support set is constructed using annotated training videos, while in zero-shot action recognition, the cache can be progressively built during inference, based on test samples and pseudo-labels. The prediction from the cache can be calculated as

$$\boldsymbol{p}_{\text{cache}}(\boldsymbol{v}_i) = \mathcal{A}(\boldsymbol{F}_{\text{cache}}^{\mathrm{T}} \boldsymbol{f}_{v_i}) \boldsymbol{L}_p, \tag{2}$$

where $\mathcal{A}(x) = \exp(-\beta(1-x))$ represents the scaling function with a smoothing scalar β, $\boldsymbol{F}_{\text{cache}}$ denotes visual features of samples in the cache, and $\boldsymbol{L}_p$ is the corresponding pseudo-labels in the form of one-hot vectors.

4 Reliable and Diverse (R&D) Hierarchical Adapter

Inspired by Training-free Dynamic Adapter (TDA) [Karmanov *et al.*, 2024], this work aims to adapt VLMs during inference by leveraging knowledge from the cache. However, the cache in TDA adopts a simplistic approach to storing support samples, failing to capture complicated action semantics. Moreover, the entropy-based cache update strategy in TDA cannot ensure the diversity of support samples, which is contradictory to the inherently complex and diverse nature of motions in real-world scenarios. To this end, we propose a reliable and diverse hierarchical adapter to enable efficient and effective test-time adaptation with VLMs. As shown in Figure 2, R&D Hierarchical Adapter is mainly composed of Frame-level Cache Refiner (FCR) and Video-level Cache Updater (VCU). In FCR, each video sample is refined at frame level based on prediction entropy and temporal probability difference before entering the corresponding cache. In VCU, the cache is updated based on the principle of diversity.

4.1 Frame-level Cache Refiner

To enhance video representations, we propose a Frame-level Cache Refiner (FCR), which discards low-confidence frames based on prediction entropy and temporal probability difference. In TDA [Karmanov *et al.*, 2024], a single entropy-based criterion is employed to measure sample reliability. Given a test video v, the prediction entropy of v is calculated as $\boldsymbol{e} = -\sum_{c=1}^{C} \boldsymbol{p}(\hat{\boldsymbol{y}} = c|\boldsymbol{v}) \log \boldsymbol{p}(\hat{\boldsymbol{y}} = c|\boldsymbol{v})$. Although frames with lower prediction entropy have a lower likelihood of causing error, entropy is not always reliable as a confidence metric under biased scenarios. To avoid selecting overconfident samples based on incorrect cues, DeYO [Lee *et al.*, 2024] further introduce a probability difference metric to ensure sample reliability, which we refer to as spatial probability difference (SPD). SPD quantifies the influence of reliable static cues, such as structure information, on inference by measuring the pseudo-label probability difference between the original frame and its spatial-shuffled variant independently.

Given a test video v and corresponding spatial-shuffled video v_s, SPD is calculated in a parallelized manner as

$$\boldsymbol{d}_{\text{spatial}} = |\boldsymbol{p}(\hat{\boldsymbol{y}} = c^*|\boldsymbol{v}) - \boldsymbol{p}(\hat{\boldsymbol{y}} = c^*|\boldsymbol{v}_s)|, \tag{3}$$

where $c^* = \arg\max_{\mathrm{c}} \boldsymbol{p}(\hat{\boldsymbol{y}} = c|\boldsymbol{v})$ is the pseudo-label of $\boldsymbol{v}$ predicted by the VLM. The model is expected to be sensitive to discriminative factors, for which frames with higher $\boldsymbol{d}_{\text{spatial}}$ are more reliable.

Algorithm 1 Reliable and Diverse Cache Update

Input: test video $\boldsymbol{v}$, cache, pseudo-label of test video c^*
Parameter: cache size n, similarity threshold τ,
Output: cache updated

```
if The cache of class $c^*$ is not full then
    Add new sample to the corresponding cache.
else
    for $i = 1$ to $n$ do
        Calculate similarity between $\boldsymbol{v}$ and $i$-th sample in the cache.
    end for
    if $\text{similarity}_{\max} > \tau$ then
        Sample fusion based on momentum update.
    else
        Remove the sample with the lowest confidence.
    end if
end if
return cache updated
```

To further assess the model's sensitivity to temporal dynamic information, we design a temporal probability difference (TPD), which is shown in Figure 2 (b). TPD measures the influence of temporal dynamic cues on inference by calculating the pseudo-label probability difference between the original video and its temporal-shuffled version. Predictions with larger TPD are more likely to rely on temporal dynamic cues, indicating that the pseudo-labels are more reliable.

Given a test video v and its temporal-shuffled variant v_t, TPD is obtained as

$$\boldsymbol{d}_{\text{temp}} = |\boldsymbol{p}(\hat{\boldsymbol{y}} = c^*|\boldsymbol{v}) - \boldsymbol{p}(\hat{\boldsymbol{y}} = c^*|\boldsymbol{v}_t)|, \tag{4}$$

where $c^* = \arg\max_{\mathrm{c}} \boldsymbol{p}(\hat{\boldsymbol{y}} = c|\boldsymbol{v})$ is the predicted category of the test video $\boldsymbol{v}$.

Taking into account that VLMs exhibit bias in per-class accuracy, we select top-K confident frames for each video, instead of keeping all frames with confidence scores higher than a pre-defined threshold, which is a commonly adopted strategy [Lee *et al.*, 2024]. For a test video $\boldsymbol{v} \in R^{T \times d}$, the refined video $\boldsymbol{v''} \in R^{K_2 \times d}$ is obtained by

$$\boldsymbol{v'} = \texttt{TopK\_Selection}(\boldsymbol{v}, K_1, -\boldsymbol{e}), \tag{5}$$

$$\boldsymbol{v''} = \texttt{TopK\_Selection}(\boldsymbol{v'}, K_2, \boldsymbol{d}_{\text{temp}}), \tag{6}$$

where $\boldsymbol{v'} \in R^{K_1 \times d}$ serves as an intermediate result and is not utilized in the following stages.

4.2 Video-level Cache Updater

The cache stores prototypes of different activities as a database. During inference, the test video is used as a query to aggregate information from the cache via similarity-based retrieval. However, in zero-shot video classification, the model cannot access the ground truth labels of historical samples and must rely on pseudo-labels to construct the cache. This inevitably introduces noise, which negatively impacts the model's performance. Therefore, it is necessary to update the cache progressively during inference. In TDA, the cache is implemented as a priority queue, where entropy serves as the criterion for prioritization.

The semantics of the same activity are diverse in visual space, but samples with high reliability are often visual-similar, which hinders the cache in capturing the underlying data manifold. To this end, we propose Video-level Cache Updater (VCU) to maintain a diverse cache.

Specifically, when the test video v is added to the cache, the prototype most similar to v is updated as

$$\boldsymbol{q} \leftarrow \mu \boldsymbol{q} + (1-\mu)\boldsymbol{f}_v, \tag{7}$$

where $\mu \in [0,1]$ is the momentum coefficient, $\boldsymbol{f}_v$ is video embeddings of $\boldsymbol{v}$, and $\boldsymbol{q}$ is the prototype most similar to $\boldsymbol{v}$.

In the image domain, the similarity between two images can be measured by calculating the cosine similarity. However, in the video domain, temporal sequence matching is required to evaluate the similarity between videos. We follow [Haresh *et al.*, 2021] to use dynamic time warping (DTW) to compute the temporal sequence similarity. Given two video features $\boldsymbol{f}_v$ and $\boldsymbol{f}_w$ to be matched, we can obtain the similarity matrix $\boldsymbol{S} \in R^{n \times m}$ based on cosine similarity, where $\boldsymbol{S}(i,j) = \boldsymbol{f}_{v_i} \cdot \boldsymbol{f}_{w_j}$. DTW can adaptively find the path with the highest similarity in $\boldsymbol{S}$. Let the similarity between the first i frames of $\boldsymbol{f}_v$ and the first j frames of $\boldsymbol{f}_w$ be denoted as $\boldsymbol{S}_{\text{seq}}(i,j)$, then the following state transition equation can be established.

$$\begin{aligned}\boldsymbol{S}_{\text{seq}}(i,j) = \boldsymbol{S}(i,j) + \min\{&\boldsymbol{S}_{\text{seq}}(i-1,j), \boldsymbol{S}_{\text{seq}}(i,j-1),\\ &\boldsymbol{S}_{\text{seq}}(i-1,j-1)\},\end{aligned} \tag{8}$$

the similarity between $\boldsymbol{f}_v$ and $\boldsymbol{f}_w$ is given by $\boldsymbol{S}_{\text{seq}}(n,m)$.

For clarity, we provide the whole cache update process in Algorithm 1 in the form of pseudo-code.

5 Experiments

5.1 Datasets

HMDB-51 [Kuehne *et al.*, 2011] is a small-scale action recognition dataset. It contains around 7,000 labeled videos sourced from YouTube, covering 51 activity categories. **UCF-101** [Soomro, 2012] consists of 13,320 videos covering 101 categories, which can be further grouped into five main categories: Body motion, Human-human interactions, Human-object interactions, Playing instruments, and Sports. **Kinetics-600** [Carreira *et al.*, 2018] is a large-scale video dataset, containing 600 human action classes, with at least 600 video clips for each action. Each video is collected and annotated from YouTube and lasts approximately 10 seconds. **ActivityNet-200** [Fabian Caba Heilbron and Niebles, 2015] is also a large-scale action recognition benchmark, but it provides about 20k untrimmed videos of 5 to 10 minutes from 200 activity categories.

5.2 Implementation Details

We utilize a pre-trained ViT-B/16 of CLIP as the foundation model, and the model is not fine-tuned on extra large video datasets. In test-time adaption, we sample $T = 32$ frames from each test video. We use top-1 accuracy(%) as our evaluation metric. We perform a search for hyperparameter on the validation set of Kinetics-400. In FCR, we select 8 frames based on prediction entropy, and subsequently select 5 frames based on TPD to construct refined video embeddings. When calculating TPD, each frame is divided into 7×7 image patches, and temporal shuffling is applied between adjacent 2 frames. In Algorithm 1, cache size n is set as 10 and similarity threshold τ is 0.95. In Eq. 2, β is 8 according to TDA, and in Eq. 7, μ is set to 0.5. All the experiments are conducted using a single NVIDIA 3090 24GB GPU.

5.3 Comparison With Other Methods

We conduct a comprehensive comparison of the proposed Hierarchical Adapter with popular zero-shot video classification approaches spanning various methodological categories. Specifically, uni-modal zero-shot video recognition models are trained on video data with elaborated representation engineering. Adapting pre-trained CLIP involves additional temporal learners or vision-language prompting techniques without training the encoders while tuning pre-trained CLIP means fully fine-tuning the CLIP model via video data. Following [Rasheed *et al.*, 2023], we report the mean and standard variance of the results.

As we can see in Table 1, our method surpasses conventional uni-modal zero-shot video recognition methods, *e.g.,* ER-ZSAR [Chen and Huang, 2021], JigsawNet [Qian *et al.*, 2022], and ResT [Lin *et al.*, 2022], by a significant margin on all benchmarks. Our approach also outperforms models such as Vita-CLIP [Wasim *et al.*, 2023] and VicTR [Kahatapitiya *et al.*, 2024] that adapt pre-trained CLIP. Compared with ViFi-CLIP [Rasheed *et al.*, 2023], which serves as a baseline for our method and fine-tunes the pre-trained CLIP on Kinetics-400, our method is also superior.

5.4 Ablation Study

Component analysis. To verify the effectiveness of the proposed FCR and VCU, we conduct ablation experiments on HMDB-51 and Kinetics-600 benchmarks. As shown in Table 2, FCR brings 2.7% and 1.9% performance gains on HMDB-51 and Kinetics-600, respectively, which indicates that FCR can focus on reliable frames. In addition, VCU improves on the two datasets by 2.3% and 0.9%, respectively, which suggests that VCU can retain diverse video samples in the cache. Moreover, by combining the two modules, our full model achieves better results, confirming the complementarity and effectiveness of the proposed framework.

Reliability metric. We examine the impact of the proposed temporal probability difference by contrasting it with the model with prediction and spatial probability difference or with prediction entropy only. A shown in Table 3, temporal probability difference achieves better performance, while spatial probability difference has almost no effect. This study confirms the complementarity between prediction entropy and temporal probability difference.

Diversity strategy. To investigate the effectiveness of diversity criterion, we compare it with the cache updater without feature fusion. As shown in Table 4, our proposed feature fusion rule brings 0.9% and 0.6% performance gains on HMDB-51 and Kinetics-600, respectively, indicating that the cache we construct better represents motion semantics.

Method	Encoder	HMDB-51	UCF-101	Kinetics-600	ActivityNet-200
Uni-modal zero-shot video recognition models					
E2E [Brattoli *et al.*, 2020]	R(2+1)D	29.8	44.1	–	26.6
ER-ZSAR [Chen and Huang, 2021]	TSM	35.3 ± 4.6	51.8 ± 2.9	42.1 ± 1.4	–
JigsawNet [Qian *et al.*, 2022]	R(2+1)D	38.7 ± 3.7	56.0 ± 3.1	–	–
ResT [Lin *et al.*, 2022]	Resnet-101	41.1 ± 3.7	58.7 ± 3.3	–	32.5
Adapting pre-trained CLIP					
Vanilla CLIP [Radford *et al.*, 2021]	ViT-B/16	46.2 ± 0.2	63.1 ± 0.5	64.1 ± 0.8	73.9 ± 0.6
ActionCLIP [Wang *et al.*, 2023]	ViT-B/16	40.8 ± 5.4	58.3 ± 3.4	66.7 ± 1.1	–
A5 [Ju *et al.*, 2022]	ViT-B/16	44.3 ± 2.2	69.3 ± 4.2	55.8 ± 0.7	–
Vita-CLIP [Wasim *et al.*, 2023]	ViT-B/16	48.6 ± 0.6	75.0 ± 0.6	67.4 ± 0.5	–
XCLIP [Ni *et al.*, 2022]	ViT-B/16	44.6 ± 5.2	72.0 ± 2.3	65.2 ± 0.4	–
VicTR [Kahatapitiya *et al.*, 2024]	ViT-B/16	51.0 ± 1.3	72.4 ± 0.3	–	–
Tuning pre-trained CLIP					
ViFi-CLIP [Rasheed *et al.*, 2023]	ViT-B/16	53.9 ± 0.7	76.2 ± 0.8	67.3 ± 1.0	80.6 ± 0.8
BIKE [Wu *et al.*, 2023]	ViT-B/16	49.1 ± 0.5	77.4 ± 1.0	66.1 ± 0.6	75.2 ± 1.1
ViFi-CLIP + Hierarchical Adapter	ViT-B/16	**54.9 ± 0.1**	**77.6 ± 0.2**	**69.0 ± 0.1**	**81.8 ± 0.2**

Table 1: Comparisons with state-of-the-art methods for zero-shot video classification.

FCR	VCU	HMDB-51	Kinetics-600
✗	✗	46.2	64.1
✓	✗	48.9	66.0
✗	✓	48.8	65.0
✓	✓	**51.2**	**67.4**

Table 2: Effectiveness of different components in our method.

Metric	HMDB-51	Kinetics-600
Ent	50.7	66.8
Ent & SPD	50.8	66.8
Ent & TPD (**Ours**)	**51.2**	**67.4**

Table 3: Performance comparison using different confidence metric. Ent represents prediction entropy, SPD is spatial probability difference, and TPD (§4.1) denotes temporal probability difference.

Rule	HMDB-51	Kinetics-600
Feature Concatenate	50.3	66.8
Momentum Update (**Ours**)	**51.2**	**67.4**

Table 4: Effect of different feature fusion rules. Feature Concatenate refers to directly appending new video embeddings to the cache.

Strategy	HMDB-51	Kinetics-600
Max	49.6	66.2
Mean	49.4	66.9
Diagonal	49.4	66.2
DTW (**Ours**)	**51.2**	**67.4**

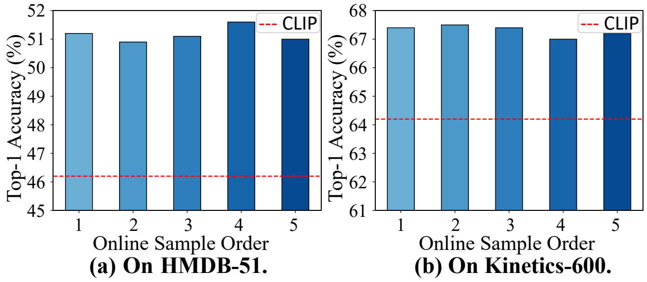
Table 5: Different implementation of temporal sequence matching strategies.

Pre-trained VLM	HMDB-51	Kinetics-600
CLIP with ViT-B/32	40.4	60.7
+ Hierarchical Adapter	**45.3**	**63.6**
CLIP with ViT-B/16	46.2	64.1
+ Hierarchical Adapter	**51.2**	**67.4**
CLIP with ViT-L/14	50.9	72.1
+ Hierarchical Adapter	**54.3**	**75.4**
ViFi-CLIP with ViT-B/16	53.9	67.3
+ Hierarchical Adapter	**54.9**	**69.0**

Table 6: Top-1 accuracy(%) on HMDB-51 and Kinetics-600 using different VLMs, *i.e.*, Vanilla CLIP [Radford *et al.*, 2021] and ViFi-CLIP [Rasheed *et al.*, 2023].

Temporal sequence matching. To validate the superiority of Dynamic Time Warping, we compare it with various temporal sequence matching strategies. In Max strategy, the maximum of the similarity matrix $\boldsymbol{S}$ is taken as the final result. In Mean setting, the average of $\boldsymbol{S}$ is the similarity between the two videos. In Diagonal strategy, the diagonal of the similarity matrix $\boldsymbol{S}$ is involved in temporal sequence matching. The optimal path from the top-left corner to the bottom-right corner of $\boldsymbol{S}$ is found by DTW to measure the similarity between the two videos. The results are provided in Table 5, which indicates Dynamic Time Warping outperforms the other three strategies.

Generalization to different pre-trained VLMs. We use Vanilla CLIP with ViT-B/16 as the VLM of choice throughout our ablation studies. In Table 6, we demonstrate results when our proposed Hierarchical Adapter is applied on top of ViFi-CLIP and three versions of Vanilla CLIP. Hierarchical Adapter improves model performance on both HMDB-51 and Kinetics-600, indicating the proposed framework can be applied to an arbitrary VLM.



(a) On HMDB-51. **(b) On Kinetics-600.**

Figure 3: Sensitivity to test-time sample order on HMDB-51 (a) and Kinetics-600 (b). Vanilla CLIP [Radford *et al.*, 2021] with ViT-B/16 is used as the VLM of choice.

Sensitivity to test time sample order. As Hierarchical Adapter classifies videos online, the performance of the model is inevitably influenced by the order of the test videos. To investigate this influence, we conduct five rounds of experiments on HMDB-51 and Kinetics-600, with each round employing a different sample order. As we can see in Figure 3, the model's performance is slightly influenced by the sample order fluctuations on both datasets. Regardless of the sample order, the performance of our proposed model consistently surpasses Vanilla CLIP.

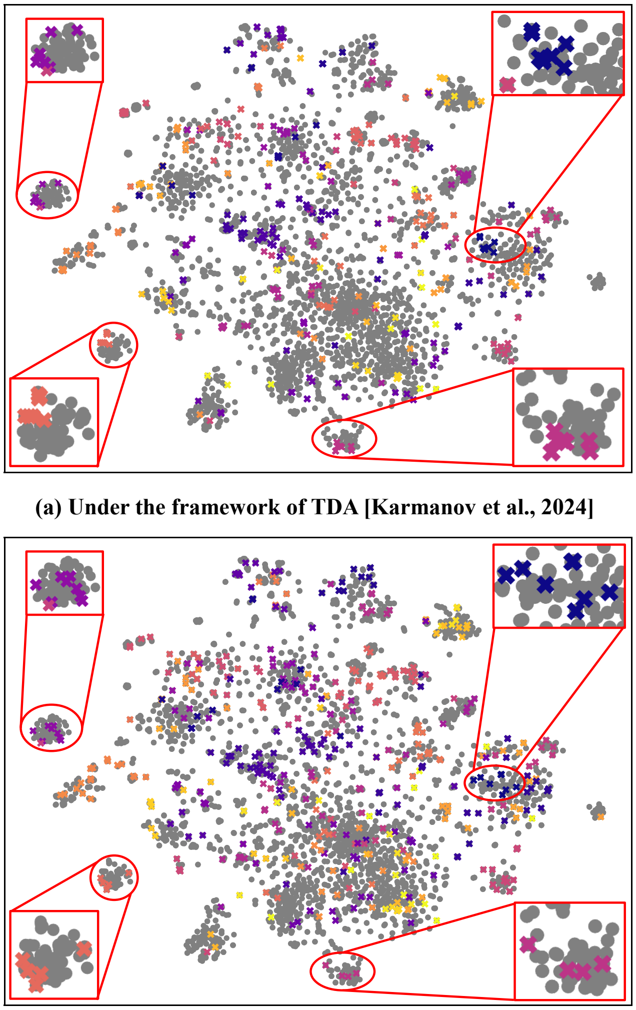
Visualization. In Figure 4, we apply t-SNE to visualize the stored video features in the cache under the framework of TDA [Karmanov *et al.*, 2024] and our proposesd Hierarchical Adapter on the HMDB-51 [Kuehne *et al.*, 2011] dataset. The stored video features are highlighted using different colors while the others are marked in gray. The visualization results indicate that Hierarchical Adapter is able to construct and update reliable and diverse prototypes to represent motion semantics.

6 Conclusion

In this work, we propose an effective and efficient Hierarchical Adapter, which is the first training-free test-time adapter for zero-shot video classification. This framework aims to select reliable and diverse visual features at frame level and video level, which consists of two core modules: 1) Frame-level Cache Refiner for selecting rich and effective motion features; 2) Video-level Cache Updater for capturing the underlying data manifold. Experimental results on four video classification benchmarks demonstrate the superiority of our Hierarchical Adapter against existing methods. In future work, exploring the potential of leveraging the rich text representations provided by large language models to enrich the cache holds promising prospects.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (Grant No. 62222207, 62427808, 62472208), and the open Foundation of the Key lab (center) of Anhui Jianzhu University Anhui Province Key Laboratory of Intelligent Building & Building Energy Saving (Grant No. IBES2024KF02). Wenxuan Ge and Peng Huang equally contributed to this work, Xiangbo Shu and Rui Yan are the equal corresponding authors.



(a) Under the framework of TDA [Karmanov et al., 2024]

(b) Under the framework of Hierarchical Adapter (ours)

Figure 4: t-SNE visualizations of the stored video features in the cache under the framework of TDA (left) and our proposed Hierarchical Adapter (right).

References

[Abdul Samadh *et al.*, 2024] Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. *NIPS*, 36, 2024.

[Brattoli *et al.*, 2020] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *CVPR*, pages 4613–4623, 2020.

[Carreira *et al.*, 2018] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

[Chen and Huang, 2021] Shizhe Chen and Dong Huang. Elaborative rehearsal for zero-shot action recognition. In *ICCV*, pages 13638–13647, 2021.

[Fabian Caba Heilbron and Niebles, 2015] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.

[Feng *et al.*, 2023] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *ICCV*, pages 2704–2714, 2023.

[Gan *et al.*, 2016a] Chuang Gan, Ming Lin, Yi Yang, Gerard Melo, and Alexander G Hauptmann. Concepts not alone: Exploring pairwise relationships for zero-shot video activity recognition. In *AAAI*, volume 30, 2016.

[Gan *et al.*, 2016b] Chuang Gan, Tianbao Yang, and Boqing Gong. Learning attributes equals multi-source domain generalization. In *CVPR*, pages 87–97, 2016.

[Gao *et al.*, 2019] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *AAAI*, volume 33, pages 8303–8311, 2019.

[Grave *et al.*, 2017] Edouard Grave, Moustapha M Cisse, and Armand Joulin. Unbounded cache model for online language modeling with open vocabulary. *NIPS*, 30, 2017.

[Haresh *et al.*, 2021] Sanjay Haresh, Sateesh Kumar, Huseyin Coskun, Shahram N Syed, Andrey Konin, Zeeshan Zia, and Quoc-Huy Tran. Learning by aligning videos in time. In *CVPR*, pages 5548–5558, 2021.

[Jain *et al.*, 2015] Mihir Jain, Jan C Van Gemert, Thomas Mensink, and Cees GM Snoek. Objects2action: Classifying and localizing actions without any video example. In *ICCV*, pages 4588–4596, 2015.

[Ju *et al.*, 2022] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *ECCV*, pages 105–124. Springer, 2022.

[Kahatapitiya *et al.*, 2024] Kumara Kahatapitiya, Anurag Arnab, Arsha Nagrani, and Michael S Ryoo. Victr: Video-conditioned text representations for activity recognition. In *CVPR*, pages 18547–18558, 2024.

[Karmanov *et al.*, 2024] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *CVPR*, pages 14162–14171, 2024.

[Kay *et al.*, 2017] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[Kuehne *et al.*, 2011] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[Lee *et al.*, 2024] Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. *ICLR*, 2024.

[Li *et al.*, 2025] Pengpeng Li, Xiangbo Shu, Chun-Mei Feng, Yifei Feng, Wangmeng Zuo, and Jinhui Tang. Surgical video workflow analysis via visual-language learning. *npj Health Systems*, 2(1):5, 2025.

[Lin *et al.*, 2022] Chung-Ching Lin, Kevin Lin, Lijuan Wang, Zicheng Liu, and Linjie Li. Cross-modal representation learning for zero-shot action recognition. In *CVPR*, pages 19978–19988, 2022.

[Liu *et al.*, 2011] Jingen Liu, Benjamin Kuipers, and Silvio Savarese. Recognizing human actions by attributes. In *CVPR*, pages 3337–3344. IEEE, 2011.

[Mandal *et al.*, 2019] Devraj Mandal, Sanath Narayan, Sai Kumar Dwivedi, Vikram Gupta, Shuaib Ahmed, Fahad Shahbaz Khan, and Ling Shao. Out-of-distribution detection for generalized zero-shot action recognition. In *CVPR*, pages 9985–9993, 2019.

[Merity *et al.*, 2016] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

[Ni *et al.*, 2022] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV*, pages 1–18. Springer, 2022.

[Qian *et al.*, 2022] Yijun Qian, Lijun Yu, Wenhe Liu, and Alexander G Hauptmann. Rethinking zero-shot action recognition: Learning from latent atomic actions. In *ECCV*, pages 104–120. Springer, 2022.

[Qin *et al.*, 2017] Jie Qin, Li Liu, Ling Shao, Fumin Shen, Bingbing Ni, Jiaxin Chen, and Yunhong Wang. Zero-shot action recognition with error-correcting output codes. In *CVPR*, pages 2833–2842, 2017.

[Qu *et al.*, 2024] Hongyu Qu, Rui Yan, Xiangbo Shu, Hailiang Gao, Peng Huang, and Guo-Sen Xie. Mvp-shot: Multi-velocity progressive-alignment framework for few-shot action recognition. *arXiv preprint arXiv:2405.02077*, 2024.

[Qu *et al.*, 2025a] Hongyu Qu, Jianan Wei, Xiangbo Shu, and Wenguan Wang. Learning clustering-based prototypes for compositional zero-shot learning. In *ICLR*, 2025.

[Qu *et al.*, 2025b] Hongyu Qu, Ling Xing, Rui Yan, Yazhou Yao, Guo-Sen Xie, and Xiangbo Shu. Hierarchical relation-augmented representation generalization for few-shot action recognition. *arXiv preprint arXiv:2504.10079*, 2025.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Rasheed *et al.*, 2023] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *CVPR*, pages 6545–6554, 2023.

[Shen and Tang, 2024] Fei Shen and Jinhui Tang. Imagpose: A unified conditional framework for pose-guided person generation. *NIPS*, 37:6246–6266, 2024.

[Shen *et al.*, 2025] Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinhui Tang. Imagdressing-v1: Customizable virtual dressing. In *AAAI*, volume 39, pages 6795–6804, 2025.

[Shu *et al.*, 2019] Xiangbo Shu, Jinhui Tang, Guo-Jun Qi, Wei Liu, and Jian Yang. Hierarchical long short-term concurrent memory for human interaction recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):1110–1118, 2019.

[Shu *et al.*, 2021] Xiangbo Shu, Liyan Zhang, Guo-Jun Qi, Wei Liu, and Jinhui Tang. Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3300–3315, 2021.

[Shu *et al.*, 2022a] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *NIPS*, 35:14274–14289, 2022.

[Shu *et al.*, 2022b] Xiangbo Shu, Binqian Xu, Liyan Zhang, and Jinhui Tang. Multi-granularity anchor-contrastive representation learning for semi-supervised skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7559–7576, 2022.

[Soomro, 2012] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[Tang *et al.*, 2019] Jinhui Tang, Xiangbo Shu, Rui Yan, and Liyan Zhang. Coherence constrained graph lstm for group activity recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(2):636–647, 2019.

[Udandarao *et al.*, 2023] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *ICCV*, pages 2725–2736, 2023.

[Wang *et al.*, 2023] Mengmeng Wang, Jiazheng Xing, Jianbiao Mei, Yong Liu, and Yunliang Jiang. Actionclip: Adapting language-image pretrained models for video action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Wasim *et al.*, 2023] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Vita-clip: Video and text adaptive clip via multimodal prompting. In *CVPR*, pages 23034–23044, 2023.

[Wu *et al.*, 2023] Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, and Wanli Ouyang. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In *CVPR*, pages 6620–6630, 2023.

[Xu *et al.*, 2017] Xun Xu, Timothy Hospedales, and Shaogang Gong. Transductive zero-shot action recognition by word-vector embedding. *International Journal of Computer Vision*, 123:309–333, 2017.

[Yan *et al.*, 2020] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Higcin: Hierarchical graph-based cross inference network for group activity recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):6955–6968, 2020.

[Yan *et al.*, 2023] Rui Yan, Lingxi Xie, Xiangbo Shu, Liyan Zhang, and Jinhui Tang. Progressive instance-aware feature learning for compositional action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10317–10330, 2023.

[Yan *et al.*, 2024] Rui Yan, Hongyu Qu, Xiangbo Shu, Wenbin Li, Jinhui Tang, and Tieniu Tan. Dts-tpt: Dual temporal-sync test-time prompt tuning for zero-shot activity recognition. In *IJCAI*, 2024.

[Yan *et al.*, 2025] Rui Yan, Jin Wang, Hongyu Qu, Xiaoyu Du, Dong Zhang, Jinhui Tang, and Tieniu Tan. Test-v: Test-time support-set tuning for zero-shot video classification. *arXiv preprint arXiv:2502.00426*, 2025.

[Zhang *et al.*, 2022] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *ECCV*, pages 493–510. Springer, 2022.

[Zhang *et al.*, 2024a] Jingyi Zhang, Jiaxing Huang, Xiaoqin Zhang, Ling Shao, and Shijian Lu. Historical test-time prompt tuning for vision foundation models. *arXiv preprint arXiv:2410.20346*, 2024.

[Zhang *et al.*, 2024b] Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. Boostadapter: Improving vision-language test-time adaptation via regional bootstrapping. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Zhou *et al.*, 2022a] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022.

[Zhou *et al.*, 2022b] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.