DTS-TPT: Dual Temporal-Sync Test-time Prompt Tuning for Zero-shot Activity Recognition

Rui Yan[1] , **Hongyu Qu**[2] , **Xiangbo Shu**[2] , **Wenbin Li**[1] , **Jinhui Tang**[2] and **Tieniu Tan**[1]
[1]Nanjing University
[2]Nanjing University of Science and Technology
{ruiyan, liwenbin, tnt}@nju.edu.cn, {quhongyu, shuxb, jinhuitang}@njust.edu.cn

Abstract

Finetuning the large vision-language models on video data with a set of learnable prompts has shown promising performance on zero-shot activity recognition but still requires extra video data and expensive training costs. Inspired by recent Test-time Prompt Tuning (TPT) on the image domain, this work attempts to extend TPT to video data for zero-shot activity recognition. However, monotonous spatial augmentation and short class names cannot meet the need to capture diverse and complicated semantics of human behavior during prompt tuning. To this end, this work proposes a Dual Temporal-Sync Test-time Prompt Tuning (DTS-TPT) framework for zero-shot activity recognition. DTS-TPT tunes the learnable prompts appended to text inputs on video feature sequences of different temporal scales in multiple steps during test time. In each tuning step, we minimize the semantic consistency among the predictions from video feature sequences randomly augmented via AugMix with both original class names and the corresponding description generated through LLM. Compared with the state-of-the-art methods, the proposed method improves the zero-shot top-1 accuracy by approximately $2\% \sim 5\%$ on popular benchmarks. The code is available at https://github.com/quhongyu/DTS-TPT.

1 Introduction

Since the era of deep learning, activity recognition technology has developed rapidly and achieved significant improvements in predefined categories such as K400 [Kay *et al.*, 2017] and K600 [Carreira *et al.*, 2018]. However, due to the diversity of human activity, it is difficult for humans to list all human activity categories in the real world completely. Therefore, it is urgent to develop zero-shot algorithms to predict unseen but known activity categories, *i.e.*, zero-shot activity recognition [Brattoli *et al.*, 2020; Liu *et al.*, 2011] via matching the video and class name feature directly. This research has made progress in recent years but is still insufficient, mainly hindered by challenges such as i) **Intra-modal Distribution Shift** [Lin *et al.*, 2022]: Within each modality, visual/textual features of each semantic should be clear and distinguishable; ii) **Inter-modal Semantic Gap** [Lin *et al.*, 2022]: Between two modalities, the feature distribution of visual samples and labels with the same semantics should be more aligned/consistent.

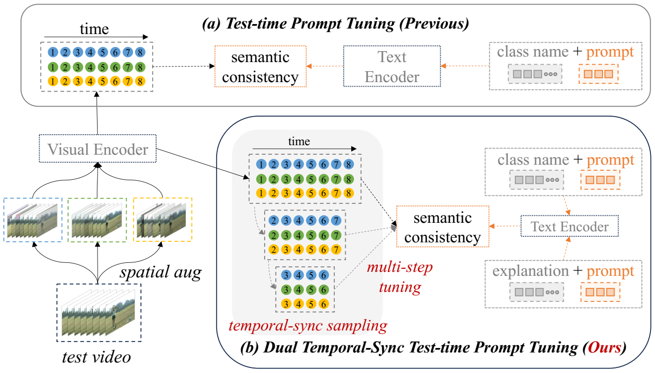


Figure 1: Motivation of this work. Standard Test-time Prompt Tuning adopts single-scale features for tuning with a single semantic consistency. To capture the diverse temporal semantics underlying human activities, we build multiple temporal scale video features for multi-step tuning supervised by the semantic consistency on both class names and explanations.

Existing solutions can be divided into two types, *i.e.*, training-based and training-free. The **training-based** methods tune an off-the-shelf visual encoder on a certain scale of video data for adapting to zero-shot activity recognition. To enhance the discriminability of features from each modality (visual or text), previous works build fine-grained motion features from the objects or attributes in videos [Gao *et al.*, 2019; Liu *et al.*, 2011], or enhance activity class name embeddings with object tags [Gao *et al.*, 2019] or elaborative descriptions [Chen and Huang, 2021]. To facilitate video-text semantic alignment, some recent works [Lin *et al.*, 2022; Wang and Chen, 2017] try to convert text features of unknown class names into the video feature space according to the knowledge association between known and unknown class names.

However, with the rise of large Vision-Language models technology, intra-modal representations are becoming more robust, and the inter-modal semantic gap (between visual and text) is gradually narrowing. Thanks to this, many **training-free** methods built on pre-trained large visual and text models have recently emerged in zero-shot visual classification. The training-free method usually fixes the visual embedding, and then i) replaces the fixed class embedding with a more flexible support set [Zhang *et al.*, 2022b; Udandarao *et al.*, 2023]; or ii) tunes the class embedding in a self-supervised manner on a single sample during the test time for better alignment (Test-time Prompt Tuning, TPT [Shu *et al.*, 2022]), as shown in Figure 1 (a).

Without a doubt, we can apply these works to zero-shot activity recognition but will encounter the following issues. i) The samples of the support set are retrieved from the large database or generated via large generative models according to class names, in which the quality of visual samples is limited and the cost of video data generation is expensive; ii) TPT relies heavily on high-quality augmentation views to calibrate the embedding of class labels. However, on the one hand, it is difficult to mine rich and effective motion semantics via directly extending spatial augmentation to video data with high redundancy; on the other hand, simple class names are insufficient to describe human activity semantics, which may limit the effectiveness of the semantic consistency objective.

To this end, we propose a Dual Temporal-Sync Test-time Prompt Tuning framework that tunes the class embedding with different temporal structures in multiple steps supervised by dual semantic consistency, as shown in Figure 1 (b). This framework consists of the following two modules. **Temporal-Sync Test-time Prompt Tuning (TS-TPT)**: To capture rich and effective motion features, we append learnable prompts to the original class embedding and tune these prompts with different temporal scale video sequences in multiple steps. Specifically, we build a set of video augmentations via spatial AugMix [Hendrycks *et al.*, 2019] and then select high-confident augmentations for further tuning. We sample frames from selected video augmentations synchronously at the same scales in a single step but at different scales in different steps. Details are in Sec 4.1. **Dual Semantic Consistency (DSC)**: To more accurately calibrate the motion semantic in the language space, we adopt both original class names of human activities and the corresponding explanations enquired from LLMs for semantic consistency. Learnable prompts are equipped and optimized via different backward propagation solutions including Parallel, Hybrid, and Switch Backward. Details are in Sec 4.2. We extensively evaluate the proposed framework on various benchmarks and quantitative and qualitative results show its substantial improvements and strong interpretability.

2 Related Work

2.1 Zero-shot Activity Recognition

Zero-shot action recognition aims to identify novel action classes that do not appear during model training, which is more appropriate for practical applications. The typical pipeline of zero-shot matching [Brattoli *et al.*, 2020] extracts the video and class text features and then computes the distance between them in the feature space for final prediction. To build robust representation, earlier works represent the video via handcraft attributes [Liu *et al.*, 2011] or building knowledge graphs on object tags [Gao *et al.*, 2019], and enhance the class text features with the help of object tags [Gao *et al.*, 2019] or elaborative descriptions [Chen and Huang, 2021] retrieved from dictionaries or Wikipedia according to class names. Some recent works adapt the large pre-trained model (*e.g.*, CLIP) to zero-shot activity recognition via fully tuning (*e.g.*, ViFi-CLIP [Rasheed *et al.*, 2023]) or partial visual prompt tuning (*i.e.*, ActionClip [Wang *et al.*, 2021]) or partial visual/text prompt tuning (*i.e.*, XCLIP [Ni *et al.*, 2022]). The above works still need to be fine-tuned on lots of video samples during the training phase. However, this work is the first to append several learnable prompts to the label and tune them via a single sample during inference.

2.2 Prompt Tuning for VLMs

In recent years, it has become a trend to adapt Visual-Language Models (VLMs) pre-trained on large-scale visual-text corpora to various tasks in computer vision [Radford *et al.*, 2021; Chen *et al.*, 2020]. As a heuristic way, prompt tuning appends some learnable prompts to inputs and then tuning them on the few training samples (training-time tuning) [Zhou *et al.*, 2022c; Zhou *et al.*, 2022b; Gao *et al.*, 2023] or a single test sample without annotation (test-time tuning) [Shu *et al.*, 2022]. For instance, CoOp [Zhou *et al.*, 2022c] inserts a set of learnable vectors to class embedding in different positions as a prompt context which is directly optimized by the classification loss. To avoid the over-fitting issue, CoCoOp [Zhou *et al.*, 2022b] is proposed to learn a lightweight network to make the prompt conditioned on model inputs. Inspired by this, recent work [Sun *et al.*, 2022] decomposes backbone parameters into successive matrices and efficiently tunes part of them for few-shot visual recognition, achieving promising results. To get rid of annotation cost, some recent works [Huang *et al.*, 2022; Zhou *et al.*, 2022a] optimize the prompts in an unsupervised manner. However, this line of work still requires downstream training data, which is not compatible with the zero-shot setting. To this end, Test-time Prompt Tuning (TPT) [Shu *et al.*, 2022] is proposed to learn adaptive text prompts dynamically with a single test sample during inference, which reduces the cost of labeling and calculation. In this work, we extend TPT to the video domain, taking into account the diversity of motion semantics, for zero-shot activity recognition.

2.3 Test-time Tuning

Test-time Tuning (TTT) seeks to leverage unlabeled test data to enhance the model's generalization capabilities for the target task in the presence of data distribution shifts, which has been well explored for several applications [Shocher *et al.*, 2018; Nitzan *et al.*, 2022; Xie *et al.*, 2023]. The key to TTT is to formulate an effective test-time objective. In the image domain, previous works apply rotation prediction [Sun *et al.*, 2020; Liu *et al.*, 2021] or image reconstruction [Gandelsman *et al.*, 2022; Wang *et al.*, 2023] as the self-supervision task to optimize the model during inference. Besides, TENT [Wang *et al.*, 2020] utilize entropy minimization [Roy *et al.*, 2019; Saito *et al.*, 2019] to tune the parameters of BN layers to overcome the distribution shift [Zhang *et al.*, 2022a]. To extend

TENT [Wang *et al.*, 2020] to scenarios involving a single test sample, MEMO [Zhang *et al.*, 2022a] proposes utilizing data enhancement methods to generate diverse enhanced views from the single test sample for test-time tuning. In the video domain, recent works attempt to extend TTT to video data via self-supervised dense tracking [Azimi *et al.*, 2022] or segmentation on video streams [Wang *et al.*, 2023]. However, the optimization cost of tracking and segmentation is higher than entropy minimization (on class predictions). Therefore, this work applies entropy minimization as the objective function for prompt tuning but optimizes it at different scale video sequences in multiple steps.

3 Preliminary

3.1 Problem Statement

Given a set of N class labels $Y_{\text{u}} = \{y_0, y_1, \cdots, y_N\}$, the goal of zero-shot activity recognition is to predict the label of a testing video sequence $\boldsymbol{V}_{\text{test}}$ as $\boldsymbol{y}_i \in Y_{\text{u}}$. Specifically, a zero-shot pipeline i) adopts pre-trained models $\mathcal{F}_{\text{vis}}$ and $\mathcal{F}_{\text{txt}}$ respectively to encode the input video and class labels into visual and textual features with the same dimension, ii) then calculates the distance among them in feature space, and the nearest category is regarded as the prediction result.

Notably, the conventional methods finetunes $\mathcal{F}_{\text{vis}}$ and $\mathcal{F}_{\text{txt}}$ on an extra large video dataset (*e.g.*, K400) with the labels of Y_{s} and have the constraint of $Y_{\text{s}} \cap Y_{\text{u}} = \varnothing$. However, inspired by [Shu *et al.*, 2022], this work focuses on directly applying an arbitrary VLM pre-trained model (such as CLIP [Radford *et al.*, 2021] and BIKE [Wu *et al.*, 2023]) to identify the human activity in downstream tasks, which gets rid of training data and is more practical.

3.2 Test-time Prompt Tuning

Zero-shot Matching. Given a test video $\boldsymbol{V}_{\text{test}}$ and all class labels Y_u, we first encode them as $\boldsymbol{f}^{\text{vis}} = \mathcal{F}_{\text{vis}}(\boldsymbol{v})$ and $\boldsymbol{f}_i^{\text{txt}} = \mathcal{F}_{\text{txt}}(\boldsymbol{y}_i)$. For zero-shot visual recognition, we calculate the probability that the test video $\boldsymbol{v}$ belongs to category $\boldsymbol{y}_i$ as

$$p(\boldsymbol{y}_i|\boldsymbol{V}_{\text{test}}) = \frac{\exp(\text{sim}(\boldsymbol{f}^{\text{vis}}, \boldsymbol{f}_i^{\text{txt}})/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\boldsymbol{f}^{\text{vis}}, \boldsymbol{f}_j^{\text{txt}})/\tau)}. \quad (1)$$

Here, $\text{sim}(\cdot)$ is usually implemented by computing cosine similarity between the two feature vectors, N is the total number of class labels, and τ is the temperature parameter.

Prompt Tuning. Large pre-trained models have strong semantic representation capabilities but still need to be further generalized to downstream tasks that may contain out-of-distribution samples. For zero-shot testing, a common method is to add learnable prompts to class labels and tune these prompts with only a single sample in a self-supervised manner, which is called Test-time Prompt Tuning [Shu *et al.*, 2022]. The optimization objective can be formulated as

$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{p}} \mathcal{L}(\mathcal{F}, \boldsymbol{p}, \boldsymbol{v}_{\text{test}}), \quad (2)$$

where $\mathcal{F}$ is a pre-trained model consisting of visual and textual encoders. $\boldsymbol{p}$ is the learnable prompts appended to the tokenized text labels as $\{\boldsymbol{p}, \boldsymbol{y}_i\}$.

In the zero-shot inference phase, only one unlabeled test video is given to match all class labels, thus we can only optimize the above objective function in a self-supervised manner. Hence, Shu *et al.* implemented the loss function $\mathcal{L}$ via minimizing the averaged entropy of predictions from K augmented views as follows.

$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{p}} -\sum_{i=1}^{N}\hat{p}_{\boldsymbol{p}}(\boldsymbol{y}_i|\boldsymbol{V}_{\text{test}})\log\hat{p}_{\boldsymbol{p}}(\boldsymbol{y}_i|\boldsymbol{V}_{\text{test}}), \quad (3)$$

$$\hat{p}_{\boldsymbol{p}}(\boldsymbol{y}_i|\boldsymbol{V}_{\text{test}}) = \frac{1}{K}\sum_{j=1}^{K}p_{\boldsymbol{p}}(\boldsymbol{y}_i|\texttt{Aug}_j^{\text{S}}(\boldsymbol{V}_{\text{test}})). \quad (4)$$

Here, $\texttt{Aug}_j^{\text{S}}$ denotes the random spatial augmentation and $p_{\boldsymbol{p}}(\boldsymbol{y}_i|\texttt{Aug}_j(\boldsymbol{V}_{\text{test}}))$ is the prediction probabilities calculated between the j-th video augmentation and i-th class label with the prompt $\boldsymbol{p}$.

To reduce the noise from random augmentations, Shu *et al.* [Shu *et al.*, 2022] further select confident samples whose prediction entropy is lower than the threshold θ from K augmentation views as

$$\hat{p}_{\boldsymbol{p}}(\boldsymbol{y}_i|\boldsymbol{V}_{\text{test}}) = \tfrac{1}{\rho K}\textstyle\sum_{i=1}^{K}\mathbb{1}[\mathcal{H}(p_i) \le \theta]p_{\boldsymbol{p}}(\boldsymbol{y}_i|\texttt{Aug}_i^{\text{S}}(\boldsymbol{V}_{\text{test}})). \quad (5)$$

Here, $\mathcal{H}$ calculates the self-entropy of the prediction on an augmented view and $\mathbb{1}[\cdot]$ is the indicator function.

4 Methodology

Inspired by Tese-time Prompt Tuning (TPT), this work aims to tune the embedding of activity labels for zero-shot activity recognition during test time. However, monotonous spatial augmentation and short class names used in the standard TPT framework are insufficient for tuning complicated motion semantics in the language space. To this end, this work proposes a Dual Temporal-Sync Test-time Prompt Tuning (DTS-TPT) framework consisting of two core modules (*i.e.*, Temporal-Sync Test-time Prompt Tuning (TS-TPT) and Dual Semantic Consistency (DSC)) as shown in Figure 2. TS-TPT builds multi-scale video features from all spatial-augmented videos synchronously and then tunes the prompts with video features with different temporal scales in multiple steps. In each tuning step, DSC minimizes the average entropy of predictions from all augmented samples with both class names and descriptions generated by LLMs [OpenAI, 2023].

4.1 Temporal-Sync Test-time Prompt Tuning

As mentioned above, the key to TPT is augmenting the test visual sample into multiple views for tuning class embedding. For the sake of clarity, we firstly rewrite the standard TPT (*i.e.*, Eq. (2)) as

$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{p}, \mathcal{F}(\texttt{Aug}^{\text{S}}(\boldsymbol{v}_{\text{test}}))), \quad (6)$$

where $\mathcal{F}(\texttt{Aug}^{\text{S}}(\boldsymbol{v}_{\text{test}}))$ denotes that TPT performs spatial augmentation $\texttt{Aug}^{\text{S}}$ on the test video and extracts the visual features from them via the pre-trained model $\mathcal{F}$.

Based on this, how to augment the video data is the difficulty in adapting TPT to zero-shot activity recognition. A

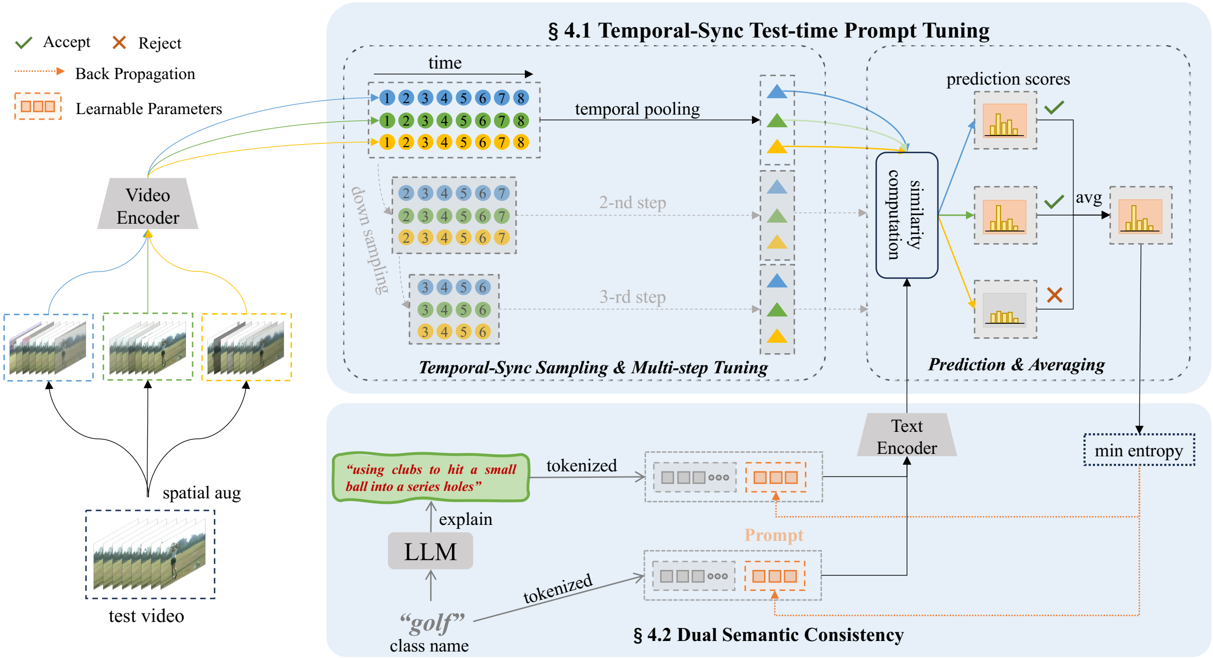


Figure 2: Overview of the proposed framework. We first construct multiple spatial augmentation views from the given video sequence via AugMix (only three views are shown for simplicity). The framework consists of two core modules, **i) Temporal-Sync Test-time Prompt Tuning**: builds multiple temporal-scale video features, and then calculates the averaged predictions in multiple steps for subsequent tuning; **ii) Dual Semantic Consistency**: In each tuning step, both the original class names and the LLM-generated description are wrapped with learnable prompts for similarity calculations with all spatial-augmented videos. After that, we minimize the averaged entropy from confident prediction scores, following [Shu *et al.*, 2022].

straightforward approach is to perform spatial ($\mathtt{Aug}^{\mathrm{S}}$) and temporal ($\mathtt{Aug}^{\mathrm{T}}$) augmentation on the video data and combine them for a single-step optimization of TPT as

$$\boldsymbol{p}^{*} = \arg\min_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{p}, \mathcal{F}(\mathtt{Aug}^{\mathrm{T}}(\mathtt{Aug}^{\mathrm{S}}(\boldsymbol{v}_{\text{test}})))). \tag{7}$$

However, spatio-temporal augmented videos are usually different from the original video and they should not be directly used for consistency constraints, as shown in Table 5.

To this end, we designed a multi-step optimization strategy with temporal synchronization, *i.e.*, the same temporal structure should be used in single-step optimization, but different temporal structures should be used in different steps.

Single-step Tuning. Formally, we sample the video features extracted by $\mathcal{F}(\mathtt{Aug}^{\mathrm{S}}(\boldsymbol{v}_{\text{test}}))$ with different temporal scales and tune the prompt in a single step as

$$\boldsymbol{p}^{*} = \arg\min_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{p}, \mathtt{Aug}^{\mathrm{T}}(\mathcal{F}(\mathtt{Aug}^{\mathrm{S}}(\boldsymbol{v}_{\text{test}})))). \tag{8}$$

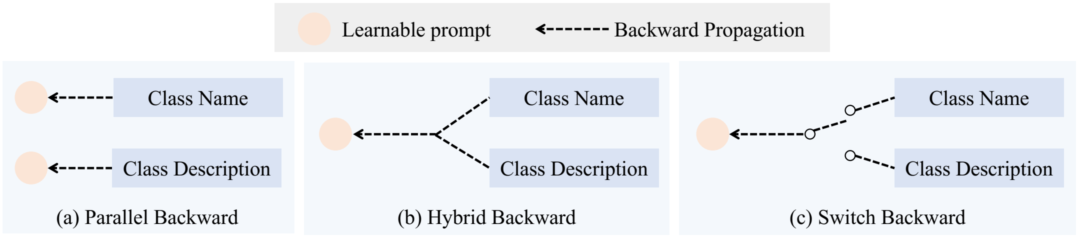
Here, $\mathtt{Aug}^{\mathrm{T}}$ aims to further augment the given spatial views in the temporal domain, and can essentially be viewed as a Temporal Sampling function. We implement $\mathtt{Aug}^{\mathrm{T}}$ via Video-version AugMix [Hendrycks *et al.*, 2019] which simply extends the spatial AugMix technology to the spatial-temporal data.

Multiple-step Tuning. Given a test video, we aim to tune the learnable prompt multiple times by video sequences with different temporal structures augmented from the original test video. Thus, there are two core design points, i) Tuning Times: How many times should we tune? ii) Tuning Structure: What temporal structure (*i.e.*, $\mathtt{Aug}^{\mathrm{T}}$) should be used each time?

Algorithm 1 Pseudocode of DTS-TPT

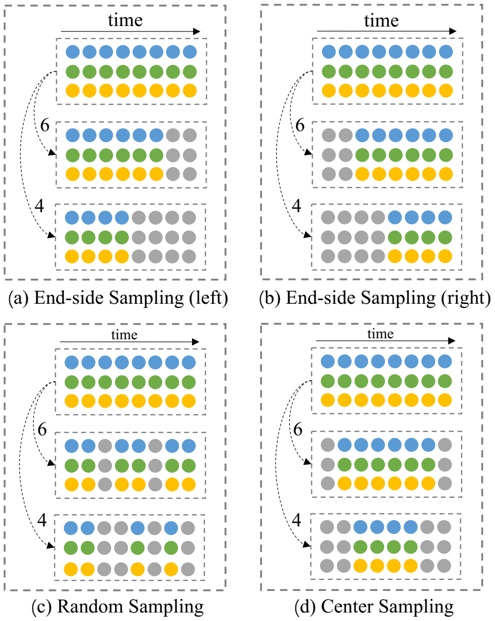
```
Input: a single test video $v_{\text{test}}$, class names $\{y_j\}$, learnable prompts $\boldsymbol{p}$, video encoder $\mathcal{F}_{\text{video}}$, and text encoder $\mathcal{F}_{\text{text}}$.
Perform spatial augmentation on the test video $K$ times and extract the visual features as $\boldsymbol{x}_k^{\mathrm{S}} = \mathcal{F}_{\text{video}}(\mathtt{Aug}_k^{\mathrm{S}}(\boldsymbol{v}_{\text{test}}))$, thus we get a set of augmented video features $\boldsymbol{X}^{\mathrm{S}} = \{\boldsymbol{x}_1^{\mathrm{S}}, \boldsymbol{x}_2^{\mathrm{S}}, \boldsymbol{x}_K^{\mathrm{S}}\}$.
Extract text features $\boldsymbol{x}_j^{\text{text}} = \mathcal{F}_{\text{text}}(\{\boldsymbol{p}, y_j\})$.
for $m = 1, 2, \cdots, M$ do
    Temporal augmentation $\boldsymbol{X}_m^{\mathrm{T}} = \mathtt{Aug}_i^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{S}}) \in \mathbb{R}^{K \times d}$;
    Calculate the predictions $p_{\boldsymbol{p}} \in \mathbb{R}^{K \times N}$ based on $\boldsymbol{X}_m^{\mathrm{T}}$
    Select confident ones $\hat{p}_{\boldsymbol{p}} \in \mathbb{R}^{K' \times N}$ with Eq (5);
    Minimize the averaged entropy of $\hat{p}_{\boldsymbol{p}}$ with Eq (3)
    Perform backward propagation for $\boldsymbol{p}$.
end for
```

For tuning structure, we gradually sample frame-level features from the sequence via different strategies, including *Random Sampling*, *Center Sampling*, *Left-side Sampling*, and *Right-side Sampling*, as shown in Figure 4. Based on this, we try different tuning times in the experiment, details in Sec-



(a) Parallel Backward (b) Hybrid Backward (c) Switch Backward

Figure 3: Different backward paradigms for dual semantic consistency.



(a) End-side Sampling (left) (b) End-side Sampling (right)

(c) Random Sampling (d) Center Sampling

Figure 4: Different methods of Temporal-Sync Sampling.

tion 5.4. To facilitate the understanding, we present the whole tuning process in Algorithm 1 in the form of pseudo-code.

4.2 Dual Semantic Consistency

Since the semantics of the same behavior are diverse in visual space, it is insufficient to describe human behaviors with short class names. A straightforward solution is to use dictionaries or handcrafted templates to empirically enhance these class names, but this will ultimately be limited by the size of the predefined knowledge database. This work aims to apply Large Language Models to explain the meaning of each human behavior in detail, which will greatly facilitate cross-modal alignment of behavioral semantics.

Specifically, we ask LLMs the question of "If you're an expert in video classification. Please answer the following questions in one or two sentences. What does [CLASS NAME] mean?". After that, we clean up the noisy words in the answers from LLMs and limit the descriptions to 77 words (max-length limit of Tokenizer). The explanation of each CLASS NAME generated by LLMs can be found in Supplementary Materials.

Moreover, how to optimize the tuning objective defined in Eq. (8) with the class name and the corresponding explanations becomes an issue. In this work, we provide three different backward propagation for semantic consistency, including Parallel, Hybrid, and Switch Backward. Notably, Parallel Backward tunes two sets of learnable prompts separately, but Hybrid and Switch Backward tunes the same prompts serially or alternately, as shown in Figure 3. Comparison results can be found in Table 6.

5 Experiments

5.1 Datasets

HMDB-51 [Kuehne *et al.*, 2011] contains 51 activity categories and $6,766$ manually annotated videos extracted from various sources ranging from digitized movies to YouTube. **UCF-101** [Soomro *et al.*, 2012] collects a total of $13,320$ real activity videos in 101 categories from YouTube, totaling more than 27 hours. All categories can be classified into 5 types: Body motion, Human-human interactions, Human-object interactions, Playing instruments, and Sports. **Kinetics-600** [Carreira *et al.*, 2018] is an extension of the Kinetics-400 [Kay *et al.*, 2017] and consists of approximately 480K videos covering 600 categories. Each video is a 10-second activity clip annotated from the original YouTube video. **ActivityNet** [Caba Heilbron *et al.*, 2015] provides a large-scale video dataset covering the 200 actions most relevant to human daily life. It contains a total of $19,994$ untrimmed videos, with 137 videos per category.

5.2 Setup

We sample $T = 8$ frames from test video and augment them $K - 1$ ($K = 32$) times using AugMix [Hendrycks *et al.*, 2019]. We create 3 learnable tokens as text prompts and initialize them as "an action of". We employ CLIP with ViT-B/16 as the visual encoder and use the corresponding textual encoder of CLIP for text encoding. Our model does not need any training data such as K400. For each inference, we compute predictions based on a batch of 32 augmented views (including the original one) and then select top 20% confident predictions for further optimization. We adopt the AdamW optimizer with a learning rate of 0.001.

5.3 Comparisons with State-of-the-art Methods

We compare the proposed DTS-TPT with existing zero-shot video activity recognition methods on four benchmarks with 8 frames, as shown in Table 1. Notably, DTS-TPT does not need training data but still achieves better generalization when utilizing a pre-trained model such as BIKE [Wu *et*

Method	**Encoder**	**HMDB-51**	**UCF-101**	**Kinetics-600**	**ActivityNet**
Uni-modal zero-shot video recognition models					
ER-ZSAR [Chen and Huang, 2021]	TSM	35.3 ± 4.6	51.8 ± 2.9	42.1 ± 1.4	–
JigsawNet [Le and Li, 2019]	R(2+1)D	38.7 ± 3.7	56.0 ± 3.1	–	–
E2E [Brattoli *et al.*, 2020]	R(2+1)D	29.8	44.1	–	26.6
ResT [Lin *et al.*, 2022]	Resnet-101	41.1 ± 3.7	58.7 ± 3.3	–	32.5
Adapting pre-trained CLIP					
ActionCLIP [Wang *et al.*, 2021]	ViT-B/16	40.8 ± 5.4	58.3 ± 3.4	66.7 ± 1.1	–
Vita-CLIP [Wasim *et al.*, 2023]	ViT-B/16	48.6 ± 0.6	75.0 ± 0.6	67.4 ± 0.5	–
A5 [Ju *et al.*, 2022]	ViT-B/16	44.3 ± 2.2	69.3 ± 4.2	55.8 ± 0.7	–
XCLIP [Ma *et al.*, 2022]	ViT-B/16	44.6 ± 5.2	72.0 ± 2.3	65.2 ± 0.4	–
Tuning pre-trained CLIP					
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 [Rasheed *et al.*, 2023]	ViT-B/16	50.9 ± 0.7	74.9 ± 0.6	67.7 ± 1.1	–
BIKE + DTS-TPT (Ours)	ViT-B/16	**54.4 ± 0.8**	**79.4 ± 0.7**	**68.2 ± 1.2**	**76.8 ± 0.8**

Table 1: Comparisons with the state-of-the-art methods. Top-1 accuracy using single-view inference is reported.

al., 2023]. Following [Ni *et al.*, 2022; Rasheed *et al.*, 2023; Lin *et al.*, 2023], we report the mean and standard deviation of results on three official validation sets.

Our approach outperforms uni-modal zero-shot activity recognition methods by a significant margin, such as ER-ZSAR and JigsawNet, which train on Kinetics-400 and need crawled descriptions of activity classes with manual correction. Moreover, DTS-TPT is superior to recent methods, either adapted directly from CLIP or tuned from CLIP with K400 for zero-shot recognition. As a baseline for our approach, BIKE fine-tunes the pre-trained CLIP model with video attributes and activity categories. Our approach achieves state-of-the-art results on HMDB51, UCF101, Kinetices-600, and ActivityNet in terms of top-1 accuracy, surpassing BIKE by a significant margin. Besides, the proposed framework is compatible with different VLMs with different backbones, as shown in Table 3.

5.4 Ablation Study

Component Analysis. To illustrate the effectiveness of each component of our approach, we conduct ablation experiments on HMDB51 and UCF101 datasets. As shown in Table 2, the proposed modules (TS-TPT and DSC) bring obvious gains respectively, and it is better to use them together, which indicates that they are complementary.

TS-TPT	**DSC**	**HMDB-51**	**UCF-101**
✗	✗	49.1	77.4
✓	✗	51.6	78.0
✗	✓	52.4	79.1
✓	✓	**54.4**	**79.4**

Table 2: Effect of different components of our approach.

Temporal-Sync Sampling. As shown in Figure 4, we have designed four different temporal sampling strategies for TS-TPT. For a fair comparison, all the variants employ Vanilla CLIP as a visual encoder and sample 6 frames out of 8, and perform TS-TPT twice. From Table 4, we find that the results of three non-random sampling methods are slightly better than random sampling. It indicates that TS-TPT is not affected by the temporal sampling strategy.

Backbone	**Method**	**HMDB-51**	**UCF-101**
ViT-B/16	Vanilla CLIP	41.2	69.9
	+ TPT	44.0	70.8
	+ DTS-TPT (Ours)	**46.0**	**72.3**
	BIKE	49.1	77.4
	+ TPT	50.6	77.7
	+ DTS-TPT (Ours)	**54.4**	**79.4**
ViT-L/14	Vanilla CLIP	46.3	77.3
	+ TPT	47.9	80.4
	+ DTS-TPT (Ours)	**49.3**	**81.0**
	BIKE	58.9	85.5
	+ TPT	55.6	84.9
	+ DTS-TPT (Ours)	**60.4**	**87.9**

Table 3: Adaption to different pre-trained models.

Sampling Strategy	**HMDB-51**	**UCF-101**
Random	43.8	70.9
Left-side	44.2	71.3
Right-side	**44.8**	**71.5**
Center	44.5	71.3

Table 4: Effect of different temporal sampling strategies.

Comparison with TPT. We compare the proposed TS-TPT with the previous standard TPT in Table 5. Based on the results, we can conclude the following: i) Applying simple spatio-temporal augmentation (ST-TPT) cannot bring improvements, compared with TPT. ii) The performance of the TS-TPT is gradually improved with the number of optimiza-

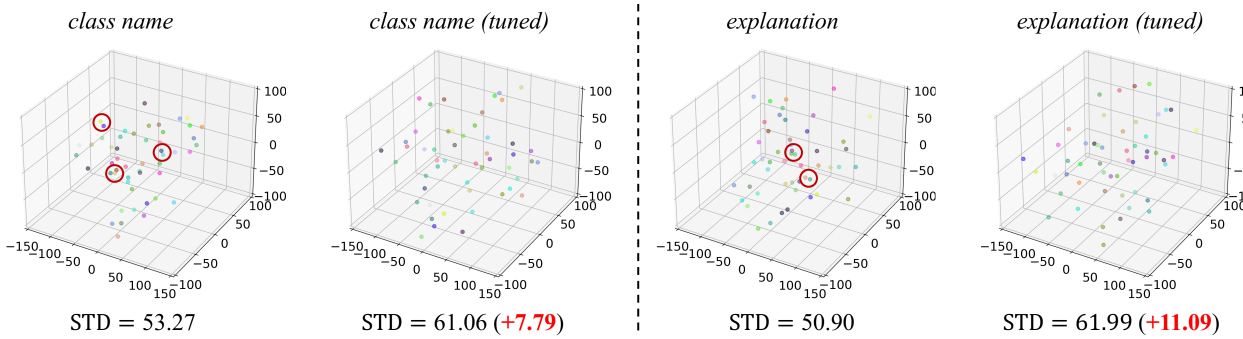


Figure 5: Visualization of the feature distribution of class names and the corresponding explanation.

Dataset	Method	Steps			
		1	2	3	4
UCF-101	TPT	71.2	71.2	70.8	70.8
	ST-TPT	71.2	71.3	71.2	70.7
	Temp.-Sync TPT	**71.2**	**71.5**	**71.7**	**71.0**
HMDB-51	TPT	44.2	44.0	44.0	43.4
	ST-TPT	44.3	44.0	43.8	43.2
	Temp.-Sync TPT	**44.2**	**44.8**	**45.6**	**44.9**

Table 5: Comparison with TPT. "TPT" applies AugMix to each frame independently and "ST-TPT" applies it to all frames.

Semantic Consistency		HMDB-51	UCF-101
Single	class name	51.6	78.0
	explanation	**53.0**	**78.9**
Dual	switch	52.9	78.9
	parallel	54.4	79.2
	hybrid	**54.6**	**79.4**

Table 6: Different Implementation of Dual Semantic Consistency.

tion steps increasing from 1 to 3. This indicates that tuning prompts collaboratively with video features at different temporal scales is more likely to approximate the latent semantic distribution of labels. iii) Depending on the original design, TPT can be optimized repeatedly using the same scale video features, but more update steps do not lead to better performance, and sometimes even worse.

Context for Text Prompts. We investigate the effect of different context (class name and explanation) for text prompts in Table 6. The explanation generated through LLM significantly improves performance compared with using class names alone, suggesting that fine-grained text context helps prompt tuning of diverse motion semantics.

Explanation from Different LLMs. We study the effect of description generated by different LLMs in Table 7. We observed that LLM with a larger number of parameters (*i.e.*, Gemini and GPT 3.5) resulted in a more significant performance improvement. Notably, the description generated by GPT-3.5 brings the best overall performance, thus we used it for the final model.

Text branch	Params	HMDB-51	UCF-101
Baseline: class name	-	51.6	78.0
+Llama-2 [Touvron *et al.*, 2023]	13B	49.5	78.1
+Vicuna [Chiang *et al.*, 2023]	13B	50.4	77.9
+Gemini [Team *et al.*, 2023]	1800B	52.7	78.9
+GPT 3.5 [OpenAI, 2023]	175B	**53.0**	**78.9**

Table 7: Effect of different large language models.

Backward Types of DSC. We have designed three backward types of dual semantic consistency (as shown in Figure 3). As reported in Table 6, Switch Backward cannot combine the benefit from two types of semantic consistency (class name and explanation), but optimizing the Dual Semantic Consistency via Parallel and Hybrid Backward has a better performance compared to single semantic consistency, Therefore, we used Hybrid Backward for the final model.

Visualization. We apply t-SNE to visualize the original features and tuned features of class names and their corresponding explanation generated by GPT 3.5, as shown in Figure 5. At the same time, we calculated the standard deviation of the data points in different feature distributions. We found that the standard variance of the features increases significantly and some confused features (in red circles) no longer exist after tuning, indicating that divergent feature distribution is more conducive to cross-modal alignment.

6 Conclusion

This work has managed to adapt Test-time Prompt Tuning (TPT) to zero-shot activity recognition by fully exploiting the temporal characteristics and semantic diversity of human behaviors. We developed the Dual Temporal-Sync Test-time Prompt Tuning framework which tunes the prompts in multi-steps on different temporal scales video features with dual semantic consistency. Extensive experiments on various benchmarks demonstrated that DTS-TPT can effectively improve the generalization ability of the pre-trained model on video data without training data/annotation and is compatible with various VLMs. In the future, it is critical to improve the efficiency of multi-step tuning and class name explanation in practical applications.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (Grant No. 62302208, 62222207, 62072245, and 61925204), the Natural Science Foundation of Jiangsu Province (Grant No. BK20211520), the Postdoctoral Fellowship Program of CPSF (Grant No. GZB20230302), the China Postdoctoral Science Foundation (Grant No. 2023TQ0151 and 2023M731596), the Jiangsu Funding Program for Excellent Postdoctoral Talent (Grant No. 2023ZB256), and the Fundamental Research Funds for the Central Universities (Grant No. 14380109 and 14913211). Rui Yan and Hongyu Qu equally contributed to this work, and Wenbin Li is the corresponding author.

References

[Azimi *et al.*, 2022] Fatemeh Azimi, Sebastian Palacio, Federico Raue, Jörn Hees, Luca Bertinetto, and Andreas Dengel. Self-supervised test-time adaptation on video data. In *WACV*, pages 3439–3448, 2022.

[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.

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

[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.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020.

[Chiang *et al.*, 2023] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna.lmsys.org (accessed 14 April 2023)*, 2023.

[Gandelsman *et al.*, 2022] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-time training with masked autoencoders. In *NeurIPS*, pages 29374–29385, 2022.

[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.

[Gao *et al.*, 2023] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *IJCV*, pages 1–15, 2023.

[Hendrycks *et al.*, 2019] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

[Huang *et al.*, 2022] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022.

[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, 2022.

[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 *ICCV*, pages 2556–2563, 2011.

[Le and Li, 2019] Canyu Le and Xin Li. Jigsawnet: Shredded image reassembly using convolutional neural network and loop-based composition. *TIP*, 28(8):4000–4015, 2019.

[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.

[Lin *et al.*, 2023] Wei Lin, Leonid Karlinsky, Nina Shvetsova, Horst Possegger, Mateusz Kozinski, Rameswar Panda, Rogerio Feris, Hilde Kuehne, and Horst Bischof. Match, expand and improve: Unsupervised finetuning for zero-shot action recognition with language knowledge. In *ICCV*, pages 2851–2862, October 2023.

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

[Liu *et al.*, 2021] Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? In *NeurIPS*, pages 21808–21820, 2021.

[Ma *et al.*, 2022] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACMMM*, pages 638–647, 2022.

[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, 2022.

[Nitzan *et al.*, 2022] Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar

Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. *TOG*, 41(6):1–10, 2022.

[OpenAI, 2023] OpenAI. Chatgpt, 2023. https://openai.com/blog/chatgpt/, Last accessed on 2024-01-13.

[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 *ICML*, pages 8748–8763, 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.

[Roy *et al.*, 2019] Subhankar Roy, Aliaksandr Siarohin, Enver Sangineto, Samuel Rota Bulo, Nicu Sebe, and Elisa Ricci. Unsupervised domain adaptation using feature-whitening and consensus loss. In *CVPR*, pages 9471–9480, 2019.

[Saito *et al.*, 2019] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *ICCV*, pages 8050–8058, 2019.

[Shocher *et al.*, 2018] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *CVPR*, pages 3118–3126, 2018.

[Shu *et al.*, 2022] 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. In *NeurIPS*, pages 14274–14289, 2022.

[Soomro *et al.*, 2012] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[Sun *et al.*, 2020] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, pages 9229–9248, 2020.

[Sun *et al.*, 2022] Yanpeng Sun, Qiang Chen, Xiangyu He, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Jian Cheng, Zechao Li, and Jingdong Wang. Singular value fine-tuning: Few-shot segmentation requires few-parameters fine-tuning. In *NeurIPS*, pages 37484–37496, 2022.

[Team *et al.*, 2023] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[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 and Chen, 2017] Qian Wang and Ke Chen. Zero-shot visual recognition via bidirectional latent embedding. *IJCV*, 124:356–383, 2017.

[Wang *et al.*, 2020] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

[Wang *et al.*, 2021] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

[Wang *et al.*, 2023] Renhao Wang, Yu Sun, Yossi Gandelsman, Xinlei Chen, Alexei A Efros, and Xiaolong Wang. Test-time training on video streams. *arXiv preprint arXiv:2307.05014*, 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.

[Xie *et al.*, 2023] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *PAMI*, 2023.

[Zhang *et al.*, 2022a] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *NeurIPS*, pages 38629–38642, 2022.

[Zhang *et al.*, 2022b] 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, 2022.

[Zhou *et al.*, 2022a] Chunting Zhou, Junxian He, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Prompt consistency for zero-shot task generalization. *arXiv preprint arXiv:2205.00049*, 2022.

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