# Test-Time Retrieval-Augmented Adaptation for Vision-Language Models

Xinqi Fan[1], Xueli Chen[2], Luoxiao Yang[3], Chuin Hong Yap[1], Rizwan Qureshi[4],
Qi Dou[5], Moi Hoon Yap[1], Mubarak Shah[4]
[1]Manchester Metropolitan University [2]Hong Kong Metropolitan University
[3]Xi'an University of Technology [4]University of Central Florida [5]Chinese University of Hong Kong

## Abstract

*Vision-language models (VLMs) have shown promise in test-time adaptation tasks due to their remarkable capabilities in understanding and reasoning about visual content through natural language descriptions. However, training VLMs typically demands substantial computational resources, and they often struggle to adapt efficiently to new domains or tasks. Additionally, dynamically estimating the test distribution from streaming data at test time remains a significant challenge. In this work, we propose a novel test-time retrieval-augmented adaptation (TT-RAA) method that enables VLMs to maintain high performance across diverse visual recognition tasks without the need for task-specific training or large computational overhead. During inference, TT-RAA employs a streaming mixture of Gaussian database (SMGD) to continuously estimate test distributions, requiring minimal storage. Then, TT-RAA retrieves the most relevant information from the SMGD, enhancing the original VLM outputs. A key limitation of CLIP-based VLMs is their inter-modal vision-language optimization, which does not optimize vision-space similarity, leading to larger intra-modal variance. To address this, we propose a multimodal retrieval augmentation module that transforms the SMGD into a unified multimodal space, enabling retrieval that aligns both vision and language modalities. Extensive experiments across both cross-domain and out-of-distribution benchmarks comprising fourteen datasets demonstrate TT-RAA's superior performance compared to state-of-the-art methods. Ablation studies and hyperparameter analyses further validate the effectiveness of the proposed modules. The source code of our work is available at https://github.com/xinqi-fan/TT-RAA.*

## 1. Introduction

Vision-language models (VLMs) have significantly advanced visual understanding by leveraging natural language supervision with contrastive language-image pre-training (CLIP) notably pioneering this field [34]. The zero-shot capability enables this model to be deployed in various domains. This foundational work has catalyzed a wave of architectural innovations, including domain-specific variants and architectural enhancements that build upon CLIP's core principles [52]. Some variants have focused on scaling up capacities, learning from noisy text supervision, and fine-grained alignment [16, 50]. However, these models consistently encounter challenges when deployed in specialized domains, where the distribution of images often differs substantially from their web-scale training data [11, 33]. This challenge has become even more prominent with the recent introduction of more sophisticated large VLMs such as GPT-4V [31], LLaVA [25], and Llama 3.2 Vision [1]. While these advanced models excel at integrating visual and textual data, the high computational cost of fine-tuning them for domain-specific tasks poses a major challenge, especially in specialized visual domains that differ significantly from general web images.

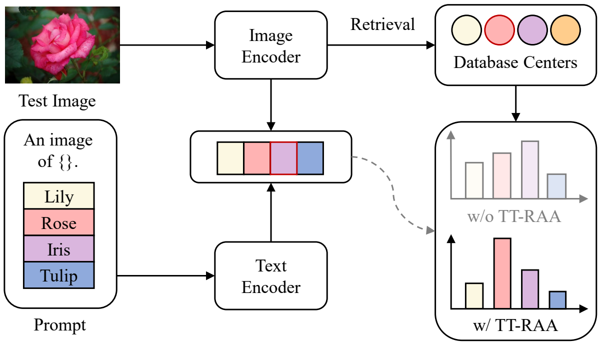


Figure 1. Illustration of our proposed test time retrieval augmented adaptation (TT-RAA). Without TT-RAA, the CLIP prediction will be the final prediction. With TT-RAA, the image feature will be used to retrieve the most similar database centers which will be used to improve the final prediction in a training-free manner.

Test-time adaptation (TTA) has emerged as a promising solution that enables models to adapt to new domains without accessing source domain training data [23]. Traditional TTA approaches include self-training-based methods that leverage pseudo-labels [22, 41, 49], self-supervised learning-based methods that incorporate auxiliary tasks [42], batch normalization [28, 37], and entropy minimization-based methods [30, 46]. For VLMs, recent works have explored prompt-based adaptation methods [10, 38] to bridge the domain gap between web-scale training data and specific downstream tasks. However, these methods often incur substantial computational overhead, highlighting the need for more efficient approaches [10, 17]. Advancements in training-free methods have explored the use of cache mechanisms [17, 55] and distribution-based methods [51, 58] to improve the efficiency. Despite these advances, dynamically estimating the test distribution from streaming data at test time remains a significant challenge, due to the test sample arriving one at a time.

In this work, we introduce test-time retrieval-augmented adaptation (TT-RAA), a novel training-free method that allows VLMs to sustain strong performance across various visual recognition tasks without requiring task-specific training or high computational costs (Fig. 1). TT-RAA is a training-free framework, which uses a streaming mixture of Gaussian database (SMGD) to dynamically capture test distribution characteristics with minimal storage requirements at inference time. TT-RAA retrieves relevant information from the SMGD to enhance the performance of the original VLM. Furthermore, an inherent limitation of CLIP-based VLMs is their lack of optimization for intramodal similarity, which does not optimize similarity in the vision space, resulting in greater intramodal variance [24]. To mitigate this, we introduce a multimodal retrieval augmentation (MRA) module that projects the SMGD into a shared multimodal embedding space, facilitating retrieval that harmonizes vision and language modalities. Extensive testing across ten cross-domain datasets and four out-of-distribution datasets confirms that TT-RAA delivers superior performance relative to state-of-the-art methods, with additional ablation studies and hyperparameter analyses supporting the effectiveness of the proposed modules. The key contributions of this work include:

- We introduce test-time retrieval-augmented adaptation (TT-RAA), an efficient, training-free framework that significantly improves domain adaptability without additional fine-tuning.
- We propose streaming mixture of Gaussian database (SMGD) for dynamic, efficient, and robust estimation of test domain statistics during inference.
- We propose multimodal retrieval augmentation (MRA), a novel mechanism addressing CLIP's inherent intra-modal similarity limitation, aligning both vision and language modalities effectively.

The rest of this paper is organized as follows: Section 2 provides some related works, Section 3 describes the proposed method. Experimental results and performance evaluation are presented in Section 4. Finally, Section 5 concludes the article and outlines future research directions.

## 2. Related Work

### 2.1. Vision Language Model

The development of vision language models (VLMs) can be traced back to earlier works that explored vision-language pre-training for specific tasks, where VisualBERT [21] and UNITER [5] pioneered joint representation learning between visual and textual modalities for tasks like visual question answering. The field witnessed a paradigm shift with CLIP [34], which demonstrated impressive zero-shot capabilities by learning transferable visual representations through contrastive learning between images and texts at web scale. ALIGN [16] further scaled up this approach with 1.8 billion noisy image-text pairs and introduced noise-robust training strategies. FILIP [50] enhanced CLIP by introducing fine-grained region-word alignment, enabling more detailed vision-language correspondences. FLAVA [39] introduced a single foundation model that can tackle vision, language, and multimodal tasks. More recently, the emergence of multimodal large language models has pushed the boundaries of vision-language understanding, where GPT-4V [31] extends GPT-4's language capabilities to handle visual inputs, LLaVA [25] achieves strong visual conversation abilities by aligning vision encoders with language models through minimal task-specific training, and Llama 3.2 [1] demonstrates remarkable vision and language capabilities with open models. Other notable VLMs include VinVL [53], which enhanced visual understanding by incorporating object detection and segmentation, and VisualGPT [4], which demonstrated zero-shot transfer to various vision-language tasks. These models have continued to push the boundaries of multimodal representation learning, enabling a wide range of applications from intelligent assistants to creative tools. While these newer models demonstrate impressive capabilities in combining visual and textual understanding, they face similar domain adaptation challenges but at an even larger scale. The computational demands for fine-tuning these models on domain-specific tasks present a significant barrier, particularly for specialized applications where the visual domain differs markedly from general web imagery.

### 2.2. Test Time Adaptation

Normal domain adaptation techniques require access to both source and target domain data, but test-time adaptation (TTA) adapts models to the target domain without source data, making it particularly useful for scenarios with privacy or legal restrictions [23]. To adapt a pre-trained model to a target domain, methods like self-training refine model predictions using pseudo-labels. Approaches such as centroid-based [22], neighbor-based [49], and optimization-based [41] strategies enhance the quality of these pseudo-labels. Self-supervised methods introduce auxiliary tasks to align features, exemplified by test-time training (TTT) [42] with a self-supervised rotation prediction task. Batch normalization [28, 37] techniques and entropy minimization strategies [30, 46] have also been adopted in TTA tasks.

Recent advancements focus on VLMs like CLIP, using domain-specific prompts from test data. Test-time prompt tuning (TPT) [38] fine-tunes a learnable prompt with each testing sample, while DiffTPT [10] uses pre-trained diffusion models to enhance test data diversity. Although these methods are promising, they still require a large amount of computational resources for their training-based adaptation, making it hard to adapt large VLMs. To reduce costs, a training-free dynamic adapter (TDA) [17] has been introduced, using a lightweight dynamic key-value cache without backpropagation. However, its cache only captures the streaming test data with a limited cache capacity without the ability to estimate the full spectrum of the test data.

### 2.3. Retrieval Augmented Generation

Retrieval-augmented generation (RAG) has emerged as a promising solution to enhance large language models (LLMs) by incorporating external knowledge during inference, effectively addressing challenges like hallucination and outdated knowledge [20]. In its naive form, RAG follows a three-stage process: indexing, where documents are chunked and encoded into vectors; retrieval, where relevant documents are retrieved based on semantic similarity with the input query; and generation, where retrieved documents along with the query are fed to LLMs for response generation [12]. Recent efforts have extended RAG to VLMs, including methods such as I2I/T2I adaptors [27], RA-TTA [19], and neural priming [45]. These methods rely on additional databases constructed from large-scale datasets like LAION-2B/5B, which often incur high computational and storage costs during both construction and inference. In contrast, our work focuses on constructing a database dynamically from incoming test samples, without relying on additional large-scale databases. In addition, we employed a multimodal retrieval augmentation module that transforms our database into a unified multimodal space, enabling retrieval that aligns both vision and language modalities.

## 3. Method

In this work, we introduce test-time retrieval-augmented Adaptation (TT-RAA) to adapt vision-language models (VLMs) to new domains (Fig. 2). TT-RAA leverages a streaming mixture of Gaussian database (SMGD) to model test data statistics in real time. As test data streams in, we dynamically update the SMGD with Gaussian centers representing the distribution of vision features. Each test sample's vision features query the SMGD to retrieve relevant Gaussian centers, providing contextual information. Since CLIP's joint vision-text optimization does not explicitly optimize vision-space similarity, we transform the Gaussian centers into a multimodal space using text embeddings. This allows retrieval in a space that is aligned with both vision and language modalities. Final predictions combine CLIP outputs with retrieval results from both vision and multimodal spaces, enhancing accuracy and robustness. TT-RAA thus provides a powerful framework for adapting VLMs to new domains by leveraging both vision and language information.

### 3.1. Contrastive Language-Image Pretraining

Contrastive language-image pre-training (CLIP) [34] is a vision-language model that has two separate encoders for processing images and text, respectively. These encoders are trained to project their respective inputs into a shared embedding space where related images and texts are positioned close to each other. Consider a $K$-class classification problem where the CLIP's objective in the zero-shot setting is to match images with their most relevant textual descriptions using an image encoder $E_I$ and a text encoder $E_T$. To obtain the textual descriptions $T$, $K$-class names are concatenated with hand-crafted prompts and then mapped into the $D$-dimensional text embeddings $Z$ using the text encoder $E_T$ as

$$Z = E_T(T) \in \mathbb{R}^{D \times K}. \tag{1}$$

Given the $t$-th test samples $x^t$, CLIP vision encoder produces image features

$$f^t = E_I(x^t) \in \mathbb{R}^{D}. \tag{2}$$

CLIP was trained using contrastive loss to minimize the distance between corresponding text and image embeddings. When making predictions, the CLIP output can be obtained as

$$P^t_{\text{CLIP}} = Z^T f^t \in \mathbb{R}^{K}. \tag{3}$$

Then, we can obtain the final prediction for test sample $x^t$ as $\hat{y}^t = \underset{k}{\operatorname{argmax}}\, P^t_{\text{CLIP}}(k)$, where $k \in \{1, 2, ..., K\}$.

### 3.2. Streaming Mixture of Gaussian Database

Our work addresses a challenging training-free test-time adaptation problem where models must adapt to streaming test data under two constraints: (1) Both input and output distributions of the source and target domains may be subject to distribution shifts. (2) Test samples arrive sequentially, one at a time, rather than in batches, reflecting real-world deployment. This setting is demanding because it requires training-free, zero-shot adaptation without access to batch statistics. Therefore, it is important to estimate the statistics of the new domain by capturing its full spectrum, and based on this we can retrieve the most relevant information to enhance the final prediction. In this work, we propose a Streaming Mixture of Gaussian Database (SMGD)

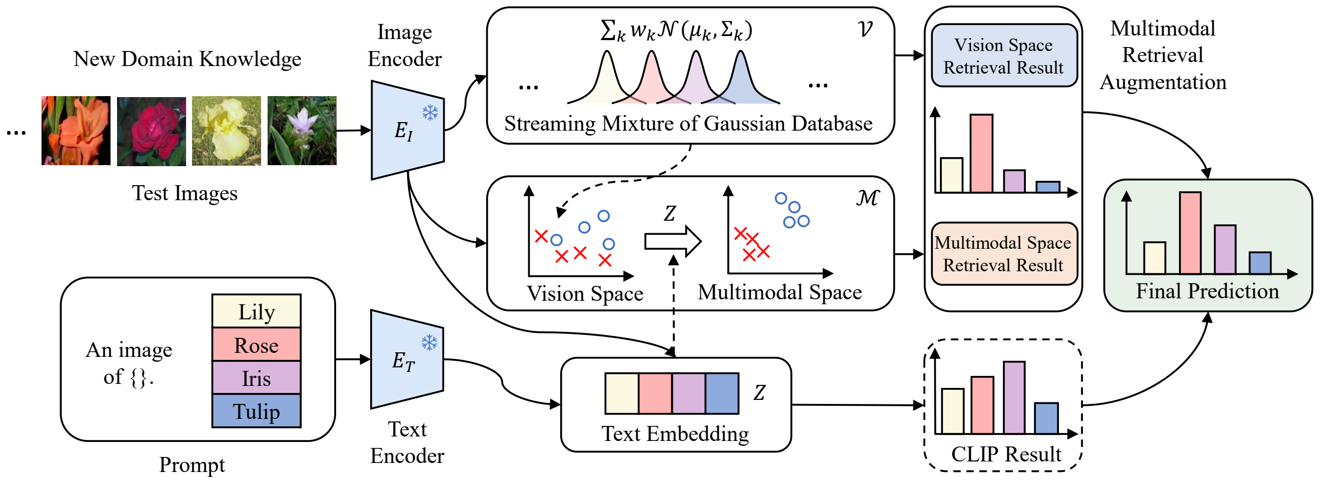


Figure 2. Overview of the proposed test-time retrieval-augmented adaptation (TT-RAA) method. TT-RAA leverages a streaming mixture of gaussian database (SMGD) to dynamically model the distribution of streaming test data in vision space $\mathcal{V}$. The multimodal retrieval augmentation (MRA) module projects these Gaussian centers into a unified multimodal space $\mathcal{M}$, enabling effective retrieval aligned across both vision and language modalities. Combining retrieval results from vision and multimodal spaces with CLIP's original predictions, TT-RAA significantly enhances VLM performance without additional training.

module by using a Gaussian mixture model to estimate the dynamic statistics from the test time streaming data.

We assume that feature embedding $f^t$ of the $t$-th sample $x_t$ is drawn from a mixture of Gaussian distributions

$$f^t \sim \sum_k w_k \mathcal{N}(\mu_k, \Sigma_k), \ \forall k = 1, 2, ..., K, \tag{4}$$

where $\mu_k \in \mathbb{R}^D$ and $\Sigma_k \in \mathbb{R}^{D\times D}$ are the mean and covariance of the $k$-th Gaussian component $\mathcal{N}$ repectively. The mixture weights $w_k \geq 0$, $\sum_k w_k = 1$, and $K$ is the number of classes. For simplicity, we assume the weights $w_k = p(y = k) = 1/K$. Here, SMGD is a key-value database, where the keys are Gaussian mean vectors $\{\mu_k\}_{k=1}^K$ and the values are the corresponding pseudo labels of these Gaussian mean vectors. The covariance matrices $\{\Sigma_k\}_{k=1}^K$ are also recorded which will involve in the retrieval during test time adaptation. Given the test time setting, all samples come in a stream. Our goal is to estimate the optimal mean $\mu_k$ and covariance $\Sigma_k$ for the class $k$.

To estimate the optimal value of the mean $\mu_k$ and covariance $\Sigma_k$, we draw inspiration from streaming data statistics estimation [35]. Given a test sample $x^t$ at time $t$, its pseudo label $\hat{l}_t$ is determined by its CLIP prediction as $\hat{l}_t = \underset{k}{\mathrm{argmax}}\, P^t_{\mathrm{CLIP}}(k)$. Initially, we set the mean vector $\mu_k$ to the first test sample with its pseudo label as $k$ from the streaming test data. We set the initial covariance $\Sigma_k = \sigma^2\mathbf{I}$ with $\sigma = 0.1$. We also record the corresponding entropy $h_k$ as the entropy of the first test sample's CLIP logit. Once the mean vector $\mu_k$ and covariance matrix $\Sigma_k$ are initialized, we progressively update them using selected test samples with high-quality pseudo labels, as measured by entropy. If the test sample $x^t$ at time $t$ has a pseudo label $\hat{l}^t = k$ and an entropy smaller than $h_k^{t-1}$, the mean $\mu_k^t$ and the covariance $\Sigma_k^t$ at time $t$ are updated as

$$\mu_k^t = (1-\eta)\mu_k^{t-1} + \eta f^t, \tag{5}$$

$$\Sigma_k^t = (1-\eta)\Sigma_k^{t-1} + \eta(f^t - \mu_k^t)(f^t - \mu_k^t)^T, \tag{6}$$

where $\eta$ is the update coefficient. Otherwise, $\mu_k^t = \mu_k^{t-1}$ and $\Sigma_k^t = \Sigma_k^{t-1}$ for $k \in \{1, 2, \ldots, K\}$, meaning the mean vectors and the covariance matrices remain unchanged if there is no test sample with pseudo label $k$ or if the test sample does not have a high-quality pseudo label. A high-quality pseudo label is identified when its entropy is lower than that of the Gaussian mean vector. Once the mean and covariance are updated according to Eqs. (5) and (6), the corresponding entropy is updated accordingly as

$$h_k^t = (1-\eta)h_k^{t-1} + \eta H(P^t_{\mathrm{CLIP}}), \tag{7}$$

where $\mathrm{H}(P^t_{\mathrm{CLIP}}) = -\sum_{k=1}^K P^t_{\mathrm{CLIP}}(k)\log(P^t_{\mathrm{CLIP}}(k))$ is the entropy of the CLIP logits $P^t_{\mathrm{CLIP}}$.

In this way, we keep updating a high-quality database with high confidence as measured by the entropy of test-time streaming data's pseudo labels. The SMGD $S_G$ is composed of the following:

$$S_G = \{G, L\}, \tag{8}$$

$$G = [\mu_0^t, \mu_1^t, \ldots, \mu_K^t], \tag{9}$$

where $G \in \mathbb{R}^{D\times K}$ represents the Gaussian centers of the SMGD, and each column of $L \in \mathbb{R}^{K\times K}$ denotes a $K$-dimensional one-hot pseudo-label vector. Along with SMGD, we have the covariances $\{\Sigma_0^t, \Sigma_1^t, \ldots, \Sigma_K^t\}$.

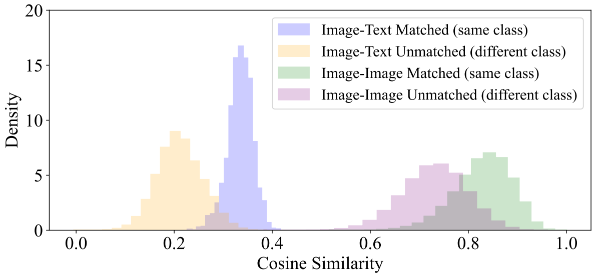


Figure 3. Motivation of multimodal retrieval augmentation. Cosine similarity distributions of matched and unmatched image-text pairs (inter-modal) exhibit less overlap than those of matched and unmatched image-image pairs (intra-modal). This suggests that samples are more easily distinguishable in the multimodal space than in the vision space.

### 3.3. Multimodal Retrieval Augmentation

We can directly retrieve the most similar data center from the SMGD to improve the final prediction. However, the CLIP model was optimized to reduce the inter-modal (vision-text) similarities rather than the intra-modal (vision-vision or text-text) similarities [24]. This optimization process leaves a disadvantage that similar images in the vision (image) feature space are not well clustered [43]. As shown in Fig. 3, the cosine similarity distributions of matched and unmatched image-text pairs (inter-modal) exhibit less overlap than those of matched and unmatched image-image pairs (intra-modal). This indicates that matched and unmatched pairs are more easily distinguishable in the multimodal space than in the vision-only space, a distinction shaped by CLIP's image-text contrastive learning. Therefore, in addition to the vision space retrieval augmentation, we also added a multimodal space retrieval augmentation, which compares the similarities of the samples in the CLIP text-image space rather than the vision space.

**Vision Space Retrieval Augmentation.** To fully leverage the available information from the vision space, we include two terms: similarity retrieval and discriminant analysis.

*Similarity Retrieval.* This term treats the test sample as a query and retrieves the most similar class center from the SMGD based on their similarities within the vision space. Given the SMGD $S_G = \{G, L\}$ and test image features $f^t$ generated from the CLIP's image encoder, the prediction can be calculated as follows

$$P_{sim}(f^t) = LA(G^T f^t) \in \mathbb{R}^K, \tag{10}$$

where $A(z) = \exp(-\beta(1-z))$ is an adaptation function with a sharpness ratio $\beta$ [54].

*Discriminant Analysis.* This term leverages probabilistic information derived from the Gaussian means and covariances of the SMGD by using a closed-form expression from Gaussian discriminant analysis [2]. Under the assumption that all Gaussian components share an identical covariance $\Sigma^t$, the discriminant score can be computed as

$$\Omega_{disc}(f^t) = G^T {\Sigma^t}^{-1} f^t - \frac{1}{2}\text{diag}(G^T {\Sigma^t}^{-1} G) + \log \frac{1}{K}\mathbf{1}_K \in \mathbb{R}^K, \tag{11}$$

where $\Sigma^t$ is the mean of $K$ covariances $\{\Sigma_k^t\}_{k=1}^K$. $\Omega_{disc}(f^t)$ contains scores that reflect how well $f^t$ matches different classes, considering both the distance from the class mean and the spread of the class distribution. Therefore, the discriminant analysis-based term can be given as

$$P_{disc}(f^t) = L\Omega_{disc}(f^t) \in \mathbb{R}^K. \tag{12}$$

After that, the prediction from the vision space $\mathcal{V}$ retrieval augmentation can then be calculated as

$$P_{\text{R}}^{\mathcal{V}}(f^t) = P_{sim}(f^t) + P_{disc}(f^t) \in \mathbb{R}^K, \tag{13}$$

**Multimodal Space Retrieval Augmentation.** During test-time adaptation, multimodal space retrieval augmentation treats the test sample as a query and retrieves the most similar Gaussian center from the SMGD by evaluating their distances within CLIP's multimodal space. This process involves measuring the Kullback–Leibler (KL) divergence between the image features of the test sample or Gaussian centers and the textual features of the categories in CLIP's embedding space.

Given the SMGD $S_G = \{G, L\}$ and the text embeddings $Z \in \mathbb{R}^{D\times K}$ generated from CLIP's text encoder, we transform Gaussian centers $G$ of the SMGD from the vision space $\mathcal{V}$ to the multimodal space $\mathcal{M}$ as $\Psi$:

$$\Psi = \sigma(Z^T G) \in \mathbb{R}^{K\times K}, \tag{14}$$

where $\sigma(\cdot)$ is a sigmoid function. We also project the test sample $f^t$ to the multimodal space $\mathcal{M}$ as

$$\psi = \sigma(Z^T f^t) \in \mathbb{R}^K. \tag{15}$$

We use Kullback–Leibler (KL) divergence to compare the similarities of the two distributions in the multimodal space as

$$\Phi_k = KL(\psi || \Psi_k), \tag{16}$$

where $\Psi_k$ is the $k$-th column vector of $\Psi$, and $KL(P|Q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right)$. Therefore, $\Phi = [\Phi_1, \Phi_2, \ldots, \Phi_K] \in \mathbb{R}^K$ consists of KL-divergence values. Since the most similar samples have low KL-divergence values (close to 0), they receive small weights. To address this, we negate the values in $\Phi$. The prediction of the multimodal space $\mathcal{M}$ retrieval augmentation will be given as

$$P_R^{\mathcal{M}} = -L\Phi \in \mathbb{R}^K. \tag{17}$$

Then, we can obtain the output of the multimodal retrieval augmentation (MRA) module $P_{\text{R}}$ as a combination of the

| **Method** | **Aircraft** | **Caltech101** | **Cars** | **DTD** | **EuroSAT** | **Flower102** | **Food101** | **Pets** | **SUN397** | **UCF101** | **Average** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 | 18.47 | 93.70 | 64.51 | 41.92 | 46.39 | 68.71 | 85.30 | 89.14 | 64.15 | 66.55 | 63.88 |
| CoCoOp | 22.29 | 93.79 | 64.90 | 45.45 | 39.23 | 70.85 | 83.97 | **90.46** | 66.89 | 68.44 | 64.63 |
| TPT | 24.78 | 94.16 | 66.87 | 47.75 | 42.44 | 68.98 | 84.67 | 87.79 | 65.50 | 68.04 | 65.10 |
| DiffTPT | **25.60** | 92.49 | 67.01 | 47.00 | 43.13 | 70.10 | **87.23** | 88.22 | 65.74 | 62.67 | 65.47 |
| MTA | 25.32 | 94.13 | **68.05** | 45.59 | 38.71 | 68.26 | 84.95 | 88.22 | 64.98 | 68.11 | 64.63 |
| DN | 24.30 | 93.60 | 64.00 | 45.70 | 53.30 | 68.00 | 86.00 | 87.70 | 66.50 | 68.40 | 65.75 |
| ZERO | 25.21 | 93.66 | 68.04 | 46.12 | 34.33 | 67.68 | 86.53 | 87.75 | 65.03 | 67.77 | 67.72 |
| DMN | 24.84 | 94.12 | 65.64 | 44.39 | 47.77 | 71.38 | 84.48 | 89.07 | 66.28 | 66.75 | 65.47 |
| TDA | 23.91 | **94.24** | 67.28 | 47.40 | 58.00 | 71.42 | 86.14 | 88.63 | 67.62 | 70.66 | 67.53 |
| **TT-RAA (ours)** | 25.38 | 94.08 | 66.42 | **47.99** | **66.12** | **72.68** | 86.09 | 89.83 | **67.69** | **71.29** | **68.76** |

Table 1. Results on the cross-domain benchmark. Our TT-RAA is compared with several state-of-the-art methods designed for vision-language models: the baseline method CLIP, four training-based adaptation methods (CoOp, CoCoOp, TPT, and DiffTPT), and five training-free adaptation methods (MTA, DN, DMN, ZERO, and TDA).

vision space retrieval argumentation result $P_{\mathrm{R}}^{\mathcal{V}}$ and multi-modal space augmentation result $P_{\mathrm{R}}^{\mathcal{M}}$ as

$$P_{\mathrm{R}} = P_{\mathrm{R}}^{\mathcal{V}} + P_{\mathrm{R}}^{\mathcal{M}}. \tag{18}$$

After that, we compute our final TT-RAA logits by combining the original CLIP logits $P_{\mathrm{CLIP}}$ with the MRA prediction $P_{\mathrm{R}}$ using a weighting factor $\alpha$ as

$$P_{\mathrm{TT-RAA}} = P_{\mathrm{CLIP}} + \alpha P_{\mathrm{R}}. \tag{19}$$

In the end, we can obtain the final prediction of TT-RAA as

$$\hat{y} = \underset{k}{\operatorname{argmax}}\, P_{\mathrm{TT-RAA}}(k), \tag{20}$$

where $k \in \{1, 2, ..., K\}$.

# 4. Experiments

## 4.1. Dataset

To validate the method's test time adaptation capability across different domains and variations within a familiar domain, we conducted experiments on two benchmarks: a cross-domain (CD) benchmark comprising 10 datasets and an out-of-distribution (OOD) benchmark comprising 4 datasets. The CD benchmark primarily assesses transferability by testing on domains that diverge significantly from the original training distribution. Specifically, it includes Aircraft [26], Caltech101 [9], Cars [18], DTD [6], EuroSAT [13], Flower102 [29], Food101 [3], Pets [32], SUN397 [48], and UCF101 [40]. In contrast, the OOD benchmark focuses on robustness to distribution shifts within the same general domain. Specifically, it includes ImageNet-A [15], ImageNet-V2 [36], ImageNet-R [14], and ImageNet-S [47].

## 4.2. Implementation Details

We implement our method based on the pre-trained CLIP architecture [34], which comprises a vision transformer (ViT) based image encoder [7] and a transformer-based text encoder [44]. Our experimental setup follows the challenging single-image test-time adaptation scenario, processing samples sequentially with a batch size of 1. To maintain computational efficiency and eliminate the need for backpropagation during inference, we utilize the hand-crafted prompts as proposed in [34] rather than learnable prompts. The experiments are conducted using the PyTorch framework, with evaluations performed on a single NVIDIA RTX A5000 GPU. Following standard practice in similar tasks, we employ top-1 accuracy as our evaluation metric.

## 4.3. Comparisons with State-of-the-art Methods

We conduct extensive experiments to evaluate TT-RAA against state-of-the-art approaches on the CD benchmark in Table 1 and OOD benchmark in Table 2. The competing methods include the CLIP baseline [34], training-based adaptation methods, and training-free adaptation methods.

**Discussions on Competing Methods.** In the field of efficient domain adaptation, several methods require training, either in the training stage or at the test time, to align the model more closely with the target distribution. Context optimization (CoOp) [57], conditional context optimization (CoCoOp) [56] both involve learning prompts during a training phase. CoOp focuses on optimizing prompt tokens that are fixed for each domain, while CoCoOp extends this to generate a conditional prompt token for each image. Test-time prompt tuning (TPT) [38] and diffusion model for TPT (DiffTPT) [10], on the other hand, adapt the model at test time by tuning prompt embeddings. TPT performs direct gradient-based prompt tuning using test data, while DiffTPT uses new data generated by diffusion models.

In addition to training-based methods, we also compare with several training-free adaptation methods, including two distribution-based methods, one entropy-based method, and two cache-based methods. Distribution-based methods evaluated are MeanShift for Test-time Augmentation (MTA) [51] and Distribution Normalization (DN) [58].

| Method | ImageNet-A | ImageNet-V2 | ImageNet-R | ImageNet-S | Average |
|---|---|---|---|---|---|
| CLIP-ViT-B/16 | 49.89 | 61.88 | 77.65 | 48.24 | 59.42 |
| CoOp | 49.71 | 64.20 | 75.21 | 47.99 | 59.28 |
| CoCoOp | 50.63 | 64.07 | 76.18 | 48.75 | 59.91 |
| TPT | 54.77 | 63.45 | 77.06 | 47.94 | 60.81 |
| DiffTPT | 55.68 | 65.10 | 75.00 | 46.80 | 60.52 |
| MTA | 57.41 | 63.61 | 76.92 | 48.58 | 61.63 |
| DN | 58.71 | 62.89 | 80.20 | 48.94 | 62.69 |
| ZERO | 59.61 | 64.16 | 77.22 | 48.40 | 62.35 |
| DMN | 58.28 | **65.17** | 78.55 | **53.20** | 63.80 |
| TDA | 60.11 | 64.67 | 80.24 | 50.54 | 63.89 |
| **TT-RAA (Ours)** | **60.59** | 64.69 | **80.58** | 49.98 | **63.96** |

Table 2. Results on the out-of-distribution benchmark. Our TT-RAA is compared with several state-of-the-art methods designed for vision-language models: the baseline method CLIP, four training-based adaptation methods (CoOp, CoCoOp, TPT, and DiffTPT), and five training-free adaptation methods (MTA, ZERO, DN, DMN, and TDA).

They improve the test data distribution estimation via MeanShift algorithm and distribution normalization, respectively. Cache-based methods include the Dual Memory Network (DMN) [55] and TDA [17]. DMN comprises a dynamic and a static cache, while TDA employs a positive and a negative cache. For a fair comparison, we reproduced DMN's zero-shot results using our standardized prompts instead of DMN's original LLM-generated prompts, and we reduced the batch size to 1. We also compare against ZERO [8], an entropy-based method. ZERO sets the temperature of most confident predictions as zero to approximate marginal entropy minimization.

**Results on the Cross-domain Benchmark.** Despite having the advantage of task-specific training, training-based adaptation methods, such as CoOp, CoCoOp, TPT, and DiffTPT, show limited generalization capabilities. TT-RAA outperforms these methods by significant margins (about 3%-4%). On fine-grained recognition tasks such as Flowers102, TT-RAA achieves 72.68% accuracy, surpassing CoCoOp (70.85%) by 1.83%. This gap widens further on specialized datasets like SUN397, where TT-RAA (67.69%) demonstrates superior feature adaptation compared to CoCoOp (64.15%). These results suggest that static prompt learning, even with input-conditional mechanisms, may not be sufficient enough for handling significant domain shifts.

In comparison to existing training-free adaptation approaches, TT-RAA demonstrates consistent superiority across various domain shifts. On remote sensing data (EuroSAT), TT-RAA (66.12%) significantly outperforms TDA (58.00%) by 8.12%. This substantial gain can be attributed to our SMGD effectively capturing the unique characteristics of satellite imagery. Similarly, for action recognition (UCF101), TT-RAA (71.29%) surpasses TDA (70.66%) by 0.63%, demonstrating robust feature adaptation. The performance advantages extend to other specialized domains as well. This consistent improvement across diverse domains validates the effectiveness of our method in bridging domain gaps in test-time settings.

**Results on the Out-of-distribution Benchmark.** The experimental results in Table 2 demonstrate that our proposed TT-RAA method achieves state-of-the-art performance on the OOD benchmark with an average accuracy of 63.96% across all ImageNet variants, outperforming both training-based adaptation methods (CoOp, CoCoOp, TPT, DiffTPT) and training-free approaches (MTA, DN, ZERO, DMN, TDA). Notably, TT-RAA exhibits good performance on ImageNet-R (80.58%) and ImageNet-A (60.59%), highlighting its robust generalization capabilities under diverse distribution shifts. When examining performance across different distribution shifts, we observe that most methods achieve their highest accuracy on ImageNet-R and ImageNet-V2. ImageNet-S with sketches poses the greatest challenge for all methods, indicating that semantic variations remain difficult to address even with advanced adaptation techniques.

## 4.4. Ablation Studies

We conduct ablation studies on the CD benchmark to systematically evaluate the effectiveness of each proposed component in TT-RAA. Figure 4 presents the quantitative results of our analysis with both ResNet-50 and ViT-B/16 backbones. We establish our baseline using the TDA method with one-shot capacity, which achieves 59.82% (ResNet-50) and 67.04% (ViT-B/16) accuracy.

**Streaming Mixture of Gaussian Database Module.** When integrating the SMGD module with the baseline, we observe a performance improvement of about 1.4% with ViT-B/16 and 1.2% with ResNet-50. This enhancement demonstrates that our SMGD effectively captures the dynamic distribution of test samples while maintaining computational efficiency. Its adaptive nature in updating class statistics online is useful for handling distribution shifts.

**Multimodal Retrieval Augmentation Module.** Incorporating the MRA module also improved the performance by

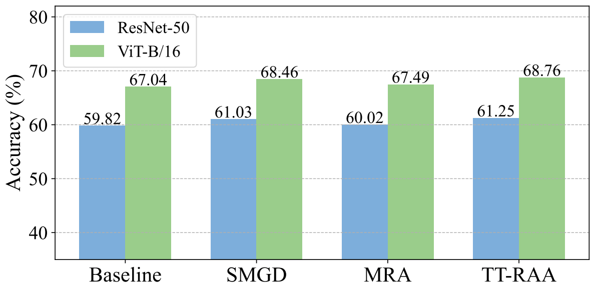


Figure 4. Ablation studies of TT-RAA. The SMGD and MRA modules independently improve performance over the baseline. Their combination yields further performance gains, demonstrating complementary benefits.

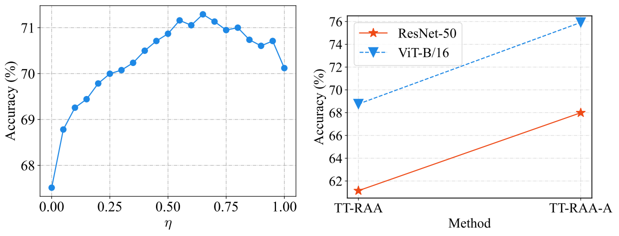
about 0.4% and 0.2% with ViT-B/16 and ResNet-50, respectively. This gain supports our hypothesis regarding vision space misalignment in CLIP, which stems from its vision-text joint optimization objective. The performance improvement demonstrates that projecting both the SMGD database and test samples from vision space to multimodal space effectively addresses this misalignment.

**Complementary Benefits of Proposed Modules.** When combining both SMGD and MRA modules (denoted as TT-RAA), we achieve the best performance of 68.76% with ViT-B/16 and 61.25% with ResNet-50, representing a substantial improvement of 1.72% and 1.43% over the baseline. These gains suggest that SMGD and MRA offer complementary and synergistic benefits.

**Analysis of Feature Backbone Impact.** The effectiveness of our approach is further validated across two vision ecoders: ViT-B/16 and ResNet-50. These consistent improvements across different backbones suggest that our adaptation strategy effectively addresses fundamental domain shift challenges rather than exploiting architecture-specific characteristics.

## 4.5. Hyperparameter Analysis

The hyperparameter analysis of the update coefficient $\eta$ on the UCF101 is shown in Fig. 5a. When $\eta = 0$, the model maintains a static Gaussian center initialized at the start, resulting in poor adaptation. From 0 to 0.55, accuracy increases sharply, highlighting the importance of incorporating new information. However, the performance deterioration beyond 0.65 reveals a critical threshold where excessive emphasis on new samples compromises the model's stability. This observation aligns with theoretical expectations, as larger $\eta$ values reduce the influence of historical information, potentially leading to over-adaptation to new samples. The empirical optimal value of $\eta = 0.65$ suggests that maintaining approximately 35% of historical information while incorporating 65% of new information achieves the best trade-off.



(a) (b)

Figure 5. (a) Hyperparameter analysis of the update coefficient $\eta$. It shows that the best coefficient is 0.65. (b) Experiment with additional data. The method can accept additional data, which improve the performance significantly.

## 4.6. Experiments with Access to Additional Data

In the standard test-time adaptation setting, access to the target domain's training set is typically restricted. However, in many cases, the training set data might be available. Fine-tuning large-scale vision-language models in such scenarios remains computationally expensive. This raises an important question: how can performance be improved using the training data without computationally expensive fine-tuning? With a slight modification to TT-RAA, we propose leveraging the target domain training set (if accessible prior to test time) to construct a database and calculate class centers using a Gaussian mixture model. By doing so, we perform the same MRA process as in TT-RAA. This modified approach, referred to as TT-RAA-A, demonstrates significant performance improvements while maintaining a low computational requirement. The corresponding results are presented in Fig. 5b, where TT-RAA-A outperforms TT-RAA by a large margin, highlighting the benefits of incorporating target training data in a training-free manner, which open a new direction on adaptation without training.

# 5. Conclusion

In this paper, we introduced TT-RAA, a novel test-time retrieval augmented adaptation method that effectively addresses the computationally expensive domain adaptation challenges faced by vision-language models. Experimental results validated the effectiveness of SMGD by continuously estimating the test distribution and compressing all useful information into a single representation for each class without discarding old samples. By transforming Gaussian centers from vision space into a shared multimodal space, MRA enables more effective retrieval that leverages both visual and textual features. The success of TT-RAA opens up new possibilities for deploying vision-language models in real-world applications where computational resources are limited and rapid adaptation to new domains is crucial. Further exploration on adapting our method to larger multimodal foundation and large language models is promising.

## References

[1] Meta AI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Report*, 2024. 1, 2

[2] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006. 5

[3] Lukas Bossard, Matthieu Guillaumin, et al. Food-101: Mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 6

[4] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. VisualGPT: Data-efficient adaptation of pretrained language models for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040, 2022. 2

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. *European Conference on Computer Vision*, pages 104–120, 2020. 2

[6] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 6

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 6

[8] Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. *Advances in Neural Information Processing Systems*, 2024. 7

[9] 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 *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2004. 6

[10] Chun-Mei Feng, Kai Yu, Yong Liu, et al. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *IEEE/CVF International Conference on Computer Vision*, pages 2704–2714, 2023. 2, 3, 6

[11] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, et al. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024. 1

[12] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, et al. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2024. 3

[13] Patrick Helber, Benjamin Bischke, et al. 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*, 2019. 6

[14] Dan Hendrycks, Steven Basart, Norman Mu, 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*, 2021. 6

[15] 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

[16] 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. *International Conference on Machine Learning*, pages 4904–4916, 2021. 1, 2

[17] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3, 7

[18] Jonathan Krause, Michael Stark, et al. 3D object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops*, 2013. 6

[19] Youngjun Lee, Doyoung Kim, Junhyeok Kang, Jihwan Bang, Hwanjun Song, et al. RA-TTA: Retrieval-augmented test-time adaptation for vision-language models. In *International Conference on Learning Representations*, 2025. 3

[20] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, and Vladimir others Karpukhin. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, pages 9459–9474, 2020. 3

[21] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[22] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pages 6028–6039. PMLR, 2020. 1, 2

[23] Jian Liang, Ran He, et al. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, pages 1–34, 2024. 1, 2

[24] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 2, 5

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[26] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, et al. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[27] Yifei Ming and Yixuan Li. Understanding retrieval-augmented task adaptation for vision-language models. In *International Conference on Machine Learning*, pages 35719–35743, 2024. 3

[28] Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, et al. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020. 2, 3

[29] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 6

[30] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable

test-time adaptation in dynamic wild world. In *Internetional Conference on Learning Representations*, 2023. 2, 3

[31] OpenAI. Gpt-4v(ision) system card. In *OpenAI Technical Report*, 2023. 1, 2

[32] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 6

[33] Rizwan Qureshi, Ranjan Sapkota, Abbas Shah, et al. Thinking beyond tokens: From brain-inspired intelligence to cognitive foundations for artificial general intelligence and its societal impact. *arXiv preprint arXiv:2507.00951*, 2025. 1

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 1, 2, 3, 6

[35] Aditi Raghunathan, Prateek Jain, et al. Learning mixture of gaussians with streaming data. *Advances in Neural Information Processing Systems*, 30, 2017. 4

[36] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, et al. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 2019. 6

[37] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, et al. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 2020. 2, 3

[38] 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*, 2022. 2, 3, 6

[39] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 2

[40] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 6

[41] Tao Sun, Cheng Lu, and Haibin Ling. Prior knowledge guided unsupervised domain adaptation. In *European Conference on Computer Vision*. Springer, 2022. 1, 2

[42] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, pages 9229–9248. PMLR, 2020. 2, 3

[43] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. SuS-X: Training-free name-only transfer of vision-language models. In *IEEE/CVF International Conference on Computer Vision*, pages 2725–2736, 2023. 5

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 6

[45] Matthew Wallingford, Vivek Ramanujan, Alex Fang, Aditya Kusupati, Roozbeh Mottaghi, Aniruddha Kembhavi, et al. Neural priming for sample-efficient adaptation. *Advances in Neural Information Processing Systems*, 2023. 3

[46] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. 2, 3

[47] Haohan Wang, Songwei Ge, et al. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, 2019. 6

[48] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 6

[49] Shiqi Yang, Joost Van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 34:29393–29405, 2021. 1, 2

[50] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *International Conference on Learning Representations*, 2022. 1, 2

[51] 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. 2, 6

[52] Jingyi Zhang, Jiaxing Huang, et al. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[53] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. VinVL: Revisiting visual representations in vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 2

[54] 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. 5

[55] 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. 2, 7

[56] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 6

[57] 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. 6

[58] Yifei Zhou, Juntao Ren, Fengyu Li, Ramin Zabih, and Ser Nam Lim. Test-time distribution normalization for contrastively learned visual-language models. *Advances in Neural Information Processing Systems*, pages 47105–47123, 2023. 2, 6