Test-Time Prototype Evolution for Generalizable Vision-Language Models

Ce Zhang [1] **Simon Stepputtis** [1] **Katia Sycara** [1] **Yaqi Xie** [1]

Abstract

Test-time adaptation, which enables models to generalize to diverse data during testing, holds significant value in real-world scenarios. Recently, researchers have applied this setting to advanced pre-trained vision-language models (VLMs), developing approaches such as test-time prompt tuning to further extend their practical applicability. However, these methods typically focus solely on adapting VLMs from a single modality and fail to accumulate task-specific knowledge as more samples are processed. To address this, we introduce Dual Prototype Evolving (DPE), a novel test-time adaptation approach for VLMs that effectively *accumulates* task-specific knowledge from *multi-modalities*. Specifically, we create and evolve two sets of prototypes—textual and visual—to progressively capture more accurate multi-modal representations for target classes during test time. Moreover, to promote consistent multi-modal representations, we introduce and optimize learnable residuals for each test sample to align the prototypes from both modalities. Extensive experimental results on 15 benchmark datasets demonstrate that our proposed DPE consistently outperforms previous state-of-the-art methods while also exhibiting competitive computational efficiency.

1. Introduction

Recently, large-scale vision-language models (VLMs), such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), have garnered increasing attention in the research community. These models, pre-trained on massive web-scale datasets, exhibit remarkable zero-shot capabilities and open-world visual understanding (Radford et al., 2021; Yu et al., 2022; Zhai et al., 2022; Li et al., 2022). While the large-scale pre-trained (source) datasets like LAION-5B (Schuhmann et al., 2022) are accessible, it is impractical for individuals to train on them due to their immense size. Consequently, adapting VLMs to downstream tasks via efficient fine-tuning with limited annotated samples from the target domain has become a focus of recent research (Zhou et al., 2022b;a; Zhang et al., 2022b; Yu et al., 2023). However, although these methods have proven effective, they pose a significant limitation: they assume the availability of annotated samples from the target domain, which is often not practical in real-world scenarios. This constraint hinders the broader deployment of VLMs in diverse and dynamic environments.

To address the label scarcity problem in practice, a number of approaches apply the *test-time adaptation* setting to the domain of adapting VLMs to downstream tasks, as shown in Figure 1. Specifically, Shu *et al*. (Shu et al., 2022) propose test-time prompt tuning to learn an adaptive prompt for each individual sample in the test data stream to enhance CLIP's zero-shot generalizability to out-of-distribution domains. Building on TPT, DiffTPT (Feng et al., 2023) incorporates diffusion-based data augmentations to facilitate more effective prompt tuning during test time. More recently, Karmanov *et al*. (Karmanov et al., 2024) propose an alternative training-free dynamic adapter approach to establish dynamic visual caches with the unlabeled test samples.

However, we recognize that existing works overlook the following inherent properties of *test-time adaptation* in VLMs: (1) *Cumulative.* We expect that with more seen samples, the performance should improve as task-specific knowledge accumulates (Mirza et al., 2022; Sun et al., 2020). However, test-time prompt tuning methods (Shu et al., 2022; Feng et al., 2023) treat each test instance independently, resetting to the original model for each new sample, failing to extract historical knowledge from previous test samples. (2) *Multi-modal.* Effective adaptation of VLMs benefits from leveraging knowledge from both textual and visual modalities (Khattak et al., 2023; Lin et al., 2023b). However, previous works only capture domain-specific knowledge from a single modality, adapting CLIP based solely on textual (Shu et al., 2022; Feng et al., 2023) or visual (Karmanov et al., 2024) feature refinement.

To this end, we propose Dual Prototype Evolving (DPE), a novel test-time VLM adaptation approach that effectively *accumulates* task-specific knowledge from *multi-modalities*.

[1]All the authors are with the School of Computer Science, Carnegie Mellon University. Correspondence to: Yaqi Xie <yaqix@cs.cmu.edu>.

Published at ICML 2024 Workshop on Foundation Models in the Wild.

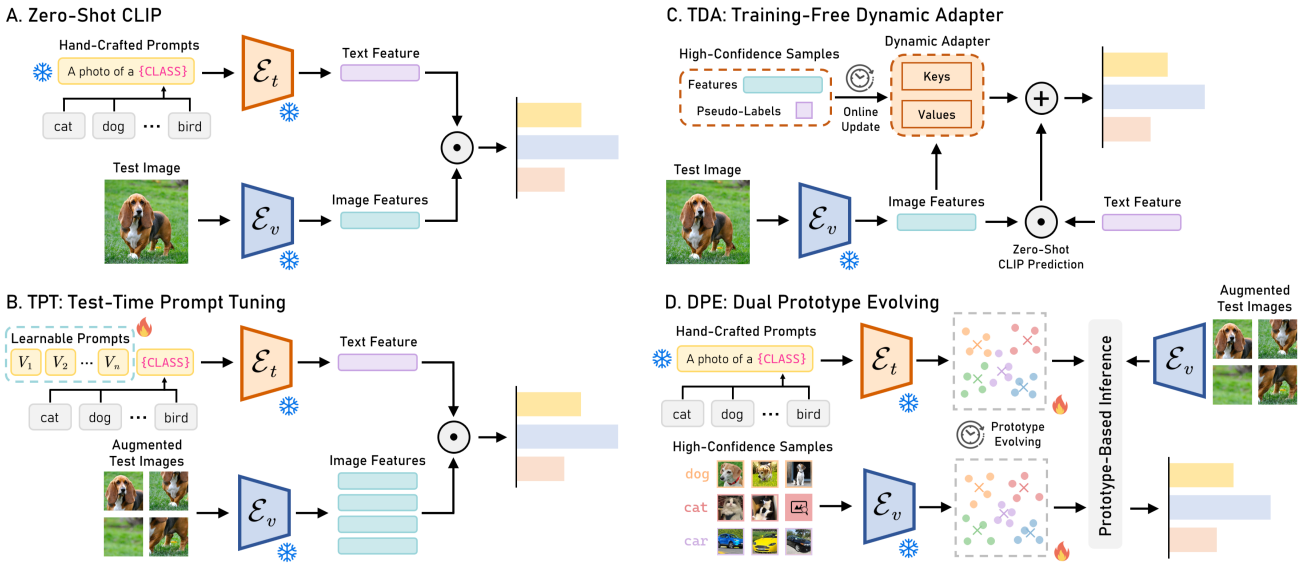

Figure 1. **Comparison of our DPE with zero-shot CLIP (Radford et al., 2021), TPT (Shu et al., 2022), and TDA (Karmanov et al., 2024).** We denote CLIP's parallel textual and visual encoders as $\mathcal{E}_t$ and $\mathcal{E}_v$, respectively. While previous methods solely adapt the CLIP model from a single modality, we design our DPE to evolve prototypes from both textual and visual modalities to progressively capture more accurate multi-modal representations for target classes during test time.

Unlike previous methods that focus on adapting VLMs from a single modality, we create and evolve two sets of prototypes—textual and visual—progressively capturing more accurate multi-modal representations for target classes during test time. To extract historical knowledge from previous test samples, we update these two sets of prototypes online using cumulative average and priority queue strategies, respectively. We further optimize these multi-modal prototypes by introducing learnable residual parameters for each individual test sample to enhance the zero-shot generalization capability of our model. Specifically, rather than solely relying on the entropy minimization objective (Wang et al., 2021; Zhang et al., 2022a), our DPE also accounts for the alignment between multi-modal prototypes to ensure consistent multi-modal representations. Notably, our DPE requires only the optimization of multi-modal prototypes in the embedding space during test time, eliminating the need to back-propagate gradients through the textual encoder of CLIP, as required in TPT (Shu et al., 2022) and DiffTPT (Feng et al., 2023).

The test-time generalization capabilities of our proposed DPE method are extensively evaluated across 15 diverse recognition datasets in two scenarios: natural distribution shifts and cross-dataset generalization. The experimental results validate the superior performance of our DPE, which achieves an average improvement of 3.55% and 4.30% over the state-of-the-art TPT (Shu et al., 2022) method in these scenarios. Moreover, our proposed DPE achieves this performance while also exhibiting $5\times$ and over $10\times$ test-time efficiency compared to TPT (Shu et al., 2022) and DiffTPT (Feng et al., 2023), respectively.

2. Method

We introduce Dual Prototype Evolving (DPE) as illustrated in Figure 2, to enhance CLIP's zero-shot generalization capabilities across diverse distributions during test time. Unlike previous methods that focus solely on one modality, we design two sets of prototypes, textual and visual, which are progressively updated using the unlabeled test dataset.

2.1. Preliminaries

Zero-Shot CLIP. CLIP (Radford et al., 2021) utilizes two pre-trained parallel encoders: a visual encoder $\mathcal{E}_v(\cdot)$ and a textual encoder $\mathcal{E}_t(\cdot)$, which embed images and text descriptions into a shared embedding space $\mathbb{R}^d$. For a C-class classification task, CLIP performs zero-shot predictions by computing the similarities between the extracted image feature and the C candidate text features, written as

$$f_v = \mathcal{E}_v(X_{\texttt{test}}), \quad f_{t_c} = \mathcal{E}_t(\mathcal{T}_c), \tag{1}$$

$$\mathbb{P}_{\texttt{CLIP}}(y = y_c | X_{\texttt{test}}) = \frac{\exp\left(\text{sim}\left(f_{t_c}, f_v\right)/t\right)}{\sum_{t'} \exp\left(\text{sim}\left(f_{t'}, f_v\right)/t\right)}, \tag{2}$$

where $X_{\texttt{test}} \in \mathcal{D}_{\texttt{test}}$ denotes the input test image, and $\mathcal{T}_c$ represents the the class-specific description input for class y_c. The pairwise similarities $\text{sim}(\cdot, \cdot)$ are calculated using cosine similarity, and t represents the temperature parameter in the softmax function.

Test-Time Prompt Tuning. To enhance the zero-shot generalizability of CLIP, TPT (Shu et al., 2022) proposes learning an adaptive prompt using the test stream samples. Specifically, for each test sample $X_{\texttt{test}}$, TPT generates N aug-

mented views $\{\mathcal{A}_n(X_{\texttt{test}})\}_{n=1}^{N}$ and averages the top ρ-percentile confident predictions based on an entropy threshold τ to obtain the final prediction:

$$\mathbb{P}_{\texttt{TPT}}(X_{\texttt{test}}) = \frac{1}{\rho N}\sum_{n=1}^{N}\mathbb{1}[\mathcal{H}(\mathbb{P}(\mathcal{A}_n(X_{\texttt{test}}))\leq\tau]\,\mathbb{P}(\mathcal{A}_n(X_{\texttt{test}})). \quad (3)$$

Here, $\mathcal{H}(p) = -\sum_{i=1}^{C} p_i \log p_i$ calculates the self-entropy of the prediction p. The objective of TPT is to optimize the learnable prompt to minimize the self-entropy of the final prediction, *i.e.*, $\min \mathcal{H}(\mathbb{P}_{\texttt{TPT}}(X_{\texttt{test}}))$.

2.2. Dual Prototype Evolving

In our DPE method, we construct and iteratively evolve two sets of class-specific prototypes from both visual and textual modalities to achieve a more precise representation of each class over time.

Textual Prototype Evolution. In this work, we follow CLIP (Radford et al., 2021) to use multiple context prompt templates for prompt ensembling. Specifically, for each class c, we generate a total of S text descriptions, denoted as $\{\mathcal{T}_c^{(i)}\}_{i=1}^{S}$. The prototypes of these descriptions in the embedding space are calculated as $\mathbf{t}_c = \frac{1}{S}\sum_i \mathcal{E}_t(\mathcal{T}_c^{(i)})$. To further improve the quality of these prototypes over time, we design them to be updated online through a cumulative average with each individual sample $X_{\texttt{test}}$ in the test stream. The update rule is given by:

$$\mathbf{t} \leftarrow \frac{(k-1)\mathbf{t} + \mathbf{t}^*}{\|(k-1)\mathbf{t} + \mathbf{t}^*\|}, \quad k \leftarrow k+1, \quad (4)$$

where $\mathbf{t} = [\mathbf{t}_1\,\mathbf{t}_2\,\cdots\,\mathbf{t}_C]^\top \in \mathbb{R}^{C\times d}$ is the online updated prototype set, and $\mathbf{t}^* \in \mathbb{R}^{C\times d}$ is the optimized textual prototypes for each individual sample $X_{\texttt{test}}$ in Eq. (8). To ensure stable online updates, we set an entropy threshold τ_t to filter out low-confidence samples (for which $\mathcal{H}(\mathbb{P}_{\texttt{CLIP}}(X_{\texttt{test}})) < \tau_t$) from updating the online prototypes, and maintain a counter k for tracking confident samples.

Visual Prototype Evolution. Inspired by TDA (Karmanov et al., 2024), we recognize that the historical image features of test images can also be utilized to enhance CLIP's discrimination capability. Therefore, we design a priority queue strategy to store the top-M image features for each class and symmetrically compute a set of visual prototypes that evolve over time. Note that since we cannot access the labels of the test samples, we assign the image features to the queue according to their predicted pseudo-labels. The priority queue for each class c is initialized as empty, denoted as $q_c = \varnothing$. As test samples arrive, we store the image features f_c and the corresponding self-entropy h_c in the priority queue, represented as $q_c = \{(f_c^{(m)}, h_c^{(m)})\}_m$. The elements

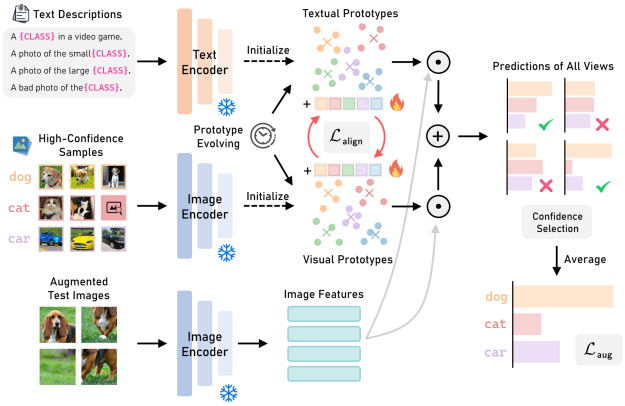

Figure 2. **An overview of our proposed DPE method**. We introduce two sets of prototypes from both textual and visual modalities and enable prototype-based inference with CLIP. For each test sample, we optimize the both prototypes using two sets of residual parameters with alignment loss $\mathcal{L}_{\text{align}}$ and self-entropy loss $\mathcal{L}_{\text{aug}}$. These prototypes are also progressively evolved over time to capture more accurate multi-modal representations for target classes.

are sorted by self-entropy $h_c^{(m)}$ such that $h_c^{(m)} < h_c^{(>m)}$. Using this priority queue, the class-specific visual prototype is obtained by: $\mathbf{v}_c = \frac{1}{S_c}\sum_m f_c^{(m)}$, where $S_c \leq M$ denotes the total number of image features stored in the queue.

The priority queues are updated during testing by replacing low-confidence image features with high-confidence ones. Specifically, for each individual test sample $X_{\texttt{test}}$, we first predict the pseudo-label ℓ and compute the self-entropy h as:

$$\ell = \arg\max_{y_c} \mathbb{P}(y = y_c|X_{\texttt{test}}), \quad h = \mathcal{H}(\mathbb{P}(X_{\texttt{test}})). \quad (5)$$

Then, we consider the following two scenarios to iteratively update the priority queue q_ℓ for class ℓ: (1) If the priority queue is not full, we directly add the pair $(\mathcal{E}_v(X_{\texttt{test}}), h)$ to the pqueue; (2) If the priority queue is full and the entropy h of the new sample is lower than the highest entropy value (the last element) currently in the queue, we replace the highest-entropy element with the new feature and self-entropy $(\mathcal{E}_v(X_{\texttt{test}}), h)$. If f is not lower, we discard the new sample and leave the queue unchanged. After each update, we re-sort the priority queue based on the self-entropy values and re-compute the visual prototypes $\mathbf{v} = [\mathbf{v}_1\,\mathbf{v}_2\,\cdots\,\mathbf{v}_C]^\top \in \mathbb{R}^{C\times d}$.

Prototype-Based Inference. Based on our two sets of multi-modal prototypes $\{\mathbf{t}_c\}_{c=1}^{C}$ and $\{\mathbf{v}_c\}_{c=1}^{C}$, the final prediction for input image feature f_v is given by

$$\mathbb{P}_{\texttt{Proto}}(y = y_c|X) = \frac{\exp\left(\left(f_v^\top \mathbf{t}_c + \mathcal{A}(f_v^\top \mathbf{v}_c)\right)/t\right)}{\sum_{c'}\exp\left(\left(f_v^\top \mathbf{t}_{c'} + \mathcal{A}(f_v^\top \mathbf{v}_{c'})\right)/t\right)}, \quad (6)$$

Here, t represents the temperature parameter, and $\mathcal{A}(x) = \alpha\exp\left(-\beta\,(1-x)\right)$ is the affinity function, where α is a balance hyperparameter and β is a sharpness ratio.

Table 1. **Performance comparisons on robustness to natural distribution shifts**. We present top-1 accuracy (%) results for all evaluated methods employing both ResNet-50 and ViT-B/16 visual backbones of CLIP. The best results are highlighted in **bold**.

Method	ImageNet	ImageNet-A	ImageNet-V2	ImageNet-R	ImageNet-S	Average	OOD Average
CLIP-ResNet-50 (Radford et al., 2021)	58.16	21.83	51.41	56.15	33.37	44.18	40.69
Ensemble	59.81	23.24	52.91	60.72	35.48	46.43	43.09
CoOp (Zhou et al., 2022b)	63.33	23.06	55.40	56.60	34.67	46.61	42.43
TPT (Shu et al., 2022)	60.74	26.67	54.70	59.11	35.09	47.26	43.89
DiffTPT (Feng et al., 2023)	60.80	**31.06**	55.80	58.80	37.10	48.71	45.69
TDA (Karmanov et al., 2024)	61.35	30.29	55.54	62.58	38.12	49.58	46.63
Ours	**63.41**	30.15	**56.72**	**63.72**	**40.03**	**50.81**	**47.66**
CLIP-ViT-B/16 (Radford et al., 2021)	66.73	47.87	60.86	73.98	46.09	59.11	57.20
Ensemble	68.34	49.89	61.88	77.65	48.24	61.20	59.42
CoOp (Zhou et al., 2022b)	71.51	49.71	64.20	75.21	47.99	61.72	59.28
TPT (Shu et al., 2022)	68.98	54.77	63.45	77.06	47.94	62.44	60.81
DiffTPT (Feng et al., 2023)	70.30	55.68	65.10	75.00	46.80	62.28	60.52
TDA (Karmanov et al., 2024)	69.51	**60.11**	64.67	80.24	50.54	65.01	63.89
Ours	**71.91**	59.63	**65.44**	**80.40**	**52.26**	**65.93**	**64.43**

2.3. Prototype Residual Learning

To further enhance the test-time generalization capabilities of VLMs, we update the multi-modal prototype sets for each test sample. Specifically, after being evolved with the last test sample, the dual sets of multi-modal prototypes, denoted as $\mathbf{t} = [\mathbf{t}_1\, \mathbf{t}_2\, \cdots\, \mathbf{t}_C]^\top \in \mathbb{R}^{C\times d}$ and $\mathbf{v} = [\mathbf{v}_1\, \mathbf{v}_2\, \cdots\, \mathbf{v}_C]^\top \in \mathbb{R}^{C\times d}$, are considered as the initialization for updating with the current test sample. We introduce two sets of learnable residual parameters $\hat{\mathbf{t}} = [\hat{\mathbf{t}}_1\, \hat{\mathbf{t}}_2\, \cdots\, \hat{\mathbf{t}}_C]^\top \in \mathbb{R}^{C\times d}$ and $\hat{\mathbf{v}} = [\hat{\mathbf{v}}_1\, \hat{\mathbf{v}}_2\, \cdots\, \hat{\mathbf{v}}_C]^\top \in \mathbb{R}^{C\times d}$. These parameters are initialized to zero and are used to optimize the prototypes for each given test input $X_{\texttt{test}}$, denoted as

$$\mathbf{t}_c \leftarrow \frac{\mathbf{t}_c + \hat{\mathbf{t}}_c}{\|\mathbf{t}_c + \hat{\mathbf{t}}_c\|}, \quad \mathbf{v}_c \leftarrow \frac{\mathbf{v}_c + \hat{\mathbf{v}}_c}{\|\mathbf{v}_c + \hat{\mathbf{v}}_c\|}. \tag{7}$$

Similar to TPT (Shu et al., 2022), we optimize these residual s to promote consistent predictions across a total of n different augmented views $\mathcal{A}_n(X_{\texttt{test}})$ of the given test image using the unsupervised entropy minimization objective $\mathcal{L}_{\texttt{aug}} = \mathcal{H}(\mathbb{P}_{\texttt{DPE}}(X_{\texttt{test}}))$.

However, researchers have shown that focusing solely on reducing entropy can lead the model to make overconfident predictions (Yoon et al., 2024). To address this, we apply an additional constraint to align the multi-modal prototypes during optimization, explicitly enforcing consistent multi-modal representations between dual sets of prototypes. Specifically, we introduce a self-supervised alignment loss that utilizes the contrastive InfoNCE loss (Oord et al., 2018) to bring prototypes from the same class closer together while pushing prototypes from different classes further apart:

$$\mathcal{L}_{\texttt{align}} = \frac{1}{C}\sum_{c=1}^{C}\left(-\log\frac{\exp(\mathbf{t}_c{}^\top \mathbf{v}_c)}{\sum_{c'}\exp(\mathbf{t}_c{}^\top \mathbf{v}_{c'})} - \log\frac{\exp(\mathbf{t}_c{}^\top \mathbf{v}_c)}{\sum_{c'}\exp(\mathbf{t}_{c'}{}^\top \mathbf{v}_c)}\right).$$

In summary, the final objective for optimizing the multi-modal prototypes $\mathbf{t}, \mathbf{v}$ is

$$\mathbf{t}^*, \mathbf{v}^* = \arg\min_{\mathbf{t},\mathbf{v}} \left(\mathcal{L}_{\texttt{aug}} + \lambda \mathcal{L}_{\texttt{align}}\right), \tag{8}$$

where λ is a scale factor to balance the contribution of the alignment loss.

After optimizing the prototypes for each test sample, we evolve the online textual prototypes $\mathbf{t}$ as described in Eq. (4), and also update the priority queues to re-compute the visual prototypes $\mathbf{v}$. The evolved prototype sets then serve as the initialization for the next test sample, progressively enhancing generalization capability during test-time adaptation.

3. Experiments

In this section, we evaluate our DPE on robustness to natural distribution shifts and cross-datasets generalization across 15 datasets. Specifically, we follow the experimental settings in Appendix A.1 to conduct these experiments.

Datasets. We follow previous work (Shu et al., 2022; Feng et al., 2023) to evaluate our method on two benchmarking scenarios, namely, robustness to natural distribution shifts and cross-datasets generalization. (1) For the evaluation of robustness to natural distribution shifts, we assess the performance of our method using the ImageNet (Deng et al., 2009) dataset alongside its variant out-of-distribution datasets, including ImageNet-A (Hendrycks et al., 2021b), ImageNet-V2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021a), and ImageNet-Sketch (Wang et al., 2019). (2) For cross-datasets generalization tasks, we conduct comprehensive assessments across 10 diverse recognition datasets, including FGVCAircraft (Maji et al., 2013), Caltech101 (Fei-Fei et al., 2007), StandfordCars (Krause

Table 2. **Performance comparisons on cross-datesets generalization.** We also present top-1 accuracy (%) for all methods on two backbones of CLIP. The best results are highlighted in **bold**.

Method	Aircraft	Caltech	Cars	DTD	EuroSAT	Flower	Food101	Pets	SUN397	UCF101	Average
CLIP-ResNet-50	15.66	85.88	55.70	40.37	23.69	61.75	73.97	83.57	58.80	58.84	55.82
Ensemble	16.11	87.26	55.89	40.37	25.79	62.77	74.82	82.97	60.85	59.48	56.63
CoOp (Zhou et al., 2022b)	15.12	86.53	55.32	37.29	26.20	61.55	75.59	**87.00**	58.15	59.05	56.18
TPT (Shu et al., 2022)	17.58	87.02	58.46	40.84	28.33	62.69	74.88	84.49	61.46	60.82	57.66
DiffTPT (Feng et al., 2023)	17.60	86.89	**60.71**	40.72	41.04	63.53	**79.21**	83.40	62.72	62.67	59.85
TDA (Karmanov et al., 2024)	17.61	89.70	57.78	43.74	**42.11**	**68.74**	77.75	86.18	62.53	**64.18**	61.03
Ours	**19.80**	**90.83**	59.25	**50.18**	41.67	67.60	77.83	85.97	**64.23**	61.98	**61.93**
CLIP-ViT-B/16	23.67	93.35	65.48	44.27	42.01	67.44	83.65	88.25	62.59	65.13	63.58
Ensemble	23.22	93.55	66.11	45.04	50.42	66.99	82.86	86.92	65.63	65.16	64.59
CoOp (Zhou et al., 2022b)	18.47	93.70	64.51	41.92	46.39	68.71	85.30	89.14	64.15	66.55	63.88
TPT (Shu et al., 2022)	24.78	94.16	66.87	47.75	42.44	68.98	84.67	87.79	65.50	68.04	65.10
DiffTPT (Feng et al., 2023)	25.60	92.49	67.01	47.00	43.13	70.10	**87.23**	88.22	65.74	62.67	65.47
TDA (Karmanov et al., 2024)	23.91	94.24	67.28	47.40	**58.00**	71.42	86.14	88.63	67.62	**70.66**	67.53
Ours	**28.95**	**94.81**	**67.31**	**54.20**	55.79	**75.07**	86.17	**91.14**	**70.07**	70.44	**69.40**

et al., 2013), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), OxfordPets (Parkhi et al., 2012), SUN397 (Xiao et al., 2010), and UCF101 (Soomro et al., 2012). These datasets offer a comprehensive benchmark for evaluating the robustness of various methods across different distributional variations.

Robustness to Natural Distribution Shifts. In Table 1, we compare the performance of our method with other state-of-the-art methods on in-domain ImageNet and its 4 out-of-distribution variants. Specifically, our method outperforms existing state-of-the-art prompt tuning methods, surpasses TPT (Shu et al., 2022) by 3.55% and 3.49% and DiffTPT (Feng et al., 2023) by 2.10% and 3.65% on average when using ResNet-50 and ViT-B/16 backbones, respectively. The experimental results demonstrate that our method achieves superior zero-shot generalization performance across various out-of-distribution datasets compared to other approaches.

Cross-Datasets Generalization. In Table 2, we further assess the generalizability of our proposed method against other state-of-the-art methods on 10 fine-grained recognition datasets. Given the significant distributional differences, methods may exhibit variable performance across these datasets. Notably, our method, which is not trained on any annotated data, significantly outperforms CoOp (Zhou et al., 2022b) by average margins of 5.75% and 5.52% on two respective backbones. Compared to other test-time adaptation methods, our method achieves the best performance on 7 out of 10 datasets and surpasses other methods by notable average margins ranging from 1.87% to 4.30% using the ViT-B/16 backbone. These results demonstrate the superior robustness and adaptability of our method in transferring to diverse domains during test time, which is crucial for real-world deployment scenarios.

Table 3. Efficiency comparison on ImageNet (Deng et al., 2009).

Method	Testing Time	Accuracy	Gain
CLIP (Radford et al., 2021)	9 min	59.81	-
TPT (Shu et al., 2022)	9 h 15 min	60.74	+0.93
DiffTPT (Feng et al., 2023)	$>$ 20 h	60.80	+0.99
Ours	1 h 50 min	**63.41**	**+3.60**

Efficiency Comparison. In Table 3, we compare the efficiency of our method with other test-time prompt tuning methods on the ImageNet (Deng et al., 2009) dataset using a single 48GB NVIDIA RTX 6000 Ada GPU. Our proposed method is $5\times$ faster than TPT (Shu et al., 2022) and over $10\times$ faster than DiffTPT (Feng et al., 2023), as it requires only learning the prototype residues without the need to back-propagate gradients through the textual encoder.

4. Conclusion

In this work, we introduce Dual Prototype Evolving (DPE), a novel and effective approach for enhancing the zero-shot generalizability of VLMs during test time. Unlike previous methods that only focus on adapting the VLMs from one modality, we create and evolve two sets of prototypes—textual and visual—progressively capturing more accurate multi-modal representations for target classes during test time. Further, we also introduce prototype residual learning to optimize the dual prototype sets for each individual test sample, which further enhances the test-time generalization capabilities of VLMs. Through comprehensive experiments, we demonstrate that our proposed DPE achieves state-of-the-art performance while also exhibiting competitive test-time efficiency.

Acknowledgements. The project has been funded by ARL award W911NF–2320007, Department of Agriculture award 20236702139073, ONR award N00014-23-1-2840 and AFRL/AFOSR award number FA9550-18-1-0251.

References

Abdul Samadh, J., Gani, M. H., Hussein, N., Khattak, M. U., Naseer, M. M., Shahbaz Khan, F., and Khan, S. H. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. *Advances in Neural Information Processing Systems*, 36, 2023.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, pp. 446–461. Springer, 2014.

Boudiaf, M., Mueller, R., Ben Ayed, I., and Bertinetto, L. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8344–8353, 2022.

Cho, E., Kim, J., and Kim, H. J. Distribution-aware prompt tuning for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22004–22013, 2023.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Deng, Z., Chen, Z., Niu, S., Li, T., Zhuang, B., and Tan, M. Efficient test-time adaptation for super-resolution with second-order degradation and reconstruction. *Advances in Neural Information Processing Systems*, 36, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Du, Y., Liu, Z., Li, J., and Zhao, W. X. A survey of vision-language pre-trained models. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 5436–5443, 2022.

Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

Feng, C.-M., Yu, K., Liu, Y., Khan, S., and Zuo, W. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2704–2714, 2023.

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., and Qiao, Y. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132:581–595, 2024.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

He, W., Jamonnak, S., Gou, L., and Ren, L. Clip-s4: Language-guided self-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11207–11216, 2023.

Hegde, D., Valanarasu, J. M. J., and Patel, V. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2028–2038, 2023.

Helber, P., Bischke, B., Dengel, A., and Borth, D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b.

Hu, M., Song, T., Gu, Y., Luo, X., Chen, J., Chen, Y., Zhang, Y., and Zhang, S. Fully test-time adaptation for image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pp. 251–260. Springer, 2021.

Hu, X., Zhang, C., Zhang, Y., Hai, B., Yu, K., and He, Z. Learning to adapt clip for few-shot monocular depth estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5594–5603, 2024.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916, 2021.

Kan, Z., Chen, S., Zhang, C., Tang, Y., and He, Z. Self-correctable and adaptable inference for generalizable human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5537–5546, 2023.

Karmanov, A., Guan, D., Lu, S., Saddik, A. E., and Xing, E. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., and Khan, F. S. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

Li, X., Lian, D., Lu, Z., Bai, J., Chen, Z., and Wang, X. Graphadapter: Tuning vision-language models with dual knowledge graph. *Advances in Neural Information Processing Systems*, 36, 2023.

Li, Y., Hao, M., Di, Z., Gundavarapu, N. B., and Wang, X. Test-time personalization with a transformer for human pose estimation. *Advances in Neural Information Processing Systems*, 34:2583–2597, 2021.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=zq1iJkNk3uN.

Liang, J., He, R., and Tan, T. A comprehensive survey on test-time adaptation under distribution shifts. *arXiv preprint arXiv:2303.15361*, 2023.

Lin, Y., Chen, M., Wang, W., Wu, B., Li, K., Lin, B., Liu, H., and He, X. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15305–15314, 2023a.

Lin, Z., Yu, S., Kuang, Z., Pathak, D., and Ramanan, D. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19325–19337, 2023b.

Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., and Zhou, J. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Ma, X., Zhang, J., Guo, S., and Xu, W. Swapprompt: Test-time prompt adaptation for vision-language models. *Advances in Neural Information Processing Systems*, 36, 2023.

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., and Vedaldi, A. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Mirza, M. J., Micorek, J., Possegger, H., and Bischof, H. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14765–14775, 2022.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729. IEEE, 2008.

Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., and Tan, M. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=g2YraF75Tj.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. Cats and dogs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012.

Pratt, S., Covert, I., Liu, R., and Farhadi, A. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15691–15701, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

Shen, S., Yang, S., Zhang, T., Zhai, B., Gonzalez, J. E., Keutzer, K., and Darrell, T. Multitask vision-language prompt tuning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5656–5667, 2024.

Shin, I., Tsai, Y.-H., Zhuang, B., Schulter, S., Liu, B., Garg, S., Kweon, I. S., and Yoon, K.-J. Mm-tta: Multi-modal test-time adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16928–16937, 2022.

Shocher, A., Cohen, N., and Irani, M. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3118–3126, 2018.

Shu, M., Nie, W., Huang, D.-A., Yu, Z., Goldstein, T., Anandkumar, A., and Xiao, C. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35: 14274–14289, 2022.

Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Sui, E., Wang, X., and Yeung-Levy, S. Just shift it: Test-time prototype shifting for zero-shot generalization with vision-language models. *arXiv preprint arXiv:2403.12952*, 2024.

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., and Hardt, M. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, pp. 9229–9248. PMLR, 2020.

Tang, Y., Zhang, C., Xu, H., Chen, S., Cheng, J., Leng, L., Guo, Q., and He, Z. Neuro-modulated hebbian learning for fully test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3728–3738, 2023.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, volume 32, pp. 10506–10518, 2019.

Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., and Liu, T. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11686–11695, 2022.

Wei, Y., Hu, H., Xie, Z., Liu, Z., Zhang, Z., Cao, Y., Bao, J., Chen, D., and Guo, B. Improving clip fine-tuning performance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5439–5449, 2023.

Wu, X., Zhu, F., Zhao, R., and Li, H. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7031–7040, 2023.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010.

Yoon, H. S., Yoon, E., Tee, J. T. J., Hasegawa-Johnson, M. A., Li, Y., and Yoo, C. D. C-TPT: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jzzEHTBFOT.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=Ee277P3AYC.

Yu, T., Lu, Z., Jin, X., Chen, Z., and Wang, X. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10899–10909, 2023.

Zanella, M. and Ayed, I. B. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18123–18133, 2022.

Zhang, C., Stepputtis, S., Sycara, K., and Xie, Y. Negative yields positive: Unified dual-path adapter for vision-language models. *arXiv preprint arXiv:2403.12964*, 2024a.

Zhang, J., Huang, J., Jin, S., and Lu, S. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024b.

Zhang, M., Levine, S., and Finn, C. Memo: Test time robustness via adaptation and augmentation. *Advances in Neural Information Processing Systems*, 35:38629–38642, 2022a.

Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., and Li, H. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, pp. 493–510. Springer, 2022b.

Zhang, Y., Borse, S., Cai, H., and Porikli, F. Auxadapt: Stable and efficient test-time adaptation for temporally consistent video semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2339–2348, 2022c.

Zhang, Y., Zhang, C., Liao, Z., Tang, Y., and He, Z. Bdc-adapter: Brownian distance covariance for better vision-language reasoning. In *British Machine Vision Conference*. BMVA, 2023. URL https://papers.bmvc2023.org/0182.pdf.

Zhang, Y., Zhang, C., Yu, K., Tang, Y., and He, Z. Concept-guided prompt learning for generalization in vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7377–7386, 2024c.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022a.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

Zhu, B., Niu, Y., Han, Y., Wu, Y., and Zhang, H. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15659–15669, 2023.

Test-Time Prototype Evolving for Generalizable Vision-Language Models

In this supplementary document, we provide additional details and experimental results to enhance understanding and insights into our method. This supplementary document is organized as follows:

- We specify all the experimental settings in Section A.1.
- Full numerical results on robustness to natural distribution shifts are detailed in Section A.2.
- We conduct ablation studies and present the results in Section A.3.
- We discuss releted works in Section B.
- Detailed statistics for all utilized datasets are provided in Section C.1.
- We present the specific positive and negative prompts we used for each dataset in Section C.2.
- We list the license information for all used assets in Section D.
- Finally, we discuss the limitations and broader impacts of this work in Section E.

A. Additional Experimental Details

A.1. Experimental Settings

Datasets. We follow previous work (Shu et al., 2022; Feng et al., 2023) to evaluate our method on two benchmarking scenarios, namely, robustness to natural distribution shifts and cross-datasets generalization. (1) For the evaluation of robustness to natural distribution shifts, we assess the performance of our method using the ImageNet (Deng et al., 2009) dataset alongside its variant out-of-distribution datasets, including ImageNet-A (Hendrycks et al., 2021b), ImageNet-V2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021a), and ImageNet-Sketch (Wang et al., 2019). (2) For cross-datasets generalization tasks, we conduct comprehensive assessments across 10 diverse recognition datasets, including FGVCAircraft (Maji et al., 2013), Caltech101 (Fei-Fei et al., 2007), StandfordCars (Krause et al., 2013), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), OxfordPets (Parkhi et al., 2012), SUN397 (Xiao et al., 2010), and UCF101 (Soomro et al., 2012). These datasets offer a comprehensive benchmark for evaluating the robustness of various methods across different distributional variations.

Implementation Details. We follow previous works (Shu et al., 2022; Feng et al., 2023) to adopt ResNet-50 (He et al., 2016) and ViT-B/16 (Dosovitskiy et al., 2020) backbones as the visual encoder of CLIP. In Appendix C.2, we detail the specific hand-crafted prompts utilized for each dataset. Following TPT (Shu et al., 2022), we generate 63 augmented views for each test image using random resized cropping to create a batch of 64 images. We learn the prototype residual parameters using AdamW (Loshchilov & Hutter, 2019) optimizer with a learning rate of 0.0005 for a single step. In default, the scale factor λ in Eq. (8) is set to 0.5, the normalized entropy threshold τ_t is set to 0.1, and the queue size M is set to 3. For the affinity function in Eq. (6), we set $\alpha = 6.0$ and $\beta = 5.0$, respectively. All experiments are conducted on a single 48GB NVIDIA RTX 6000 Ada GPU. To ensure the reliability of our results, we perform each experiment three times using different initialization seeds and report the mean accuracy achieved. We will make the source code publicly available upon acceptance to facilitate reproducibility.

Baselines. We compare our method with established test-time adaptation approaches for CLIP: (1) TPT (Shu et al., 2022), a prompt tuning method which aims to minimize self-entropy across predictions of multiple augmented views; (2) DiffTPT (Feng et al., 2023), an enhanced version of TPT that utilizes diffusion-based augmentations to optimize prompts; (3) TDA (Karmanov et al., 2024), a training-free, adapter-based method which constructs positive and negative caches during test time. Additionally, we present the zero-shot performance of CLIP using the simple prompt "`a photo of {CLASS}`" as well as the results from prompt ensembling to show the absolute performance improvements. We also report the performance of CoOp (Zhou et al., 2022b), a train-time adaptation method, using 16-shot annotated samples per class on ImageNet. For a fair comparison, we directly report the results of these baselines from their respective original papers.

Note that in the DiffTPT (Feng et al., 2023) paper, the results are based on a subset of the datasets containing 1,000 test samples. This limited sample size may introduce potential imprecision in the reported results.

A.2. Full Results on Robustness to Natural Distribution Shifts

In Table A1, we compare the performance of our method with other state-of-the-art methods on in-domain ImageNet and its 4 out-of-distribution variants. Specifically, we demonstrate that our DPE can also be applied to prompts learned using CoOp (Zhou et al., 2022b) with a 16-shot ImageNet setup. Our methods also demonstrates competitive performance compared to other methods. It is also important to notice that, our proposed method accumulates task-specific knowledge over time, therefore can achieve higher performance gain on a larger test set (*e.g.*, ImageNet-R and ImageNet-S).

Table A1. **Performance comparisons on robustness to natural distribution shifts**. We present top-1 accuracy (%) results for all evaluated methods employing both ResNet-50 and ViT-B/16 visual backbones of CLIP. Additionally, we assess the performance using prompts learned by CoOp (Zhou et al., 2022b) with 16-shot training data per class on ImageNet (Deng et al., 2009). The best results are highlighted in **bold**.

Method	ImageNet	ImageNet-A	ImageNet-V2	ImageNet-R	ImageNet-S	Average	OOD Average
CLIP-ResNet-50 (Radford et al., 2021)	58.16	21.83	51.41	56.15	33.37	44.18	40.69
Ensemble	59.81	23.24	52.91	60.72	35.48	46.43	43.09
TPT (Shu et al., 2022)	60.74	26.67	54.70	59.11	35.09	47.26	43.89
DiffTPT (Feng et al., 2023)	60.80	**31.06**	55.80	58.80	37.10	48.71	45.69
TDA (Karmanov et al., 2024)	61.35	30.29	55.54	62.58	38.12	49.58	46.63
Ours	**63.41** (± 0.23)	30.15 (± 0.41)	**56.72** (± 0.22)	**63.72** (± 0.20)	**40.03** (± 0.11)	**50.81** (± 0.21)	**47.66** (± 0.22)
CoOp (Zhou et al., 2022b)	63.33	23.06	55.40	56.60	34.67	46.61	42.43
TPT + CoOp (Shu et al., 2022)	64.73	30.32	57.83	58.99	35.86	49.55	45.75
DiffTPT + CoOp (Feng et al., 2023)	64.70	**32.96**	**61.70**	58.20	36.80	**50.87**	**47.42**
Ours + CoOp	**64.86** (± 0.18)	30.08 (± 0.27)	57.96 (± 0.31)	**59.78** (± 0.19)	**37.80** (± 0.17)	50.10 (± 0.22)	46.41 (± 0.23)
CLIP-ViT-B/16 (Radford et al., 2021)	66.73	47.87	60.86	73.98	46.09	59.11	57.20
Ensemble	68.34	49.89	61.88	77.65	48.24	61.20	59.42
TPT (Shu et al., 2022)	68.98	54.77	63.45	77.06	47.94	62.44	60.81
DiffTPT (Feng et al., 2023)	70.30	55.68	65.10	75.00	46.80	62.28	60.52
TDA (Karmanov et al., 2024)	69.51	**60.11**	64.67	80.24	50.54	65.01	63.89
Ours	**71.91** (± 0.09)	59.63 (± 0.18)	**65.44** (± 0.17)	**80.40** (± 0.24)	**52.26** (± 0.11)	**65.93** (± 0.16)	**64.43** (± 0.18)
CoOp (Zhou et al., 2022b)	71.51	49.71	64.20	75.21	47.99	61.72	59.28
TPT + CoOp (Shu et al., 2022)	73.61	57.95	**66.83**	77.27	49.29	64.99	62.83
DiffTPT + CoOp (Feng et al., 2023)	**75.00**	58.09	66.80	73.90	49.50	64.12	61.97
Ours + CoOp	73.67 (± 0.14)	**59.43** (± 0.36)	66.38 (± 0.32)	78.49 (± 0.06)	**50.78** (± 0.08)	**65.75** (± 0.23)	**63.77** (± 0.26)

A.3. Ablation Studies

Different Textual Prototype Evolution Rules. In Table A2, we report the performance on ImageNet (Deng et al., 2009) using different textual prototype evolution rules. We have the following key observations: (1) Fully updating our textual prototypes $\mathbf{t}$ to the optimized prototypes $\mathbf{t}^*$ for each individual test image results in collapsed performance; (2) Compared to not evolving the textual prototypes, using an exponential moving average update rule with a decay rate of 0.99 leads to a slight performance improvement of 0.18%; however, setting a lower decay rate of 0.95 decreases the performance by 0.36%. (3) Our cumulative average update rule yields the highest performance, achieving a 0.48% improvement compared to no update on ImageNet (Deng et al., 2009).

Table A2. Performance comparison using different textual prototype evolution rules on ImageNet.

Update Rule	Formula	Accuracy
No Update	$\mathbf{t} \leftarrow \mathbf{t}$	62.93
Full Update	$\mathbf{t} \leftarrow \mathbf{t}^*$	21.83
Exponential Avg.	$\mathbf{t} \leftarrow 0.99\mathbf{t} + 0.01\mathbf{t}^*$	63.11
Exponential Avg.	$\mathbf{t} \leftarrow 0.95\mathbf{t} + 0.05\mathbf{t}^*$	62.57
Cumulative Avg.	$\mathbf{t} \leftarrow ((k-1)\mathbf{t} + \mathbf{t}^*)/k$	**63.41**

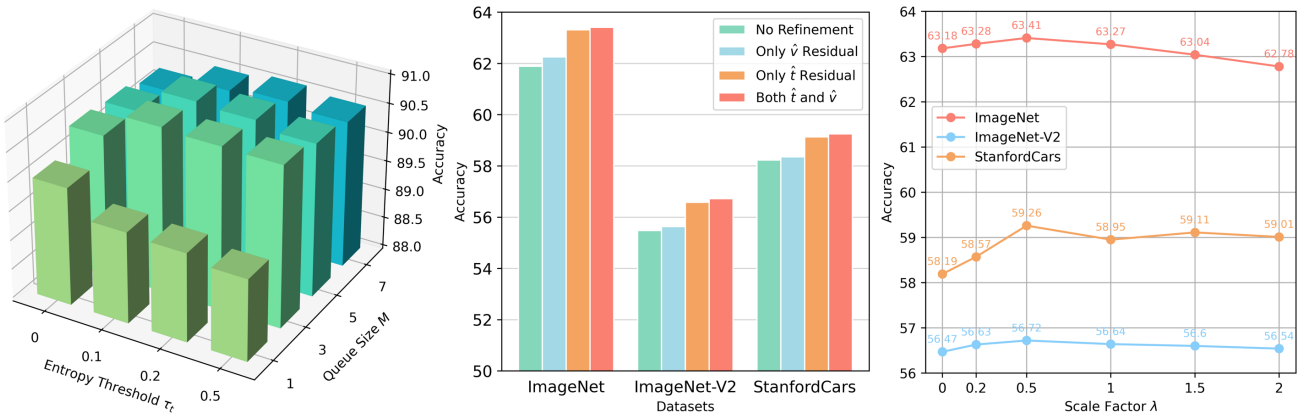

Figure A1. **Ablation studies**. (*Left*) Sensitivity analysis of τ_t and M on Caltech101 (Fei-Fei et al., 2007); (*Middle*) Analysis of the performance contributions from various learnable parameter settings across three datasets; (*Right*) Performance on three datasets with varying scale factor λ in Eq. (8).

Hyperparameters for Dual Prototype Evolution. We provide a sensitivity analysis for the hyperparameters τ_t and M on the Caltech101 (Fei-Fei et al., 2007) dataset in Figure A1 (*Left*). Specifically, τ_t represents the normalized entropy threshold for evolving our textual prototypes. When $\tau_t = 0$, our method does not evolve the textual prototypes, leading to a significant performance decrease, as shown in Figure A1 (*Left*). Moreover, setting $\tau_t = 0.1$ results in the highest performance, whereas a higher threshold leads to a slight decrease in performance. Additionally, the queue size M acts as a soft threshold hyperparameter for evolving the visual prototypes. Our setting of $M = 3$ consistently yields the highest performance. Lowering M causes the visual prototypes to fail in capturing the diversity of test samples from the same class, while increasing M introduces additional low-confidence noisy samples that hinder discrimination among target classes.

More Sensitivity Analyses of Hyper-Parameters. In our experiments on ImageNet (Deng et al., 2009), we set the hyperparameters α and β as defined in Eq. (6) to 6.0 and 5.0, respectively, as detailed in the implementation section. To thoroughly examine the impact of different hyperparameters, we performed a sensitivity analysis by varying each hyperparameter individually and assessing the performance on ImageNet with a ResNet-50 backbone, as shown in Table A3. The results show that our selected values of $\alpha = 6.0$ and $\beta = 5.0$ provide the best performance.

Table A3. **Sensitivity of hyper-parameters**. All the results are reported on ImageNet (Deng et al., 2009) using ResNet-50 backbone.

α	2.5	4.0	5.0	**6.0**	7.5	10.0
	62.83	63.17	63.28	**63.41**	63.07	62.43
β	2.0	3.0	4.0	**5.0**	6.0	7.0
	62.85	63.02	63.30	**63.41**	63.37	63.29

Effects of Different Learnable Modules. Recall that in our DPE method, we optimize our multi-modal prototypes by introducing two sets of learnable residual parameters $\hat{\mathbf{t}}$ and $\hat{\mathbf{v}}$ for each individual test image. In Figure A1 (*Middle*), we ablate the effects of each set of learnable residual parameters and report the performance across three datasets. Specifically, on ImageNet (Deng et al., 2009), optimizing only the textual prototypes for individual samples results in a 1.40% improvement, while optimizing only the visual prototypes yields a non-trivial 0.36% improvement, compared to keeping both $\hat{\mathbf{t}}$ and $\hat{\mathbf{v}}$ fixed. Optimizing both sets of residual parameters leads to a further performance increase, *e.g.*, by 1.52% on ImageNet (Deng et al., 2009). This indicates both learnable modules contribute to the overall effectiveness of DPE.

Scaling the Alignment Loss. Finally, we ablate the effect of the alignment loss by varying the scale factor λ in Figure A1 (*Right*). Compared to optimizing solely using entropy minimization loss (*i.e.*, $\lambda = 0$) during test-time adaptation, applying the additional alignment loss results in a performance improvement of 0.23% to 1.07% across three different datasets. However, there is a trade-off between prototype alignment and self-entropy minimization: setting λ too high leads to a performance drop. Our experiments show that our setting of $\lambda = 0.5$ yields the highest performance.

B. Related Work

Vision-Language Models. Leveraging vast image-text pairs from the Internet, recent large-scale vision-language models (VLMs), such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), have shown remarkable and transferable visual knowledge through natural language supervision (Zhang et al., 2024b; Du et al., 2022). These VLMs enable a "pre-train,

fine-tune" paradigm for performing downstream visual tasks, such as recognition (Radford et al., 2021; Hegde et al., 2023; Liu et al., 2024), segmentation (Wang et al., 2022; Lin et al., 2023a; He et al., 2023) and detection (Wu et al., 2023; Wei et al., 2023). To effectively transfer VLMs to these downstream tasks, researchers have developed two primary methods for adapting the model with few-shot data: prompt learning methods (Zhou et al., 2022b;a; Khattak et al., 2023; Shen et al., 2024; Zhu et al., 2023; Cho et al., 2023; Zhang et al., 2024c; Hu et al., 2024) and adapter-based methods (Zhang et al., 2022b; Gao et al., 2024; Zhang et al., 2024a; Yu et al., 2023; Li et al., 2023; Zhang et al., 2023). For instance, CoOp (Zhou et al., 2022b) and CoCoOp (Zhou et al., 2022a) explores input prompt learning with few-shot downstream data supervision, while Tip-Adapter (Zhang et al., 2022b) and TaskRes (Yu et al., 2023) directly modify the extracted visual or textual representations. However, these approaches often assume the availability of labeled samples from the target domain, which can limit their effectiveness in real-world scenarios. In this work, we focus on the test-time adaptation setting, where we have no access to any training samples. Our goal is to adapt the model during test time without any ground-truth labels.

Test-Time Adaptation. To effectively transfer a model trained on the source domain to the target domain, test-time adaptation methods (Wang et al., 2021; Zhang et al., 2022a; Tang et al., 2023; Boudiaf et al., 2022) aim to adjust the model online using a stream of unlabeled test samples. These methods enable the deployment of well-trained models in various out-of-distribution scenarios, thereby enhancing the applicability and reliability of machine learning models in real-world applications (Liang et al., 2023; Koh et al., 2021; Niu et al., 2023). Researchers have applied test-time adaptation techniques successfully across various machine learning tasks, including semantic segmentation (Hu et al., 2021; Shin et al., 2022; Zhang et al., 2022c), human pose estimation (Li et al., 2021; Kan et al., 2023), and image super-resolution (Shocher et al., 2018; Deng et al., 2023).

Recently, increasing research efforts have focused on adapting large-scale VLMs during test time (Ma et al., 2023; Sui et al., 2024; Abdul Samadh et al., 2023; Zanella & Ayed, 2024). As the seminal work, Shu *et al*. (Shu et al., 2022) firstly propose test-time prompt tuning (TPT), which enforces consistency across different augmented views of each test sample. Building on this approach, several subsequent studies have sought to further enhance TPT. For instance, DiffTPT (Shu et al., 2022) utilizes diffusion-based augmentations to increase the diversity of augmented views, while C-TPT (Yoon et al., 2024) addresses the rise in calibration error during test time prompt tuning. Unlike these approaches, which treat each test sample independently, TDA (Karmanov et al., 2024) establishes positive and negative visual caches during test time, enhancing model performance as more samples are processed. However, these methods solely adapt the model from a single modality perspective, limiting their effectiveness in capturing task-specific knowledge from out-of-distribution domains. Given this, we design DPE to evolve two sets of prototypes from both textual and visual modalities to progressively capture more accurate multi-modal representations for target classes during test time.

C. Additional Implementation Details

C.1. Dataset Details

In Table C4, we present the detailed statistics of each dataset we used in our experiments, including the number of classes, the sizes of training, validation and testing sets, and their original tasks.

C.2. Textual Prompts Used in Experiments

In Table C5, we detail the specific hand-crafted prompts utilized for each dataset.

D. License Information

Datasets. We list the known license information for the datasets below:

- MIT License: ImageNet-A (Hendrycks et al., 2021b), ImageNet-V2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021a), and ImageNet-Sketch (Wang et al., 2019).
- CC BY-SA 4.0 License: OxfordPets (Parkhi et al., 2012).
- Research purposes only: ImageNet (Deng et al., 2009), StandfordCars (Krause et al., 2013), DTD (Cimpoi et al., 2014), FGVCAircraft (Maji et al., 2013), SUN397 (Xiao et al., 2010).

Code. In this work, we also use some code implementations from existing codebase: CLIP (Radford et al., 2021), CoOp (Zhou et al., 2022b), TPT (Shu et al., 2022), and TDA (Karmanov et al., 2024). The code used in this paper are all

Table C4. **Detailed statistics of datasets used in experiments**. Note that the last 4 ImageNet variant datasets are designed for evaluation and only contain the test sets.

Dataset	Classes	Training	Validation	Testing	Task
Caltech101 (Fei-Fei et al., 2007)	100	4,128	1,649	2,465	Object recognition
DTD (Cimpoi et al., 2014)	47	2,820	1,128	1,692	Texture recognition
EuroSAT (Helber et al., 2019)	10	13,500	5,400	8,100	Satellite image recognition
FGVCAircraft (Maji et al., 2013)	100	3,334	3,333	3,333	Fine-grained aircraft recognition
Flowers102 (Nilsback & Zisserman, 2008)	102	4,093	1,633	2,463	Fine-grained flowers recognition
Food101 (Bossard et al., 2014)	101	50,500	20,200	30,300	Fine-grained food recognition
ImageNet (Deng et al., 2009)	1,000	1.28M	-	50,000	Object recognition
OxfordPets (Parkhi et al., 2012)	37	2,944	736	3,669	Fine-grained pets recognition
StanfordCars (Krause et al., 2013)	196	6,509	1,635	8,041	Fine-grained car recognition
SUN397 (Xiao et al., 2010)	397	15,880	3,970	19,850	Scene recognition
UCF101 (Soomro et al., 2012)	101	7,639	1,898	3,783	Action recognition
ImageNet-V2 (Recht et al., 2019)	1,000	-	-	10,000	Robustness of collocation
ImageNet-Sketch (Wang et al., 2019)	1,000	-	-	50,889	Robustness of sketch domain
ImageNet-A (Hendrycks et al., 2021b)	200	-	-	7,500	Robustness of adversarial attack
ImageNet-R (Hendrycks et al., 2021a)	200	-	-	30,000	Robustness of multi-domains

Table C5. **Textual prompts used in experiments**. In addition to these prompts, we also employ CuPL (Pratt et al., 2023) prompts to further enhance performance.

Dataset	Prompts
ImageNet (Deng et al., 2009) ImageNet-V2 (Recht et al., 2019) ImageNet-Sketch (Wang et al., 2019) ImageNet-A (Hendrycks et al., 2021b) ImageNet-R (Hendrycks et al., 2021a)	"itap of a {CLASS}." "a bad photo of the {CLASS}." "a origami {CLASS}." "a photo of the large {CLASS}." "a {CLASS} in a video game." "art of the {CLASS}." "a photo of the small {CLASS}."
Caltech101 (Fei-Fei et al., 2007)	"a photo of a {CLASS}."
DTD (Cimpoi et al., 2014)	"{CLASS} texture."
EuroSAT (Helber et al., 2019)	"a centered satellite photo of {CLASS}."
FGVCAircraft (Maji et al., 2013)	"a photo of a {CLASS}, a type of aircraft."
Flowers102 (Nilsback & Zisserman, 2008)	"a photo of a {CLASS}, a type of flower."
Food101 (Bossard et al., 2014)	"a photo of {CLASS}, a type of food."
OxfordPets (Parkhi et al., 2012)	"a photo of a {CLASS}, a type of pet."
StanfordCars (Krause et al., 2013)	"a photo of a {CLASS}."
SUN397 (Xiao et al., 2010)	"a photo of a {CLASS}."
UCF101 (Soomro et al., 2012)	"a photo of a person doing {CLASS}."

under the MIT License.

E. Further Discussions

Limitations. While our proposed DPE method effectively adapts CLIP to out-of-distribution domains during test time, we identify two potential limitations: (1) It still requires gradient back-propagation to optimize the multi-modal prototypes. This optimization process introduces additional computational complexity compared to zero-shot CLIP (Radford et al., 2021), which may affect its real-time performance in real-world deployment scenarios. (2) Since DPE needs to maintain priority queues to evolve the visual prototypes, it increases the memory cost during inference.

Broader Impacts. In this work, we aim to build more reliable machine learning systems by leveraging the extensive knowledge of current foundational models. Specifically, we follow TPT (Shu et al., 2022) to apply the test-time adaptation setting to vision-language models to align with real-world scenarios. By employing our DPE approach, the CLIP model can adapt itself to diverse domains during test time, which enhances its practical applicability in real-world deployment scenarios. We hope this work inspires future studies to focus on the generalization and robustness of pre-trained large-scale foundation models.