DART: Dual-Modal Adaptive Online Prompting and Knowledge Retention for Test-Time Adaptation

Zichen Liu, Hongbo Sun, Yuxin Peng, Jiahuan Zhou*

Wangxuan Institute of Computer Technology, Peking University
lzc20180720@stu.pku.edu.cn, {sunhongbo, pengyuxin, jiahuanzhou}@pku.edu.cn

Abstract

As an up-and-coming area, CLIP-based pre-trained vision-language models can readily facilitate downstream tasks through the zero-shot or few-shot fine-tuning manners. However, they still face critical challenges in test-time generalization due to the shifts between the training and test data distributions, hindering the further improvement of the performance. To address this crucial problem, the latest works have introduced Test-Time Adaptation (TTA) techniques to CLIP which dynamically learn text prompts using only test samples. However, their limited learning capacity due to the overlook of visual modality information, and the underutilization of knowledge in previously seen test samples result in reduced performance. In this paper, we propose a novel **D**ual-modal **A**daptive online prompting and knowledge **ReT**ention method called **DART** to overcome these challenges. To increase the learning capacity, DART captures knowledge from each test sample by learning class-specific text prompts and instance-level image prompts. Additionally, to fully leverage the knowledge from previously seen test samples, DART utilizes dual-modal knowledge retention prompts to adaptively retain the acquired knowledge, thereby enhancing the predictions on subsequent test samples. Extensive experiments on various large-scale benchmarks demonstrate the effectiveness of our proposed DART against state-of-the-art methods.

Introduction

Recently, the emergence of CLIP-based pre-trained vision-language models (Radford et al. 2021; Zhou et al. 2022b,a) has significantly propelled the advancement in computer vision. Through the exploration of appropriately designed text prompts, the pre-trained CLIP can be adapted to various downstream tasks for test-time inference (Luo et al. 2022; Wang et al. 2022a,c; Gal et al. 2022). However, manual crafting of task-specific prompts requires linguistic expertise and is time-consuming. A naive solution, that fine-tuning the entire pre-trained CLIP model on downstream tasks, will inevitably incur a large computational overhead and hinder the generalization ability of CLIP. An alternative approach involves adjusting the text prompts through few-shot fine-tuning, allowing the model to adapt to downstream tasks (Zhou et al. 2022b,a; Khattak et al. 2023). However, due to the distribution shifts between the training and test data, such methods may encounter severe performance limitations. To address this issue, the latest approaches (Shu et al. 2022; Niu et al. 2023) focus on enhancing the performance of pre-trained models during the test phase by adapting them to fit the distribution of test data, which is known as Test-Time Adaptation (TTA).

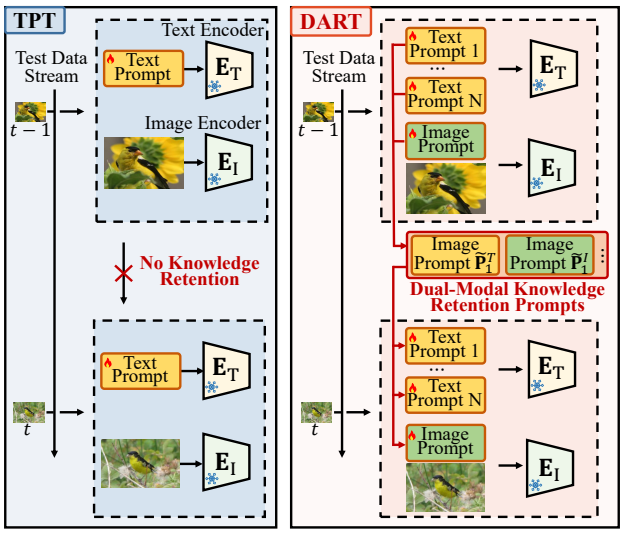

Figure 1: The latest CLIP-based test-time adaptation method TPT (Shu et al. 2022) learns an independent instance-level text prompt for each sample. In contrast, our proposed DART utilizes dual-modal instance-level prompts and knowledge retention prompts to comprehensively capture sample-specific knowledge and retain the acquired knowledge, enhancing the model's adaptability.

Facing the practical challenge of the small batch size of test-time inference, or even one individual sample per batch, various TTA methods have been proposed (Wang et al. 2021; Shu et al. 2022; Niu et al. 2022; Döbler, Marsden, and Yang 2023; Niu et al. 2023). These methods primarily adopt three strategies, batch normalization calibration (Schneider et al. 2020; Wang et al. 2021; Niu et al.

*Corresponding author

2023), consistency regularization (Yuan, Xie, and Li 2023; Döbler, Marsden, and Yang 2023), and anti-forgetting regularization (Niu et al. 2022; Shu et al. 2022; Song et al. 2023) to align the pre-trained model with the distribution of test data. However, the above methods neither fully leverage the dual-modality knowledge contained in the pre-trained CLIP nor effectively utilize the knowledge of previously seen test samples. Specifically, the recent work (Shu et al. 2022) preserves the knowledge of training data by freezing the backbone and adapts the model during test time through learning an instance-level unified text prompt for all classes. However, the learning of instance-level text prompts for each test sample is independent and the historical knowledge from the test samples that have already been encountered can not be utilized. Moreover, the overlook of the image modality counterpart severely limits the cross-modality capability of CLIP.

With the breakthrough of prompt learning technology in the field of natural language processing (Tsimpoukelli et al. 2021; Li and Liang 2021), various methods migrate prompt learning to the field of computer vision (Jia et al. 2022; Gao et al. 2022; Wang et al. 2022e), which adapt the pre-trained model to downstream tasks by learning few additional parameters. Therefore, in this paper, we propose a novel CLIP-based **D**ual-modal **A**daptive online prompting method for knowledge **R**e**T**ention in TTA, called **DART**. As demonstrated in Figure 1, DART involves both text and visual prompts to effectively capture the individual knowledge of each test sample, with the goal to enhance the model's prediction accuracy. Furthermore, to fully leverage the knowledge contained in historical test samples, dual-modal knowledge retention prompts are designed to adaptively retain the historical knowledge, and facilitate predictions on subsequent test samples. As a result, our proposed DART can effectively and adaptively tackle unseen test instances to improve the overall performance via TTA. In summary, the contributions of this paper are three-fold:

- To enhance the test-time generalization ability of CLIP and mitigate the severe training-test distribution shifting challenge, the proposed DART utilizes dual-modal online prompts to thoroughly capture information from each individual test sample, thereby improving its prediction accuracy.
- To fully leverage the knowledge from historical test samples, DART proposes dual-modal knowledge retention prompts to adaptively retain the knowledge from the historical test samples and benefit the predictions of subsequent test samples.
- Extensive experiments on various large-scale benchmarks demonstrate the superiority of our proposed DART against the state-of-the-art TTA methods.

Related Work

Pre-trained Vision-Language Models

The recent pre-trained vision-language models including CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021), have presented a surprising ability to learn general visual representations for downstream tasks in a zero-shot manner through proper prompts. Specifically, CLIP aimed to use "a photo of a CLS" as a prompt on the language side for zero-shot image classification. However, its classification performance heavily depends on the elaborately designed text prompts, which requires time-consuming prompt engineering by experts. Therefore, few-shot approaches like CoOp (Zhou et al. 2022b) and CoCoOp (Zhou et al. 2022a) are proposed to regard the text prompt as learnable parameters and employ a small amount of downstream task data for prompt training. Additionally, MaPLe (Khattak et al. 2023) proposes to train cross-modal prompts for the adaptation of CLIP. However, all the above methods have to rely on the collection of training data from downstream tasks. Due to the variations between training and test data, they could not generalize well on unseen test data from shifted data distributions which severely limits their practical effectiveness.

Test-Time Adaptation

In realistic scenarios, the test data always undergo natural variations or corruptions, resulting in data distribution shifting between the training and test phases (Hendrycks and Dietterich 2019; Koh et al. 2021). Therefore, even the large-scale pre-trained models are also difficult to generalize well on test data when domain shifting occurs (Recht et al. 2018). Recently, various Test-Time Adaptation (TTA) approaches are proposed to adaptively adjust the pre-trained models in the test phase to fit the distribution of test data (Schneider et al. 2020; Sun et al. 2020; Wang et al. 2021).

In terms of model parameter optimization, several methods (Schneider et al. 2020; Wang et al. 2021; Shu et al. 2022; Niu et al. 2023) propose to capture domain variations in test data by optimizing the batch normalization layers but are severely limited by the model architecture. Besides, various approaches (Yuan, Xie, and Li 2023; Döbler, Marsden, and Yang 2023) utilize consistency regularization to ensure stable model predictions when the data are perturbed slightly, and the model is optimized by using unlabeled test samples. Typically, these methods employ a teacher-student network architecture, where different augmented samples are fed into the two models whose outputs are constrained to be as close as possible. However, these methods usually update the entire model and can not retain the knowledge of training data well to assist in the prediction of the current data. Based on this, recent works (Wang et al. 2022b; Niu et al. 2022; Shu et al. 2022) propose using anti-forgetting techniques to preserve the knowledge of the training data during TTA, aiding the predictions of test samples. However, they do not effectively explore the historical knowledge from the seen test data, still resulting in limited performance on unseen test samples.

Prompt Learning

Prompts are initially applied in the field of natural language processing (NLP) (Ponti et al. 2020; Brown et al. 2020) by manually designing text prompts to make adaptive adjustments to downstream tasks. Although manually designed prompt templates are intuitive and promising, they require tremendous human effort and specific expertise which are

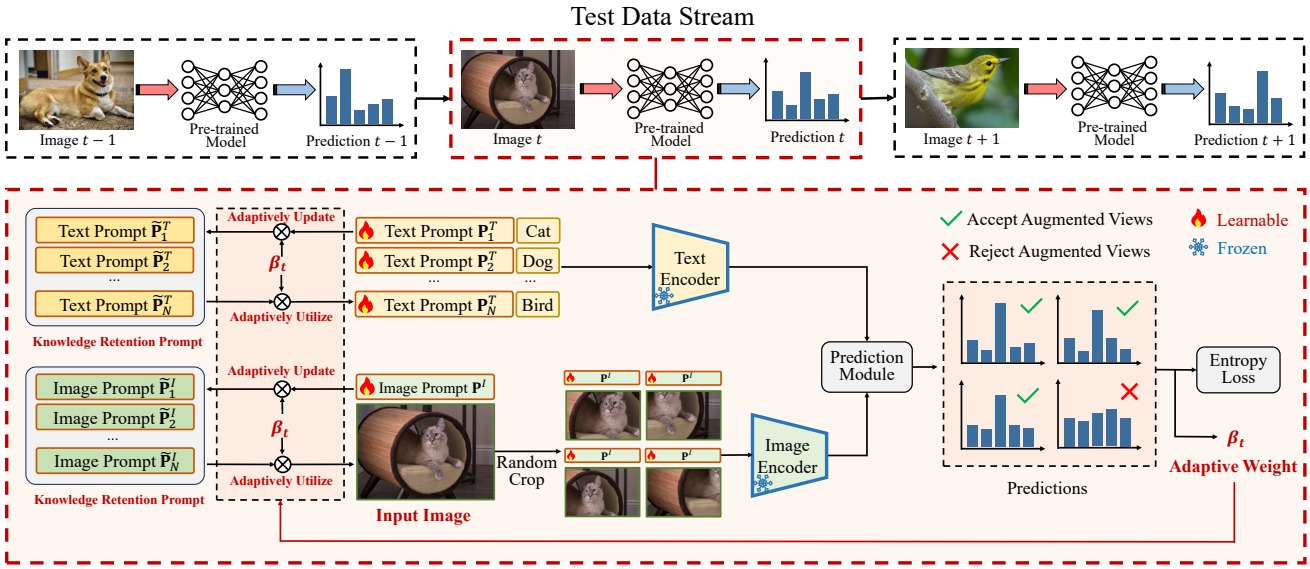

Figure 2: The overall pipeline of our proposed Dual-modal Adaptive online prompting and knowledge ReTention (DART) method. For each test sample, DART utilizes dual-modal instance-level prompts to capture its specific information. Additionally, dual-modal knowledge retention prompts are designed to adaptively retain informative knowledge of seen test samples to benefit the prediction of subsequent test instances.

costly. To address this issue, numerous NLP methods (Li and Liang 2021; Tsimpoukelli et al. 2021) no longer focus on designing human-interpretable natural language prompt templates but treat prompts as learnable parameters, which greatly increase the flexibility and diversity of prompts.

Recently, there are various methods (Jia et al. 2022; Gao et al. 2022; Chen et al. 2022) migrating prompt learning to the field of computer vision. The main idea is concentrated on adopting the pre-trained vision transformer (Dosovitskiy et al. 2020) to downstream tasks by training a small number of prompt parameters. After the emergence of VPT (Jia et al. 2022) which initially migrates prompt learning to the field of image recognition, prompting in vision has been rapidly spread to various tasks such as image recognition (Chen et al. 2022), incremental learning (Wang et al. 2022e,d; Wang, Huang, and Hong 2022). However, the aforementioned methods need to utilize sufficient training data to train prompts before they can be applied to downstream tasks, thus they can hardly adjust prompts for online test data during test-time inference.

Therefore, few works have been proposed recently to focus on the important and promising direction of test-time online prompting. DePT (Gao et al. 2022) proposes a hierarchical self-training model to dynamically train the learnable prompts and classifier of the visual model at test time to cope with variations of test data. Moreover, TPT (Shu et al. 2022) is designed as a TTA method for the pre-trained CLIP model, which tunes the instance-level text prompt by minimizing the entropy loss. However, for a vision-language model like CLIP, TPT only concentrates on the prompt on the text side but lacks the utilization of the important image side, which greatly limits the multi-modality ability of the CLIP model to tackle the test data.

The Proposed Method

Problem Setting and Notations

In this work, we focus on the test-time adaptation (TTA) scenario, where the distribution of data in the test phase differs from the training phase. The training data are denoted as $\mathcal{X}_s$, and the test data during the online test-time are donated as $\mathcal{X} = \{x_i\}_{i=1}^n$ which exhibits a different distribution against $\mathcal{X}_s$. For each x_i, we first learn dual-modal instance-level prompts $\boldsymbol{\mathcal{P}} = \left\{\mathbf{P}^T, \mathbf{P}^I\right\}$ for it, then apply $\boldsymbol{\mathcal{P}}$ into the pre-trained vision-language model θ to obtain the predicted category $y_i = \theta(x_i, \boldsymbol{\mathcal{P}}), y_i \in \mathcal{Y}$ of input sample x_i.

Test-Time Adaptation for CLIP

A pre-trained CLIP model $\theta = \{\mathbf{E}_\mathrm{T}, \mathbf{E}_\mathrm{I}\}$ consists of two encoders, one for the text modality $\mathbf{E}_\mathrm{T}$ and the other one for the image modality $\mathbf{E}_\mathrm{I}$. These two encoders separately encode text and image inputs into the text-image cross-modality representation space. Generally, the architecture of the text encoder $\mathbf{E}_\mathrm{T}$ is a Transformer model (Vaswani et al. 2017), as well as the image encoder is a CNN (He et al. 2016) or a ViT (Dosovitskiy et al. 2020). The CLIP model is trained by a contrastive loss with the goal of maximizing the cosine similarity of the matched text-image pairs and minimizing the cosine similarity of the unmatched pairs.

For a test image $x \in \mathbb{R}^{C\times H\times W}$, the representation of image x can be obtained as $\boldsymbol{r}^I = \mathbf{E}_\mathrm{I}(x)$. For the labels, a hand-crafted prompt $\mathbf{P}$ = "a photo of a" combined with all

class names in $\mathcal{Y}$, i.e. $\{[\mathbf{P};\mathrm{c}_1],[\mathbf{P};\mathrm{c}_2],\cdots,[\mathbf{P};\mathrm{c}_N]\}$ are fed into the text encoder $\mathbf{E}_\mathrm{T}$ to obtain the label representations $\left\{\boldsymbol{r}_1^T,\boldsymbol{r}_2^T,\cdots,\boldsymbol{r}_N^T\right\}$, where $\boldsymbol{r}_i^T=\mathbf{E}_\mathrm{T}(\mathbf{P},\mathrm{c}_i)$. Then the prediction probability $P(y=j|x)$ can be obtained as:

$$P(y=j|x)=\frac{\exp(\boldsymbol{s}_j/\tau)}{\sum_{i=1}^{N}\exp(\boldsymbol{s}_i/\tau)}, \tag{1}$$

where $\boldsymbol{s}_i:=\cos(\boldsymbol{r}^I,\boldsymbol{r}_i^T)$ and τ is the temperature of the softmax function. Although classification prediction can be obtained via Eq. 1, its performance can be severely limited when test data suffers from significant data distribution shifting against training samples. Therefore, we propose a dual-modal adaptive online prompting and knowledge retention (**DART**) method for test-time adaptation to improve the online generalization ability of CLIP.

Dual-Modal Online Prompting in DART

As mentioned above, the text input for the inference of CLIP is $\{[\mathbf{P};\mathrm{c}_1],[\mathbf{P};\mathrm{c}_2],\cdots,[\mathbf{P};\mathrm{c}_N]\}$, which means all classes share a same prompt. The latest work e.g. TPT (Shu et al. 2022) followed this setting and their performance is limited by using a single prompt for different classes.

In order to mitigate the discrepancy between different classes during online test time, we propose to adopt a class-specific text prompt $\mathbf{p}_i^T$ for each class c_i. Then we initialize a group of class-specific text prompts as below:

$$\mathbf{P}^T=\left\{\mathbf{P}_1^T,\mathbf{P}_2^T,\cdots,\mathbf{P}_N^T\right\}. \tag{2}$$

With $\mathbf{P}^T$, the text representation of a class c_i can be computed as follows:

$$\boldsymbol{r}_i^T=\mathbf{E}_\mathrm{T}(\mathbf{P}_i^T,\mathrm{c}_i). \tag{3}$$

The instance-level class-specific text prompts $\mathbf{P}^T$ helps the text encoder $\mathbf{E}_\mathrm{T}$ to explore the semantic information in the text side related with the current image x.

Unlike the existing CLIP-based few-shot or test-time prompt learning methods (Zhou et al. 2022b,a; Shu et al. 2022) only adjusting the text prompts, our DART elaborately integrates visual prompts $\mathbf{p}^I$ into the image encoder $\mathbf{E}_\mathrm{I}$ via an instance-level adjustment for different test samples as below:

$$\boldsymbol{r}^I=\mathbf{E}_\mathrm{I}(x,\mathbf{P}^I), \tag{4}$$

The instance-level visual prompt $\mathbf{P}^I$ assists the image encoder $\mathbf{E}_\mathrm{I}$ in utilizing the intrinsic semantic information.

Adaptive Knowledge Retention in DART

Through dual-modal online prompting, DART comprehensively captures knowledge from individual test samples. To retain and harness the knowledge unearthed from individual samples, aiding the test of subsequent samples, we first design dual-modal knowledge retention prompts for each category to retain knowledge as below:

$$\tilde{\mathbf{P}}^T=\left\{\tilde{\mathbf{P}}_1^T,\tilde{\mathbf{P}}_2^T,\cdots,\tilde{\mathbf{P}}_N^T\right\}, \tag{5}$$

$$\tilde{\mathbf{P}}^I=\left\{\tilde{\mathbf{P}}_1^I,\tilde{\mathbf{P}}_2^I,\cdots,\tilde{\mathbf{P}}_N^I\right\}. \tag{6}$$

According to Eq. 1, by using the dual-modal prompts $\mathbf{P}^T$, $\mathbf{P}^I$ of t-th test sample x_t, we can obtain its predicted class $j=\hat{y}_t\in\mathcal{Y}$ and similarity with corresponding text prompt $\boldsymbol{s}_j=\cos(\boldsymbol{r}^I,\boldsymbol{r}_j^T)$. An intuitive idea is that when the similarity $\boldsymbol{s}_j$ and prediction confidence $P(y=j|x_t)$ are higher, the dual-modal prompts $\mathbf{P}^T$, $\mathbf{P}^I$ contains more useful knowledge. So a fusing weight β_t is calculated as bellow:

$$\beta_t=1-e^{-\boldsymbol{s}_j/h}, \tag{7}$$

where h is a temperature hyper-parameter. To adaptively retain knowledge from seen test samples, then we merge $\mathbf{P}^T$, $\mathbf{P}^I$ with the corresponding class-specific knowledge retention prompts $\tilde{\mathbf{P}}_j^T$, $\tilde{\mathbf{P}}_j^I$ using the calculated weight β_t to preserve the learned knowledge:

$$\tilde{\mathbf{P}}_j^T\leftarrow\tilde{\mathbf{P}}_j^T\cdot(1-\beta_t)+\mathbf{P}_j^T\cdot\beta_t, \tag{8}$$

$$\tilde{\mathbf{P}}_j^I\leftarrow\tilde{\mathbf{P}}_j^I\cdot(1-\beta_t)+\mathbf{P}^I\cdot\beta_t, \tag{9}$$

where $\leftarrow$ denotes updating the value of a variable.

When the next test sample x_{t+1} is coming, for the text modality, $\tilde{\mathbf{P}}^T$ is used for initialization to leverage the knowledge from past test samples:

$$\mathbf{P}^T\leftarrow\mathbf{P}^T\cdot(1-w_T)+\tilde{\mathbf{P}}^T\cdot w_T, \tag{10}$$

where w_T is a hyper-parameter. For the imaging modality, since only one prompt is used and the category of the image cannot be known in advance, the information from previous samples can only be utilized in the test-time inference phase. Assuming that in the test-time training phase, the class j will get the highest confidence, then the past sample's knowledge is utilized as follows:

$$\mathbf{P}^I\leftarrow\mathbf{P}^I\cdot(1-w_I)+\tilde{\mathbf{P}}_j^I\cdot w_I, \tag{11}$$

where w_I is a hyper-parameter. In summary, our DART can adaptively retain the knowledge learned from high-confidence samples through dual-modal knowledge prompts and utilize them to assist in predicting subsequent unseen samples.

The Optimization of DART

As shown in Figure 2, the introduced prompts in DART can be readily optimized in an online learning manner where the pre-trained CLIP is frozen. Follow the protocol of TPT (Shu et al. 2022), for each test image x, we first augment it to $\{x_1^a,x_2^a,\cdots,x_B^a\}$ where B is the batch size in training. Then, all the B augmented images are fed into the CLIP model to get the prediction probability distribution of x_i^a:

$$\left\{P_{\boldsymbol{\mathcal{P}}}(y=j|x_1^a),P_{\boldsymbol{\mathcal{P}}}(y=j|x_2^a),\cdots,P_{\boldsymbol{\mathcal{P}}}(y=j|x_B^a)\right\}, \tag{12}$$

where $\boldsymbol{\mathcal{P}}=\left\{\mathbf{P}^T,\mathbf{P}^I\right\}$ is the designed dual-modal instance-level prompts of image x. To reduce the noise interference caused by some unsuitable augmentations, we eliminate the predictions with low self-confidence. The self-confidence of one augmented view x_i^a is computed as below:

$$\boldsymbol{H}(x_i^a)=\sum_{j=1}^{N}P_{\boldsymbol{\mathcal{P}}}(y=j|x_i^a)\log P_{\boldsymbol{\mathcal{P}}}(y=j|x_i^a). \tag{13}$$

Then we select the top ρ ratio samples of high self-confidence, where ρ is a pre-defined hyper-parameter. Finally, we optimize the prompts $\mathcal{P}$ by minimizing the entropy of average prediction distribution over the selected confident samples, i.e.

$$\arg\min_{\mathcal{P}} -\sum_{j=1}^{N} \overline{P}_{\mathcal{P}}(y=j|x) \log \overline{P}_{\mathcal{P}}(y=j|x), \quad (14)$$

where

$$\overline{P}_{\mathcal{P}}(y=j|x) = \frac{1}{\rho B} \sum_{i=1}^{\rho B} P_{\mathcal{P}}(y=j|x_i^a). \quad (15)$$

Experiments

We first introduce the benchmarks used for evaluating our DART and the compared state-of-the-art methods, then the implementation details are demonstrated accordingly. Finally, extensive experiment results and analyses are further presented along with discussions about the ablation study of our proposed method. Moreover, more experimental results and analyses are included in our Supplementary.

Methods	Publication	I-A	I-R	I-S	Average
CLIP	ICML 2021	47.87	73.98	46.09	55.98
Ensemble	ICML 2021	49.89	77.65	48.24	58.59
CoOp	IJCV 2022	49.71	75.21	47.99	57.64
CoCoOp	CVPR 2022	50.63	76.18	48.75	58.52
MaPLe	CVPR 2023	50.90	76.98	49.15	59.01
DART	**This Paper**	**60.56**	**79.56**	**49.76**	**63.29**

Table 1: The Acc@1 comparison results against CLIP and the latest few-shot fine-tuning methods on three benchmark datasets. The I-A, I-R, and I-S represent ImageNet-A, ImageNet-R, and ImageNet-Sketch respectively.

Datasets

Since the data distribution shifting will inevitably occur in real-world scenarios, the experiments are conducted on three large-scale benchmarks, ImageNet-A (Hendrycks et al. 2021b), ImageNet-R (Hendrycks et al. 2021a), and ImageNet-Sketch (Wang et al. 2019) which are variants of the ImageNet (Deng et al. 2009) dataset to evaluate the performance of different methods for improving the test-time generalization ability of CLIP. These benchmarks have been considered as out-of-distribution data for ImageNet previously (Radford et al. 2021), and we follow the same setting in our experiments.

- **ImageNet-A** (Hendrycks et al. 2021b) is a natural adversarial image dataset that contains natural images misclassified by ResNet-50 (He et al. 2016) in ImageNet (Deng et al. 2009). In total, it contains 7,500 images of 200 categories. These misclassified images in ImageNet-A usually suffer from various distribution shifting which poses critical challenges to the test-time generalization ability of models.
- **ImageNet-R** (Hendrycks et al. 2021a) is a multi-domain (e.g. art, cartoon, painting) image dataset consisting of 30,000 images and 200 categories of ImageNet. All the images in ImageNet-R are collected from 15 different style domains, thus there exist drastic domain gaps between different images.
- **ImageNet-Sketch** (Wang et al. 2019) is a sketch image dataset consisting of 50000 images, 50 images for each of the 1000 ImageNet classes. The images in ImageNet-Sketch are all black and white, and their distribution differs significantly from the training data of CLIP.

Comparison Methods

In the experiments, we compared our proposed DART with state-of-the-art TTA and few-shot fine-tuning methods designed for CLIP. In detail, TPT (Shu et al. 2022) is a test-time prompt-tuning approach focusing on fine-tuning a learnable text prompt for CLIP. Tent (Wang et al. 2021), EATA (Niu et al. 2022), SAR (Niu et al. 2023) and RMT (Döbler, Marsden, and Yang 2023) are general TTA methods. CoOp (Zhou et al. 2022b) and CoCoOp (Zhou et al. 2022a) are few-shot prompt-tuning methods aiming at fine-tuning the text prompts. MaPLe (Khattak et al. 2023) is a few-shot method training cross-modal prompts on each dataset. Following the same protocol in (Zhou et al. 2022b,a), 16-shot extra training images of each category are provided for fine-tuning. For test-time inference, once the text prompts are learned, the aforementioned methods are directly used in the same way as CLIP does. In addition, since the pre-trained CLIP can be directly applied to downstream classification tasks in a zero-shot manner, we also regard it as a baseline method. Two different text prompt settings of CLIP are evaluated, one is the default "a photo of a" and the other one is the ensemble of 80-hand-crafted prompts from (Radford et al. 2021).

Implementation Details

The pre-trained CLIP model with ViT-B/16 is used as our backbone (Radford et al. 2021). For each test image, we initialize all the text prompts in our DART as "a photo of a". The image prompts are initialized with a uniform distribution of $(-1, 1)$ following the previous visual prompting methods (Wang et al. 2022e,d). The length of image prompts is set to 2, and they are added to the second layer of the CLIP image encoder. The hyper-parameters h, w_T, and w_I of dual-modal knowledge retention prompts are set to 5000, 0.1, and 0.1 respectively. For the learning of DART, we use randomly resized crops to augment the single test sample to obtain a batch of $B = 64$ images, and the confidence threshold ρ follows the same setting in (Shu et al. 2022). An Adam optimizer with a learning rate of 0.003 is used to optimize the prompts $\mathcal{P}$. All experiments are implemented on a single NVIDIA 4090 GPU.

Comparison with State-of-the-arts

The overall comparison results against the state-of-the-art TTA and few-shot fine-tuning methods on ImageNet-A, ImageNet-R, and ImageNet-Sketch are reported in Table 1

	Methods	Publication	ImageNet-A		ImageNet-R		ImageNet-Sketch		Average	
			Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
ViT	Tent	ICLR 2021	38.91	69.40	61.95	76.01	48.26	72.17	49.71	72.53
	EATA	ICML 2022	38.05	67.48	59.08	73.35	44.18	67.62	47.10	69.48
	SAR	ICLR 2023	37.71	69.72	61.13	75.83	**49.82**	73.71	49.55	73.09
	RMT	CVPR 2023	30.71	65.83	59.91	74.02	46.41	70.75	45.68	70.20
CLIP	Tent	ICLR 2021	48.44	78.91	74.61	91.87	46.82	74.13	56.62	81.64
	TPT	NIPS 2022	54.77	81.52	77.06	92.08	47.94	74.78	59.92	82.79
	EATA	ICML 2022	49.91	79.23	74.54	91.48	46.93	73.89	57.13	81.53
	SAR	ICLR 2023	48.89	78.92	75.81	91.85	47.59	74.14	57.43	81.64
	RMT	CVPR 2023	48.28	78.61	74.47	91.30	47.34	74.20	56.70	81.37
	DART	**This Paper**	**60.56**	**82.59**	**79.56**	**93.27**	49.76	**75.73**	**63.29**	**83.86**

Table 2: The comparison results against state-of-the-art TTA methods on three benchmark datasets. ViT represents the ViT-B/16 model pre-trained on ImageNet, and CLIP represents the pre-trained CLIP model with ViT-B/16 architecture.

Components in DART				ImageNet-A	
$\mathbf{P}^T$	$\tilde{\mathbf{P}}^T$	$\mathbf{P}^I$	$\tilde{\mathbf{P}}^I$	Acc@1	Acc@5
✗	✗	✗	✗	47.87	79.09
✓	✗	✗	✗	57.12	81.59
✗	✗	✓	✗	54.24	80.63
✓	✓	✗	✗	58.04	81.45
✗	✗	✓	✓	55.26	80.51
✓	✓	✓	✓	**60.56**	**82.59**

Table 3: Ablation study about the different components of DART. $\mathbf{P}^T$ and $\mathbf{P}^I$ represent the class-specific text prompts and image prompts respectively. $\tilde{\mathbf{P}}^T$ and $\tilde{\mathbf{P}}^I$ represent knowledge retention text prompts and knowledge retention image prompts respectively. ✓ and ✗ represent without or with this component. When none of the components is used, the model degenerates to the baseline CLIP.

Learnable Text Prompt	Acc@1
Unified Text Prompt in TPT (Shu et al. 2022)	57.27
[“a photo of a” + CLS]	59.53
[“a photo of a”] in DART	**60.56**

Table 4: Ablation study about the influence of different learnable text prompts.

and Table 2 respectively. Compared to few-shot fine-tuning methods, DART outperforms the second-best player MaPLe by 4.28% at average Acc@1 over all three datasets. Even though these few-shot methods utilize extra labeled data from ImageNet for fine-tuning, they still struggle to address the issue of test data distribution shifting. In contrast, our DART employs dual-modal online prompting to dynamically adapt the pre-trained CLIP model to handle various test data from different distributions.

As for the TTA methods including Tent, EATA, SAR, and RMT, they originally employed the pre-trained ViT on ImageNet as the backbone. Considering the distinct training data distributions and generalization capability between the pre-trained ViT and CLIP, to ensure a fair and equitable comparison, we conduct additional experiments with all comparison methods and DART, utilizing the pre-trained CLIP model as the same backbone. As demonstrated in Table 2, DART outperforms the second-best player TPT by 3.37% at average Acc@1 and 1.07% at average Acc@5 on all three datasets. Specifically, our proposed DART significantly outperforms TPT by 5.79% at Acc@1 on ImageNet-A. Since ImageNet-A consists of natural images misclassified by ResNet-50, this result verifies that our DART can well handle natural distribution shifting in the test phase. The same conclusion can also be confirmed by the experimental results on ImageNet-R and ImageNet-Sketch. Notably, when SAR employs ViT as the backbone, it achieves comparable performance as our DART at Acc@1 on ImageNet-Sketch. However, when using the same CLIP backbone, DART significantly outperforms SAR by 2.17%. This is credited to DART's dual-modal online prompting and knowledge retention prompts, which effectively tap into and utilize information from samples during test time, even in cases of significant style variations in the samples (e.g. cartoon).

Ablation Studies and Analyses

The Influence of Different Components in DART. To verify the effectiveness of the proposed dual-modal adaptive online prompting and knowledge retention components in our DART, an ablation experiment is conducted on ImageNet-A. As demonstrated in Table 3, utilizing either the instance-level text prompts $\mathbf{P}^T$ or the image prompts $\mathbf{P}^I$ consistently enhance the robustness against the distribution shifting compared to the naive zero-shot CLIP. Moreover, with the integration of the text knowledge retention prompts $\tilde{\mathbf{P}}^T$ or image knowledge retention prompts $\tilde{\mathbf{P}}^I$, performance is further elevated beyond the utilization of single-modal prompts $\mathbf{P}^T$, $\mathbf{P}^I$ alone. This improvement can be attributed to the inherent ability of knowledge retention prompts to adaptively retain knowledge. Notably, employing the whole dual-modal adaptive online prompting and knowledge retention components yields the best results, as it efficiently captures information from each individual test sample and retains the knowledge from previously seen test samples, thereby facilitating the performance of the pre-trained CLIP.

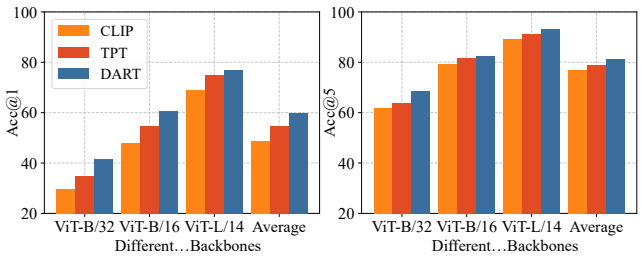

Figure 3: Ablation study about the generalization across different backbones on ImageNet-A.

The Influence of Different Learnable Text Prompts in DART. Different from TPT which learns a unified text prompt for all categories, our proposed DART proposes to utilize class-specific text prompts initialized as "a photo of a" for different categories during the test time. From the results on ImageNet-A in Table 4, our proposed class-specific text prompts outperform the unified text prompt by a margin of 3.29% at Acc@1. The reason is that our class-specific text prompts are more flexible to capture and highlight the class-specific information of the test sample. Moreover, an extra experiment that simultaneously updates our class-specific text prompts and the CLS token at test time reports inferior performance than ours. This is mainly because the important semantic information contained by the CLS token may be hindered during online learning.

The Generalization across Different Backbones. To validate the generalization ability of DART, across different backbones, we conduct experiments using pre-trained CLIP models with various backbones. As shown in Figure 3, when employing the ViT-B/32, DART exhibits significant improvements at Acc@1 compared to CLIP and TPT, with increments of 11.81% and 6.79% respectively. For the higher-parameter ViT-L/14, DART demonstrates enhancements of 8.2% and 2.14% at Acc@1 compared to CLIP and TPT respectively. Furthermore, across different backbones, DART consistently exhibits further advancements at Acc@5 over CLIP and TPT. This can be credited to DART's dual-modal online prompting, which introduces additional learnable parameters as the backbone is frozen. This enables DART to adapt the pre-trained CLIP model across both modalities. Moreover, the dual-modal knowledge retention prompts effectively preserve and leverage knowledge learned from seen test samples, resulting in superior performance.

The Influence of Different Hyper-parameters in DART. There are several hyper-parameters in our DART. We initially conduct experiments to investigate the influence of the hyper-parameter h in Eq. 7, which is responsible for generating the adaptive weight β_t. As illustrated in Figure 4, DART demonstrates insensitivity to variations in h. Notably, DART shows remarkable performance when h falls within the range of 4500 to 6000. This is attributed to a favorable balance achieved between the retention of newly acquired knowledge and historical knowledge. Subsequently, we explore the impact of the fusion coefficients w_T and w_I

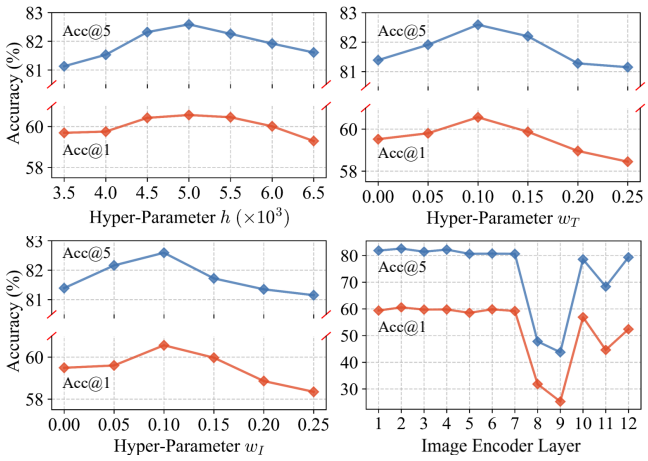

Figure 4: Ablation study about the influence of different hyper-parameters on ImageNet-A.

employed in Eq. 10 and Eq. 11. These coefficients exhibit a similar trend, with their optimal performance observed within the range of 0.05 to 0.15. This outcome demonstrates that while knowledge retained through retention prompts provides auxiliary support, newly captured knowledge remains more crucial and tailored for accurate predictions on the current samples. Then we conduct experiments to investigate which layer of the image encoder the proposed image prompts $\mathbf{p}^I$ should be added. Considering the limited learning condition (only one test sample available) during online test time, we propose to add the prompts to only one layer. As presented in Figure 4, adding the prompts to the second layer of the image encoder performs the best.

Conclusion

In conclusion, we propose a novel dual-modal adaptive online prompting and knowledge retention method, named DART, for test-time adaptation of CLIP-based pre-trained vision-language models. Specifically, our approach involves learning class-specific text prompts and instance-level image prompts for each test sample, effectively capturing the knowledge within an individual test sample to enhance the model's prediction accuracy. Moreover, we design text and image knowledge retention prompts to adaptively retain and utilize the knowledge from previously seen test samples, to facilitate the predictions of subsequent test samples. This enables our DART to adapt to new test instances and improve overall performance during test-time adaptation. Extensive experiments on various large-scale benchmarks demonstrate the effectiveness of DART against state-of-the-art approaches. Our work investigates a promising direction, addressing the challenging problem of training-test data distribution shifting in pre-trained vision-language models using dual-modal prompting and knowledge retention.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62376011, 61925201, 62132001).

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901.

Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; and Luo, P. 2022. Adaptformer: Adapting vision transformers for scalable visual recognition. arXiv:2205.13535.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Döbler, M.; Marsden, R. A.; and Yang, B. 2023. Robust mean teacher for continual and gradual test-time adaptation. In *CVPR*, 7704–7714.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929.

Gal, R.; Patashnik, O.; Maron, H.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4): 1–13.

Gao, Y.; Shi, X.; Zhu, Y.; Wang, H.; Tang, Z.; Zhou, X.; Li, M.; and Metaxas, D. N. 2022. Visual prompt tuning for test-time domain adaptation. arXiv:2210.04831.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; et al. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *CVPR*, 8340–8349.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. arXiv:1903.12261.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021b. Natural adversarial examples. In *CVPR*, 15262–15271.

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

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *ECCV*, 709–727.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *CVPR*, 19113–19122.

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. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, 5637–5664.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. arXiv:2101.00190.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2022. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304.

Niu, S.; Wu, J.; Zhang, Y.; Chen, Y.; Zheng, S.; Zhao, P.; and Tan, M. 2022. Efficient test-time model adaptation without forgetting. In *ICML*, 16888–16905.

Niu, S.; Wu, J.; Zhang, Y.; Wen, Z.; Chen, Y.; Zhao, P.; and Tan, M. 2023. Towards stable test-time adaptation in dynamic wild world. arXiv:2302.12400.

Ponti, E. M.; Glavaš, G.; Majewska, O.; Liu, Q.; Vulić, I.; and Korhonen, A. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. arXiv:2005.00333.

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

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2018. Do cifar-10 classifiers generalize to cifar-10? arXiv:1806.00451.

Schneider, S.; Rusak, E.; Eck, L.; Bringmann, O.; Brendel, W.; and Bethge, M. 2020. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33: 11539–11551.

Shu, M.; Nie, W.; Huang, D.-A.; Yu, Z.; Goldstein, T.; Anandkumar, A.; and Xiao, C. 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. arXiv:2209.07511.

Song, J.; Lee, J.; Kweon, I. S.; and Choi, S. 2023. EcoTTA: Memory-efficient continual test-time adaptation via self-distilled regularization. In *CVPR*, 11920–11929.

Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A.; and Hardt, M. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 9229–9248.

Tsimpoukelli, M.; Menick, J. L.; Cabi, S.; Eslami, S.; Vinyals, O.; and Hill, F. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34: 200–212.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30: 5998–6008.

Wang, C.; Chai, M.; He, M.; Chen, D.; and Liao, J. 2022a. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *CVPR*, 3835–3844.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *ICLR*, 1–15.

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

Wang, Q.; Fink, O.; Van Gool, L.; and Dai, D. 2022b. Continual test-time domain adaptation. In *CVPR*, 7201–7211.

Wang, Y.; Huang, Z.; and Hong, X. 2022. S-Prompts learning with pre-trained transformers: An Occam's razor for domain incremental learning. arXiv:2207.12819.

Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; and Liu, T. 2022c. Cris: Clip-driven referring image segmentation. In *CVPR*, 11686–11695.

Wang, Z.; Zhang, Z.; Ebrahimi, S.; Sun, R.; Zhang, H.; Lee, C.-Y.; Ren, X.; Su, G.; Perot, V.; Dy, J.; et al. 2022d. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *ECCV*, 631–648.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022e. Learning to prompt for continual learning. In *CVPR*, 139–149.

Yuan, L.; Xie, B.; and Li, S. 2023. Robust test-time adaptation in dynamic scenarios. In *CVPR*, 15922–15932.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *CVPR*, 16816–16825.

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