Condensed Test-Time Adaptation of VLMs for Action Recognition

Supplementary Material

7. Overview of Supplementary Material

In the supplementary material, we provide additional details in the following sections:

- Sec. 9 provides a pseudo-code description of Adaptive Tube Construction module.
- Sec. 10 describes in detail the three short-term video datasets, two long-term video datasets, and two egocentric video datasets involved in our experiments.
- Sec. 11 conducts further experiments to validate the effectiveness of our method. Specifically, Sec. 11.1 compares our proposed CONDA with other test-time adaptation methods on additional VLMs, while Sec. 11.3 analyzes the sensitivity of CONDA to key hyperparameters.

8. Additional Related Work

Test-Time Training. Test-Time Training (TTT) [58] mitigates distribution shifts by executing self-supervised updates during inference, a paradigm that necessitates a dual-branch architecture during pre-training to jointly optimize the primary task with auxiliary tasks such as rotation prediction [58] or contrastive learning [32]. Additionally, TTT-MAE [15] incorporates Masked Autoencoders as the auxiliary task to extract robust adaptive representations through pixel reconstruction. Distinct from TTT's requirement for training-phase intervention and architectural modifications, our work focuses on Test-Time Adaptation (TTA), which offers the superior advantage of plug-and-play deployment flexibility by operating directly on off-the-shelf pre-trained models without access to the original training pipeline.

9. Pseudo-code of Adaptive Tube Construction

To enrich the spatial-temporal representations of video patches, we design the Adaptive Tube Construction (ATC) module, which consists of spatial expansion and temporal expansion. Spatial expansion is employed to capture spatial structural information, while temporal expansion is used to capture temporal motion dynamics. In Sec. 4.2, we provide a detailed description of the motivation and design of Adaptive Tube Construction. To facilitate the understanding, we present the whole construction process in Algorithm 1 in the form of pseudo-code. Given the local feature of the test video $\boldsymbol{f_v} \in \mathbb{R}^{T\times H\times W\times d}$, the selected patch $\boldsymbol{p} \in \mathbb{R}^d$, and its position (t, h, w) in $\boldsymbol{f_v}$, the objective of Algorithm 1 is to adaptively retrieve patches from $\boldsymbol{f_v}$ based on $\boldsymbol{p}$ to construct spatial-temporal video tube.

Algorithm 1 Adaptive Tube Construction

Input: local feature of test video $\boldsymbol{f_v} \in \mathbb{R}^{T\times H\times W\times d}$, selected patch $\boldsymbol{p} \in \mathbb{R}^d$, and its position (t, h, w) in $\boldsymbol{f_v}$
Parameter: random spatial expansion range (α, β), and similarity threshold τ

 # Spatial Expansion:
$s_h \sim U(\alpha, \beta), s_w \sim U(\alpha, \beta)$
$\boldsymbol{f_{\text{region}}} = \texttt{Crop}(\boldsymbol{f_v}[t], (h, w), (s_h \cdot H, s_w \cdot W)) \in \mathbb{R}^{r\times d}$
 # Temporal Expansion:
$\boldsymbol{f_{\text{anchor}}} = \texttt{Avg}(\boldsymbol{f_{\text{region}}}) \in \mathbb{R}^d$
$\boldsymbol{f_{\text{tube}}} = \texttt{[]}$
for $j = 1$ to T **do**
  $\boldsymbol{R} = \boldsymbol{f_v}[j]\boldsymbol{f_{\text{anchor}}} \in \mathbb{R}^{H\times W}$
  $r_{\max}, (h', w') = \texttt{Max}(\boldsymbol{R})$
  if $r_{\max} > \tau$ **then**
    $s'_h \sim U(\alpha, \beta), s'_w \sim U(\alpha, \beta)$
    $\boldsymbol{f'_{\text{region}}} = \texttt{Crop}(\boldsymbol{f_v}[j], (h', w'), (s'_h \cdot H, s'_w \cdot W))$
    $\boldsymbol{f_{\text{tube}}}.\texttt{append}(\boldsymbol{f'_{\text{region}}})$
  end if
end for
return $\boldsymbol{f_{\text{tube}}}$

10. Details of Datasets

HMDB-51 [27] is a widely used small-scale short-term action recognition benchmark, which consists of approximately 6,766 annotated videos covering 51 activities. Each video is collected from diverse, publicly accessible sources, predominantly from movies, as well as from public video archives such as YouTube and Google.

UCF-101 [55] consists of about 13,320 short-term videos covering 101 categories, which can be grouped into five categories: Body motion, Human-human interactions, Human-object interactions, Playing instruments, and Sports. The videos are sourced from YouTube and depict real-world, unconstrained human actions.

Kinetics-600 [2] is a large-scale video dataset, containing 600 human action classes. Each video is collected and annotated from YouTube and lasts approximately 10 seconds. Kinetics-600 is an extension of Kinetics-400 [26], expanding the number of action classes to provide a more diverse set of human activities.

ActivityNet-200 [12] is a long-term video dataset comprising approximately 4926 untrimmed videos, each lasting 5-10 minutes and spanning 200 activity categories. The videos are sourced from the web, primarily YouTube, and are verified by crowdsourced workers to ensure they contain the intended activities.

Method	Venue	Encoder	HMDB-51	UCF-101	K600(Top-1)	K600(Top-5)
ViCLIP [63]	ICLR'24	ViT-B/16	46.4	75.9	69.8	90.1
+ TPT [50]	NeurIPS'22	ViT-B/16	47.1	76.7	70.1	90.3
+ TDA [25]	CVPR'24	ViT-B/16	47.4	76.7	70.9	90.5
+ DPE [72]	NeurIPS'24	ViT-B/16	47.6	77.1	70.8	90.6
+ Point-Cache [57]	CVPR'25	ViT-B/16	47.5	76.8	70.3	90.4
+ CONDA (Ours)	**Ours**	ViT-B/16	48.4	77.6	72.1	91.8
ViFi-CLIP [44]	CVPR'23	ViT-B/16	52.3	75.0	60.9	85.6
+ TPT [50]	NeurIPS'22	ViT-B/16	52.4	75.3	61.2	85.3
+ TDA [25]	CVPR'24	ViT-B/16	52.6	76.2	61.7	86.0
+ DPE [72]	NeurIPS'24	ViT-B/16	52.9	77.0	62.1	86.2
+ Point-Cache [57]	CVPR'25	ViT-B/16	52.4	76.5	61.8	85.8
+ CONDA (Ours)	**Ours**	ViT-B/16	53.6	77.4	62.8	86.4
OST [6]	CVPR'24	ViT-B/16	54.6	78.2	74.6	92.0
+ TPT [50]	NeurIPS'22	ViT-B/16	54.7	78.8	74.7	91.5
+ TDA [25]	CVPR'24	ViT-B/16	54.7	79.9	74.8	92.0
+ DPE [72]	NeurIPS'24	ViT-B/16	54.9	80.6	74.9	92.2
+ Point-Cache [57]	CVPR'25	ViT-B/16	55.0	80.2	74.7	91.9
+ CONDA (Ours)	**Ours**	ViT-B/16	**55.2**	**81.5**	**75.1**	**92.2**

Table 6. Comparisons with state-of-the-art TTA methods on various VLMs, *i.e.* ViCLIP [63], ViFi-CLIP [44], and OST [6]. Following [36, 61], we report top-1 accuracy (%) on HMDB-51 and UCF-101, and both top-1 accuracy (%) and top-5 accuracy (%) on K600.

Strategy	HMDB-51	K600	COIN
CAM	47.6	71.0	65.3
Grad-CAM	48.0	71.8	66.2
SPA	**48.4**	**72.1**	**67.4**

Table 7. Comparison between semantic patch activation and CAM/Grad-CAM.

COIN [60] consists of about 3k long-term videos with the average length of 2.36 minutes, collected from YouTube. Each video belongs to one of 180 distinct procedural tasks. COIN employs a three-level hierarchical structure to organize the video, ensuring that each video is accurately associated with a domain, a task, and specific steps, facilitating the analysis of complex instructional videos.

EPIC-KITCHENS-100 [11] is one of the most popular large-scale egocentric datasets with about 9.5k action clips collected in 45 kitchen environments using the head-mounted cameras. It collects the participant's variety of daily activities in the kitchen with a first-person view.

EGTEA [28] is a large-scale egocentric dataset, which is collected using gaze tracking. It consists of about 6k video clips and is labeled with 106 activity categories.

11. Further Experiments

11.1. Comparison with Other TTA Methods

In Sec. 5.3, we compare the proposed CONDA with three categories of approaches: uni-modal zero-shot video recognition models, methods adapting pre-trained CLIP, and methods tuning pre-trained CLIP. The results show that CONDA consistently outperforms these zero-shot action recognition approaches. In addition, we further compare our method with state-of-the-art test-time adaptation (TTA) methods on ViCLIP [63]. The results demonstrate that CONDA surpasses existing SOTA TTA methods when adapting to ViCLIP.

To further validate the superiority of our proposed CONDA over existing TTA methods, we additionally compare CONDA with SOTA TTA approaches on ViFi-

(α, β)	HMDB-51	K600	COIN
$(0.1, 0.9)$	48.3	71.8	67.2
$(0.2, 0.8)$	48.0	72.0	67.2
$\mathbf{(0.3, 0.7)}$	**48.4**	**72.1**	**67.4**
$(0.4, 0.6)$	48.2	71.9	67.3

Table 8. Hyperparameter study on the range of random spatial expansion (α, β). Without loss of generality, we assume that β is greater than α.

τ	HMDB-51	K600	COIN
0.1	47.9	71.7	66.9
0.3	48.2	71.8	67.2
0.5	**48.4**	**72.1**	**67.4**
0.7	48.2	71.9	66.9
0.9	47.7	71.6	66.5

Table 9. Hyperparameter study on the similarity threshold τ in temporal expansion.

CLIP [44] and OST [6] in Tab. 6. The results indicate that our method not only outperforms existing TTA approaches but also generalizes effectively to arbitrary VLMs.

11.2. Ablation Study

Semantic Patch Activation. In Sec. 5.4, we conduct a component analysis of Semantic Patch Activation (SPA). To further demonstrate its effectiveness, we compare SPA with CAM and Grad-CAM. While Grad-CAM measures *static representation sensitivity* via token embeddings, SPA computes gradients over the attention map to capture *dynamic token interactions*. This allows SPA to better identify human-environment interactions essential for action recognition, rather than individual token features. As shown in Tab. 7, replacing SPA with Grad-CAM for patch selection leads to a 1.4% performance drop. This validates that interaction-based activation provides more robust spatial-temporal cues than traditional individual feature activations.

11.3. Hyperparameter Study

In Sec. 5.4, we analyze the effect of the number of selected patches k on CONDA, and find that the performance is not significantly affected by k, with k finally set to 10.

We further conduct a sensitivity analysis of CONDA with respect to other hyperparameters, *i.e.* the range of random spatial expansion (α, β) and the similarity threshold τ in temporal expansion. As shown in Tab. 8 and Tab. 9, the results similarly indicate that our method is not significantly affected by these hyperparameters, and we set (α, β) to $(0.3, 0.7)$ and τ to 0.5.