TEST-V: TEst-time Support-set Tuning for Zero-shot Video Classification

Rui Yan[1,2] , **Jin Wang**[1] , **Hongyu Qu**[1] , **Xiaoyu Du**[1] , **Dong Zhang**[3] , **Jinhui Tang**[1] and **Tieniu Tan**[2]
[1]Nanjing University of Science and Technology
[2]Nanjing University
[3]Hong Kong University of Science and Technology
{ruiyan, wangjin-andy, quhongyu, duxy, jinhuitang}@njust.edu.cn, dongz@ust.hk, tnt@nju.edu.cn

Abstract

Recently, adapting Vision Language Models (VLMs) to zero-shot visual classification by tuning class embedding with a few prompts (Test-time Prompt Tuning, TPT) or replacing class names with generated visual samples (support-set) has shown promising results. However, TPT cannot avoid the semantic gap between modalities while the support-set cannot be tuned. To this end, we draw on each other's strengths and propose a novel framework, namely **TE**st-time **S**upport-set **T**uning for zero-shot **V**ideo Classification (**TEST-V**). It first dilates the support-set with multiple prompts (Multi-prompting Support-set Dilation, MSD) and then erodes the support-set via learnable weights to mine key cues dynamically (Temporal-aware Support-set Erosion, TSE). Specifically, **i) MSD** expands the support samples for each class based on multiple prompts inquired from LLMs to enrich the diversity of the support-set. **ii) TSE** tunes the support-set with factorized learnable weights according to the temporal prediction consistency in a self-supervised manner to dig pivotal supporting cues for each class. **TEST-V** achieves state-of-the-art results across four benchmarks and shows good interpretability.

1 Introduction

In the past few decades, the research on behavior recognition has made rapid progress and has been successfully applied in many fields, including intelligent robotics, security surveillance, and automatic driving. However, the number of behaviors existing methods can identify is still limited, which makes it difficult to meet the growing needs of practical applications. In recent years, the powerful ability of large-scale pre-trained Vision-Language models (VLMs) in zero-shot generalization has promoted the rapid development of zero-shot/general behavior recognition. VLMs bring hope for breaking through the perceptual boundaries of existing behavior recognition technologies.

Recently, it has become a trend to project the test video and class names into the joint feature space based on the pre-trained models, and then assign a label to the video according to its feature similarities. Although the amount of pre-training data is huge, the out-of-distribution issue cannot be thoroughly eliminated in the test environment. Hence, it is necessary to further tune the parameters or prompts for the pre-trained model from different modal perspectives.

To address this challenge, existing solutions can be roughly divided into the following two categories. i) **Training-based**: Conventional methods fine-tune *part of parameters* belonging/appended to the image-based pre-trained model (*e.g.*, CLIP [Radford *et al.*, 2021]) or *extra learnable prompts* appended to (visual/text) inputs, based on a mass of video data. ii) **Training-free**: To avoid high training costs, some recent Test-time Prompt Tuning (TPT)-based methods [Shu *et al.*, 2022; Yan *et al.*, 2024] append learnable prompts to visual/text input sequence and tune prompts in a self-supervised manner (*e.g.*, multi-view prediction consistency) based on the single test video during test time (Figure 1 (a)). Beyond that, some recent works [Udandarao *et al.*, 2023; Zhang *et al.*, 2021] build a video support-set according to unseen class names via retrieval or generation(*i.e.*, converts class names to videos) and performs intra-modality (video-video) alignment (Figure 1 (b)). TPT freezes the visual space and only tunes the text space, which cannot reduce the modality gap. On the contrary, the support-set-based methods reduce the gap via inter-modality alignment, but the supporting samples are fixed.

To this end, we proposed a novel framework namely **TE**st-time **S**upport-set **T**uning for **V**ideo classification (**TEST-V**) in Figure 1 (c). It builds a semantically diverse support-set via off-the-shelf text-video generation models and tunes learnable weights to select effective support samples from the set via two core modules. **i) Multi-prompting Support-set Dilation (MSD)**: generate multiple text prompts from the class names via LLMs and feed them into the text-video generation model (*i.e.*, LaVie [Wang *et al.*, 2023b]) for building a semantically diverse support set. **ii) Temporal-aware Support-set Erosion (TSE)**: tunes a few learnable weights to dynamically reshape the contribution of each video frame in a self-supervised manner, according to the multi-scale semantic consistency hypothesis in the temporal domain. Extensive experimental results show that **TEST-V** improves the

*Extended version on arxiv: https://arxiv.org/pdf/2502.00426

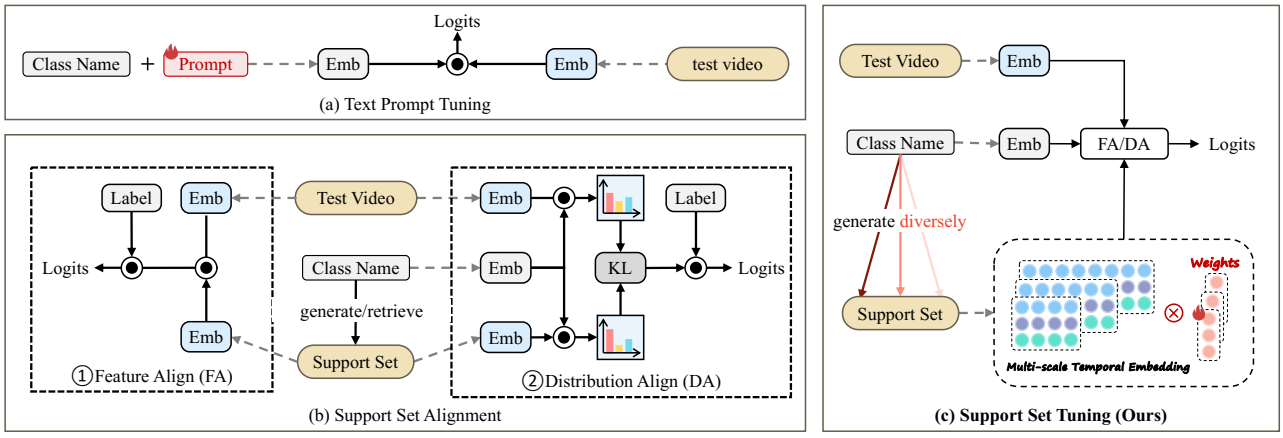

Figure 1: Innovation of the zero-shot activity recognition framework. a) Tuning the text input via the given test video in a self-supervised manner. b) Aligning the test video with the label based on the support set from feature similarity or predicted distribution similarity. c) This work combines the above thoughts to construct the support set diversely and tunes this set in a self-supervised manner to mine high-quality support samples.

state-of-art pre-trained VLMs, namely CLIP [Radford *et al.*, 2021], BIKE [Wu *et al.*, 2023] and ViFi-CLIP [Rasheed *et al.*, 2023] by 2.98%, 2.15%, and 1.83% absolute average accuracy respectively across four benchmarks.

Our main contributions can be summarized as follows:

- We propose a novel framework, namely **TE**st-time **S**upport-set **T**uning, which first dilates and then erodes the visual support-set to enhance the zero-shot generalization on activity recognition during test time.
- To ensure the diversity of the support-set, we propose the Multi-prompting Support-set Dilation (MSD) to generate videos via a text-to-video model as supporting samples for each class according to multiple descriptions.
- To mine critical supporting cues from the support-set, we proposed a Temporal-aware Support-set Erosion (TSE) module to tune factorized learnable weights with different temporal-scale features in multiple steps supervised by the prediction consistency.

2 Related Work

2.1 Zero-shot Activity Recognition

Zero-Shot Activity Recognition (ZSAR) aims to recognize actions that are not observed by the model during training, which is a non-trivial task yet useful for practical applications. Early work learn a joint embedding space by aligning video representation with corresponding textual semantic representation, *e.g.*, manually defined attributes [Liu *et al.*, 2011; Zellers and Choi, 2017], word embeddings of category names [Qin *et al.*, 2017; Shao *et al.*, 2020], and elaborative action descriptions [Chen and Huang, 2021; Qian *et al.*, 2022]. With the advent of deep neural networks, many works [Chen and Huang, 2021; Lin *et al.*, 2019; Lin *et al.*, 2022] make use of various video architectures (*e.g.*, I3D [Carreira and Zisserman, 2017] and TSM [Lin *et al.*, 2019]) to obtain high-level visual representation from videos and further project such visual representation into semantic embedding space. Recently, large-scale Vision-Language models (*e.g.*, CLIP [Radford *et al.*, 2021]) have shown great zero-shot transfer ability. Thus, many ZSAR solutions attempt to take advantage of CLIP to align video features and textual features, showing impressive performance in ZSAR. ActionCLIP [Wang *et al.*, 2021b], A5 [Ju *et al.*, 2022], and XCLIP [Ni *et al.*, 2022] adapt CLIP for videos by training additional components, *e.g.*, temporal transformer blocks and text prompts. ViFi-CLIP [Rasheed *et al.*, 2023], BIKE [Wu *et al.*, 2023], and MAXI [Lin *et al.*, 2023] fully fine-tune the encoder of CLIP to learn video-specific inductive biases, without introducing any additional learnable parameters. Unlike current approaches that rely heavily on many labeled video samples to fine-tune network parameters, our work pioneers the idea of building an efficient visual support set for zero-shot classification without training.

2.2 Adaptation of Visual-Language Models

Recently, pre-trained Vision-Language Models (VLMs) (*e.g.*, CLIP [Radford *et al.*, 2021], ALIGN [Jia *et al.*, 2021]) have shown impressive zero-shot transfer ability in recognizing novel categories. Thus, many works attempt to adapt VLMs for downstream tasks through prompt tuning or training-free methods. CoOp [Zhou *et al.*, 2022b] fine-tunes the learnable text prompts on downstream training data instead of using handcrafted templates [Zhou *et al.*, 2022b], effectively improving CLIP zero-shot performance. CoCoOp [Zhou *et al.*, 2022a] extends CoOp to learn input-conditioned prompts, leading to better generalization to unseen classes and different domains. Despite impressive, these works need access to labeled samples from the target distribution, and further train prompts over these samples. In addition, recent training-free methods such as TIP-Adapter [Zhang *et al.*, 2021] leverage a few labeled training samples as a key-value cache to make zero-shot predictions during inference, avoiding the conventional model parameter fine-tuning. Yet these methods still

rely on samples from the target distribution. To discard costly labeled data and training, SuS-X [Udandarao *et al.*, 2023] utilizes category name and text-to-image generation model to construct a visual support-set, thus enhancing zero-shot transfer abilities. Our test-time support-set tuning shares a similar spirit of pursuing a visual support-set curation strategy to adapt VLMs, but we strive to construct a diverse video support set guided by elaborated descriptions of each class name, and make full use of each sample in the support-set by dynamically adjusting the contribution of each frame.

2.3 Test-time Tuning

Test-time Tuning (TTT) [Shocher *et al.*, 2018; Nitzan *et al.*, 2022; Xie *et al.*, 2023] aims to improve the generalization capabilities of pre-trained models by utilizing test samples to fine-tune the model, especially when faced with data distribution shifts. One of the key challenges in TTT is designing effective optimization objectives. Early works [Sun *et al.*, 2020; Liu *et al.*, 2021] enhance the training process by integrating a self-supervised multitask loss. Besides, TENT [Wang *et al.*, 2021a] minimizes predictions' entropy on target data to improve generalization capability, yet requiring multiple test samples. To address this issue, MEMO [Zhang *et al.*, 2022] and TPT [Shu *et al.*, 2022] utilize multiple augmented views generated from a single test sample through data augmentation for further test-time tuning. Recent methods extend TTT to the video domain by self-supervised dense tracking [Azimi *et al.*, 2022] or aligning video features with different sampling rates [Zeng *et al.*, 2023]. Inspired by this, we tune the support-set via learnable parameters with temporal prediction consistency loss for mining reliable supporting cues.

3 Preliminary and Definition

3.1 Problem Statement

Zero-Shot Activity Recognition (ZSAR) aims to recognize unseen actions during model testing. Specifically, given a test video $\boldsymbol{V}_{\text{test}}$ and a set of class names $\boldsymbol{Y} = \{y_0, y_1, \cdots, y_c\}$, we employ the pre-trained visual ($\mathcal{F}_{\text{vis}}$) and text ($\mathcal{F}_{\text{txt}}$) encoders *e.g.*, CLIP [Radford *et al.*, 2021], to encode them and then calculate the feature similarity for prediction as follows,

$$\begin{aligned} \boldsymbol{f} = \mathcal{F}_{\text{vis}}(\boldsymbol{V}_{\text{test}}), \boldsymbol{W} = \mathcal{F}_{\text{txt}}(\boldsymbol{Y}), \\ \boldsymbol{Z}_{\text{ZS}} = \boldsymbol{f} \cdot \boldsymbol{W}^{\text{T}}, \boldsymbol{f} \in \mathbb{R}^{d}, \boldsymbol{W} \in \mathbb{R}^{C \times d}. \end{aligned} \tag{1}$$

Here, $\boldsymbol{f}$ and $\boldsymbol{W}$ denote visual and text features, respectively, and "$\cdot$" calculates their cosine similarity.

Although the pre-trained representation is robust enough, it cannot avoid the inherent semantic gap between visual and text features (*a.k.a*, modality gap). To this end, some recent methods (*e.g.*, TIP-Adapter [Zhang *et al.*, 2021], SuS-X [Udandarao *et al.*, 2023]) attempt collecting a set of visual samples (support-set) as the replacement of class names and perform zero-shot matching in the visual space for reduce such gap. We first briefly review such support-set-based zero-shot matching methods as below.

3.2 Support-Set-based Zero-shot Prediction

Support-Set Construction. Given a set of class name $\boldsymbol{Y}$, the text-to-image (`T2I`) generation model (*e.g.*, Stable Diffusion [Rombach *et al.*, 2022]) is employed to generate K images for each category to construct the support set. To obtain more diverse generated samples, SuS-X [Udandarao *et al.*, 2023] explains each class name in detail through LLMs (*e.g.*, GPT-3 [Brown *et al.*, 2020]) as input of `T2I` model following CuPL [Pratt *et al.*, 2023]. Thus, this support-set consists of $C \times K$ visual samples where C and K denote the number of classes and the number of generated samples for each class.

TIP-Adapter Inference. Based on the above support-set, CLIP's visual encoder $\boldsymbol{E}_{\text{v}}$ is applied to extract its visual features $\boldsymbol{F} \in \mathbb{R}^{CK \times d}$, and its labels are converted into the one-hot vector $\boldsymbol{L} \in \mathbb{R}^{CK \times C}$. TIP-Adapter aims to calculate the distance between the test video feature $\boldsymbol{f}$ and support videos' feature $\boldsymbol{F}$, and then make predictions ($\boldsymbol{Z}_{\text{TA}}$) with the help of label information L as follows,

$$\boldsymbol{Z}_{\text{TA}} = \exp(-\beta(1 - \boldsymbol{f}\boldsymbol{F}^{\text{T}}))\boldsymbol{L}. \tag{2}$$

Here, β adjusts the sharpness and affinities calculated from $\boldsymbol{f}\boldsymbol{F}^{\text{T}}$ is used as attention weights over $\boldsymbol{L}$ to achieve logits.

TIP-X Inference. To avoid the uncalibrated intra-modal feature distances in TIP-Adapter, TIP-X [Udandarao *et al.*, 2023] uses text modality as a bridge between visual modalities. TIP-X calculates the distance between the test video feature f and text features $\boldsymbol{W}$, and between support videos' feature F and text feature W, respectively. TIP-X applies KL-divergence to construct the affinity matrix for final prediction ($\boldsymbol{Z}_{\text{TX}}$) as follows,

$$\boldsymbol{Z}_{\text{TX}} = \psi(-\texttt{KL}(\boldsymbol{f}\boldsymbol{W}^{\text{T}} || \boldsymbol{F}\boldsymbol{W}^{\text{T}}))\boldsymbol{L}. \tag{3}$$

Here, $\boldsymbol{f}\boldsymbol{W}^{\text{T}}$ and $\boldsymbol{F}\boldsymbol{W}^{\text{T}}$ are normalized via softmax, $\texttt{KL}(\cdot)$ computes the similarity between two probabilities ($\texttt{KL}(\boldsymbol{P} \parallel \boldsymbol{Q}) = \sum_i \boldsymbol{P}_i \log \frac{\boldsymbol{P}_i}{\boldsymbol{Q}_i}$) and ψ is the scaling factor.

4 Methodology

4.1 Multi-prompting Support-set Dilation (MSD)

As shown in Figure 3, the performance of the support-set based on SuS-X is extremely saturated when the number of videos per category increases, suggesting that the generated video samples' representations are highly overlapping/similar in each class. This inspires this work to improve the diversity of generated samples to refine the semantic boundary of each class within the support-set. To this end, we propose the Multi-prompting Support-set Dilation (MSD) module to generate diverse video samples via a video-text generation model with different prompts.

Diversified Motion Description. Given a set of action classes $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^{C}$, we employ a large language model (LLM) (*e.g.*, ChatGPT [OpenAI, 2023]) to generate M different prompts for each class as

$$\boldsymbol{d}_i = \texttt{LLM}(\texttt{query\_template}(\boldsymbol{y}_i, M)). \tag{4}$$

Here, $\texttt{query\_template}(\cdot)$ is a crafted text template used to wrap each action class into a text query as the input of LLM. After traversing the entire action class set Y, we can achieve

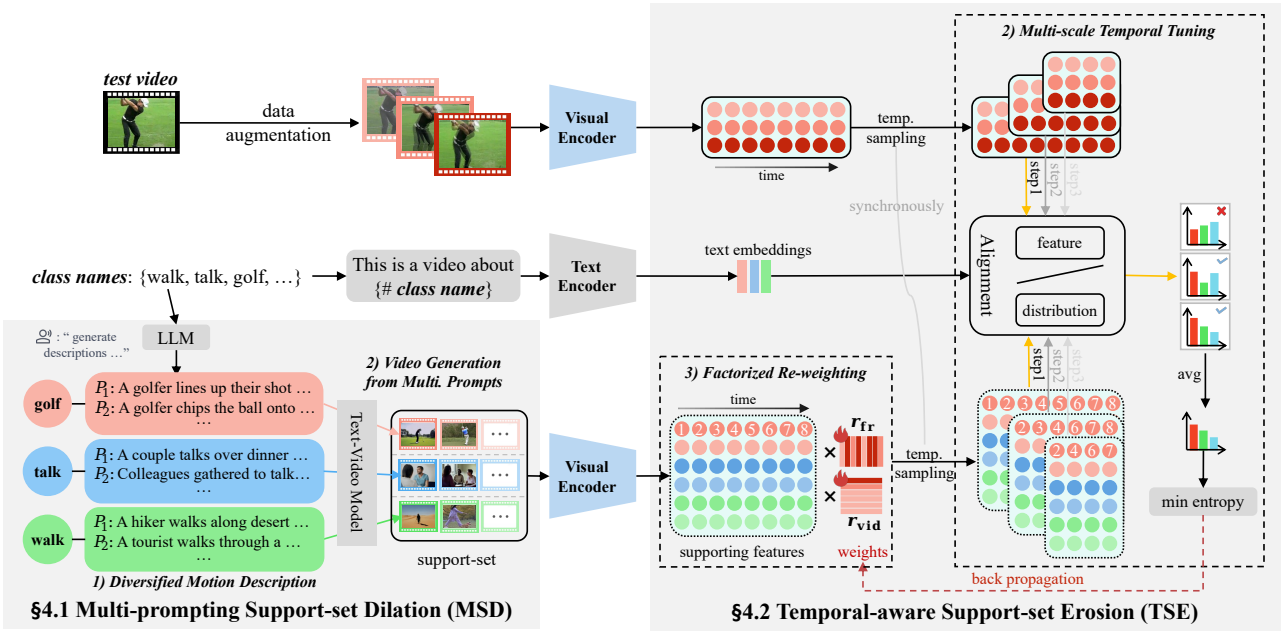

Figure 2: Overview of the proposed framework **TEST-V** which *first dilates and then erodes* the support set for zero-shot video classification. i) **Multi-prompting Support-set Dilation (MSD)**: It builds diversified motion description for each class name via the LLM and then generates video samples with these elaborate descriptions via the text-to-video generation model for constructing a diverse support set. ii) **Temporal-aware Support-set Erosion (TSE)**: Based on the visual feature of the given test video $\boldsymbol{f}$ and support set $\boldsymbol{F}$, it applies factorized weights $\boldsymbol{r}_*$ to mine critical supporting cues from the support set and tunes the weights with prediction consistency at multiple temporal scales.

a diversified motion description set $\boldsymbol{D} = \{\boldsymbol{d}_i\}_{i=1}^{C}$ in which each class corresponds to M different explanations and C is the number of action classes.

Video Generation from Diversified Prompts. For the i-th action class, we generate $K = m \times n$ video samples via the text-to-video model $\texttt{T2V}(\cdot,\cdot)$ according to text prompts stored in $\boldsymbol{d}_i$ as follows,

$$\boldsymbol{s}_i = \texttt{T2V}(\texttt{Sample}(\boldsymbol{d}_i, m), n). \tag{5}$$

Here, we $\texttt{Sample}$ m prompts for each class from $\boldsymbol{d}_i$ and repeatedly generate n videos corresponding to each prompt. *K is set to be always factorable for ease of implementation, two hyper-parameters (i.e., m and n) are ablated in the experiments (as shown in Figure 3).* After traversing the entire description set $\boldsymbol{D}$, we can construct the video support-set $\boldsymbol{S} = \{s_i\}_{i=1}^{C}$ in which each class has K support videos.

Different from previous methods (SuS-X [Udandarao *et al.*, 2023], CaFo [Zhang *et al.*, 2023b], and DMN [Zhang *et al.*, 2024]) which explain class names repeatedly, MSD constructs diverse descriptions for each class name with different context/setting/perspectives.

4.2 Temporal-aware Support-set Erosion (TSE)

The support set generated via MSD has a greater diversity in each category (as shown in Figure 4). MSD refines the boundary of each class but still contains many invalid samples (*i.e.*, outliers or duplications), which motivates us to further adjust the support-set, *i.e.*, "**eroding**" some invalid information from the set. A straightforward solution is to apply learnable weights on video samples and tune them with a self-supervision loss, inspired by TPT [Shu *et al.*, 2022]. However, video-level tuning can miss the potential fine-grained supporting cues hidden in video data. Hence, it is necessary to tune the weights on frame-level or patch-level visual features, but patch-level introduces so many parameters that test-time tuning techniques cannot accomplish optimization. To this end, we proposed a novel Temporal-aware Support-set Erosion module to reweigh the importance of each frame from supporting videos via factorized video-frame weights, which is tuned via different temporal-scale features in multiple steps during test time.

Factorized Re-weighting. Given the support set $\boldsymbol{S}$ generated from MSD, we utilize the pre-trained visual encoder ($\mathcal{F}_{\text{vis}}$) from popular VLMs (*e.g.*, CLIP) to extract visual features and concatenate them together as follows,

$$\begin{aligned} \boldsymbol{F}_i &= \mathcal{F}_{\text{vis}}(\boldsymbol{s}_i), i \in [1, CK], \boldsymbol{F}_i \in \mathbb{R}^{T\times d}, \\ \boldsymbol{F} &= \texttt{Concat}([\boldsymbol{F}_1, \cdots, \boldsymbol{F}_{CK}]), \boldsymbol{F} \in \mathbb{R}^{CK\times T\times d}. \end{aligned} \tag{6}$$

Here, C, K, T, and d represent the number of classes, supporting samples of each class, frames, and channels, respectively. After that, we reweight the support-set feature $\boldsymbol{F}$ in video and frame level via factorized learnable weights as follows,

$$\boldsymbol{F}' = (\boldsymbol{F} \odot \boldsymbol{r}_{\text{vid}}) \odot \boldsymbol{r}_{\text{fr}}, \boldsymbol{F}' \in \mathbb{R}^{CK\times T\times d}. \tag{7}$$

Here, "$\odot$" denotes element-wise matrix multiplication (*i.e.*, Hadamard product). $\boldsymbol{r}_{\text{vid}} \in \mathbb{R}^{CK\times 1\times 1}$ and $\boldsymbol{r}_{\text{fr}} \in \mathbb{R}^{1\times T\times 1}$ are frame-level and video-level weights, respectively.

Multi-scale Temporal Tuning. Inspired by Test-time Prompt Tuning (TPT [Shu *et al.*, 2022]), we tune learnable

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 [Qian *et al.*, 2022]	R(2+1)D	38.7 ± 3.7	56.0 ± 3.1	–	–
Adapting pre-trained CLIP					
Vanilla CLIP [Radford *et al.*, 2021]	ViT-B/16	40.8 ± 0.3	63.2 ± 0.2	59.8 ± 0.3	–
Vanilla CLIP*(reproduce)	ViT-B/16	39.2 ± 0.2	61.7 ± 0.5	58.4 ± 0.8	68.8 ± 0.6
Vanilla CLIP +TEST-V (Ours)	ViT-B/16	$\mathbf{44.3 \pm 0.6}$	$\mathbf{65.5 \pm 0.3}$	$\mathbf{60.1 \pm 0.5}$	$\mathbf{70.1 \pm 0.4}$
ActionCLIP [Wang *et al.*, 2021b]	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	–
XCLIP [Ni *et al.*, 2022]	ViT-B/16	44.6 ± 5.2	72.0 ± 2.3	65.2 ± 0.4	–
Vita-CLIP [Wasim *et al.*, 2023]	ViT-B/16	48.6 ± 0.6	75.0 ± 0.6	67.4 ± 0.5	–
VicTR [Kahatapitiya *et al.*, 2024]	ViT-B/16	51.0 ± 1.3	72.4 ± 0.3	–	–
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
BIKE + DTS-TPT [Yan *et al.*, 2024]	ViT-B/16	$\mathbf{51.6 \pm 0.5}$	$\mathbf{78.0 \pm 0.6}$	$\mathbf{67.0 \pm 0.5}$	$\mathbf{75.8 \pm 0.6}$
BIKE + TEST-V (Ours)	ViT-B/16	$\mathbf{52.9 \pm 0.6}$	$\mathbf{78.6 \pm 0.5}$	$\mathbf{68.4 \pm 0.4}$	$\mathbf{76.5 \pm 0.5}$
ViFi-CLIP [Rasheed *et al.*, 2023]	ViT-B/16	51.3 ± 0.7	76.8 ± 0.8	71.2 ± 1.0	76.9 ± 0.8
ViFi-CLIP +TEST-V (Ours)	ViT-B/16	$\mathbf{53.0 \pm 0.9}$	$\mathbf{78.3 \pm 0.3}$	$\mathbf{73.8 \pm 0.6}$	$\mathbf{78.4 \pm 0.5}$

Table 1: Comparisons with state-of-the-art methods for zero-shot activity recognition.

weights ($\boldsymbol{r}$) via minimizing the prediction consistency between re-weighted support-set features and the augmented test video features as follows,

$$\min \mathcal{L}(\texttt{Pred}(\boldsymbol{F}', \{\boldsymbol{f}_i^{\mathrm{S}}\}_{i=1}^{n})). \tag{8}$$

Here $\boldsymbol{f}_*^S = \mathcal{F}_{\text{vis}}(\texttt{Aug}_n^{\mathrm{S}}(\boldsymbol{V}_{\text{test}}))$ denotes the augmented test video features and $\texttt{Aug}_n^{\mathrm{S}}(\cdot)$ performs spatial augmentation n times. $\texttt{Pred}(\cdot)$ computes multi-view predictions from the support-set features and augmented test video features according to Eq. (3). $\mathcal{L}$ calculates the averaged entropy from high-confident predictions [Shu *et al.*, 2022].

However, the temporal evolution rate of different human activities is variable, and building multi-scale representations from the temporal domain has proven beneficial in video representation [Feichtenhofer *et al.*, 2019; Yan *et al.*, 2024]. Therefore, Eq. (8) can be rewritten as follows,

$$\min \mathcal{L}(\texttt{Pred}(\texttt{Aug}_m^{\mathrm{T}}(\boldsymbol{F}', \{\boldsymbol{f}_i^{\mathrm{S}}\}_{i=1}^{n}))). \tag{9}$$

Here, $\texttt{Aug}_m^{\mathrm{T}}(\cdot)$ augments two feature sets synchronously via m time sampling at different temporal scales. Different sampling strategies are compared in Table 4 and discussed in Sec. 5.2. *Notably, this loss function is optimized in m steps with different temporal-scale features.*

5 Experiments

We evaluate the effectiveness of the proposed method on four popular video benchmarks, *i.e.*, HMDB-51 [Kuehne *et al.*, 2011], UCF-101 [Soomro *et al.*, 2012], Kinetics-600 [Carreira *et al.*, 2018], ActivityNet [Fabian Caba Heilbron and Niebles, 2015].

5.1 Comparison with State-of-the-Arts

We evaluate the effectiveness of **TEST-V**with existing state-of-the-art methods on four popular action recognition benchmarks in Table 1. Existing methods can be categorized into three types: i) Uni-modal zero-shot video recognition models: they are trained on video data with elaborate representation engineering; ii) Adapting pre-trained CLIP: they adapt CLIP to video data via additional temporal learners or VL prompting techniques; iii) Tuning pre-trained CLIP: they fully fine-tune the CLIP model via video data.

We apply off-the-shelf pre-trained visual/text encoders from VLMs, *e.g.*, BIKE [Wu *et al.*, 2023] and ViFi-CLIP, to extract features and tune the proposed **TEST-V** during test time without any training data. Our method outperforms the conventional uni-modal zero-shot video recognition models by a large margin ($15\% \sim 20\%$). Moreover, compared with the methods adapted from pre-trained CLIP with video data, our method achieve consistent improvements across all benchmarks without training. Furthermore, our method can significantly improve some recent methods fully-tuned from CLIP, thanks to the reduction of modality gaps through the tuned support set. In general, the proposed **TEST-V** achieves state-of-the-art performance on four popular benchmarks in a training-free manner.

5.2 Ablation Study

Choice of LLM and T2V for MSD. As described in Sec. 4.1, MSD builds multiple descriptions (*i.e.*, prompts) for each class via the Large Language Model (LLM) and then generates video samples via the Text-to-Video (T2V) model according to multiple prompts. There are many options for LLM and T2V, thus we compare them in Table 2 and 3, and make the following analysis. **i) LLMs**: We compare different Large Language Models (LLMs) for motion description generation defined in Eq. (4) in Table 2. In general, our approach works well on all LLMs, and new versions of the same model perform better. ChatGPT achieves the best

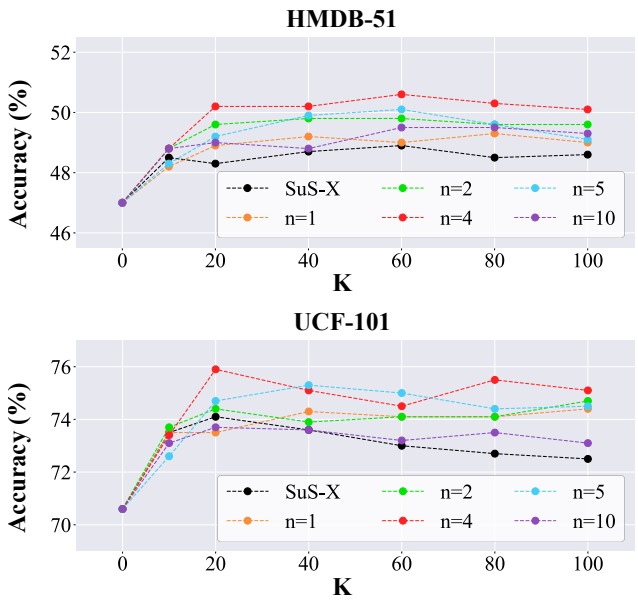

Figure 3: Effect of support hyper-parameters K and n ($K = m \times n$ defined in the equation 5) with a single prompt (SuS-X) and multiple prompts (our MSD) on HMDB-51 and UCF-101. Top-1 zero-shot recognition accuracy is reported.

performance and is selected as the default LLM. **ii) T2Vs**: The proposed method is compatible with various Text-to-Video (T2V) generation models as shown in Table 3. Benefited from the extra training videos data, LaVie is slightly better than others and is adopted by default.

LLM	Version	HMDB-51	UCF-101
Gemini [Team *et al.*, 2024]	1.0 pro	49.3	74.5
	1.5 pro	50.8	76.5
Llama-3 [AI@Meta, 2024]	8B	50.9	76.1
	70B	51.2	77.5
Claude-3 [Anthropic, 2024]	haiku	50.0	75.5
	sonnet	51.0	77.1
	opus	51.1	77.4
ChatGPT [OpenAI, 2023]	**3.5-turbo**	**50.1**	**77.4**
	4-turbo	**51.3**	**77.8**

Table 2: Effect of different LLMs used in MSD.

T2V Model	PT data	HMDB-51 / UCF-101
Show-1	WV10M	50.2 / 77.1
HiGen	WV10M + Inter. data20M	50.3 / 77.3
TF-T2V	WV10M + LAION5B + Inter. data10M	49.8 / 77.3
ModelScopeT2V	WV10M + LAION5B	51.0 / 77.6
LaVie	WV10M + LAION5B + Vimeo25M	**51.3 / 77.8**

Table 3: Effect of different T2V models[1] used in MSD.

[1]including Show-1 [Zhang *et al.*, 2023a], HiGen [Qing *et al.*, 2024], TF-T2V [Wang *et al.*, 2024], ModelScopeT2V [Wang *et al.*, 2023a], and LaVie [Wang *et al.*, 2023b]

Importance of Diversity of the Support-set To figure out the key factor of the quality of the support-set, we ablate the support-set hyper-parameters of the proposed MSD, *i.e.*, the number of supporting videos for each class K and the number of videos repeatedly generated for each prompt n defined in Equation (5), as shown in Figure 3. The performance of all models (baseline SuS-X and the proposed MSD with different n) gradually plateaus as the support-set size (K) increases across two benchmarks. Besides, the proposed multi-prompting method (MSD) outperforms the single-prompting baseline (SuS-X) when $K \neq 0$, demonstrating the importance of sample diversity in the support-set. Meanwhile, we ablate the number of repeatedly generated videos n when the total number of videos K is fixed. ($K = m \times n$). We found that $n = 4$ brings stabilized gains compared to the baseline and **set $n = 4$ for all benchmarks**.

Dataset	Strategy	single	multiple		
		8	6, 8	4, 6, 8	2, 4, 6, 8
HMDB-51	random	52.0	52.0	52.1	51.8
	top	52.0	52.6	**52.9**	52.5
UCF-101	random	77.8	77.7	77.8	78.0
	top	77.8	78.4	**78.6**	78.4

Table 4: Effect of different sampling methods for multi-scale tuning in TSE. Each video contains a total of 8 frames.

Multi-scale Temporal Tuning. To find the suitable temporal scale used in TSE, we try different sampling strategies and report the performance in Table 4. "random" sampling frames from the feature sequence randomly, while "top" selects the top k features with higher frame-level weights (r_{f}). We observed that multi-scale tuning with the "random" strategy does not work compared with the single-scale tuning. Besides, multi-scale tuning with the "top" strategy of "4,6,8" achieves state-of-the-art results.

Module		HMDB-51	UCF-101
MSD	TSE		
✗	✗	50.1	76.8
✓	✗	51.3	77.8
✗	✓	51.9	78.1
✓	✓	**52.9**	**78.6**

Table 5: Effect of different components of our approach.

Component Analysis. We explore the effect of each module designed in **TEST-V** in Table 5. Compared with the baseline, MSD brings approximately 1% improvements on two benchmarks, indicating that the diversity of supporting samples in each class is beneficial in enhancing the quality of the support set. Meanwhile, TSE demonstrates the ability to select critical visual cues from highly redundant video data, which improves the baseline by approximately $1\% \sim 2\%$. Furthermore, the combination of MSD and TSE yields even

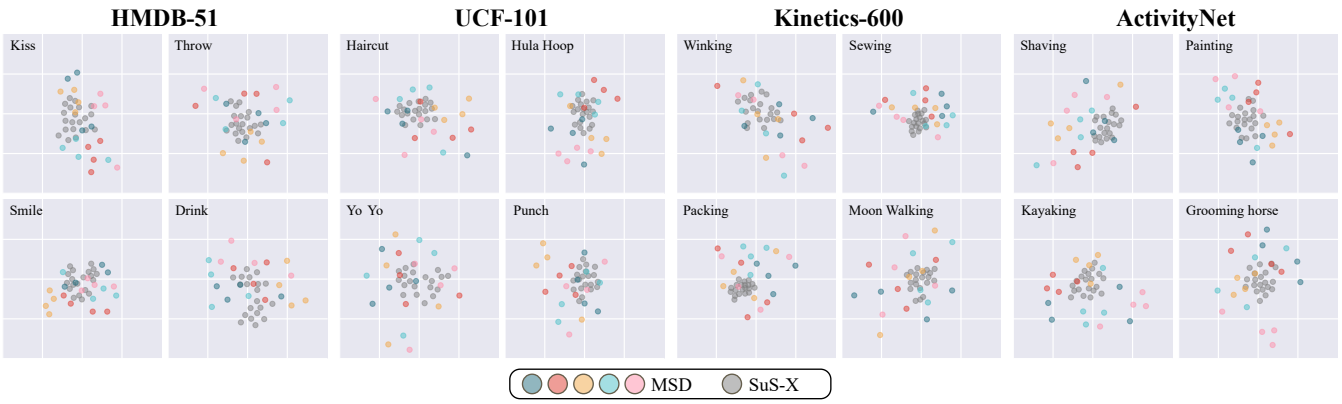

Figure 4: Feature distribution of supporting samples generated with multiple prompts (MSD) and single prompt (SuS-X) on different benchmarks. Multi-prompting and single-prompting samples are shown in color and grey, respectively.

Learnable weights		HMDB-51	UCF-101
r_{vid}	r_{fr}		
✗	✗	51.3	77.8
✓	✗	52.1	78.3
✗	✓	52.4	78.0
✓	✓	**52.9**	**78.6**

Table 6: Effect of different components in weights module.

more significant performance gains, confirming the complementarity of the two modules.

As defined in Eq. (7), learnable weights r_{vid} and r_{fr} are designed to adjust the contribution of video-level and frame-level features, respectively. As reported in Table 6, each of them brings performance gains, and they are complementary.

Backbone	Method	HMDB-51	UCF-101
ViT-B/16	Vanilla CLIP	39.2	61.7
	+ SuS-X	41.7	64.5
	+ TEST-V (Ours)	**44.3**	**65.5**
	BIKE	47.0	70.6
	+ SuS-X	50.1	76.8
	+ TEST-V (Ours)	**52.9**	**78.6**
ViT-L/14	Vanilla CLIP	45.9	69.8
	+ SuS-X	46.8	73.6
	+ TEST-V (Ours)	**48.2**	**75.7**
	BIKE	53.2	80.5
	+ SuS-X	56.7	85.0
	+ TEST-V (Ours)	**59.7**	**87.5**

Table 7: Generalization to different pre-trained VLMs.

Generalization to Different VLMs. We also extract visual/text features via different pre-trained VLMs, *i.e.*, Vanilla CLIP [Radford *et al.*, 2021], and BIKE [Wu *et al.*, 2023], with various network backbones, and report top-1 zero-shot accuracy on two benchmarks in Table 7 compared with SuS-X [Udandarao *et al.*, 2023]. Our approach shows consistent improvement across two benchmarks with different models and backbones, demonstrating good generalization across different video representations.

Visualization. We randomly sample 20 videos from the support-sets generated with multiple prompts and the single prompt for four benchmarks, respectively, and encode them via the pre-trained visual encoder from BIKE [Wu *et al.*, 2023] with 4 frames. Feature distribution of sampled videos in each benchmark is visualized via t-SNE [Van der Maaten and Hinton, 2008] in Figure 4. We observed that single-prompting samples are more likely to cluster together, whereas multi-prompting samples are more dispersed. It indicates that the multi-prompting strategy can refine the supporting boundary to enhance the quality of the support-set.

6 Conclusion

This work presents a novel framework, namely **TEST-V** which `dilates` and `erodes` the support set to enhance the zero-shot generalization capability of video classification methods during test time. **TEST-V** applies multiple prompts to enrich the support set (`Dilation`) and then mines critical supporting samples from the support set with learnable spatio-temporal weights (`Erosion`). Our method demonstrated superior performance to existing state-of-arts on four benchmarks. While **TEST-V** effectively adapts pre-trained models (i.e., CLIP) to out-of-distribution domains during test time for video data, it still has two potential limitations: **i)** extra spatio-temporal tuning costs may increase with the length of the video (i.e., long video). **ii)** relying on the quality of LLMs and text-to-video generation models.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (Grant No. 62472208 and 62172226), the Postdoctoral Fellowship Program of CPSF (Grant No. GZB20230302). Rui Yan and Jin Wang contributed equally to this work, and Xiaoyu Du is the corresponding author.

References

[AI@Meta, 2024] AI@Meta. Llama 3. https://llama.meta.com/llama3/, 2024. Accessed: 2024-08-11.

[Anthropic, 2024] Anthropic. Claude 3 haiku: our fastest model yet. https://www.anthropic.com/news/claude-3-haiku, 2024. Accessed: 2024-08-11.

[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.

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *NeurIPS*, pages 1877 – 1901, 2020.

[Carreira and Zisserman, 2017] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017.

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

[Feichtenhofer *et al.*, 2019] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019.

[Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021.

[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.

[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.

[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.

[Lin *et al.*, 2019] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, pages 7083–7093, 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, 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.

[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. https://openai.com/blog/chatgpt/, 2023. Accessed: 2024-01-13.

[Pratt *et al.*, 2023] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, pages 15691–15701, 2023.

[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, 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.

[Qing *et al.*, 2024] Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie Wei, Yingya Zhang, Changxin Gao, and Nong Sang. Hierarchical spatio-temporal decoupling for text-to-video generation. In *CVPR*, pages 6635–6645, 2024.

[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.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn

Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[Shao *et al.*, 2020] Hao Shao, Shengju Qian, and Yu Liu. Temporal interlacing network. In *AAAI*, pages 11966–11973, 2020.

[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, 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.

[Team *et al.*, 2024] 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, 2024.

[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.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(86):2579–2605, 2008.

[Wang *et al.*, 2021a] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization, 2021.

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

[Wang *et al.*, 2023a] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report, 2023.

[Wang *et al.*, 2023b] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. Lavie: High-quality video generation with cascaded latent diffusion models, 2023.

[Wang *et al.*, 2024] Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. A recipe for scaling up text-to-video generation with text-free videos. In *CVPR*, pages 6572–6582, 2024.

[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*, 45(7):9004–9021, 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*, pages 1534–1542, 2024.

[Zellers and Choi, 2017] Rowan Zellers and Yejin Choi. Zero-shot activity recognition with verb attribute induction. In *EMNLP*, pages 946–958, 2017.

[Zeng *et al.*, 2023] Runhao Zeng, Qi Deng, Huixuan Xu, Shuaicheng Niu, and Jian Chen. Exploring motion cues for video test-time adaptation. In *ACMMM*, pages 1840–1850, 2023.

[Zhang *et al.*, 2021] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling, 2021.

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

[Zhang *et al.*, 2023a] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation, 2023.

[Zhang *et al.*, 2023b] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *CVPR*, pages 15211–15222, 2023.

[Zhang *et al.*, 2024] Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. Dual memory networks: A versatile adaptation approach for vision-language models. In *CVPR*, pages 28718–28728, 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. *IJCV*, 130(9):2337–2348, 2022.