ROSE-TTA: RELIABLE ONLINE STRUCTURAL ENHANCEMENT FOR TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale vision-language models like CLIP exhibit remarkable zero-shot generalization but suffer significant performance degradation under real-world distribution shifts. Although recent cache-based test-time adaptation (TTA) methods mitigate the issues, they are limited by: (i) unreliability in cache construction, as entropy-based sample selection is insufficient under distribution shifts; and (ii) incomplete cache information at inference, with both imbalanced category information caused by sequential online updates and insufficient sample-specific information for next online instance. To address these limitations, we propose ***ROSE-TTA*** (Reliable Online Structural Enhancement for Test-Time Adaptation), a unified framework that enhances both cache construction and utilization for more reliable and stable adaptation. For construction, we introduce a noise-aware uncertainty measure that combines entropy with perturbation-based prediction stability to robustly select cache entries. To complete the cache information for utilization, we develop a graph-based structural completion strategy, which effectively mitigates class imbalance and completes global information by transferring information between text embeddings and cached features. Additionally, we introduce a sample-specific refinement mechanism to dynamically update cache features and incorporate local information of each online test sample. Experiments on 15 widely used datasets demonstrate the effectiveness of our method.

1 INTRODUCTION

The emergence of large-scale vision-language models (VLMs), such as CLIP (Radford et al., 2021), has profoundly reshaped the landscape of multi-modal learning. By effectively aligning visual and textual modalities through contrastive pretraining on web-scale datasets, these models demonstrate remarkable zero-shot transfer capabilities across diverse downstream tasks. However, deploying CLIP in real world remains challenging, especially on specific data with distribution shifts between the test and pretrained distributions (Shu et al., 2022; Han et al., 2024; Karmanov et al., 2024).

A common approach to address distribution shift at test time is test-time adaptation (TTA) (Sun et al., 2020; Wang et al., 2021; Chen et al., 2022), which has also been leveraged to improve CLIP's zero-shot capability (Shu et al., 2022; Karmanov et al., 2024). Test-time prompt tuning (TPT) (Shu et al., 2022; Feng et al., 2023; Karmanov et al., 2023; Yoon et al., 2024) fine-tunes textual prompts at inference, often guided by entropy minimization. Despite their effectiveness in addressing distribution shifts, these methods are computationally expensive due to backpropagation. To improve efficiency, cache-based TTA methods have emerged (Karmanov et al., 2024; Han et al., 2024). These approaches introduce a dynamic cache to store representative test features online, which are used to refine predictions without backpropagation, providing a lightweight online adaptation mechanism.

Despite their promising efficiency, existing cache-based TTA methods face two fundamental limitations that undermine their robustness under distribution shifts. First, cache construction is often unreliable. Selecting cache samples solely by entropy fails to effectively filter out noisy or misclassified instances (Nguyen et al., 2023a; Han et al., 2024; Shamsi et al., 2024; Zhou et al., 2025), leading to unreliable cache. Second, the available cache information at inference is inherently incomplete. On the global level, sequential online updates induce class imbalance in cache features. This imbalance causes the cache to favor the majority classes and provides insufficient guidance for the minority classes, ultimately biasing model predictions and even leading to error accumulation during contin-

uous adaptation (Zhang et al., 2023). Additionally, cached features lack local information of each incoming sample, providing limited sample-specific guidance for the online instance. These issues hinder both the stability and generalization of cache-based adaptation.

To systematically address these limitations, we propose ***ROSE-TTA*** (***R***eliable ***O***nline ***S***tructural ***E***nhancement for ***T***est-***T***ime ***A***daptation), a novel and unified framework that enhances the reliability and the stability in cache update and integrates complementary global (class level) and local (sample-specific) information for robust cache-based adaptation. In cache construction, we introduce an improved cache update mechanism to synergistically combines entropy with a noise-enhanced uncertainty measure, which evaluates the stability of test features under perturbations. This dual-criterion approach ensures that only the most reliable and consistently stable online instances are incorporated into the cache, improving the cache reliability. Moreover, to complement global semantic information in the cache, we propose a novel graph-based structural completion strategy (Li et al., 2024). The method reconstructs the categorical graph with the class information from text embeddings and cached features, mitigating class imbalance and strengthening the representation of underrepresented categories within the cache (Zhang et al., 2024a). We also design a sample-specific refinement mechanism that updates cached features on-the-fly using the information of each test sample, incorporating local information and improving alignment between the cache and instance.

We evaluate ***ROSE-TTA*** on 15 widely used datasets, covering typical evaluation scenarios such as domain generalization and cross-dataset. The experimental results demonstrate the effectiveness of our method on enhancing the reliability and adaptability of cache-based TTA.

2 Revisiting Cache-Based Test-Time Adaptation For CLIP

2.1 Preliminary

CLIP (Radford et al., 2021). CLIP is well known for the remarkable ability in vision-language representations learning through large-scale training in image-text data. The pretrained CLIP model consists of an image encoder $\mathcal{F}_{\theta_I}(\cdot)$ and a text encoder $\mathcal{F}_{\theta_T}(\cdot)$, with θ_I and θ_T denoting the model parameters, respectively. Based on a zero-shot C-class classification task, for each class $c \in \{1, \ldots, C\}$, we generate a text prompt $\boldsymbol{t}_c$ by instantiating a template such as "*a photo of a [class]*", where "*[class]*" is replaced with the name corresponding to class c. Each text prompt $\boldsymbol{t}_c$ is then encoded as $\boldsymbol{f}_c = \mathcal{F}_{\theta_T}(\boldsymbol{t}_c)$ and the image $\boldsymbol{x}$ is encoded as $\boldsymbol{f}_{\boldsymbol{x}} = \mathcal{F}_{\theta_I}(\boldsymbol{x})$. Collecting all text embeddings as the matrix $\boldsymbol{W}_C = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_C]$, CLIP seeks to associate the image feature $\boldsymbol{f}_{\boldsymbol{x}}$ with the most semantically relevant text feature from $\boldsymbol{W}_C$. The probability of $\boldsymbol{x}$ to be classified as class c is $p(\hat{y} = c \mid \boldsymbol{x}) = \frac{\exp(\cos(\boldsymbol{f}_{\boldsymbol{x}}, \boldsymbol{f}_c)/\tau)}{\sum_{k=1}^{C} \exp(\cos(\boldsymbol{f}_{\boldsymbol{x}}, \boldsymbol{f}_k)/\tau)}$, where $\cos(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature parameter. The most relevant class is obtained from CLIP by $\arg\max_c \; p(\hat{y} = c \mid \mathbf{x})$. For subsequent analysis, we denote the predicted probabilities $p(\hat{y} = c \mid \boldsymbol{x})$ over all C classes as P_{clip}.

Test-time adaption based on key-value cache for CLIP. As a training-free solution that adapts pre-trained models to test data with distributional shift (Tahir et al., 2022; Zhang et al., 2024b; Gao et al., 2025), test-time adaptation adjusts model predictions on-the-fly to better align with the test data. A prominent family of methods leverages a *key–value cache* that accumulates reliable test samples to refine CLIP's predictions (Han et al., 2024; Karmanov et al., 2024).

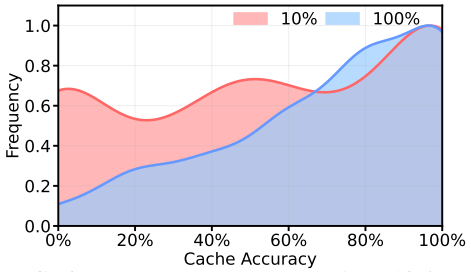
In the cache-based methods, a memory $(\boldsymbol{F}, \hat{\boldsymbol{L}})$ is introduced to store N historical features $\boldsymbol{F} \in \mathbb{R}^{N \times d}$ and their corresponding (pseudo-)labels $\hat{\boldsymbol{L}} \in \mathbb{R}^{N \times C}$. The interaction between a new feature $\boldsymbol{f}_{test}$ and the cache follows a unified paradigm:

$$P_{cache}(\boldsymbol{f}_{test}) = A\big(\boldsymbol{f}_{test}\boldsymbol{F}^T\big)\ \hat{\boldsymbol{L}}, \tag{1}$$

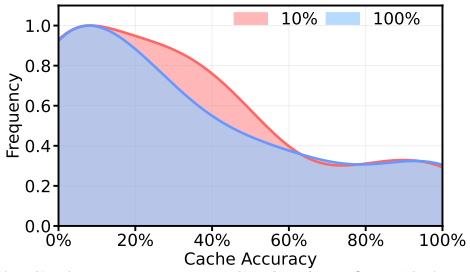
where $\boldsymbol{f}_{test}\boldsymbol{F}^T$ denotes the affinity scores (Karmanov et al., 2024) between the new feature and cached features, and $A(\cdot)$ is an activation function that maps these affinities into weights. Finally, the complete prediction combines the original CLIP output with the cache contribution:

$$P_{final} = P_{clip} + \alpha \cdot P_{cache}. \tag{2}$$

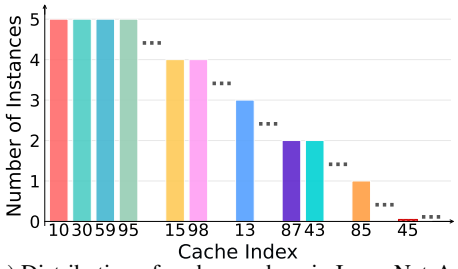
where α is a scaling factor to balance the influence of cache information. This formulation provides a standardized cache-based mechanism for refining the CLIP zero-shot predictions, where the effectiveness depends on how $(\boldsymbol{F}, \hat{\boldsymbol{L}})$ are updated during inference.



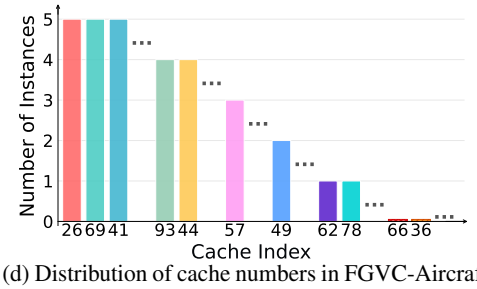
(a) Cache accuracy on ImageNet-A at 10% and 100% test stage.



(b) Cache accuracy on FGVC-Aircraft at 10% and 100% test stage.



(c) Distribution of cache numbers in ImageNet-A at 10% test stage.



(d) Distribution of cache numbers in FGVC-Aircraft at 10% test stage.

Figure 1: **Category and accuracy statistics of the entropy-based online cache.** The online updated cache can be unreliable (a and b) and imbalanced (c and d).

2.2 Issues of Cache-Based Test-Time Adaptation

Although the cache-based methods achieve efficient test-time adaptation for zero-shot classification of CLIP (Wang et al., 2021; Zhang et al., 2024b), the entropy-based cache construction and utilization are not always stable and reliable (Han et al., 2024; Zhou et al., 2025). Samples with low entropy can still be misclassified in unseen test data distributions (Lee et al., 2024). In Figures 1a and 1b, we report the precision of the features stored in the entropy-based online cache (Karmanov et al., 2024) at the beginning and end of the online test stage on two different datasets. We found that even when selecting features with minimum entropy, many of them are cached under incorrect labels, especially for unfamiliar datasets (Figure 1b). The problem is more severe at the earlier test stage (10%). These findings indicate that entropy alone is an insufficient criterion for cache construction and update.

Moreover, since the online test samples often arrive in random order in practice, the online cache update strategy naturally introduces class imbalance, especially at the early stage of the process. We counted the number of per-class features in the online updated cache again for different datasets after processing 10% of the stream. As shown in Figures 1c and 1d, the number of cache features are extremely imbalanced among categories in both cases. The class imbalance biases the predictions toward head classes with larger caches while neglecting classes with few or no entries, yielding skewed and inaccurate predictions. The inaccuracies can even accumulate during online learning, leading to progressively worse overall outcomes.

Additionally, the incoming instance in the online test stream is unpredictable, which can be a new class or a an unseen style. Although the cache is updated online, there remains a lack of local instance information of such specific test samples, leading to unstable predictions.

Therefore, in this paper, we propose an enhanced cache for online test-time adaptation by uncertainty-aware cache construction and graph-based information completion.

3 Methodology

To reduce the issues of unreliable cache construction and biased predictions caused by incomplete information, we propose ***ROSE-TTA***, a more reliable and stable cache for CLIP test-time adaptation, consisting of uncertainty-aware cache construction and graph-based information completion with sample-specific refinement for cache utilization. An illustration of our method is shown in Figure 2.

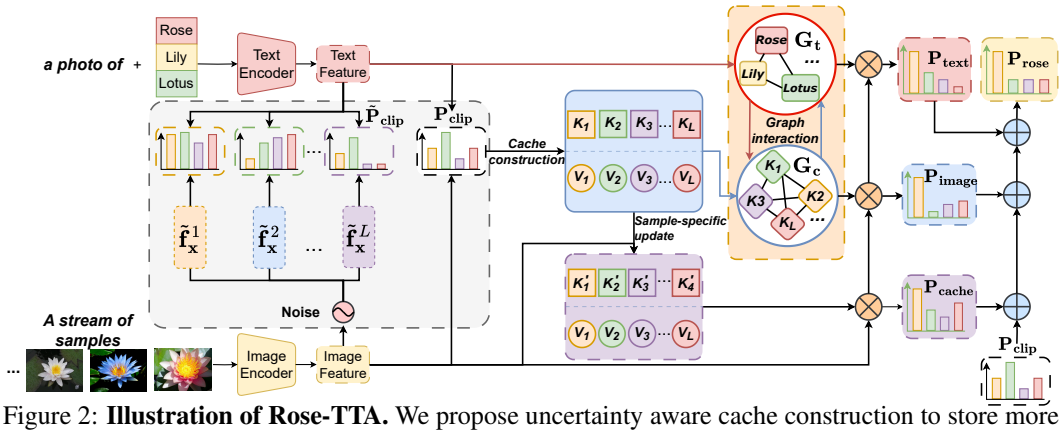


Figure 2: **Illustration of Rose-TTA.** We propose uncertainty aware cache construction to store more reliable test samples in the cache. During utilization, the cache is further enhanced by graph-based structural completion and sample-specific refinement, which inject the global category information and local instance information, respectively.

3.1 Uncertainty-aware cache construction

To directly combat the unreliability of cache construction, where sole reliance on entropy often leads to noisy or misclassified samples(Nguyen et al., 2023b; Han et al., 2024; Lee et al., 2024), we introduce an improved cache update mechanism. This mechanism ensures that only the most reliable and consistently stable features are incorporated into the cache, improving the quality of the cache construction features.

Noise-enhanced uncertainty estimation. To improve the reliability of samples stored in the cache, we propose a noise-enhanced uncertainty estimation mechanism that evaluates the stability of predictions under controlled perturbations. The mechanism is used to select the stable features for cache construction and update during online test-time adaption.

Given an input test sample $\boldsymbol{x}$, we first obtain its CLIP prediction $\hat{c} = \arg\max_c P_{clip}$ to find the corresponding class-wise cache. To access the stability of this sample, we generate n augmented features by adding calibrated Gaussian noise on the original feature:

$$\tilde{\boldsymbol{f}}_{\boldsymbol{x}}^{i} = \boldsymbol{f}_{\boldsymbol{x}} + \epsilon_i \cdot \sigma, \quad \epsilon_i \sim \mathcal{N}(0, \boldsymbol{I}), \quad i = 1, \ldots, n, \tag{3}$$

where σ controls the noise magnitude and n is the number of augmentation. With the predictions on the noise-perturbed features, we obtain the prediction stability for the sample $\boldsymbol{x}$ by:

$$\mathbf{d}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} |p(\hat{y} = \hat{c} \mid \tilde{\boldsymbol{f}}_{\boldsymbol{x}}^{i}) - p(\hat{y} = \hat{c} \mid \boldsymbol{f}_{\boldsymbol{x}})|, \tag{4}$$

where $p(\hat{y} = \hat{c} \mid \tilde{\boldsymbol{f}}_{\boldsymbol{x}}^{i})$ and $p(\hat{y} = \hat{c} \mid \boldsymbol{f}_{\boldsymbol{x}})$ denote the predicted probability on class $\hat{c}$ given the noise-perturbed feature $\tilde{\boldsymbol{f}}_{\boldsymbol{x}}^{i}$ and original feature $\boldsymbol{f}_{\boldsymbol{x}}$, respectively. The prediction stability metric measures how confidence changes under noise perturbations. Smaller $\mathbf{d}(\boldsymbol{x})$ indicates that the instance is more robust to pertubations and therefore more reliable for caching.

Doubly robust cache. To achieve a more reliable cache during online test-time adaptation, we adopt both entropy and noise-enhanced stability as our overall selection metric to construct a doubly robust cache $\mathcal{C} = (\boldsymbol{F}, \hat{\boldsymbol{L}}) = \{\mathcal{C}_c\}_{c=1}^{C} = \{(\boldsymbol{F}_c, \hat{\boldsymbol{L}}_c)\}_{c=1}^{C}$. Following TDA (Karmanov et al., 2024), $\boldsymbol{F}_c, \hat{\boldsymbol{L}}_c$ are the raw features of the historical samples and the one-hot labels of class c, respectively. When the cache $\mathcal{C}_c$ for class c has not reached its maximum capacity, the new sample is directly appended along with its entropy and stability measure. Since the cache size cannot be infinite, when the cache reaches its capacity $n_{\mathcal{C}}$, the new sample is admitted to the cache $\mathcal{C}_c$ only if it surpasses existing cached samples on both metrics. Specifically, the replacement occurs if and only if both conditions are satisfied:

$$\begin{cases} \mathcal{H}(p(y|\boldsymbol{x})) < \mathcal{H}(p(y|\boldsymbol{x}_{j^*})) & \text{(lower entropy)} \\ \mathbf{d}(\boldsymbol{x}) < \mathbf{d}(\boldsymbol{x}_{j^*}) & \text{(higher stability)}, \end{cases} \tag{5}$$

where $\mathcal{H}$ and $\mathbf{d}$ are the entropy and prediction stability of the input sample. j^* denotes the index of the weakest cached sample with the highest combined entropy and the lowest stability score. This

dual-criterion approach ensures that the cache maintains samples that are not only confident(low entropy) but also robust to input perturbations, avoiding overconfidence under distribution shifts and improving the reliability of cached features for test-time adaptation.

3.2 GLOBAL AND LOCAL COMPLETION IN CACHE UTILIZATION

Beyond improving cache reliability during construction, our method also strengthens the utilization of the test-time cache by completing both global and local information.

Graph-based structural completion. As shown in Figure 1, since the test data sequence is random and unpredictable, the online test-time cache exhibit class imbalance and incomplete global task information, especially in the early test stage. This biases the predictions towards majority classes and neglects minority ones. For example, if no sample from class c has appeared, the cache cannot provide sufficient corresponding categorical information for prediction. Consequently, when a new sample from c arrives, the cache guidance tends to be noisy or meaningless, and even degrade the final prediction.

To mitigate class imbalance and complete the global information in online cache at inference, we propose a novel graph-based structural completion strategy. The method reconstructs the categorical graph of the task by integrating class information from both text embeddings and cache features. The graph addresses the class imbalance issue by introducing the global class information from text embeddings of all categories, while preserving the test distributional information by considering the cache features. Specifically, we construct three complementary graphs to capture different aspects of the information:

1) Text-graph $\boldsymbol{G}_t \in \mathbb{R}^{C\times C}$: $\boldsymbol{G}_t = (\boldsymbol{W}_C^T \boldsymbol{W}_C)$ encodes the semantic relationships of different classes at the text level, where $\boldsymbol{W}_C$ denotes the matrix of text embeddings for all classes. By considering all classnames, $\boldsymbol{G}_t$ provides the global category information of the test task, which is leveraged to complete the insufficient class information and mitigate class imbalance.

2) Cache-graph $\boldsymbol{G}_c \in \mathbb{R}^{N\times N}$: $\boldsymbol{G}_c = (\boldsymbol{F}^T \boldsymbol{F})$ models the relationships between cached features, where $\boldsymbol{F}$ containing all cached features. $\boldsymbol{G}_c$ provides the information of the test data in both class-level and instance-level, preserving the test specific information stored in the cache. Moreover, since cached features vary in reliability, we further introduce a reliability-weighted cache graph. First, we assign each cache sample a reliability weight $w_j = \sqrt{(1-\mathbf{d}_j)/\mathcal{H}_j}$, where $\mathcal{H}_j$ and $\mathbf{d}_j$ are the entropy and stability scores of the j-th cached sample. Samples with lower entropy and higher stability have a higher w_j, indicating higher reliability and thus greater influence in the graph. The reliability weight matrix is then constructed as $\boldsymbol{W}_{jk} = \sqrt{w_j \cdot w_k}$ and used to refine the Cache-Graph by $\hat{\boldsymbol{G}}_c = (\boldsymbol{F}^T \boldsymbol{F} \odot \boldsymbol{W})$. Here $\odot$ denotes element-wise multiplication.

3) Gate-graph $\boldsymbol{M} \in \{0,1\}^{C\times N}$ is a binary mapping matrix that links cached samples to their corresponding classes. $\boldsymbol{M}_{cj}$ is set to 1 if the cached sample $\boldsymbol{x}_j$ belongs to class c, else 0.

We reconstruct the categorical graph by integrating the three graphs to incorporate both global class relations from text embeddings and cache-specific relations from cached samples:

$$\hat{\boldsymbol{G}} = \text{softmax}(\boldsymbol{G}_t \boldsymbol{M} \hat{\boldsymbol{G}}_c \boldsymbol{M}^T \boldsymbol{G}_t^T). \tag{6}$$

With the reconstructed graph $\hat{\boldsymbol{G}} \in \mathbb{R}^{C\times C}$, we derive two graph-enhanced prediction terms from the textual and visual paths for a test feature $\boldsymbol{f}_x$:

$$P_{\text{text}} = \boldsymbol{f}_{\boldsymbol{x}} (\hat{\boldsymbol{G}} \boldsymbol{W}_C^T)^T, \qquad P_{\text{image}} = \boldsymbol{f}_{\boldsymbol{x}} (\hat{\boldsymbol{G}} \boldsymbol{M} \boldsymbol{F}^T)^T. \tag{7}$$

This dual pathway expends the textual prediction to P_{text} by injecting test-specific information stored in cache. Moreover, the cache prediction P_{image} is enriched by the global class information from text embeddings, mitigating the class imbalance problem and strengthening the underrepresented categories in the cache.

Sample-specific completion. Except for the global class information, the cache can also lack local instance information since the next online test sample is usually unpredictable in real applications. If a specific test sample is different from the cached ones during online learning, the cache prediction can still be unreliable, limiting its adaptability. To incorporate local instance information in cache utilization, we propose a sample-specific cache refinement mechanism to dynamically update cached features in utilization based onincoming test samples.

Gradient-based TTA (Shu et al., 2022; Zhang et al., 2022a) usually refines the model parameters for each test sample by gradient backpropagation of entropy minimization, which, however, is computationally expensive. To achieve efficient sample-specific cache refinement, we propose a backpropogation-free method to directly infer the pseudo gradient according to the sample feature and its entropy value.

The basic refinement of the cached features for the test sample $\boldsymbol{x}$ is the feature $\boldsymbol{f_x}$. To control the update amount of each cache feature according to the cache prediction, we first obtain the averaged cache prediction $P_n = softmax\left(\frac{1}{n}\sum_{i=1}^{n} P_{cache}(\tilde{\boldsymbol{f}}_{\boldsymbol{x}}^{i})\right)$ based on the noise-augmented features in Eq. (3). We then introduce two intensity control factors $\gamma = 1 - \mathcal{H}(P_n)$ and $\zeta = P_n - 1/C$ based on the averaged cache prediction, where $\mathcal{H}(P_n)$ denotes the normalized entropy of P_n.

Based on the text feature and control factors, we perform sample-specific refinement of the cache features as:

$$\hat{\boldsymbol{F}} = \boldsymbol{F} + \eta \cdot \gamma \cdot \zeta \cdot \boldsymbol{f_x}, \tag{8}$$

where η is a pseudo-learning rate, γ and ζ represent the uncertainty and confidence of the averaged prediction, controlling the update intensity and direction. Following Eq. (1), The refined cache prediction is calculated by:

$$\hat{P}_{cache} = A\left(\boldsymbol{f_x}\hat{\boldsymbol{F}}^{T}\right)\hat{\boldsymbol{L}}, \tag{9}$$

where $A(\cdot)$ is an activation function and $\hat{\mathbf{L}}$ denotes the one-hot pseudo labels of the cache features. This sample-specific refinement enables the cache to consider local instance information for adaptation efficiently, without gradient computation and backpropogation.

Overall, our method integrates multiple complementary sources of information to produce robust test-time predictions. The final prediction is calculated by:

$$P_{rose} = P_{clip} + \alpha \cdot (\hat{P}_{cache} + P_{text} + P_{image}), \tag{10}$$

where α is hyperparameter to control the magnitude of logits.

4 Related Work

Vision-language models. The emergence of large-scale vision-language models has fundamentally transformed the landscape of multi-modal understanding. Early pioneering work CLIP (Radford et al., 2021) demonstrated that contrastive learning on web-scale image-text pairs enables powerful zero-shot transfer capabilities across diverse visual recognition tasks. Building upon this foundation, ALIGN(Jia et al., 2021) demonstrated that scale matters significantly and FILIP(Yao et al., 2022) introduced fine-grained cross-modal alignment through token-level interactions, moving beyond global image-text matching. With ongoing research, works such as InternVL3.5 (Wang et al., 2025) further advance open-source multimodal models in versatility. Concurrently, studies like Qwen-Image (Wu et al., 2025) focus on image generation and DINOv3 (Siméoni et al., 2025) continues to push the boundaries of self-supervised learning.

Test-time adaptation for vision-language models. Test-time adaptation (TTA) (Sun et al., 2020; Nado et al., 2020; Wang et al., 2021) emerged in computer vision to address distribution shift between training and test environments without requiring labeled target data, and has recently been extended to vision-language models with their dual-modal structure and rich semantics. Early works explored *gradient-based prompt tuning*, such as TPT (Shu et al., 2022),C -TPT (Yoon et al., 2024) and DynaPrompt(Xiao et al., 2025), which adapt text prompts via entropy minimization, richer augmentations, or calibration objectives. While effective, these methods incur high computational overhead and may suffer from instability. To improve efficiency, *training-free approaches* were proposed. Cache-based methods (e.g., Tip-Adapter (Zhang et al., 2022b), HisTPT (Tang et al., 2023), TDA (Karmanov et al., 2024)) leverage stored representative samples or multi-granularity knowledge to refine predictions without gradient updates, while PromptAlign (Karmanov et al., 2023) instead mitigates distribution shifts by aligning test and source statistics. Motivated by the limitations of entropy-based objectives, DOTA (Han et al., 2024) and Bayesian TTA (Zhou et al., 2025) provide principled alternatives through distribution estimation and uncertainty quantification respectively. Despite progress, current methods often over-rely on entropy-based selection, suffer from class

Table 1: **Comparisons on the cross-dataset setting.** Our method outperforms the alternatives on five of the ten datasets and achieves best overall performance based on both ResNet-50 and ViT-B/16. The best and runner-up results are bolded and underlined, respectively.

Method	Aircraft	Caltech101	Cars	DTD	EuroSAT	Flower102	Food101	Pets	SUN397	UCF101	*Mean*
CLIP-ResNet-50	16.11	87.26	55.89	40.37	25.79	62.77	74.82	82.97	60.85	59.48	56.63
TPT (Shu et al., 2022)	17.58	87.02	58.46	40.84	28.33	62.69	74.88	84.49	61.46	60.82	57.66
C-TPT (Yoon et al., 2024)	17.50	87.40	57.30	43.10	29.40	65.30	76.00	84.00	62.10	60.70	58.28
HisTPT (Tang et al., 2023)	18.10	87.20	**61.30**	41.30	42.50	67.60	**81.30**	84.90	63.50	64.10	61.18
TDA (Karmanov et al., 2024)	17.61	89.70	57.78	43.74	42.11	**68.74**	77.75	86.18	62.53	64.18	61.03
BCA (Zhou et al., 2025)	**19.89**	89.70	58.13	**48.58**	42.12	66.30	77.19	85.58	63.38	63.51	61.44
Rose-TDA (Ours)	18.46	**89.73**	58.05	45.10	**49.31**	68.17	77.78	**86.51**	**63.41**	**64.23**	**62.07**
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
TPT (Shu et al., 2022)	24.78	94.16	66.87	47.75	42.44	68.98	84.67	87.79	65.50	68.04	65.10
C-TPT (Yoon et al., 2024)	23.90	94.10	66.70	46.80	48.70	69.90	84.50	87.40	66.00	66.70	65.47
MTA (Zanella & Ben Ayed, 2024)	25.20	94.21	68.47	45.90	45.36	68.06	85.00	88.24	66.67	68.69	65.58
PromptAlign (Karmanov et al., 2023)	24.80	94.01	**68.50**	47.24	47.86	72.39	86.65	90.76	67.54	69.47	66.92
HisTPT (Tang et al., 2023)	26.90	94.50	69.20	48.90	49.70	71.20	**89.30**	89.10	67.20	70.10	67.61
TDA (Karmanov et al., 2024)	23.91	94.24	67.28	47.40	58.00	71.42	86.14	88.63	67.62	70.66	67.53
BCA (Zhou et al., 2025)	**28.59**	94.69	66.86	**53.49**	56.63	73.12	85.97	90.43	**68.41**	67.59	68.58
Rose-TDA (Ours)	25.86	**94.76**	67.06	48.00	**65.64**	**74.17**	86.20	**90.95**	68.00	**71.34**	**69.20**

imbalance where underrepresented classes receive insufficient cache guidance, and employ static feature representations that fail to adapt to evolving test distributions or fully utilize the semantic structure in text embeddings.

5 Experiments

Datasets. We evaluate on two commonly used benchmarks in TTA on zero-shot CLIP (Shu et al., 2022; Karmanov et al., 2024; Han et al., 2024). The *cross dataset* benchmark covers ten heterogeneous datasets, including Aircraft (Maji et al., 2013), Cars (Krause et al., 2013), Pets (Parkhi et al., 2012), Flower102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), Caltech101 (Fei-Fei, 2004), SUN397 (Xiao et al., 2010), DTD (Cimpoi et al., 2014)), EuroSAT (Helber et al., 2019), and UCF101 (Soomro et al., 2012). The diverse datasets allows us to assess adaptability across distinct semantic and visual domains. The *out-of-distribution(OOD)* benchmark consists of ImageNet(Deng et al., 2009) and its four variants: ImageNet-A (Hendrycks et al., 2021b), ImageNet-V2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021a), and ImageNet-Sketch (Wang et al., 2019), which provide a rigorous test of robustness under different types of distribution shift.

Implementation details. Our experiments are conducted on the pre-trained CLIP model (Radford et al., 2021), which comprises an image encoder and a text encoder. Following (Karmanov et al., 2024), we adopt ResNet-50 (He et al., 2016) and ViT-B/16 (Dosovitskiy et al., 2020) as the image encoder backbones, and a Transformer (Vaswani et al., 2017) as the text encoder. To reflect real-world online test-time adaptation scenarios, we set the test batch size to 1. For hyperparameters, we fix n / σ to 10 / 0.5 across all datasets. The cache capacity n_C is set to 5 for all dataset–backbone combinations. The coefficient α is tuned individually for each dataset to adapt to the specific scenario. We adopt the same hand-crafted prompt templates as in Karmanov et al. (2024). Top-1 accuracy is used as the evaluation metric. All experiments are performed on a single NVIDIA RTX 4090 GPU.

5.1 Comparisons

Baselines. In this section, we compare our method mainly with two kinds of methods: (1) test-time propmt tuning methods: TPT (Shu et al., 2022), C-TPT (Yoon et al., 2024), HisTPT (Tang et al., 2023), MTA (Zanella & Ben Ayed, 2024), and PromptAlign (Karmanov et al., 2023). (2) cache-based efficient test-time adaptation methods: TDA (Karmanov et al., 2024) and BCA (Zhou et al., 2025).

Results on the cross-dataset benchmark. We first compare ROSE-TTA with state-of-the-art methods on the *cross-dataset* benchmark (Table 1). For ResNet-50, ROSE-TTA achieves the best overall performance across the ten datasets compared with other state-of-the-art methods. Compared with gradient-based prompt tuning methods (Shu et al., 2022; Tang et al., 2023; Yoon et al., 2024), our method performs better on eight of the ten datasets. Moreover, since ROSE-TTA avoids gradient computation and backpropagation, it is also more efficient than the prompt tuning methods. Our method also outperforms recent cache-based efficient adaption methods (Karmanov et al., 2024; Zhou et al., 2025) on six of the ten datasets. On ViT-B/16, our method again delivers the best overall

Table 2: **Comparisons on the out-of-distribution benchmark.** ROSE-TTA achieves best overall performance and surpasses other alternatives on three of the five datasets. The best and runner-up results are bolded and underlined, respectively.

Method	ImageNet	ImageNet-V2	ImageNet-S	ImageNet-A	ImageNet-R	*Mean*
CLIP-ResNet-50 (Radford et al., 2021)	59.81	52.91	35.48	23.24	60.72	46.43
TPT (Shu et al., 2022)	60.74	54.70	35.09	26.67	59.11	47.62
C-TPT (Yoon et al., 2024)	61.20	54.80	35.70	25.60	59.70	47.40
TDA (Karmanov et al., 2024)	61.35	55.54	**38.12**	30.29	62.58	49.57
BCA (Zhou et al., 2025)	<u>61.81</u>	**56.58**	<u>38.08</u>	<u>30.35</u>	<u>62.89</u>	<u>49.94</u>
Rose-TTA(Ours)	**62.00**	<u>56.45</u>	37.96	**30.93**	**63.01**	**50.07**
CLIP-ViT-B/16	68.34	61.88	48.24	49.89	77.65	61.20
TPT (Shu et al., 2022)	68.98	63.45	47.94	54.77	77.06	62.44
C-TPT (Yoon et al., 2024)	69.30	63.40	48.50	52.90	78.00	62.42
MTA (Zanella & Ben Ayed, 2024)	70.08	64.24	49.61	58.06	78.33	64.06
PromptAlign (Karmanov et al., 2023)	-	65.29	50.23	59.37	79.33	63.56
TDA (Karmanov et al., 2024)	69.51	64.67	50.54	60.11	80.24	65.01
BCA (Zhou et al., 2025)	<u>70.22</u>	<u>64.90</u>	**50.87**	<u>61.14</u>	<u>80.72</u>	<u>65.57</u>
Rose-TDA(Ours)	**70.54**	**65.83**	<u>50.82</u>	**61.22**	**81.81**	**66.04**

results, leading on five of the ten datasets. The results demonstrate that our method effectively adapts across diverse datasets.

Results on the OOD benchmark. We also evaluate ROSE-TTA on the OOD benchmark (Table 2), with conclusions consistent with the cross-dataset results. Our method surpasses both prompt tuning approaches and recent cache-based approaches in the overall accuracy, achieving the best performance on ImageNet, ImageNet-A, ImageNet-V2, and ImageNet-R. The findings demonstrate the effectiveness of ROSE-TTA for adaptation across different data distributions.

5.2 Ablation Study

Component analysis. To validate the effectiveness of each proposed component, we perform ablation studies by systematically removing individual modules from ROSE-TTA. The experiments are conducted on four datasets in the cross-dataset benchmark based on ViT/B-16. As shown in Table 3, without the noise-based prediction stability $\mathbf{d}(\boldsymbol{x})$ for cache construction, the cache can only select samples according to the entropy value, which can be unreliable and lead to performance degradation. When removing P_{text} and P_{image} from the final prediction, there is no class information completion for the cache utilization, resulting performance degradation. Similarly, without the sample-specific refined cache features $\hat{\boldsymbol{F}}$, relying only on static cache prevents adaptation to individual test samples, highlighting the importance of our dynamic pseudo-gradient updates.

Table 3: **Ablations on components for computing cache predictions.** Removing $\mathbf{d}$ in Eq. (4), P_{text} and P_{image} in Eq. (10), or $\hat{\boldsymbol{F}}$ in Eq. (8) leads to performance degradation.

Method	Caltech101	DTD	Cars	Aircraft	*Mean*
w/o $\mathbf{d}$	93.61	47.10	65.97	24.92	57.90
w/o P_{image} & P_{text}	93.91	47.87	66.62	25.83	58.56
w/o $\hat{\boldsymbol{F}}$	94.08	47.52	66.57	25.47	58.41
Rose-TTA	**94.76**	**48.00**	**67.06**	**25.86**	**58.92**

Effect of Cache Capacity $n_{\mathcal{C}}$. To assess the impact of cache capacity, i.e., the number of historical key-value pairs per class, we conduct experiments by varying it from 1 to 20 on four datasets based on ViT/B-16. As shown in Figure 3, ROSE-TTA performs best usually with a cache size 5, while it maintaining relatively stable accuracy on most datasets. This stability suggests that our structural completion and replacement strategies effectively regulate cache content. However, some datasets like Flower102 exhibit more pronounced fluctuations, especially with larger capacities. The reason can be its fine-grained class structure, where categories share highly similar visual patterns. In such cases, an overly large cache risks accumulating redundant or noisy samples that blur decision boundaries. Moreover, under online adaptation, large capacities can also amplify the effect of early, less reliable entries, introducing additional instability.

Impact of Noise Augmentation Number n. We also ablate the number of noise augmentations used for stability estimation in cache construction. The experiments are also conducted on four datasets based on ViT/B-16. As shown in Figure 4, performance rises slightly with more augmentations, with

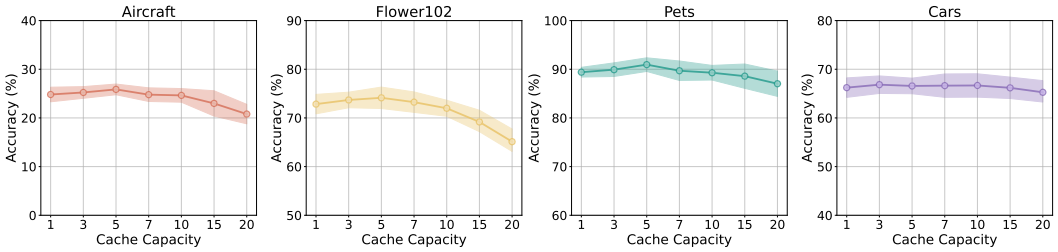


Figure 3: **Analysis on cache capacity $n_{\mathcal{C}}$.** Our method performs best with $n_{\mathcal{C}}$ as 5 for most datasets.

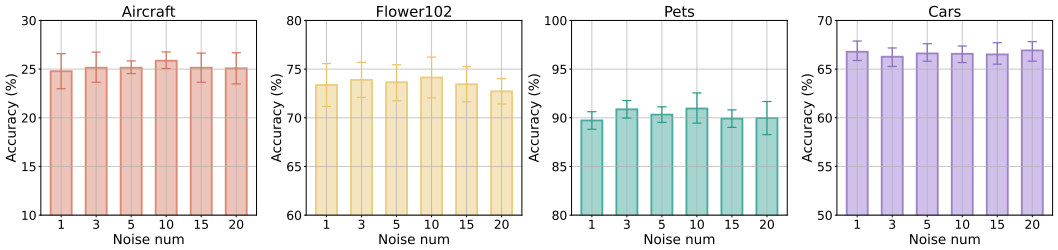


Figure 4: **Analysis on the number of noise-augmented features n.** Our performance is relatively stable across n, with the best results around 10.

a peak around 10. This indicates that the uncertainty-aware mechanism in ROSE-TTA is robust, yeilding reliable stability estimation with even small numbers of augmentations (e.g., 3). While increasing augmentations slightly improves early stability, excessive augmentation tends to yield diminishing returns and may introduce redundant signals, thereby hindering the model's ability to capture genuine data characteristics.

Progressive adaptation analysis. To further understand the dynamic behavior of ROSE-TTA during online adaptation, we evaluate the method at multiple test-time checkpoints on ImageNet and EuroSAT. The results are provided in Figure 5. Compared with TDA, our method consistently achieves higher accuracy throughout the entire process. Notably, advantage is more pronounced at the early stage (e.g., at 10%–30% online data), where TDA suffers from unstable performance while ROSE-TTA maintains good improvement. The results indicate that our information completion and uncertainty-aware cache construction provide reliable guidance from the beginning of adaptation, caching higher-quality samples and alleviating class imbalance. As adaptation progresses, the performance gap narrows but remains, suggesting that ROSE-TTA not only ensures robustness in the initial phase, but also preserves long-term effectiveness throughout the evaluation.

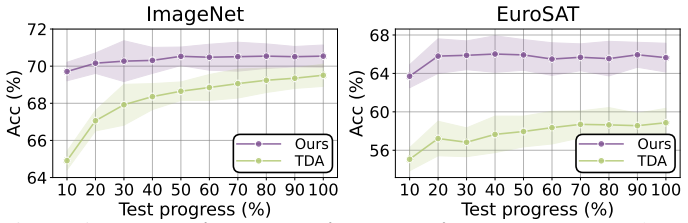


Figure 5: **Progressive adaptation analysis** on ImageNet and EuroSAT. Our method consistently outperforms TDA while more pronounced at the early stages, indicating that the method caches higher-quality samples and mitigates class imbalance by information completion.

6 Conclusion

In this work, we propose ROSE-TTA, a reliable, test-time–enhanced caching framework that improves CLIP's zero-shot performance across datasets and distribution shifts. ROSE-TTA construct a more reliable cache by selecting more stable test features via a noise-aware uncertainty strategy that integrates entropy minimization with noise-based stability. During cache utilization, we enhance its global class information through a graph-based structural completion strategy that mitigates class imbalance and strengthens cache representations. The method also injects local instance information with a sample-specific refinement module to adapt cached features to each incoming test sample. Extensive experiments on 15 datasets demonstrate the effectiveness of ROSE-TTA, providing a robust and efficient approach to test-time cache construction and utilization in practice.

USE OF LARGE LANGUAGE MODELS (LLMS).

We use LLMs (e.g., ChatGPT) only for minor language polishing. They did not contribute to research ideation, experimental design, or substantive writing.

ETHICS STATEMENT

Our contribution is primarily about the online cache for test-time adaptation in vision-language models. The method does not introduce ethical risks.

REPRODUCIBILITY STATEMENT

We provide the sufficient datasets and implementation details in Section 5 for reproducibility.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 295–305, 2022.

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

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Li Fei-Fei. Learning generative visual models from few training examples. In *Workshop on Generative-Model Based Vision, IEEE Proc. CVPR, 2004*, 2004.

Zhi-Fan Feng, Jie Zhou, Wen-Huang Li, and Zhen-Hua Zhang. Difftpt: Leveraging diffusion models for test-time prompt tuning. In *arXiv preprint arXiv:2305.13998*, 2023.

Huan-ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, et al. A survey of self-evolving agents: On path to artificial super intelligence. *arXiv preprint arXiv:2507.21046*, 2025.

Zongbo Han, Jialong Yang, Junfan Li, Qinghua Hu, Qianli Xu, Mike Zheng Shou, and Changqing Zhang. Dota: Distributional test-time adaptation of vision-language models. *arXiv preprint arXiv:2409.19375*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021a.

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*, pp. 15262–15271, 2021b.

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

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

Sarel Karmanov, Leonid Karlinsky, Shir Doveh, Assaf Arbelle, Rogerio Feris, Raja Giryes, and Alex Bronstein. Promptalign: Bridging the gap between model and human preferences via natural language feedback. *arXiv preprint arXiv:2312.01459*, 2023.

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

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *International Conference on Learning Representations*, 2024.

Yinjun Li, Han Li, Bo Li, Jialin Li, and Xiaofeng Zhu. Gita: Graph to visual and textual integration for vision-language graph reasoning. *arXiv preprint arXiv:2407.20080*, 2024.

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

Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. In *arXiv preprint arXiv:2006.16971*, 2020.

A Tuan Nguyen, Thanh Nguyen-Tang, Ser-Nam Lim, and Philip HS Torr. Tipi: Test time adaptation with transformation invariance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24162–24171, 2023a.

A. Tuan Nguyen, Thanh Nguyen-Tang, Ser-Nam Lim, and Philip H.S. Torr. Tipi: Test time adaptation with transformation invariance. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.

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

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

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

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

Afshar Shamsi, Rejisa Becirovic, Ahmadreza Argha, Ehsan Abbasnejad, Hamid Alinejad-Rokny, and Arash Mohammadi. Etage: Enhanced test time adaptation with integrated entropy and gradient norms for robust model performance. *arXiv preprint arXiv:2409.09251*, 2024.

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

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

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*, pp. 9229–9248. PMLR, 2020.

Anique Tahir, Lu Cheng, Ruocheng Guo, and Huan Liu. Distributional shift adaptation using domain-specific features. In *2022 IEEE International Conference on Big Data (Big Data)*, 2022.

Jingyi Tang, Jiaxing Wang, Jingdong Yang, Jingyao Liu, and Hongdong Li. Histpt: Historical test-time prompt tuning for vision foundation models. In *arXiv preprint arXiv:2312.09051*, 2023.

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.

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.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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

Zehao Xiao, Shilin Yan, Jack Hong, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiayi Shen, Qi Wang, and Cees GM Snoek. Dynaprompt: Dynamic test-time prompt tuning. *arXiv preprint arXiv:2501.16404*, 2025.

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. In *International Conference on Learning Representations*, 2022.

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

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 (CVPR)*, pp. 23783–23793, June 2024.

Haohan Zhang, Ximeng Liu, Peilin Liu, Chuan Lu, Xin Wang, Yan Feng, and Xiangyu Zhang. Delta: degradation-free fully test-time adaptation. *arXiv preprint arXiv:2301.13018*, 2023. URL https://arxiv.org/abs/2301.13018.

Jialong Zhang, Yu Zhou, Wenjie Sun, Yanchun Sun, and Hanjing Yu. Graph-based class-imbalance learning with label enhancement. *IEEE Transactions on Neural Networks and Learning Systems*, 35 (3):3634–3648, 2024a. URL https://ieeexplore.ieee.org/document/9656689/.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38629–38642, 2022a.

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*, 2022b.

Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. Boostadapter: Improving vision-language test-time adaptation via regional bootstrapping. *Advances in Neural Information Processing Systems*, 2024b.

Lihua Zhou, Mao Ye, Shuaifeng Li, Nianxin Li, Xiatian Zhu, Lei Deng, Hongbin Liu, and Zhen Lei. Bayesian test-time adaptation for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29999–30009, 2025.