Advancing Reliable Test-Time Adaptation of Vision-Language Models under Visual Variations

Yiwen Liang
Tsinghua University
Beijing, China
evenliang789@gmail.com

Hui Chen*
Tsinghua University
Beijing, China
jichenhui2012@gmail.com

Yizhe Xiong
Tsinghua University
Beijing, China
xiongyizhe2001@163.com

Zihan Zhou
Tsinghua University
Beijing, China
blurry.kokan@gmail.com

Mengyao Lyu
Tsinghua University
Beijing, China
mengyao.lyu@outlook.com

Zijia Lin
Tsinghua University
Beijing, China
linzijia07@tsinghua.org.cn

Shuaicheng Niu
Nanyang Technological University
Singapore, Singapore
shuaicheng.niu@ntu.edu.sg

Sicheng Zhao
Tsinghua University
Beijing, China
schzhao@gmail.com

Jungong Han
Tsinghua University
Beijing, China
jungonghan77@gmail.com

Guiguang Ding*
Tsinghua University
Beijing, China
dinggg@tsinghua.edu.cn

Abstract

Vision-language models (VLMs) exhibit remarkable zero-shot capabilities but struggle with distribution shifts in downstream tasks when labeled data is unavailable, which has motivated the development of Test-Time Adaptation (TTA) to improve VLMs' performance during inference without annotations. Among various TTA approaches, cache-based methods show promise by preserving historical knowledge from low-entropy samples in a dynamic cache and fostering efficient adaptation. However, these methods face two critical reliability challenges: (1) entropy often becomes *unreliable* under distribution shifts, causing error accumulation in the cache and degradation in adaptation performance; (2) the final predictions may be *unreliable* due to inflexible decision boundaries that fail to accommodate large downstream shifts. To address these challenges, we propose a Reliable Test-time Adaptation (ReTA)[1] method that integrates two complementary strategies to enhance reliability from two perspectives. First, to mitigate the unreliability of entropy as a sample selection criterion for cache construction, we introduce Consistency-aware Entropy Reweighting (CER), which incorporates consistency constraints to weight entropy during cache updating. While conventional approaches rely solely on low entropy for cache prioritization and risk introducing noise, our method leverages predictive consistency to maintain a high-quality cache and facilitate more robust adaptation. Second, we present Diversity-driven Distribution Calibration (DDC), which models class-wise text embeddings as multivariate Gaussian distributions, enabling adaptive decision boundaries for more accurate predictions across visually diverse content. Extensive experiments demonstrate that ReTA consistently outperforms state-of-the-art methods, particularly under real-world distribution shifts.

CCS Concepts

• **Computing methodologies** → **Transfer learning**.

Keywords

Transfer Learning, Vision-Language Models, Test-Time Adaptation

ACM Reference Format:
Yiwen Liang, Hui Chen, Yizhe Xiong, Zihan Zhou, Mengyao Lyu, Zijia Lin, Shuaicheng Niu, Sicheng Zhao, Jungong Han, and Guiguang Ding. 2025. Advancing Reliable Test-Time Adaptation of Vision-Language Models under Visual Variations. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755607

1 Introduction

Large-scale vision-language models (VLMs), such as CLIP [33] and ALIGN [18], pre-trained on massive web-scale datasets, have demonstrated impressive zero-shot capabilities and strong open-world visual understanding across a wide range of vision tasks, including classification [7, 40, 41], retrieval [17, 45], and segmentation [11, 19]. Their core mechanism involves learning a shared representation space by aligning visual and textual modalities during training, and subsequently employing similarity-based matching for classification at inference. However, VLMs often suffer from performance degradation when deployed on unlabeled test data from downstream domains that exhibit significant distribution shifts from their pre-training distribution [2, 27, 46], limiting their effectiveness in real-world scenarios with diverse visual variations.

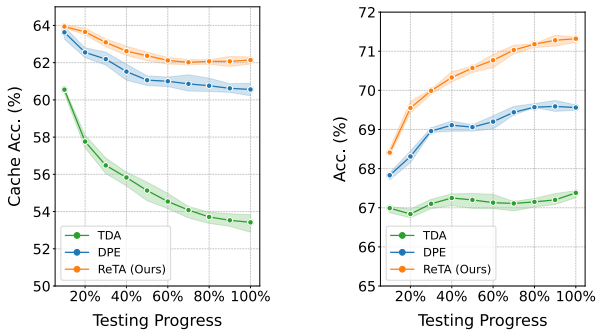
Test-time adaptation (TTA) strategies have recently emerged to boost VLMs' capacity for adaptation to downstream out-of-distribution scenarios without relying on labeled data. Initial efforts focused on prompt-based approaches [1, 10, 37], which adapt vision-language models by learning domain-specific prompts to minimize

*Corresponding author.

[1]Code is available at https://github.com/Evelyn1ywliang/ReTA.

MM '25, Dublin, Ireland.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755607



(a) (b)

Figure 1: Impact of cache reliability on test-time adaptation performance. (a): Accuracy of samples stored in the cache during testing. (b): Accuracy of test samples during testing. As adaptation progresses, baseline methods show significant degradation in cache quality, while our proposed ReTA consistently maintains more reliable samples and further improves prediction accuracy under distribution shifts via adaptive decision boundaries.

the entropy of predictions from augmented test samples. While effective, these methods typically require iterative backpropagation through the entire encoder, leading to substantial computational overhead. Recently, cache-based TTA methods [20, 49, 51, 52] have gained significant attention for their efficiency and are rapidly emerging as a dominant paradigm. These approaches typically construct a dynamic cache that is updated online as new data arrives. High-confidence samples are selected based on the entropy of VLM predictions to populate the cache, which serves as historical knowledge to calibrate model predictions either in a training-free manner or with minimal parameter adjustments. For example, TDA [20] introduces dynamic caches to generate both positive and negative predictions for precise prediction refinement. DMN[52] constructs complementary memory banks to exploit both long-term and short-term knowledge for enhanced adaptation. DPE [49] employs dual-modal residual learning to extract more accurate multi-modal representations while maintaining alignment between visual and textual prototypes. These cache-based approaches offer notable computational efficiency with minimal overhead, making them particularly well-suited for resource-constrained applications.

Despite advances in cache-based TTA methods, reliable adaptation of VLMs under significant distribution shifts—particularly visual variations such as lighting changes, style differences, or domain gaps between training and testing—remains a critical challenge. First, these approaches [20, 49, 51] generally rely on entropy for cache prioritization. However, prior studies [8, 16, 21] show that entropy-based measures become unreliable in unlabeled test-time scenarios, leading to suboptimal cache construction. Furthermore, our observations indicate that the accuracy of samples stored in the cache progressively declines as testing proceeds (as shown in Figure 1a), further validating the unreliability of entropy as a measure of sample quality. These low-quality samples inevitably introduce noise into the cache, causing the cached class-wise representations to diverge from their true distribution over time, which significantly limits performance improvements (see Figure 1b). Second, the decision boundaries of VLMs become unreliable under distribution

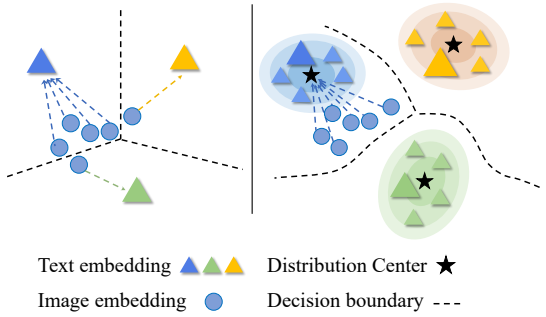


Figure 2: Decision boundary visualization. Left: Conventional cache-based TTA uses fixed text embeddings, creating rigid boundaries that struggle with distribution shifts. Right: ReTA implements adaptive distribution modeling for more flexible boundaries that better accommodate visual variations.

shifts to unlabeled target domains. Even with TTA, these models struggle to adapt effectively, as their boundaries lack the flexibility to accommodate such shifts, leading to unreliable predictions. Although some methods [49, 52] attempt to refine the decision boundaries using mean textual prototypes derived from enriched prompts, they still lack the flexibility to handle intra-class diversity shifts, exemplified by the misalignment of blue image embeddings from their corresponding blue text embeddings in Figure 2.

To address these issues, we propose Reliable Test-time Adaptation (ReTA), a method designed to improve the reliability of VLMs under distribution shifts caused by natural visual variations. ReTA integrates an entropy reweighting strategy and a distribution calibration technique to enhance adaptation performance. Specifically, we introduce Consistency-aware Entropy Reweighting (CER) to identify reliable samples for cache prioritization. CER constructs adjacent class-specific textual representations as a voting committee [28, 35, 43], evaluating prediction consistency across multiple semantic perspectives to identify reliable samples. By prioritizing semantically consistent samples during cache updates, CER promotes the retention of high-quality image features in the cache, mitigating distribution drift and enhancing adaptation stability. Furthermore, to improve the reliability of predictions under visual diversity, we propose Diversity-driven Distribution Calibration (DDC). DDC leverages adjacent text embeddings constructed by CER to extend each class representation from a single prototype to an approximate multivariate Gaussian distribution. Through residual learning and progressive updates, DDC enables more flexible decision boundaries that better accommodate visual diversity and continuously adapt to feature variations during test-time adaptation.

The contributions of this work are summarized as follows:

- We propose ReTA, a unified test-time adaptation method designed to enhance the reliability of vision-language models under significant distribution shifts in test-time adaptation.
- We introduce Consistency-aware Entropy Reweighting (CER), which constructs class-specific textual representations to evaluate prediction consistency across adjacent representations. CER identifies high-consistency samples and prioritizes them during cache updates to enhance adaptation.
- We develop Diversity-driven Distribution Calibration (DDC), which models text embeddings as approximate multivariate

Gaussian distributions with residual learning and progressive updates, enabling adaptive decision boundaries that support reliable predictions under visual variations.
- Our comprehensive evaluation across multiple benchmarks demonstrates that ReTA consistently outperforms state-of-the-art TTA methods and is effective in adapting VLMs.

2 Related Works

2.1 Vision-Language Models

Large-scale pre-trained vision-language models (VLMs), such as CLIP [33], have demonstrated remarkable capabilities in learning generalizable visual and textual representations. Numerous efforts have focused on adapting these powerful pre-trained models to downstream tasks. Early approaches like CoOp [55] and CoCoOp [54] achieved notable gains through lightweight prompt learning with limited supervision in few-shot scenarios. More recent works, such as Tip-Adapter [50] and TaskRes [48], introduced efficient adaptation strategies by leveraging feature memory constructed from a small set of labeled samples. These memory-based methods have attracted considerable attention recently for avoiding full model fine-tuning while enabling rapid domain adaptation with minimal overhead. Despite their strong performance and low data requirements, they still rely on labeled target-domain data, which limits their applicability in real-world scenarios where annotations are unavailable. In contrast, our work focuses on test-time adaptation, where models must generalize to novel domains without any labeled data during inference, better reflecting real-world constraints and advancing truly adaptive vision-language systems.

2.2 Vision-Language Test-Time Adaptation

Test-time adaptation (TTA) enhances the adaptability of vision-language models (VLMs) to target domains using only unlabeled test samples [8, 29, 37, 53]. Recent approaches primarily fall into two categories: prompt-based and cache-based methods. Prompt-based methods optimize continuous textual embeddings during inference. TPT [37] fine-tunes prompts by enforcing consistency across augmented views of the same image, while DiffTPT [10] introduces diffusion-based augmentations to increase view diversity and improve model adaptability to unseen test data. Cache-based approaches leverage historical instance-level information from test samples. TDA [20] constructs positive and negative feature caches from test samples to refine CLIP predictions via combined similarity scores. DMN [52] integrates static and dynamic memory networks to exploit both few-shot and historical unlabeled data, using a flexible interactive strategy for improved adaptation. BoostAdapter [51] incorporates regional-level bootstrapping for more precise memory construction. DPE [49] jointly evolves visual and textual prototypes to better capture domain-specific semantics. While prompt-based methods offer simplicity, cache-based strategies generally achieve stronger adaptation and maintain high computational efficiency by more effectively modeling target domain distributions through lightweight test-time memory. Our work builds on the cache-based paradigm, with advancements in cache quality via reliable sample selection and improving prediction robustness through adaptive decision boundary calibration in dynamic environments.

3 Methodology

The pipeline of our proposed method is illustrated in Figure 3. Sec. 3.1 presents preliminaries on CLIP and cache-based test-time prediction. Then, we propose Consistency-aware Entropy Reweighting (CER) in Sec. 3.2, which selects reliable samples based on prediction consistency to construct a high-quality cache. To enhance adaptation to diverse visual content, we introduce Diversity-driven Distribution Calibration (DDC) in Sec. 3.3, which models class representations as approximate Gaussian distributions. Finally, Sec. 3.4 details our unified learning process and inference algorithm, showing how they train and inference in an end-to-end manner.

3.1 Preliminaries

Zero-shot CLIP Prediction. CLIP is pre-trained on large-scale image-text pairs to align visual and textual modalities within a unified embedding space by maximizing cosine similarity through a contrastive loss. CLIP consists of a visual encoder $\mathcal{F}_V$ and a text encoder $\mathcal{F}_T$, which project inputs into a shared latent space $\mathbb{R}^d$, where d is the feature dimension, for cross-modal alignment. During inference, an image x is classified by computing similarity between the image feature $\mathbf{z} = \mathcal{F}_V(x)$, and textual embeddings $\boldsymbol{t}^c = \mathcal{F}_T(\mathcal{P}^c)$, where $\mathcal{P}^c$ denotes a text prompt describing class c, and C is the total number of classes. The zero-shot CLIP prediction probability is then given by:

$$p^c_{\text{CLIP}} = \frac{\exp\left(\mathbf{z} \cdot \boldsymbol{t}^c / \tau\right)}{\sum_{j=1}^{C} \exp\left(\mathbf{z} \cdot \boldsymbol{t}^j / \tau\right)} \tag{1}$$

where τ is the temperature parameter, typically set to 0.01.

Cache-based Test-time Prediction. Recent approaches leverage cache mechanisms at test time for efficient vision-language model adaptation. Building upon the key-value storage paradigm of TIP-Adapter [50], many cache-based TTA methods [20, 49, 51] have emerged, dynamically storing and updating high-confidence feature-label pairs during inference. The cache model functions as a priority queue, ranked by entropy derived from CLIP prediction. With $\mathbf{W}_{\text{CLIP}} = \left[\boldsymbol{t}^1, \boldsymbol{t}^2, \ldots, \boldsymbol{t}^C\right]^\top$ as stacked text embeddings serving as the classifier, each cache item is represented as:

$$\left\{\mathbf{F}_{\text{cache}}, \mathbf{L}_p, H\left(\mathbf{F}_{\text{cache}}\, \mathbf{W}^\top_{\text{CLIP}}\right)\right\}, \tag{2}$$

where $\mathbf{F}_{\text{cache}}$ is the stored image feature, $\mathbf{L}_p$ is a one-hot pseudo-label obtained from CLIP prediction, and the entropy $H(p)$ is:

$$H(p) = \left(-\sum_{c=1}^{C} p^c_{\text{CLIP}} \log p^c_{\text{CLIP}}\right). \tag{3}$$

For streaming test-time samples arriving on the fly, new items enter the corresponding cache slots based on their pseudo-labels. Once the cache reaches capacity, items with the highest entropy are replaced. At inference, a new test feature $\mathbf{z}$ retrieves relevant cache items, and the final prediction combines CLIP logits with cache-based logits:

$$\text{logits}_{\text{cls}}(\mathbf{z}) = \mathbf{z}\mathbf{W}^\top_{\text{CLIP}} + \mathcal{A}(\mathbf{z}\mathbf{F}^\top_{\text{cache}})\mathbf{L}_p \tag{4}$$

where $\mathcal{A}(x) = \alpha \exp(-\beta(1-x))$ is a modulating function with scaling factor α and sharpness ratio β. Following DPE [49], we compute $\mathbf{F}_{\text{cache}}$ as the mean of all samples within each cache slot, forming class-specific visual prototypes that become more compact and representative as more samples are accumulated.

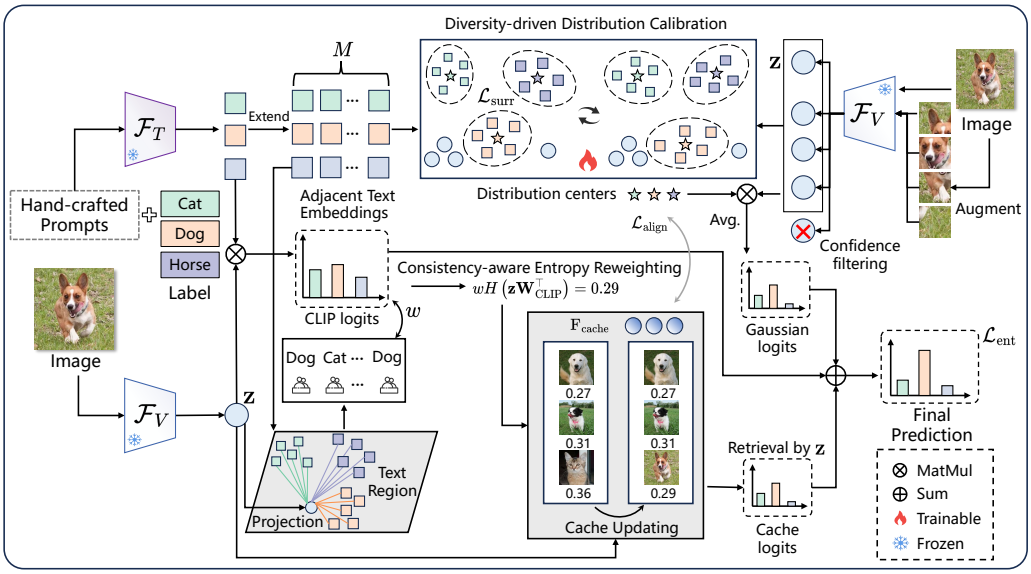


Figure 3: The overall framework of our proposed Reliable Test-time Adaptation (ReTA) approach under visual variations. ReTA leverages multiple text prompts to construct adjacent embeddings and adjust entropy-based prioritization via consistent predictions, thereby preserving more reliable samples in the cache. For subsequent robust decision-making, ReTA models multivariate Gaussian distributions that are dynamically updated using test images and their augmentations, creating flexible boundaries that better accommodate changing visual content during test time.

3.2 Consistency-aware Entropy Reweighting

Previous methods [20, 49, 51] take low-entropy cached samples as visual prototypes for feature retrieval and cache prediction computation. However, relying solely on entropy is unreliable as CLIP is usually overconfident, potentially assigning high confidence to misclassified samples [8, 21]. Since prediction confidence lacks correlation with pseudo-label accuracy [16, 47], low-entropy samples may carry incorrect labels, resulting in cached prototypes deviating from the expected distribution and degrading retrieval performance. While approaches like TDA [20] and DPE [49] have incorporated knowledge-enriched text embeddings with different prompts, they typically average them into a single representation, which fails to fully exploit their semantic diversity. To address this limitation, we propose Consistency-aware Entropy Reweighting (CER), which improves cache quality by prioritizing the storage of relatively reliable samples. Specifically, CER leverages prediction consistency across neighboring semantic representations to identify reliable and representative samples. Inspired by the committee-based strategies [28, 35, 43], we construct a semantic voting committee using multiple text embeddings to assess prediction reliability. While controversial samples expose uncertainty in committee members' judgments, CER prioritizes those with consistent agreement in voting, as they are more representative and reliable for caching.

Adjacent Class-specific Textual Embedding Construction. To measure the predictive consistency of samples, we first construct a set of adjacent class-specific textual embeddings. Following [20, 49, 52], we consider multiple class-specific prompts for each class c to build text representation sets. Formally, we define the prompt set for class c as $\mathcal{P}^c = \{\mathcal{P}_1^c, \mathcal{P}_2^c, \ldots, \mathcal{P}_K^c\}$, where $\mathcal{P}_i^c$ is the i-th prompt with corresponding text embedding $\boldsymbol{t}_i^c = \mathcal{F}_T\left(\mathcal{P}_i^c\right)$ from CLIP's text encoder. Rather than simply averaging prompts, we leverage subtle similarity differences among them to construct semantic embedding sets that form local neighborhoods in text space, enabling assessment of prediction consistency via representational neighborhood consistency [22, 25]. The relative similarity of each prompt to others within the same class is computed as:

$$sim_i^c = \sum_{j=1, j\neq i}^{K} \cos(\boldsymbol{t}_i^c, \boldsymbol{t}_j^c) \tag{5}$$

where sim_i^c represents the cumulative intra-class cosine similarity of the i-th text embedding.

We then sort text embeddings by relative cosine similarity in ascending order, arranging them from semantic outliers to centroids, and construct our final adjacent text embedding set $\{\widehat{\boldsymbol{t}}_1^c, \widehat{\boldsymbol{t}}_2^c, \ldots, \widehat{\boldsymbol{t}}_M^c\}$ with M embeddings per class using a progressive binning approach. Specifically, we define $Q_m = \lfloor \frac{mK}{M} \rfloor$ for $m = 1, 2, ..., M$ to determine embeddings for the m-th adjacent embedding. Each $\widehat{\boldsymbol{t}}_m^c$ is created by averaging the first Q_m embeddings from the sorted sequence, where embeddings are cumulatively included in progressively expanding pools. Our ascending progressive binning strategy creates a smooth semantic transition across embeddings while preserving diversity, thus forming a natural semantic neighborhood for subsequent consistency assessment. Functionally, these adjacent text embeddings operate as a semantic voting committee, where each embedding offers a diverse perspective on class prediction.

Projection for Modality Consistency. To effectively leverage textual neighborhood consistency for identifying reliable image features, we project image features into the text embedding space to mitigate the modality gap [24, 36]. Given the adjacent text embeddings $\tilde{\mathcal{T}} = \{\widehat{\boldsymbol{t}}_m^c : 1 \leq c \leq C, 1 \leq m \leq M\}$, we decompose this

collection via Singular Value Decomposition (SVD) to extract principal semantic variation directions:

$$U\Sigma V^T = \text{SVD}(\tilde{\mathcal{T}}) \tag{6}$$

where U, Σ, and V represent the left singular vectors, the singular values (in descending order), and the right singular vectors, respectively. By retaining the top-n principal components, we collect primary singular vectors $\tilde{V} \in \mathbb{R}^{n\times d}$ to capture the text feature space's intrinsic semantic structure, forming the text subspace with projection matrix:

$$\Phi_{\text{proj}} = \tilde{V}^T\tilde{V} \in \mathbb{R}^{d\times d}. \tag{7}$$

For structural coherence and compact cross-modal representations, we project each image feature into the text subspace as [56]:

$$\mathbf{z}_{\text{proj}} = \Phi_{\text{proj}}\mathbf{z} \tag{8}$$

Prediction Consistency Assessment. We assess prediction consistency by examining classification variations across semantic neighborhoods. For each test data, we create a set of pseudo-labels by computing similarity between the projected image feature and each adjacent text embedding:

$$\hat{\mathcal{Y}} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_M\}, \text{where } \hat{y}_m = \arg\max_{c\in C} (\mathbf{z}_{\text{proj}}^T \widehat{\boldsymbol{t}}_m^c) \tag{9}$$

where $\hat{y}_m$ is the pseudo-label for the m-th adjacent text embedding. We derive a **stability-consistency score** incorporating both prediction stability and consistency with the original prediction. We first determine the most frequent prediction within $\hat{\mathcal{Y}}$ as an indicator of stability:

$$y^* = \arg\max_{c\in C} \sum_{m=1}^{M} \mathbb{I}(\hat{y}_m = c) \tag{10}$$

where y^* represents the majority-voted pseudo-label and $\mathbb{I}(\cdot)$ is the indicator function. Next, we assess consistency between this voted pseudo-label and the original prediction with a consistency factor:

$$\text{Consistency: } \mathcal{R} = \begin{cases} 1, & \text{if } y^* = y \\ \gamma, & \text{if } y^* \neq y \end{cases} \tag{11}$$

where $y = \arg\max_{c\in C}(\mathbf{z}^T\widehat{\boldsymbol{t}}_M^c)$ denotes the original prediction based on $\mathbf{z}$ and the final text prototype, and $\gamma > 1$ is a penalty parameter that increases entropy, thereby reducing the priority of inconsistent predictions. Additionally, we define a stability factor $\mathcal{S}$:

$$\text{Stability: } \mathcal{S} = \frac{M}{n^*}, \quad 1 \le \mathcal{S} \le M \tag{12}$$

where n^* is the count of the most common prediction. A larger $\mathcal{S}$ indicates less stable predictions, with $\mathcal{S} = 1$ denoting perfect agreement. The stability-consistency score is calculated as:

$$w = 1 + \log(\mathcal{R}\mathcal{S}) \tag{13}$$

Entropy Reweighting for Cache Updating. Finally, we compute the reweighted entropy using our stability-consistency score:

$$H'(\mathbf{z}) = w \cdot H(\mathbf{z}) \tag{14}$$

where $H(\mathbf{z})$ is the entropy and $p = \text{Softmax}(\mathbf{z}^\top\widehat{\boldsymbol{t}}_M^c)$. The defined score w amplifies the entropy of samples with unstable or inconsistent predictions, thereby penalizing unreliable samples. This reweighted entropy replaces the standard entropy in the cache priority queue, enabling more reliable cache updates:

$$\left\{\mathbf{F}_{\text{cache}}, \mathbf{L}_p, H'(\mathbf{F}_{\text{cache}}\ \mathbf{W}_{\text{CLIP}}^\top)\right\}. \tag{15}$$

CER prioritizes features with both low entropy and high consistency. When the cache reaches capacity, items with the highest values of reweighted entropy are replaced first, filtering out unreliable samples and preserving higher-quality class prototypes for robust adaptation under visual variations.

3.3 Diversity-driven Distribution Calibration

While CER provides reliable features for cache-based adaptation, it remains challenging to make rapid adjustments to each test-time input, as the cache contains only a few high-confidence samples and may be limited in accommodating visual variations. Existing cache-based methods [20, 51, 52] attempt to compensate for this drawback by enhancing discriminability through sophisticated prompts, but their adaptation capabilities remain limited in dynamic environments due to VLMs' fixed decision boundaries, which struggle to handle distribution shifts. To overcome this challenge, we propose Diversity-driven Distribution Calibration (DDC), which leverages adjacent text embeddings from CER to construct a dynamically adjustable distribution. Drawing from recent success in modeling prompt embeddings with Gaussian distributions [3, 26], we characterize the adjacent text embeddings $\widetilde{\mathcal{T}}$ as an approximate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}^c, \Sigma^c)$, where $\boldsymbol{\mu}^c$ representing the mean embedding of class c and Σ^c denoting the corresponding covariance matrix. Building on the efficient residual learning strategy in DPE [49], we extend this idea to adapt flexible representational distributions, enabling DDC to accommodate evolving visual features.

Textual Gaussian Distribution Evolution via Residuals. For efficient adaptation while preserving well-defined semantic boundaries, we employ residual learning for text embeddings. Specifically, we evolve the class-wise Gaussian distribution by adjusting each adjacent embedding through residual updates. Residual parameters are introduced as:

$$\widehat{\boldsymbol{t}}_m^c = \frac{\widehat{\boldsymbol{t}}_m^c + \boldsymbol{r}_m^c}{||\widehat{\boldsymbol{t}}_m^c + \boldsymbol{r}_m^c||} \tag{16}$$

where $\boldsymbol{r}_m^c$ represents the learnable residual with zero initialization. Unlike DPE, which refines both textual and visual prototypes at the instance level, our DDC performs distribution-level optimization on textual embeddings, offering more adaptive decision boundaries. To continuously adapt to distribution shifts in downstream domains, we maintain a counter l for tracking confident samples and update the textual representation following the progressive strategy in [49]:

$$\widehat{\boldsymbol{t}}_m^c = \frac{(l-1)\widehat{\boldsymbol{t}}_m^c + \widehat{\boldsymbol{t}}_m^{c(*)}}{||(l-1)\widehat{\boldsymbol{t}}_m^c + \widehat{\boldsymbol{t}}_m^{c(*)}||} \tag{17}$$

where $\widehat{\boldsymbol{t}}_m^{c(*)}$ represents the newly optimized embedding from Eq. 16. To ensure reliable adaptation, we perform local residual updating (Eq. 16) only for samples deemed reliable by CER (those with $w = 1$). For global updates (Eq. 17), we apply a fixed threshold τ_c to discard updates from low-confidence samples (for which $H(\mathbf{z}) < \tau_c$), ensuring stability in the evolving global representation.

Objective for Distribution Calibration. To achieve distribution calibration, we jointly optimize textual embeddings through a comprehensive objective that integrates instance-level residual learning and distribution-level supervision:

$$\mathcal{L} = \mathcal{L}_{\text{ent}} + \lambda_1\mathcal{L}_{\text{surr}} + \lambda_2\mathcal{L}_{\text{align}} \tag{18}$$

Table 1: Performance evaluation on cross-dataset generalization. Results report top-1 accuracy (%) across different datasets for both CLIP-RN50 and CLIP-ViT-B/16.

Method	Caltech	DTD	Cars	EuroSAT	Aircraft	Flowers	Pets	UCF101	Food101	SUN397	Average
CLIP-RN50	85.88	40.37	55.70	23.69	15.66	61.75	83.57	58.84	73.97	58.80	55.82
Ensemble	87.26	40.37	55.89	25.79	16.11	62.77	82.97	59.48	74.82	60.85	56.63
CoOp [55]	86.53	37.29	55.32	26.20	15.12	61.55	87.00	59.05	75.59	58.15	56.18
TPT [37]	87.02	40.84	58.46	28.33	17.58	62.69	84.49	60.82	74.88	61.46	57.66
DiffTPT [10]	86.89	40.72	60.71	41.04	17.60	63.53	83.40	62.67	**79.21**	62.72	59.85
TDA [20]	89.70	43.74	57.78	42.11	17.61	68.74	86.18	64.18	77.75	62.53	61.03
BoostAdapter [51]	88.48	43.85	59.67	**44.40**	18.93	68.25	85.75	64.42	78.78	62.83	61.54
DPE [49]	**90.83**	50.18	59.26	41.67	19.80	67.60	85.97	61.98	77.83	64.23	61.93
ReTA	90.35	**52.46**	**61.11**	39.64	**22.62**	**70.12**	**86.90**	**66.18**	77.46	**65.11**	**63.20**
CLIP-ViT-B/16	93.35	44.27	65.48	42.01	23.67	67.44	88.25	65.13	83.65	62.59	63.58
Ensemble	93.55	45.04	66.11	50.42	23.22	66.99	86.92	65.16	82.86	65.63	64.59
CoOp [55]	93.70	41.92	64.51	46.39	18.47	68.71	89.14	66.55	85.30	64.15	63.88
TPT [37]	94.16	47.75	66.87	42.44	24.78	68.98	87.79	68.04	84.67	65.50	65.10
DiffTPT [10]	92.49	47.00	67.01	43.13	25.60	70.10	88.22	62.67	87.23	65.74	65.47
TDA [20]	94.24	47.40	67.28	58.00	23.91	71.42	88.63	70.66	86.14	67.62	67.53
Zero-Ensemble [8]	94.14	45.86	68.48	42.09	24.42	66.82	87.20	68.57	84.58	66.90	64.91
BoostAdapter [51]	94.77	45.69	**69.30**	**61.22**	27.45	71.66	89.51	71.93	87.17	68.09	68.68
DPE [49]	94.81	54.20	67.31	55.79	28.95	75.07	91.14	70.44	86.17	70.07	69.40
ReTA	**95.29**	**57.39**	69.11	58.26	**31.86**	**77.55**	**92.37**	**74.52**	**86.69**	**70.70**	**71.37**

where λ_1, λ_2 are weighting parameters. The entropy $\mathcal{L}_{\text{ent}}$ encourages confident predictions from a test sample and its N augmentations with respect to each text embedding, defined as:

$$\mathcal{L}_{\text{ent}} = H(\tilde{p}_{\text{cls}}(\mathbf{x})) = -\sum_{c=1}^{C} \tilde{p}^c_{\text{cls}} \log(\tilde{p}^c_{\text{cls}})$$
$$\text{with } \tilde{p}^{\text{cls}}_c = \frac{\sum_{i=1}^{N} \mathbb{I}[H(p^c_{\text{cls}}(\mathbf{x}_i)) < \delta] \cdot p^c_{\text{cls}}(\mathbf{x}_i)}{\sum_{i=1}^{N} \mathbb{I}[H(p^c_{\text{cls}}(\mathbf{x}_i)) < \delta]} \tag{19}$$

where p^c_{cls} is the probability derived from Eq. 4 and δ is a entropy threshold. To perform reliable distribution-level calibration, we select pseudo-labels $\tilde{y}$ corresponding to reliable samples with $w = 1$ (Eq.13) for learning, and leverage multiple augmented views to capture variations. We adopt a surrogate loss following ProDA [26]:

$$\mathcal{L}_{\text{surr}} = \mathop{\mathbb{E}}_{(\mathbf{z},\tilde{y})} \left[-\log \frac{\exp\left(\left\langle \mathbf{z}, \widehat{\boldsymbol{t}}^{\tilde{y}}_M \right\rangle / \tau\right)}{\sum_{c=1}^{C} \exp\left(\left\langle \mathbf{z}, \widehat{\boldsymbol{t}}^{c}_M \right\rangle / \tau + \mathbf{z}^T \mathbf{W}_{c,\tilde{y}} \mathbf{z} / 2\tau^2\right)} \right] \tag{20}$$

where covariance term $\mathbf{W}_{c,\tilde{y}} = \Sigma^{cc} + \Sigma^{\tilde{y}\tilde{y}} - \Sigma^{c\tilde{y}} - \Sigma^{\tilde{y}c}$ models class-wise variations. To encourage alignment between textual and visual distributions, we further introduce a cross-modal alignment loss:

$$\mathcal{L}_{\text{align}} = \frac{1}{C} \sum_{c=1}^{C} \left(-\log \frac{\exp((\widehat{\boldsymbol{t}}^c_M)^\top \mathbf{F}^c_{\text{cache}})}{\sum_{j=1}^{C} \exp((\widehat{\boldsymbol{t}}^c_M)^\top \mathbf{F}^j_{\text{cache}})} - \log \frac{\exp((\widehat{\boldsymbol{t}}^c_M)^\top \mathbf{F}^c_{\text{cache}})}{\sum_{j=1}^{C} \exp((\widehat{\boldsymbol{t}}^j_M)^\top \mathbf{F}^c_{\text{cache}})} \right) \tag{21}$$

where $\mathbf{F}^c_{\text{cache}}$ is the average visual prototype of class c from the cache. By jointly optimizing these objectives, our method calibrates class distributions to construct adaptive decision boundaries that robustly accommodate visual diversity.

3.4 Unified Learning Process and Inference

The complete ReTA test-time adaptation process is detailed in the Algorithm 1 in Appendix. For each test sample, we first apply CER to assess prediction consistency and reweight its entropy, determining its cache storage prioritization. Then, for reliable samples, DDC cumulatively updates the residuals of the text embeddings throughout the entire adaptation process.

During inference, we follow ProDA [26] by using the Gaussian mean as a refined decision boundary for classification:

$$p_{\text{gauss}} = \frac{\exp\left(\left\langle \mathbf{z}, \boldsymbol{\mu}^c \right\rangle / \tau\right)}{\sum_{c=1}^{C} \exp\left(\left\langle \mathbf{z}, \boldsymbol{\mu}^c \right\rangle / \tau\right)} \tag{22}$$

where $\boldsymbol{\mu}^c = \frac{1}{M} \sum_{m=1}^{M} \widehat{\boldsymbol{t}}^c_m$ denotes the class-wise mean of the evolved text embeddings, and p_{gauss} represents the refined probability derived from our Gaussian modeling. For final prediction, ReTA integrates the outputs from both cache-based prediction in Eq. 4 and our Gaussian modeling refined logits:

$$p_{\text{final}} = p_{\text{cls}} + \eta p_{\text{gauss}} \tag{23}$$

where $p_{\text{cls}} = \text{softmax}(\text{logits}_{\text{cls}}(\mathbf{z}))$ and η balances the contributions. In summary, by unifying CER and DDC, ReTA effectively preserves high-quality samples and dynamically refines classifier boundaries, enabling more reliable and efficient test-time adaptation.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on two widely used benchmarks, following prior works [1, 20, 37]: Cross-Datasets Generalization and Robustness to Natural Distribution Shifts. The first benchmark comprises 10 diverse image classification datasets: Aircraft [30], Caltech101 [9], Cars [23], Describable Textures (DTD) [5], EuroSAT [12], Flowers102 [31], Food101 [4], Pets [32], SUN397 [44], and UCF101 [38]. The second benchmark consists of the ImageNet [6] and its four variants: ImageNetV2 [34], ImageNet-R [13], ImageNet-Sketch [42], and ImageNet-A [15]. Following the setup of TPT [37], we report the top-1 accuracy on each dataset.

Table 2: Performance evaluation on robustness to natural distribution shifts. Results report top-1 accuracy (%) for both CLIP-RN50 and CLIP-ViT-B/16.

Method	Imagenet	-A	-V2	-R	-Sketch	Avg.	OOD Avg.
CLIP-RN50	58.16	21.83	51.41	56.15	33.37	44.18	40.69
Ensemble	59.81	23.24	52.91	60.72	35.48	46.43	43.09
CoOp [55]	63.33	23.06	55.40	56.60	34.67	46.61	42.43
TPT [37]	60.74	26.67	54.70	59.11	35.09	47.26	43.89
DiffTPT [10]	60.80	31.06	55.80	58.80	37.10	48.71	45.69
TDA [20]	61.35	30.29	55.54	62.58	38.12	49.58	46.63
TPS [39]	61.47	30.48	54.96	62.87	37.14	49.38	46.36
DMN-ZS [52]	63.87	28.57	56.12	61.44	39.84	49.97	46.49
BoostAdapter[1][51]	61.04	**35.52**	56.22	62.87	38.87	50.91	48.37
DPE [49]	63.41	30.15	**56.72**	63.72	40.03	50.81	47.66
ReTA	**63.90**	34.42	56.34	**64.10**	**40.54**	**51.86**	**48.85**
CLIP-ViT-B/16	66.73	47.87	60.86	73.98	46.09	59.11	57.20
Ensemble	68.34	49.89	61.88	77.65	48.24	61.20	59.42
CoOp [55]	71.51	49.71	64.20	75.21	47.99	61.72	59.28
TPT [37]	68.98	54.77	63.45	77.06	47.94	62.44	60.81
DiffTPT [10]	70.30	55.68	65.10	75.00	46.80	62.28	60.52
TDA [20]	69.51	60.11	64.67	80.24	50.54	65.01	63.89
TPS [39]	70.19	60.08	64.73	80.27	49.95	65.04	63.76
Zero-Ensemble [8]	70.93	64.06	65.16	80.75	50.32	66.24	65.07
DMN-ZS [52]	**72.25**	58.28	65.17	78.55	53.20	65.49	63.80
BoostAdapter[1][51]	69.39	64.06	65.19	80.77	51.51	66.19	65.39
DPE [49]	71.91	59.63	65.44	80.40	52.26	65.93	64.43
ReTA	72.18	**64.22**	**65.52**	**81.31**	**53.21**	**67.29**	**66.07**

[1] We reproduce the BoostAdapter result because the original paper does not provide ImageNet results.

Baselines. In our experiments, we compare the performance of the proposed ReTA with several state-of-the-art TTA methods, including: (1) TPT [37]; (2) DiffTPT [10]; (3) TDA [20]; (4) DMN-ZS [52]; (5) TPS [39]; (6) BoostAdapter [51]; (7) Zero [8]; and (8) DPE [49]. Results for the baselines are directly taken from their publications unless otherwise noted.

Implementation details. In our main experiments, we adopt the pre-trained CLIP-ViT-B/16 and CLIP-RN50 as the base models. The batch size is set to 1, with each test image augmented to 63 additional views through AugMix [14], with random resized cropping and random horizontal flipping as in TPT [37], resulting in 64 images per batch. The normalized entropy threshold is fixed at 0.1. We set the cache size to 3 samples per class, the number of adjacent text embeddings to $M = 3$, and retain $n = 64$ singular vectors from the SVD decomposition. The weighting factors λ_1 and λ_2 are set to 0.3 and 0.02 respectively. We conduct all experiments with three different random seeds and report the average result. Unless otherwise specified, our analytical experiments report the average accuracy on Cross-Datasets with CLIP-ViT-B/16. Additional details are provided in the Appendix.

4.2 Evaluation of Cross-Datasets Generalization

As demonstrated in Table 1, we evaluate ReTA's transferability across diverse domains using the Cross-Datasets benchmark. With the CLIP-RN50 backbone, ReTA achieves an average accuracy of 63.20%, surpassing most existing methods and demonstrating competitive performance. When using CLIP-ViT-B/16, ReTA consistently excels across the benchmark, achieving the best performance on 8 out of 10 tasksand attaining a notable 71.37% average accuracy. In particular, ReTA shows substantial improvements on more challenging datasets, such as the texture-rich DTD (57.39%) and fine-grained domains like Aircraft (31.86%) and Flowers (77.55%). These compelling gains highlight ReTA's strong generalization ability in fine-grained and diverse cross-domain scenarios.

Table 3: Ablations on ReTA components. "NDS": Natural Distribution Shifts; "CD": Cross-Datasets.

CER	DDC	NDS	CD
✗	✗	66.04	69.79
✓	✗	66.53	70.41
✗	✓	66.92	70.91
✓	✓	**67.29**	**71.37**

Table 4: Ablations on different learning mechanisms for DDC. TR: Training.

Method	Acc. (%)
w/o TR	70.43
w/ TR, w/o $\mathcal{L}_{\text{ent}}$	70.72
w/ TR, w/ only $\mathcal{L}_{\text{ent}}$	71.01
w/ TR, w/ $\mathcal{L}_{\text{ent}}$	**71.37**

Table 5: Comparison of different text embedding binning strategies.

Binning Method	Acc. (%)
Ascending progressive	**71.37**
Descending progressive	70.47
Uniform	69.41

Table 6: Comparison of projection methods.

Method	Acc. (%)
SVD	**71.37**
PCA	70.57
LDA	70.99

4.3 Evaluation of Robustness to Natural Distribution Shifts

We further evaluate ReTA on in-domain ImageNet and its four out-of-distribution variants to assess its robustness under natural distribution shifts. As shown in Table 2, ReTA consistently outperforms current state-of-the-art methods, including BoostAdapter [51] and Zero-Ensemble [8], which serve as the two strongest baselines. With CLIP-RN50, ReTA achieves an average accuracy of 51.86%, surpassing the best-performing BoostAdapter (50.91%) by 0.95%. When implemented with CLIP-ViT-B/16, ReTA demonstrates even more substantial gains, achieving the best performance across all OOD datasets. Specifically, it outperforms BoostAdapter by 1.10% and Zero-Ensemble by 1.05% on average across all five datasets, establishing a new state-of-the-art. These improvements validate the effectiveness of ReTA in achieving reliable adaptation under complex domain shift scenarios.

4.4 Experimental Analysis

Ablation on Main Components of ReTA. We conduct a comprehensive analysis to examine the individual contributions of Consistency-aware Entropy Reweighting (CER) and Diversity-driven Distribution Calibration (DDC) on both two benchmarks. As shown in Table 3, both components lead to clear performance improvements. Notably, DDC has a greater impact, yielding gains of 0.88% and 1.12% on two benchmarks respectively. This highlights the effectiveness of DDC in refining decision boundaries, thereby enabling more reliable final predictions for test-time data with significant visual variations. Additionally, integrating CER enhances the reliability of cache updating, fostering more accurate cache-based predictions. These two modules are complementary, with their combination yields synergistic improvements across both benchmarks.

Analysis of Adjacent Textual Embeddings in CER. We analyze the adjacent textual embeddings introduced in Sec. 3.2 from two perspectives: (1) the effect of varying the number of adjacent embeddings (M), and (2) the impact of different construction methods. For (1), the number of adjacent textual embeddings (M) determines both the size of the semantic voting committee in CER and the number of class-wise representations used in DDC for Gaussian modeling. As shown in Figure 5b, using a single embedding ($M = 1$)

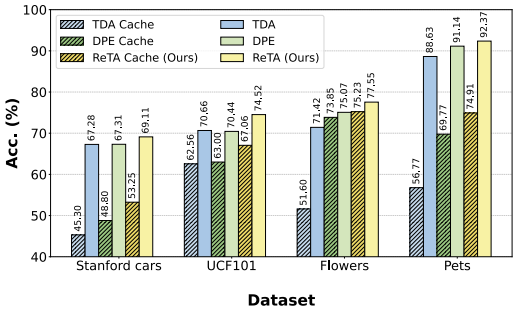


Figure 4: Final comparison of adaptation performance and cache accuracy across four datasets for cache-based methods.

reduces the method to a standard cache-based TTA setting without adjacent structure, which performs worse due to its limited ability to model visual diversity and identify reliable samples. Performance improves substantially as M increases to 3, while larger values introduce redundancy without additional gains. Furthermore, assigning unique residuals to each embedding outperforms the shared-residual variant, as it enables more expressive modeling of intra-class variations. For (2), Table 5 compares different construction methods for adjacent textual embeddings: (i) ascending progressive binning arranges text embeddings from semantic outliers to centroids based on relative cosine similarity; (ii) descending progressive binning follows the opposite order; (iii) uniform binning selects embeddings at equal intervals without considering semantic relationships. Our proposed ascending progressive binning achieves the best performance (71.37%), outperforming descending and uniform binning. We attribute these improvements to the smoother semantic transitions formed by the ascending structure, which facilitates more coherent neighborhood construction and enables more reliable prediction consistency assessment.

Analysis of Projection in CER. We analyze the effect of the number of singular vectors retained in the SVD projection, and compare different projection strategies. Figure 5a shows an inverted U-shaped accuracy curve as the number of singular vectors increases. Retaining 64 components yields the best performance by balancing semantic relevance and noise reduction. At this optimal dimension, Table 6 demonstrates that SVD projection outperforms PCA and LDA by 1.12% and 0.39%, respectively. SVD better preserves the underlying semantic structure without imposing strong class-dependent constraints. In contrast, PCA may retain excessive non-semantic variance, while LDA can be overly restrictive by focusing solely on discriminative class boundaries.

Comparison of Cache Reliability and Final Performance. Figure 4 compares cache accuracy and overall classification performance across four fine-grained datasets. ReTA shows remarkable performance in both metrics, achieving 74.52% on UCF101 (surpassing TDA by 3.86% and DPE by 4.08%) and 77.55% on Flowers (exceeding TDA by 6.13% and DPE by 2.48%). Combined with evidence from Figure 1, these results prove that ReTA achieves superior final accuracy while maintaining high cache reliability.

Effects of Learning Strategy in DDC. Table 4 investigates how different learning mechanisms within DDC affect performance. Without learning, DDC performs similarly to the CER-only baseline, indicating that static distribution modeling provides negligible benefits. Following prior methods [10, 37, 49], we adopt $\mathcal{L}_{\text{ent}}$ as

Table 7: Testing Time and performance gains on robustness to natural distribution shifts.

Method	Testing Time					Acc.	Gain
	ImageNet	-A	-R	-V	-Sk.		
CLIP	8min	1min	5min	2min	9min	59.11	-
TPT [37]	10h	26min	1h50min	2h8min	11h5min	62.44	3.33
TDA [20]	1h34min	10min	52min	16min	1h41min	65.01	5.90
DPE [49]	3h41min	34min	1h58min	44min	4h3min	65.93	6.82
BoostAdapter [51]	3h25min	13min	1h6min	51min	5h27min	66.34	7.23
ReTA	2h38min	12min	1h1min	29min	2h44min	**67.29**	**8.18**

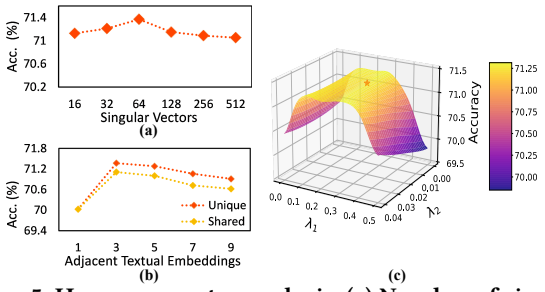


Figure 5: Hyperparameter analysis. (a) Number of singular vectors in SVD. (b) Number of adjacent textual embeddings. (c) Loss sensitivity analysis.

the primary loss. It substantially improves performance by optimizing all textual residuals for more confident predictions. While the auxiliary losses $\mathcal{L}_{\text{surr}}$ and $\mathcal{L}_{\text{align}}$ provide additional gains, properly balancing these components is crucial. As shown in Figure 5c, the best performance is obtained at λ_1=0.3 and λ_2=0.02, where a higher λ_1 is critical for effective distribution-level calibration.

Comparison of Computational Efficiency and Effectiveness. In Table 7, we evaluate the computational efficiency of ReTA and other test-time adaptation methods on ImageNet and its variants. All methods are compared under AugMix augmentation aligned with TPT [37] for fair comparison, although this operation can be time-consuming. ReTA is 1.65x faster than BoostAdapter on large ImageNet and ImageNet-Sketch datasets, despite BoostAdapter being a training-free method like TDA. However, BoostAdapter introduces an additional loop for boosting-based cache updates, incurring significant overhead, especially on larger datasets. Compared to DPE, which also employs residual learning, ReTA is 1.56x faster while achieving better accuracy gains (+8.18 vs. +6.82) due to DPE's high computational cost from visual prototype evolution. By optimizing only textual residuals and eliminating the encoder backpropagation, ReTA achieves better optimization and a favorable balance between computational cost and performance gains.

5 Conclusion

In this paper, we introduce Reliable Test-time Adaptation (ReTA) to address two reliability issues in adapting VLMs: the unreliability of cached samples and inflexible decision boundaries. ReTA comprises two components: Consistency-aware Entropy Reweighting (CER), which improves cache quality via consistency-based sample selection, and Diversity-driven Distribution Calibration (DDC), which models text embeddings as Gaussians to construct adaptive decision boundaries for reliable predictions. Experiments across multiple benchmarks show that ReTA consistently outperforms state-of-the-art methods, especially under challenging distribution shifts, offering a practical and robust solution in dynamic environments.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Nos. 62525103, 62441235, 62271281, 62021002) and Beijing Natural Science Foundation (No. L252009).

References

[1] Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H Khan. 2023. Align Your Prompts: Test-Time Prompting with Distribution Alignment for Zero-Shot Generalization. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 80396–80413.

[2] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. 2021. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818* (2021).

[3] Wentao Bao, Lichang Chen, Heng Huang, and Yu Kong. 2025. Prompting Language-Informed Distribution for Compositional Zero-Shot Learning. In *Computer Vision – ECCV 2024*. Cham, 107–123.

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101 – Mining Discriminative Components with Random Forests. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 446–461.

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing Textures in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.

[7] Zixuan Ding, Ao Wang, Hui Chen, Qiang Zhang, Pengzhang Liu, Yongjun Bao, Weipeng Yan, and Jungong Han. 2023. Exploring Structured Semantic Prior for Multi Label Recognition With Incomplete Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3398–3407.

[8] Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. [n. d.]. Frustratingly Easy Test-Time Adaptation of Vision-Language Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[9] Li Fei-Fei, R. Fergus, and P. Perona. 2004. 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*. 178–178.

[10] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. 2023. Diverse Data Augmentation with Diffusions for Effective Test-time Prompt Tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2704–2714.

[11] Tianxiang Hao, Hui Chen, Yuchen Guo, and Guiguang Ding. 2023. Consolidator: Mergeable Adapter with Grouped Connections for Visual Adaptation. arXiv:2305.00603 [cs.CV] https://arxiv.org/abs/2305.00603

[12] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. 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 (2019), 2217–2226.

[13] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. 2021. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 8340–8349.

[14] Dan Hendrycks*, Norman Mu*, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. 2020. AugMix: A Simple Method to Improve Robustness and Uncertainty under Data Shift. In *International Conference on Learning Representations*.

[15] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021. Natural Adversarial Examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15262–15271.

[16] Tony Huang, Jack Chu, and Fangyun Wei. 2022. Unsupervised Prompt Learning for Vision-Language Models. arXiv:2204.03649 https://arxiv.org/abs/2204.03649

[17] Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. 2024. Retrieval-Enhanced Contrastive Vision-Text Models. In *The Twelfth International Conference on Learning Representations*.

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 4904–4916.

[19] Jiahaoli, Yang Lu, Yuan Xie, and Yanyun Qu. 2024. Relationship Prompt Learning is Enough for Open-Vocabulary Semantic Segmentation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[20] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. 2024. Efficient Test-Time Adaptation of Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14162–14171.

[21] Zaid Khan and Yun Fu. 2024. Consistency and Uncertainty: Identifying Unreliable Responses From Black-Box Vision-Language Models for Selective Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10854–10863.

[22] Zaid Khan and Yun Fu. 2024. Consistency and Uncertainty: Identifying Unreliable Responses From Black-Box Vision-Language Models for Selective Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10854–10863.

[23] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3D Object Representations for Fine-Grained Categorization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*.

[24] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 17612–17625.

[25] Chang Liu, Lichen Wang, and Yun Fu. 2023. Rethinking Neighborhood Consistency Learning on Unsupervised Domain Adaptation. In *Proceedings of the 31st ACM International Conference on Multimedia* (Ottawa ON, Canada) *(MM '23)*. Association for Computing Machinery, New York, NY, USA, 7247–7254.

[26] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. 2022. Prompt Distribution Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5206–5215.

[27] Mengyao Lyu, Tianxiang Hao, Xinhao Xu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. 2025. Learn from the Learnt: Source-Free Active Domain Adaptation via Contrastive Sampling and Visual Persistence. In *Computer Vision – ECCV 2024*. Springer Nature Switzerland, Cham, 228–246.

[28] Mengyao Lyu, Jundong Zhou, Hui Chen, Yijie Huang, Dongdong Yu, Yaqian Li, Yandong Guo, Yuchen Guo, Liuyu Xiang, and Guiguang Ding. 2023. Box-Level Active Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 23766–23775.

[29] XIAOSONG MA, Jie ZHANG, Song Guo, and Wenchao Xu. 2023. SwapPrompt: Test-Time Prompt Adaptation for Vision-Language Models. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 65252–65264.

[30] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-Grained Visual Classification of Aircraft. arXiv:1306.5151 https://arxiv.org/abs/1306.5151

[31] Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated Flower Classification over a Large Number of Classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. 722–729.

[32] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 3498–3505.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763.

[34] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet Classifiers Generalize to ImageNet?. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 5389–5400.

[35] H. S. Seung, M. Opper, and H. Sompolinsky. 1992. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (Pittsburgh, Pennsylvania, USA) *(COLT '92)*. Association for Computing Machinery, New York, NY, USA, 287–294.

[36] Peiyang Shi, Michael C. Welle, Mårten Björkman, and Danica Kragic. 2023. Towards understanding the modality gap in CLIP. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*.

[37] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. 2022. Test-Time Prompt Tuning for Zero-Shot Generalization in Vision-Language Models. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 14274–14289.

[38] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision* 2, 11 (2012), 1–7.

[39] Elaine Sui, Xiaohan Wang, and Serena Yeung-Levy. 2025. Just Shift It: Test-Time Prototype Shifting for Zero-Shot Generalization with Vision-Language Models. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*. 825–835.

[40] Ao Wang, Hui Chen, Zijia Lin, Zixuan Ding, Pengzhang Liu, Yongjun Bao, Weipeng Yan, and Guiguang Ding. 2023. Hierarchical Prompt Learning Using CLIP for Multi-label Classification with Single Positive Labels. In *Proceedings of the 31st ACM International Conference on Multimedia* (Ottawa ON, Canada) *(MM '23)*. Association for Computing Machinery, New York, NY, USA, 5594–5604. https://doi.org/10.1145/3581783.3611988

[41] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. 2025. LSNet: See Large, Focus Small. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9718–9729.

[42] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. 2019. Learning Robust Global Representations by Penalizing Local Predictive Power. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc.

[43] Handing Wang, Yaochu Jin, and John Doherty. 2017. Committee-Based Active Learning for Surrogate-Assisted Particle Swarm Optimization of Expensive Problems. *IEEE Transactions on Cybernetics* 47, 9 (2017), 2664–2677.

[44] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 3485–3492.

[45] Chen-Wei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, and Jingren Zhou. 2023. RA-CLIP: Retrieval Augmented Contrastive Language-Image Pre-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19265–19274.

[46] Yizhe Xiong, Hui Chen, Tianxiang Hao, Zijia Lin, Jungong Han, Yuesong Zhang, Guoxin Wang, Yongjun Bao, and Guiguang Ding. 2025. PYRA: Parallel Yielding Re-activation for Training-Inference Efficient Task Adaptation. In *Computer Vision – ECCV 2024*. Springer Nature Switzerland, Cham, 455–473.

[47] Yizhe Xiong, Hui Chen, Zijia Lin, Sicheng Zhao, and Guiguang Ding. 2023. Confidence-based Visual Dispersal for Few-shot Unsupervised Domain Adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 11621–11631.

[48] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. 2023. Task Residual for Tuning Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10899–10909.

[49] Ce Zhang, Simon Stepputtis, Katia P. Sycara, and Yaqi Xie. 2024. Dual Prototype Evolving for Test-Time Generalization of Vision-Language Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[50] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2022. Tip-Adapter: Training-Free Adaption of CLIP for Few-Shot Classification. In *Computer Vision – ECCV 2022*. Springer Nature Switzerland, Cham, 493–510.

[51] Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. 2024. BoostAdapter: Improving Vision-Language Test-Time Adaptation via Regional Bootstrapping. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[52] Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. 2024. Dual Memory Networks: A Versatile Adaptation Approach for Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 28718–28728.

[53] Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. 2024. Test-Time Adaptation with CLIP Reward for Zero-Shot Generalization in Vision-Language Models. In *The Twelfth International Conference on Learning Representations*.

[54] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional Prompt Learning for Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16816–16825.

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

[56] Xingyu Zhu, Beier Zhu, Yi Tan, Shuo Wang, Yanbin Hao, and Hanwang Zhang. 2024. Selective Vision-Language Subspace Projection for Few-shot CLIP. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24)*. Association for Computing Machinery, New York, NY, USA, 3848–3857.