This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Free on the Fly: Enhancing Flexibility in Test-Time Adaptation with Online EM

Qiyuan Dai[1] Sibei Yang[2†]
[1]School of Information Science and Technology, ShanghaiTech University
[2]Sun Yat-sen University

Abstract

Vision-Language Models (VLMs) have become prominent in open-world image recognition for their strong generalization abilities. Yet, their effectiveness in practical applications is compromised by domain shifts and distributional changes, especially when test data distributions diverge from training data. Therefore, the paradigm of test-time adaptation (TTA) has emerged, enabling the use of online off-the-shelf data at test time, supporting independent sample predictions, and eliminating reliance on test annotations. Traditional TTA methods, however, often rely on costly training or optimization processes, or make unrealistic assumptions about accessing or storing historical training and test data. Instead, this study proposes FreeTTA, a training-free and universally available method that makes no assumptions, to enhance the flexibility of TTA. More importantly, FreeTTA is the first to explicitly model the test data distribution, enabling the use of intrinsic relationships among test samples to enhance predictions of individual samples without simultaneous access—a direction not previously explored. FreeTTA achieves these advantages by introducing an online EM algorithm that utilizes zero-shot predictions from VLMs as priors to iteratively compute the posterior probabilities of each online test sample and update parameters. Experiments demonstrate that FreeTTA achieves stable and significant improvements compared to state-of-the-art methods across 15 datasets in both cross-domain and out-of-distribution settings.

1. Introduction

Vision-Language models (VLMs) [2, 20, 27, 35], such as CLIP [35], pretrained on web-scale image-text pairs, encode a diverse array of visual concepts in a shared embedding space for both image and text. This enables their zero-shot generalization across various downstream tasks, including image recognition [23, 55, 58, 60, 61]. Images can be directly identified without task-specific data by aligning their features with the text embeddings of hand-crafted class prompts. However, despite its strengths, VLMs face notable challenges when downstream images demonstrate distinct domain and distribution shifts relative to their training data in the source domain. To address this, some research focuses on adapting VLMs to specific target domains and distributions using parameter-efficient fine-tuning techniques, such as adapter tuning [55] and prompt learning [23, 58, 60, 61]. Nonetheless, a key limitation of these works is their assumption of access to sufficient annotated downstream data for fine-tuning, which restricts VLMs' applicability in real-world scenarios where labeled data is unavailable, particularly in dynamic and diverse conditions.

Given this, the paradigm of utilizing off-the-shelf unlabeled data at test time, known as test-time adaption [1, 13, 21, 39, 40, 53], has emerged to improve VLMs' generalization to the target domain. Building on prompt learning in VLMs, test-time prompting [39] tuning optimizes the

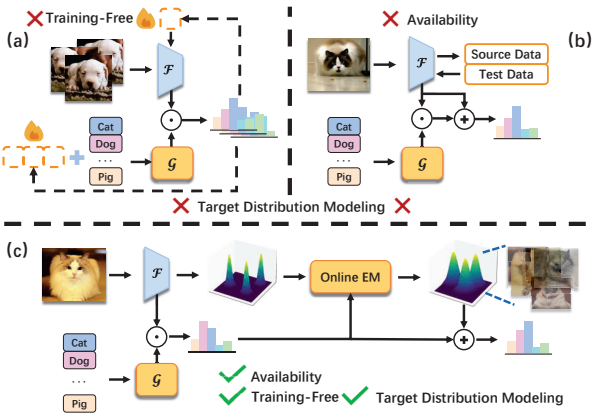


Figure 1. **Comparison of key properties**: target distribution modeling, availability, and training-free efficiency. (a) Prompt-based methods fail the training-free criterion due to lengthy backpropagation. (b) Other methods require access to source or additional test data, affecting availability. In contrast, (c) our FreeTTA satisfies all criteria, offering a universally available, training-free solution that efficiently models the target distribution without requiring additional assumptions.

[†]Corresponding author

prompt by minimizing prediction entropy across test sample augmentations. Follow-up works [1, 13, 21, 40, 52, 53] introduce new optimization objectives, such as calibration error [52], pseudo-label probability difference [26], and self-supervised learning metrics [30], or enhance sample augmentation [13]. More recently, some approaches replace learnable prompts with direct feature representation improvements via meanshift [53] or dynamic adapters [56], or with calibration of the VLMs' temperature factor [11].

However, we believe the three key intrinsic characteristics listed below, crucial for fully realizing the core advantages and applications of test-time adaptation in practical settings, are not sufficiently addressed by existing methods.

- **Target Distribution Modeling.** Explicitly estimating the target distribution can leverage intrinsic relationships among test samples to enhance individual predictions and adapt the model to the overall distribution. Current methods either treat each test sample independently [13, 39, 40] (Figure 1a) or consider only a very small set of sample correlations [21] (Figure 1b) limiting test-time adaptation's ability to fully utilize continuously incoming test samples as historical information.
- **Availability.** Assumptions regarding access to or modification of model parameters, training data, or simultaneous access to or storage of multiple test datasets should be avoided, especially in light of current API access for foundational models and associated privacy concerns. However, prompt tuning methods modify model parameters, while training-free dynamic adapters require storing a subset of test samples [56] (Figure 1b).
- **Training-Free for Efficiency and Stability.** There should be no significant increase in inference costs or prediction instability arising from the randomness of optimization and inherent biases in the optimization objective. Unfortunately, most methods significantly increase time complexity [1, 13, 39] (Figure 1a), while a reduction in entropy is demonstrated to heighten the risk of overfitting and overconfidence.

Therefore, we aim to be the first to enhance test-time adaptation to simultaneously satisfy all three aforementioned characteristics. As a starting point, we assume that the test samples for each class follow an independent Gaussian distribution, denoted as $p(x|y=i) \sim N(\mu_i, \Sigma_i)$. This assumption allows us to apply maximum likelihood estimation [14] to determine the distribution parameters (mean μ_i and variance Σ_i) for each class, and subsequently classify new samples using Bayes' theorem [3], expressed as: $\hat{y} = \arg\max_i p(y=i|x) \propto p(x|y=i)p(y=i)$, where $p(y=i)$ represents the prior probability of class i. However, applying Gaussian discriminant analysis [4] to classify test samples at test time poses new challenges: (1) The class labels of the test data are unknown, preventing direct estimation of the distribution parameters; and (2) test samples cannot be observed simultaneously, as each is predicted online and sequentially.

To address challenge (1), we adopt the Gaussian assumption in Gaussian discriminant analysis [4] and conclude that the unsupervised test samples conform to a Gaussian mixture distribution, expressed as: $p(x) = \sum_{i=1}^{I} \pi_i \mathcal{N}(x|\mu_i, \Sigma_i)$, where π_i represents the mixture weight of class i. Furthermore, we directly apply the EM algorithm [32] to iteratively predict the posterior probability γ_{ki} of each sample x_k belonging to each class i during the E-step, and update the model parameters μ_i, Σ_i, and π_i based on γ_{ki} and test samples in the M-step. To tackle challenge (2), a naive approach involves storing historical test samples during the online prediction process and using them to estimate parameters; however, this contradicts the second characteristic, *i.e.*, availability. Instead, we leverage the alignment of visual and textual features in the shared embedding space of the VLM model, using the text embeddings of each class i as the initial mean μ_i. We then extend the EM algorithm to an online version, where it sequentially estimates the posterior probability γ_{ki} of the currently arriving test sample and updates the Gaussian mixture parameters in an online manner. Additionally, leveraging the generalization capabilities of the VLM's original zero-shot branch, we use its prediction as the confidence score to weigh the influence of each incoming test sample during the online EM update process.

To evaluate the effectiveness of our FreeTTA and learning strategy, we conduct experiments on cross-domain benchmarks and out-of-distribution benchmarks. In summary, our main contributions are:

- To the best of our knowledge, we are the first to simultaneously satisfy the three intrinsic characteristics—target distribution modeling, availability, and being training-free—which enable effective and efficient test-time adaptation for VLMs.
- We introduce FreeTTA, an online EM method that iteratively predicts posterior probabilities for incoming samples across classes and updates parameters by leveraging the prior knowledge of VLMs. This approach enhances stability and incorporates uncertainty, achieving continuous online adaptation without the need to access or store past or additional data.
- Experimental results show that our FreeTTA achieves stable and significant improvements compared to state-of-the-art methods across 15 datasets in cross-domain and out-of-distribution settings, highlighting its robustness and effectiveness.

2. Related Work

Vision-Language Model. In the development of Vision-Language Models (VLMs), several landmark models [2, 20, 27, 35] continuously advance the boundaries of cross-

modal understanding between images and text. Among them, CLIP [35] leverages contrastive learning on large-scale image-text paired data to obtain cross-modal feature representations for both vision and language. In downstream tasks, these pre-trained VLMs exhibit zero-shot and few-shot learning capabilities, along with advanced semantic understanding, facilitating their broad application across diverse areas. For example, in open-world object detection [16, 25, 47, 59], zero-shot capabilities are extended to detection tasks primarily through knowledge distillation techniques; in image classification [60, 61], prompt-based few-shot adaptation is commonly employed to address domain shift; and in referring image segmentation [9, 45, 49, 50], VLMs' advanced semantic alignment facilitate cross-modal alignment at the pixel level, enabling precise segmentation of specified objects within an image. However, using VLMs in downstream tasks often requires labeled training data to bridge domain gaps. Unlike them, our focus is on test-time adaptation, which utilizes off-the-shelf unlabeled data at test time to adapt to the test domain.

Test-Time Adaptation (TTA). TTA aims to automatically adapt to new data domains or distributions during test time, enabling models to adjust to downstream tasks without requiring additional labeled data. This low-cost transfer characteristic holds practical significance and achieves success across various tasks, including image segmentation[6, 44, 51], object detection [41], action recognition [28], and image classification [1, 13, 21, 39, 40, 42, 53, 54]. In image classification, early TTA approaches utilize classifiers trained solely on the image modality, such as TENT [42], which first proposes enhancing model confidence by minimizing prediction entropy, thereby generalizing the model to the downstream domain. With the advancement of VLMs, recent TTA works [1, 13, 21, 39, 40, 53] adopt CLIP's text encoder as a classifier to leverage its generalization capabilities. Some methods integrate entropy minimization with prompt learning techniques commonly used in CLIP-based classification tasks [60, 61]. For instance, TPT [39] combines entropy minimization with prompt learning, where CLIP is fully frozen and optimizes the learnable text prompt for each test sample. Building on this, DiffTPT [13] employs a diffusion model to generate additional augmented samples, enhancing robustness. However, these approaches require gradient backpropagation for each test sample to update learnable prompts, leading to high computational costs. To enhance efficiency, TDA [21] caches historical test data to provide additional pseudo-priors for improving test-time accuracy, circumventing the need for parameter updates. However, these methods generally overlook the potential relationships among test data and often rely on backpropagation or access to additional data, limiting their practicality. In contrast, our approach models the target domain online, featuring both availability and a training-free design.

GMM & EM in Computer Vision. In computer vision, Gaussian Mixture Models (GMM) [37] and the Expectation-Maximization (EM) [**?**] algorithm are widely applied in image classification [38, 46], object detection [7], image segmentation [57], and image generation [29]. For instance, in image classification, [38] proposes a GMM-based approach for modeling facial embeddings, representing them as probability distributions to enhance robustness and accuracy in face recognition. In object detection, [7] integrates Gaussian distributions into the YOLOv3 model, modeling localization uncertainty to improve detection precision and speed. We introduce the concept of modeling the target domain with GMM into the field of TTA, leveraging an online EM approach to achieve a training-free method without the need for access to additional data.

3. Method

The framework of our proposed FreeTTA is shown in Figure 2. First, we review the zero-shot CLIP and previous TTA methods for CLIP, along with the challenges they encounter (see Sec 3.1). Next, we introduce Gaussian discriminant analysis and describe how it models data distributions, followed by a discussion of the challenges in directly applying it to TTA (see Sec 3.2). Finally, we provide a detailed introduction of our proposed FreeTTA, demonstrating how it effectively addresses these challenges (see Sec 3.3).

3.1. Preliminaries

Vision-Language Models (VLMs) [2, 20, 35] like CLIP [35] have recently shown remarkable zero-shot classification capabilities by aligning visual and linguistic modalities. Specifically, CLIP predicts the probability of an image x belonging to class i as follows:

$$P_{\text{CLIP}}(y=i \mid x)=\frac{\exp\left(\cos\left(f(x),\, g(t_i)\right)\right)}{\sum_{k=1}^{K}\exp\left(\cos\left(f(x),\, g(t_k)\right)\right)}, \quad (1)$$

where $f(\cdot)$ and $g(\cdot)$ are the image and text encoders, respectively, and $\cos(\cdot,\cdot)$ denotes the cosine similarity between features. Here, t_i is a class-specific description formulated using template prompts for the i-th class, and K denotes the number of target categories. Despite its strengths, CLIP suffers from performance degradation under domain shifts between the training and test sets. Traditional works [15, 55, 60, 61] to mitigate domain shift require fine-tuning the model with labeled data from the target domain, which incurs additional labeling and training costs, raising the application barrier for pre-trained VLMs in real-world scenarios. To address this, recent works [1, 13, 21, 39, 52] introduce test-time adaptation techniques to facilitate VLM adaptation to the target domain without additional labeled data.

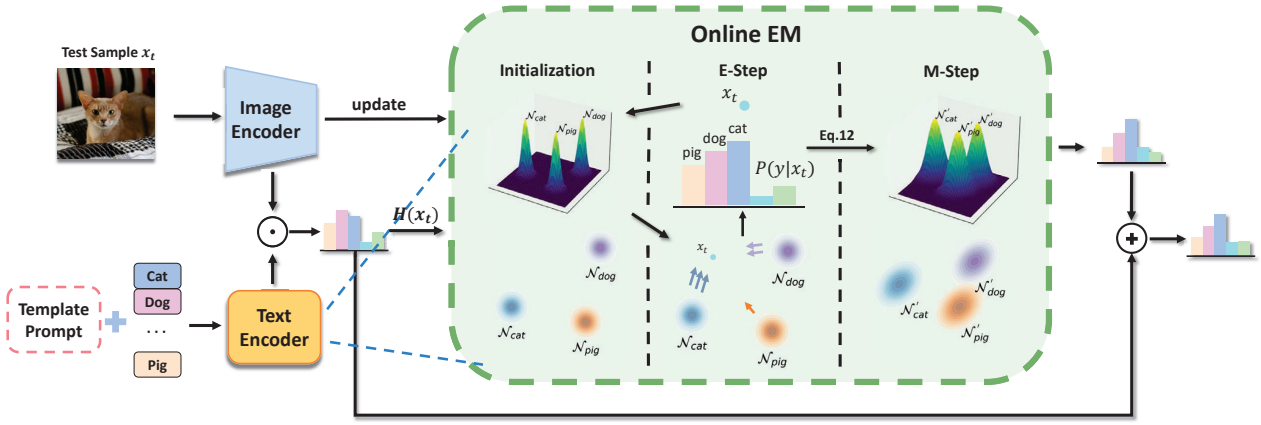


Figure 2. **The overall framework of our FreeTTA**. Given a test sample x_t, we use the frozen CLIP image encoder to extract the image feature, while the text encoder, using prompt templates, generates class feature vectors. The online EM algorithm is initialized with the text features as mean vectors and the identity matrix as the shared covariance matrix. It updates in two steps: the E-step calculates the posterior probability for each class, and the M-step updates the class-specific mean vectors and shared covariance matrix based on current prediction, leveraging CLIP priors to assess the contribution of each sample. Our FreeTTA combines CLIP's zero-shot classification results with GDA predictions to enhance stability and robustness in the target domain, explicitly modeling the target distribution without requiring time-consuming training, while meeting the availability requirement.

Test-Time Adaption (TTA) for VLMs. In the TTA setting, a VLM pre-trained on the source domain is adapted to the target domain using only the unlabeled test set $D_{\text{test}} = \{x_t\}$. This adaptation occurs either through fine-tuning the model in an unsupervised manner or by utilizing a memory of past test samples. Notably, each sample is predicted independently during this process. To enhance CLIP's classification performance on the target domain, typical TTA methods [1, 13, 39] use prompt learning to minimize prediction entropy across multiple augmentations of each test sample. This process can be formulated as follows:

$$P^*(y \mid x_t) = \frac{1}{\rho M} \sum_{m=1}^{M} \mathbb{1}\left[\mathcal{H}\left(P_{\text{CLIP}}\left(\mathcal{A}_m(x_t)\right)\right) \leq \tau\right] P_{\text{CLIP}}\left(\mathcal{A}_m(x_t)\right), \tag{2}$$

where $\mathcal{A}_m(x_t)$ denotes the m-th augmented view of t-th image x_t, M represents the number of augmented views, ρ is the proportion of high-quality augmented views selected, τ is the self-entropy threshold, and $\mathcal{H}(p) = -\sum_{k=1}^{K} P(y = k \mid x_t) \log P(y = k \mid x_t)$ denotes the self-entropy of the predicted probability distribution over K categories. The optimization objective is to minimize $\mathcal{H}(P^*(y \mid x_t))$. Additionally, other methods [21] cache high-confidence test features to provide reference information for subsequent samples, mitigating domain shift.

However, existing TTA works face several challenges: (1) they fail to model the target distribution, ignoring intrinsic relationships among test samples; (2) they require access to data beyond the current sample, conflicting with availability in practical settings; and (3) they involve time-consuming optimization, risking instability. To address these, we propose FreeTTA to overcome these limitations.

3.2. Gaussian Discriminant Analysis for TTA

We assume that test samples for each class follow an independent Gaussian distribution and employ Gaussian Discriminant Analysis (GDA) [4] for classification. Maximum likelihood estimation [14] is used to determine the mean and variance for each class, followed by classifying new samples using Bayes' theorem [3]. In this section, we first introduce GDA (Sec 3.2.1) and then discuss its challenges in applying to TTA (Sec 3.2.2).

3.2.1. Gaussian Discriminant Analysis

Assume that samples of each class y follow a multivariate normal distribution, with distinct means and covariances for each class. To simplify, according to [4], we assume that all classes share the same covariance matrix Σ, while the mean vector μ_y differs for each class. The conditional probability density function for a sample x_t from class y is given by:

$$\begin{aligned} p(x_t \mid y) &= \mathcal{N}(x_t \mid \mu_y, \Sigma) \\ &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{(x_t - \mu_y)^\top \Sigma^{-1}(x_t - \mu_y)}{2}\right), \end{aligned} \tag{3}$$

where d is the dimension of the sample x_t, μ_y is the mean vector for class y, and Σ is the shared covariance matrix. Eq. 3 represents the Mahalanobis distance between the sample x_t and the class mean μ_y, scaled by the covariance matrix Σ. This scaling accounts for feature correlations and

variance differences across dimensions, yielding a more accurate distribution estimate. Based on this assumption, the posterior probability of y can be expressed using Bayes' theorem [3] as follows:

$$P(y \mid x_t) = \frac{P(y)p(x_t \mid y)}{\sum_{y'} P(y')p(x_t \mid y')}, \tag{4}$$

where $P(y)$ represents the prior probability of class y. Substituting Eq. 3 into the expression yields:

$$P(y \mid x_t) = \frac{\exp\left(-\frac{1}{2}(x_t - \mu_y)^\top \Sigma^{-1}(x_t - \mu_y)\right)}{\sum_{y'} \exp\left(-\frac{1}{2}(x_t - \mu_{y'})^\top \Sigma^{-1}(x_t - \mu_{y'})\right)}. \tag{5}$$

Therefore, for the t-th test sample x_t, the class with the highest posterior probability is selected as the prediction:

$$y^* = \arg\max_y \left(\log P(y) - \frac{1}{2}(x_t - \mu_y)^\top \Sigma^{-1}(x_t - \mu_y)\right). \tag{6}$$

3.2.2. The Key Challenges of Applying GDA to TTA

Although GDA models the target distribution, it requires a large set of annotated samples for accurate parameter estimates. In TTA, however, test samples are unlabeled and presented sequentially, making direct application of GDA challenging.

Sequential Online Adaptation. In TTA, the model adapts in real-time based solely on incoming test samples, without access to batch data or full distribution. This "sequential online" nature makes distribution estimation from a single sample unreliable and impossible, especially with substantial shifts. Thus, methods must balance efficiency and accuracy in online updates for effective adaptation.

Unsupervised Adaptation. In TTA, where test sample labels are unavailable, the model must adapt unsupervised. However, GDA relies on labeled samples to estimate parameters. To suit the unsupervised setting, we use a Gaussian mixture model (GMM) [37] for TTA.

Uncertainty Modeling. Fortunately, we can leverage the zero-shot predictions from VLMs as priors for GDA, serving as pseudo-labels. However, these pseudo-labels carry uncertainty, and the confidence in their predictions must be accounted for.

3.3. Online EM Algorithm for TTA

To address the three challenges of applying GDA to TTA, we introduce an online Expectation-Maximization (EM) algorithm that leverages zero-shot predictions from VLMs as priors to iteratively compute the posterior probabilities of each online test sample and update the parameters. Specifically, in the E-step (Sec 3.3.2), we evaluate the posterior probability of the incoming test sample based on the distribution parameters from the previous step. In the M-step (Sec 3.3.3), we update the mean of each class and the shared covariance to estimate the dynamic changes in the distribution by using the test sample.

3.3.1. Parameter Initialization

During initialization, we use CLIP's text encoder $g(\cdot)$ to generate the class features $g(t_y)$ for each class y, which serve as the initial mean vectors μ_y for that class. We assume that the features are independent and identically distributed with unit variance, setting the shared covariance matrix Σ to the identity matrix I for a simple and unbiased starting point. The initialization formulas are:

$$\mu_y = g(t_y), \quad \Sigma = I. \tag{7}$$

3.3.2. E-Step: Computing Posterior Probabilities

In the E-step, for each new test sample x_t at time step t, we evaluate the likelihood of x_t belonging to class y by calculating its posterior probability. Based on the assumption of the normal distribution, the posterior probability $P(z_y = 1 \mid x_t)$ is given by:

$$P_{\text{GAUS}}(z_y = 1 \mid x_t) = \frac{\pi_y \cdot \mathcal{N}(x_t \mid \mu_y, \Sigma)}{\sum_j \pi_j \cdot \mathcal{N}(x_t \mid \mu_j, \Sigma)}, \tag{8}$$

where z_y is the indicator variable representing sample x_t belonging to class y, and π_y denotes the prior probability of class y. We denote the posterior probability as $\gamma_{y,t} = P(z_y = 1 \mid x_t)$.

3.3.3. M-Step: Parameters Update

In the M-step, we utilize the posterior probabilities $\gamma_{y,t}$ calculated in the E-step to update the parameters for each class, including the prior probability π_y, the mean vector μ_y, and the shared covariance matrix Σ. For each class y, the prior probability is updated as:

$$\pi'_y = \frac{N_y + \gamma_{y,t}}{n_t}, \tag{9}$$

where n_t is the total number of samples up to the t-th step, and N_y is the number of samples in class y, which is initialized to $1/\text{number of classes}$. Meanwhile, the updates to the mean vectors and the shared covariance matrix incorporate the contribution of the new sample to each class, and are formulated as follows:

$$\mu'_y = \frac{N_y \cdot \mu_y + \gamma_{y,t} \cdot x_t}{N_y + \gamma_{y,t}},$$
$$\Sigma' = \frac{(n_t - 1)\Sigma + \sum_y \gamma_{y,t}(x_t - \mu'_y)(x_t - \mu'_y)^\top}{n_t - 1}. \tag{10}$$

Notably, in statistical estimation, the unbiased estimation of the covariance matrix requires dividing by $n_t - 1$, accounting for the degrees of freedom lost due to estimating the mean. By dynamically adjusting the shared covariance matrix, the model can estimate the distribution

of each class, thereby enhancing classification ability on the target domain. After incorporating the new sample x_t, the sample count for class y is correspondingly updated as $N'_y = N_y + \gamma_{y,t}$.

Our online EM algorithm effectively addresses the distribution shift challenges in TTA for VLMs. By dynamically adjusting class-specific mean vectors, the shared covariance matrix, and prior probabilities, the model adapts to the target domain in an unsupervised, sequential online manner. Moreover, this approach eliminates the need for gradient-based optimization on individual samples, reducing computational overhead and enhancing model robustness.

3.4. Incorporating VLM Priors

During the initial stage of TTA, the model's instability leads to erratic predictions and updates, causing drift from the original semantic information. To enhance stability, we propose an adaptive update strategy that incorporates VLM priors, including zero-shot classification predictions and confidence levels. Specifically, we use the entropy of CLIP's predictions to assess each sample's confidence level and dynamically adjust its influence on parameter updates, while leveraging the classification logits to adjust the predicted probabilities.

For the t-th test sample x_t, we compute the self-entropy based on the CLIP's zero-shot predicted probabilities, denoted as $\{P_{\text{CLIP}}(z_y = 1 \mid x_t) \mid y = 1, \ldots, K\}$, then the self-entropy is defined as:

$$H(x_t) = -\sum_{y=1}^{K} P_{\text{CLIP}}(z_y = 1 \mid x_t) \log P_{\text{CLIP}}(z_y = 1 \mid x_t). \tag{11}$$

By incorporating $H(x_t)$ into the previous online EM update steps, we enable the model to adjust the influence of the single sample based on its confidence level. Specifically, we introduce the weighting function $w(h) = e^{-\beta h}$ to revise the previous Eq. 9 and Eq. 10 as follows:

$$\begin{aligned}
\pi'_y &= \frac{N_y + w(H(x_t)) \cdot \gamma_{y,t}}{t'_t + w(H(x_t))}, \\
\mu'_y &= \frac{N_y \cdot \mu_y + w(H(x_t)) \cdot \gamma_{y,t} \cdot x_t}{N_y + w(H(x_t)) \cdot \gamma_{y,t}}, \\
\Sigma' &= \frac{(n'_t - 1)\Sigma + w(H(x_t)) \sum_y \gamma_{y,t} (x_t - \mu'_y)(x_t - \mu'_y)^\top}{n'_t - 1}.
\end{aligned} \tag{12}$$

The sample count for class y is correspondingly updated as $N'_y = N_y + w(H(x_t)) \cdot \gamma_{y,t}$, and the total number of samples considering uncertainty is updated as ${n_t}' = {n_{t-1}}' + w(H(x_t))$. This integration of confidence level into the model can effectively reduce the impact of new samples with high uncertainty on the parameter estimation, thereby mitigating the noise issue. Our FreeTTA modulates the contribution of samples based on the confidence of their predictions and adjusts their influence according to self-entropy. This approach emphasizes high-confidence samples, enhancing their dynamic impact on adaptation.

The final predicted logits are a combination of the CLIP's zero-shot logits and those derived from our probabilistic generative model, expressed as:

$$\text{logits}_y = F T_y^\top + \alpha (w_y^\top F + b_y), \tag{13}$$

where $F = f(x_t)$ is the image feature, $T_y = g(t_y)$ is the text feature for class y, α is a hyper-parameter and w_y and b_y are the weight vector and bias term derived from the probabilistic generative model, with $w_y = \Sigma^{-1}\mu_y$ and $b_y = \log P(y) - \frac{1}{2}\mu_y^\top \Sigma^{-1} \mu_y$, for $y = 1, \ldots, K$.

4. Experiments

4.1. Datasets and Implementation Details

Datasets. To evaluate our method, we first conduct extensive cross-domain generalization experiments across 10 diverse datasets, encompassing image classification tasks across different domains: FGVCAircraft [31], Caltech101 [12], StanfordCars [24], DTD [8], EuroSAT [17], Flower102 [33], Food101 [5], OxfordPets [34], SUN397 [48], and UCF101 [22]. This selection ensures a comprehensive assessment of the model's adaptability across varied visual domains. To further evaluate the robustness of our method under natural distribution shifts, we utilize the ImageNet [10] dataset along with its challenging out-of-distribution (OOD) variants: ImageNet-A [19], ImageNet-V2 [36], ImageNet-R [18], and ImageNet-S [43]. These benchmarks are designed to test the model's resilience to natural variations in data distribution.

Implementation Details. We follow prior work by utilizing the pre-trained CLIP model with either ResNet-50 or ViT-B/16 as the image encoder, paired with their respective text encoders. For class labels across different datasets, we follow TDA [21], employing specific template prompts for each dataset, which are processed by the text encoder to serve as the zero-shot CLIP classifier. For our FreeTTA, we set α to 0.2 and β to 4.5. During testing, we strictly adhere to the TTA setting in [39], using a batch size of 1. We use top-1 accuracy as the evaluation metric. All experiments are conducted on an NVIDIA 3090 GPU.

4.2. Comparison with the State-of-the-Art Methods

Table 1 and Table 2 present the comparisons on the cross-domain and out-of-distribution benchmarks, respectively, against state-of-the-art methods. Among them, CoOp [61] and CoCoOp [60] are few-shot adaptation methods that require labeled data from the target domain for prompt learning. For test-time adaptation methods, we categorize them

Method	T.D	Avail.	T.F.	AIR	CAL	CAR	DTD	EUR	FLWR	FOOD	PETS	SUN	UCF	AVG
CLIP-RN50	-	-	-	16.11	87.26	55.89	40.37	25.79	62.77	74.82	82.97	60.85	59.48	56.63
CoOp [61]	✗	✗	✗	15.12	86.53	55.32	37.29	26.20	61.55	75.59	87.00	58.15	59.05	56.18
CoCoOp [60]	✗	✗	✗	14.61	87.38	56.22	38.53	28.73	65.57	76.20	**88.39**	59.61	57.10	57.23
TPT [39]	✗	✓	✗	17.58	87.02	58.46	40.84	28.33	62.69	74.88	84.49	61.46	60.82	57.66
DiffTPT [13]	✗	✓	✗	17.60	86.89	**60.71**	40.72	41.04	63.53	**79.21**	83.40	62.72	62.67	59.85
Ours	✓	✓	✓	**17.83**	**90.12**	58.01	**44.21**	**43.64**	**68.26**	77.98	86.44	**62.84**	**63.97**	**61.33**
CLIP-ViT-B/16	-	-	-	23.22	93.55	66.11	45.04	50.42	66.99	82.86	86.92	65.63	65.16	64.59
CoOp [61]	✗	✗	✗	18.47	93.70	64.51	41.92	46.39	68.71	85.30	89.14	64.15	66.55	63.88
CoCoOp [60]	✗	✗	✗	22.29	93.79	64.90	45.45	39.23	70.85	83.97	90.46	66.89	68.44	64.63
PromptAlign [1]	✗	✗	✗	24.80	94.01	68.50	47.24	47.86	72.39	86.65	90.76	67.54	69.47	66.92
TDA [21]	✗	✗	✓	23.91	94.24	67.28	47.40	58.00	71.42	86.14	88.63	67.62	70.66	67.53
TPT [39]	✗	✓	✗	24.78	94.16	66.87	47.75	42.44	68.98	84.67	87.79	65.50	68.04	65.10
DiffTPT [13]	✗	✓	✗	25.60	92.49	67.01	47.00	43.13	70.10	87.23	88.22	65.74	62.67	65.47
MTA [53]	✗	✓	✓	**25.32**	94.13	**68.05**	45.59	38.71	68.26	84.95	88.22	64.98	68.11	64.63
ZERO [11]	✗	✓	✓	24.40	93.51	67.54	45.80	39.60	67.07	84.36	86.74	64.49	67.64	64.66
Ours	✓	✓	✓	25.11	**94.63**	67.34	**46.96**	**62.93**	**71.62**	**86.62**	**90.11**	**67.76**	**71.16**	**68.42**

Table 1. **Comparison on Cross-domain Benchmark.** The best performance for each dataset is highlighted in **bold**. Methods are categorized based on three key attributes: target distribution modeling (T.D.), availability (Avail.), and training-free (T.F.) characteristics.

based on three key attributes: target distribution modeling, availability, and training-free characteristics. Compared with other methods that also possess availability and training-free characteristics, our approach achieves stable improvements across diverse datasets, with an average gain of 3.76% on cross-domain benchmark and 1.66% on out-of-distribution benchmark.

Comparison on Cross-Domain Benchmark. Recent methods, such as MTA [53] and ZERO [11], address both availability and training-free issues but fail to leverage potential relationships between test samples. Compared to them, our FreeTTA consistently leads on 8 out of 10 datasets, showing average accuracy improvements of 3.99% and 1.58%, respectively. This demonstrates our design of explicitly modeling the target distribution through online EM enables more effective adaptation to the target domain. Furthermore, even compared to other TTA methods [1, 21] that are not training-free or lack availability, we still outperform them on most datasets. Our FreeTTA achieves a 1.5% improvement in average accuracy over PromptAlign [1], which requires source domain statistics and not being training-free, underscoring the computational efficiency and adaptation capabilities of our approach. Moreover, we outperform TDA [21] on 9 out of 10 datasets with an average improvement of 0.89%, which TDA necessitates an explicit cache of test sample features and uses them solely as instance-level references. In contrast, our method eliminates the need for such storage and models the target domain using an online EM approach. Compared to TPT [39], the pioneering TTA algorithm in VLMs, our method surpasses it on both ResNet-50 and ViT-B/16, with average improvements of 3.67% and 3.32%, respectively, demonstrating the effectiveness of our approach across different backbones.

Comparison on OOD Benchmark. Furthermore, we conduct additional comparisons with other methods on OOD datasets that focus on natural distribution shifts as shown in Table 2. Our method outperforms MTA [53] and ZERO [11] across all datasets due to its target distribution modeling capabilities, achieving average accuracy increases of 2.42% and 1.56%, and OOD accuracy gains of 2.79% and 1.66%. These consistent performance improvements highlight the effectiveness of our method for target distribution modeling while maintaining the advantages of being training-free and not requiring access to the source domain. Additionally, we also achieve average OOD accuracy improvements of 0.87% and 0.53%, compared with PromptAlign [1] and TDA [21]. Even greater gains are observed when compared to TPT [39] and DiffTPT [13], with average improvements of 3.14% and 3.3% and OOD accuracy gains of 3.61% and 3.9%. Our method consistently achieves performance gains, whether compared to traditional prompt-learning-based approaches or recent methods that possess availability or training-free characteristics, which demonstrates the robustness and effectiveness of our target distribution modeling design.

4.3. Ablation Study

We conduct ablation studies to demonstrate the effectiveness of our method, as shown in Table 3, and we use zero-

Method	T.D	Avail.	T.F.	ImageNet	-A	-V2	-R	-S	Average	OOD Average
CLIP-RN50	-	-	-	59.81	23.24	52.91	60.72	35.48	46.43	43.09
CoOp [61]	✗	✗	✗	63.33	23.06	55.40	56.60	34.67	46.61	42.43
CoCoOp [60]	✗	✗	✗	62.81	23.32	55.72	57.74	34.48	46.81	42.82
Tip-Adapter	✗	✗	✓	62.03	23.13	53.97	60.35	35.74	47.04	43.30
TPT	✗	✓	✗	60.74	26.67	54.70	59.11	35.09	47.26	43.89
DiffTPT	✗	✓	✗	60.80	**31.06**	55.80	58.80	37.10	48.71	45.69
Ours	✓	✓	✓	**61.51**	30.67	**55.89**	**63.02**	**37.94**	**49.81**	**46.88**
CLIP-ViT-B/16	-	-	-	68.34	49.89	61.88	77.65	48.24	61.20	59.42
CoOp [61]	✗	✗	✗	71.51	49.71	64.20	75.21	47.99	61.72	59.28
CoCoOp [60]	✗	✗	✗	71.02	50.63	64.07	76.18	48.75	62.13	59.91
Tip-Adapter	✗	✗	✓	70.75	51.04	63.41	77.76	48.88	62.37	60.27
PromptAlign [1]	✗	✗	✗	-	59.37	65.29	79.33	50.23	-	63.55
TDA [21]	✗	✗	✓	69.51	60.11	64.67	80.24	50.54	65.01	63.89
TPT [39]	✗	✓	✗	68.98	54.77	63.45	77.06	47.94	62.44	60.81
DiffTPT [13]	✗	✓	✗	70.30	55.68	65.10	75.00	46.80	62.28	60.52
MTA [53]	✗	✓	✓	69.29	57.41	63.61	76.92	48.58	63.16	61.63
ZERO [11]	✗	✓	✓	69.06	61.35	64.13	77.28	48.29	64.02	62.76
Ours	✓	✓	✓	**70.21**	**61.41**	**64.92**	**80.49**	**50.88**	**65.58**	**64.42**

Table 2. **Comparison on OOD Benchmark.** The best performance for each dataset is highlighted in **bold**. Methods are categorized according to the same criteria as in Table 1. The OOD average reflects the mean performance across the four ImageNet variant datasets.

shot CLIP-ViT-B/16 as the baseline (row 1).

Update Mean Vectors. To verify the importance of dynamically updating the mean vectors μ_y in our approach, we perform an ablation experiment using fixed mean vectors (row 3). In this experiment, the model utilizes mean vectors initialized with CLIP text embeddings without updates during testing. The results indicate that the model with fixed mean vectors fails to adapt the class feature centers effectively under significant distribution shifts, resulting in decreased classification accuracy. This highlights that the dynamic update of mean vectors is a crucial mechanism in our method, enabling the model to leverage inter-sample relationships in the visual branch and address the generalization challenges of the CLIP on target domains.

Update Covariance Matrix. We further investigate the necessity of dynamically updating the covariance matrix Σ by conducting experiments with a fixed covariance matrix (row 4). In this setup, the covariance matrix is set as an identity matrix and remains unchanged during testing, reducing Equ 3 to a Euclidean distance metric. The results demonstrate that a fixed covariance matrix limits the model's ability to represent intra-class variability, leading to suboptimal performance when adapting to the target domain. In contrast, our method with dynamically updated covariance matrices captures class distribution variations more effectively, thereby improving classification performance at test time.

VLM priors. Finally, we evaluate the impact of incorporating VLM priors for parameter initialization and adjusting influence on parameter updates by assessing each sample's confidence level (row 5). The results show that without consideration of VLM priors, the model's performance decreases due to noise from high-uncertainty samples. Conversely, incorporating self-entropy as an uncertainty measure allows the model to leverage its knowledge as priors to assess the contribution of samples adaptively, making it more robust and stable against noise.

	Method	Average Accuracy
1	Zero-Shot CLIP	64.59
2	FreeTTA	68.42
3	2-Mean Vectors Update	64.64
4	2-Covariance Matrix Update	67.07
5	2-VLM priors	67.78

Table 3. **Ablation Study** on Cross-Domain Benchmark. The CLIP ViT-B/16 is used.

5. Conclusion

This paper introduces a novel test-time adaptation approach for VLMs, leveraging the Gaussian discriminant analysis and an adaptive online EM algorithm to improve adaptability under domain shifts. By incorporating VLM priors as uncertainty measurement, our method effectively handles varying sequential online samples and enhances model stability during adaptation. Experimental results demonstrate that our approach significantly improves performance without relying on source domain data and costly training, showcasing its robustness and efficiency.

Acknowledgment: This work was supported by the National Natural Science Foundation of China (No.62206174).

References

[1] Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 4, 7, 8

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1, 2, 3

[3] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Biometrika*, 45(3-4):296–315, 1958. 2, 4, 5

[4] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006. 2, 4

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 6

[6] Ziyang Chen, Yongsheng Pan, Yiwen Ye, Mengkang Lu, and Yong Xia. Each test image deserves a specific prompt: Continual test-time adaptation for 2d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11184–11193, 2024. 3

[7] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 502–511, 2019. 3

[8] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 6

[9] Qiyuan Dai and Sibei Yang. Curriculum point prompting for weakly-supervised referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13711–13722, 2024. 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[11] Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. *arXiv preprint arXiv:2405.18330*, 2024. 2, 7, 8

[12] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 6

[13] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2704–2714, 2023. 1, 2, 3, 4, 7, 8

[14] Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368, 1922. 2, 4

[15] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 3

[16] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 3

[17] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 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. 6

[18] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, 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*, pages 8340–8349, 2021. 6

[19] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021. 6

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1, 2, 3

[21] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14162–14171, 2024. 1, 2, 3, 4, 6, 7, 8

[22] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6

[23] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 1

[24] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 6

[25] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022. 3

[26] Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. *arXiv preprint arXiv:2403.07366*, 2024. 2

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1, 2

[28] Wei Lin, Muhammad Jehanzeb Mirza, Mateusz Kozinski, Horst Possegger, Hilde Kuehne, and Horst Bischof. Video test-time adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22952–22961, 2023. 3

[29] Yahui Liu, Marco De Nadai, Jian Yao, Nicu Sebe, Bruno Lepri, and Xavier Alameda-Pineda. Gmm-unit: Unsupervised multi-domain and multi-modal image-to-image translation via attribute gaussian mixture modeling. *arXiv preprint arXiv:2003.06788*, 2020. 3

[30] Xiaosong Ma, Jie Zhang, Song Guo, and Wenchao Xu. Swapprompt: Test-time prompt adaptation for vision-language models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[31] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[32] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996. 2

[33] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 6

[34] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 6

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3

[36] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 6

[37] Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009. 3, 5

[38] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6902–6911, 2019. 3

[39] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022. 1, 2, 3, 4, 6, 7, 8

[40] Elaine Sui, Xiaohan Wang, and Serena Yeung-Levy. Just shift it: Test-time prototype shifting for zero-shot generalization with vision-language models. *arXiv preprint arXiv:2403.12952*, 2024. 1, 2, 3

[41] Olga Veksler. Test time adaptation with regularized loss for weakly supervised salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7360–7369, 2023. 3

[42] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 3

[43] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[44] Wei Wang, Zhun Zhong, Weijie Wang, Xi Chen, Charles Ling, Boyu Wang, and Nicu Sebe. Dynamically instance-guided adaptation: A backward-free approach for test-time domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24090–24099, 2023. 3

[45] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 3

[46] Markus Weber, Max Welling, and Pietro Perona. Unsupervised learning of models for recognition. In *Computer Vision-ECCV 2000: 6th European Conference on Computer Vision Dublin, Ireland, June 26–July 1, 2000 Proceedings, Part I 6*, pages 18–32. Springer, 2000. 3

[47] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. 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*, pages 7031–7040, 2023. 3

[48] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 6

[49] Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17503–17512, 2023. 3

[50] Sibei Yang, Meng Xia, Guanbin Li, Hong-Yu Zhou, and Yizhou Yu. Bottom-up shift and reasoning for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11266–11275, 2021. 3

[51] Teresa Yeo, Oğuzhan Fatih Kar, Zahra Sodagar, and Amir Zamir. Rapid network adaptation: Learning to adapt neural networks using test-time feedback. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4674–4687, 2023. 3

[52] Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang D Yoo. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. *arXiv preprint arXiv:2403.14119*, 2024. 2, 3

[53] Maxime Zanella and Ismail Ben Ayed. 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*, pages 23783–23793, 2024. 1, 2, 3, 7, 8

[54] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022. 3

[55] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022. 1, 3

[56] Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. Boostadapter: Improving vision-language test-time adaptation via regional bootstrapping. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2

[57] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2001. 3

[58] Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. Dual memory networks: A versatile adaptation approach for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 28718–28728, 2024. 1

[59] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16793–16803, 2022. 3

[60] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 1, 3, 6, 7, 8

[61] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 3, 6, 7, 8