DANCE: Diversity-attended Dynamic Caching with Asymmetric Quantization for Test-Time Adaptation of Vision-Language Models

Shunge Zou[1], **Changhu Wang**[2], **Wei Ju**[3,*], **Ziyue Qiao**[1,*], **Xiao Luo**[4]
[1]Great Bay University [2]Fred Hutchinson Cancer Center
[3]Peking University [4]University of Wisconsin–Madison
shungezou@gmail.com, cwang3@fredhutch.org, juwei@pku.edu.cn,
zyqiao@gbu.edu.cn, xiao.luo@wisc.edu

Abstract

This paper studies the problem of test-time adaptation for vision-language models (VLMs). Recent approaches typically measure the prediction entropy to store a confident cache for logit refinement. However, these confident samples tend to approach prototypes with limited coverage of data distribution, which could result in biased predictions as the distribution evolves. Towards this end, we propose a novel approach named Diversity-attended Dynamic Caching with Asymmetric Quantization (DANCE) for test-time adaptation of VLMs. The core of our DANCE is to maintain a dynamic cache to store diversity-aware test samples, which support efficient logit adjustment via asymmetric quantization. In particular, we first generate multiple augmented views of each sample and aggregate their outputs from pretrained VLMs via a consistency-aware mechanism. More importantly, we construct a dynamic cache, which stores the most reliable and diverse samples to cover evolving test distributions. To measure the diversity efficiently, we quantize cached samples and compute the asymmetric similarity across query samples and memory samples, which guide the cache updating via replacing samples with the lowest scores iteratively. Finally, we leverage the asymmetric similarity between the quantized prototype representations from the dynamic cache to update logits under distribution shifts. Extensive experiments on various benchmark datasets validate the clear superiority of the proposed DANCE in a wide range of settings.

1 Introduction

Vision–language models (VLMs) (Radford et al., 2021; Jia et al., 2021) have become a central paradigm for jointly modeling visual perception and natural language understanding (Yu et al., 2022; Zhai et al., 2022). As a representative instance, CLIP (Radford et al., 2021) encodes images and texts into a shared feature space and aligns their embeddings via a contrastive objective. Trained on large-scale image-text pairs, these VLMs learn transferable representations, enabling zero-shot inference across various downstream tasks under diverse conditions (Li et al., 2024; Joshi et al., 2024).

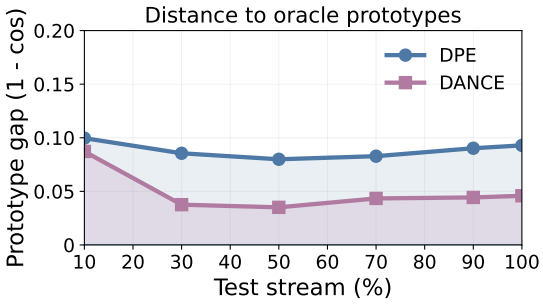

Figure 1: Prototype drift under streaming shift on DTD. We track the mean gap between each class's cache prototype and the oracle prototype (defined as the per-class mean feature) over time. Compared with the entropy-only caching method (DPE), DANCE yields smaller gaps, suggesting consistently better shift tracking over time.

However, distribution shifts between pretraining and target data can substantially degrade the zero-shot performance of VLMs in real deployments (Li et al., 2022; Liang et al., 2025a; Xiong et al., 2023). Most existing methods (Zhou et al., 2022b; Zhang et al., 2022) rely on labeled samples from the target domain to fine-tune the model or auxiliary adapters, but such approaches are impractical in deployment due to annotation and retraining costs. This motivates the use of test-time adaptation (TTA), which adapts the model on the fly using only unlabeled test data to better handle distribution shifts.

Early TTA methods focus on prompt-based strategies (Feng et al., 2023; Shu et al., 2022) that adapt the textual prompts at test time by minimizing the prediction entropy. However, they require repeated backpropagation through the en-

The official implementation of DANCE is available here.
*Corresponding Authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 37329–37343
July 2-7, 2026

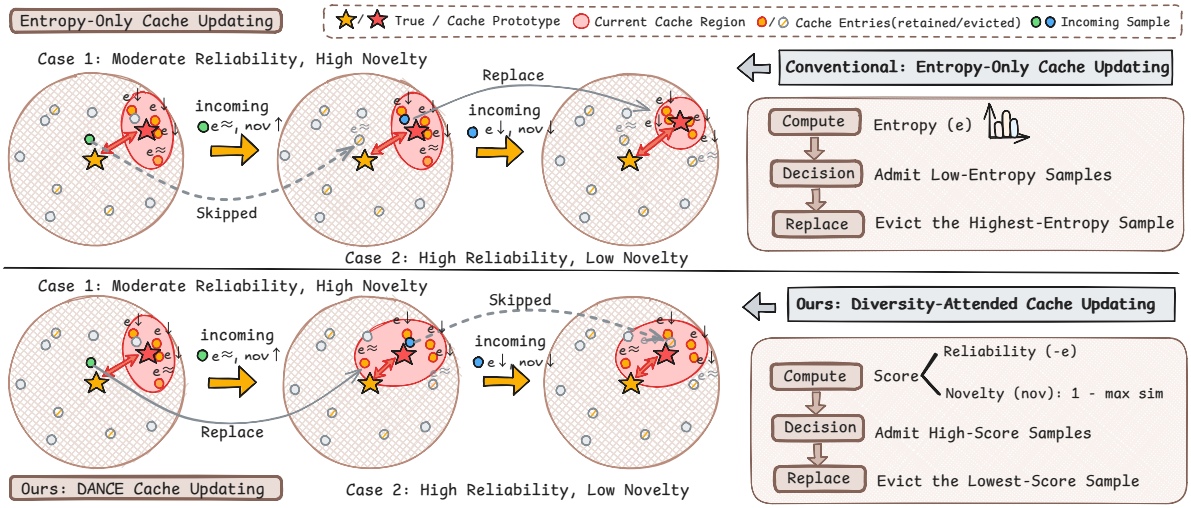

Figure 2: Mechanism schematic of cache updating. Compared with entropy-only admission, DANCE integrates an entropy-based reliability signal with diversity-oriented exploration in its cache admission and replacement rules. The upper and lower panels illustrate how the two strategies differ under samples with varying reliability and novelty.

coder, which incurs substantial computational overhead and high latency, limiting their practicality in real-time settings (Huang et al., 2025). To improve efficiency, cache-based TTA strategies (Karmanov et al., 2024; Zhang et al., 2024b,a) have gained substantial traction, shifting adaptation from per-sample backpropagation to retrieval from accumulated test-time evidence. They maintain a cache from sequential test samples, typically updated via a low-entropy filtering strategy, and query this cache at inference time to refine predictions.

However, we identify two key limitations of these TTA methods as follows. (1) Updating caches via entropy-only filtering tends to over-admit a few easy, high-confidence samples, resulting in near-duplicate entries and reduced coverage of the evolving target distribution. As illustrated in Fig. 1, this may cause visual representations to deviate from the evolving target data manifold, degrading cache quality and ultimately distorting predictions. (2) Cache retrieval based on cosine/inner-product similarity in visual feature space is often unstable. Since CLIP is trained to align visual and textual embeddings rather than to optimize a visual–visual similarity metric (Liang et al., 2022), empirical analyses (Fan et al., 2025) further show that matching and non-matching image–image pairs have heavily overlapping cosine similarity distributions. Such poor separation makes neighbor retrieval sensitive to small perturbations under shift, amplifying pseudo-label noise in cache-based corrections.

To address these challenges, we propose DANCE (Diversity-attended Dynamic Caching with Asymmetric Quantization), a novel test-time adaptation framework that jointly redesigns how caches are maintained and how they are queried under distribution shift. On the maintenance side, DANCE performs diversity-attended dynamic caching that selectively admits reliable and complementary test instances, thereby yielding representative class-wise visual prototypes of the evolving test-time distribution, as illustrated in Fig. 2. On the retrieval side, DANCE adopts an asymmetric quantization scheme, inspired by large-scale visual feature retrieval, that quantizes cached features to suppress representation noise while keeping query embeddings real-valued to preserve expressive power. This asymmetric similarity is used consistently for cache update and logits refinement, enabling more efficient and robust use of the cache under distribution shift. The core contributions of this work are as follows:

❶ *New Perspective.* We treat test time adaptation as a dynamic distribution tracking process that incorporates distributional diversity alongside entropy-based filtering to mitigate representation narrowing in streaming scenarios.

❷ *Novel Methodology.* We propose DANCE, a framework integrating stable view selection for sample denoising, diversity-attended cache updates for distribution coverage, and asymmetric quantization for robust visual retrieval.

❸ *Empirical Analysis.* We conduct extensive experiments on multiple benchmark datasets, showing that DANCE consistently improves VLMs' robustness to distribution shift and outperforms representative TTA methods.

2 Related Work

Zero-shot Inference of VLMs. Although VLMs like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) exhibit impressive zero-shot capabilities across a wide range of recognition tasks (Yu et al., 2022), their performance often degrades noticeably on target domains with significant distribution shifts (Li et al., 2022; Xiong et al., 2023). To address this issue, a common way is to adapt the model using a small amount of labeled data from the downstream task. Prompt-learning methods such as CoOp (Zhou et al., 2022b) and CoCoOp (Zhou et al., 2022a) learn continuous prompts or meta-prompts from labeled examples to better align the text branch with the target distribution. In parallel, adapter-based approaches (Zhang et al., 2022) like Tip-Adapter and TaskRes (Yu et al., 2023) construct feature memories from a few labeled instances and then refine predictions. Although these methods are effective, they depend on labeled target data and costly offline retraining, which may be unrealistic in many practical deployment scenarios and therefore motivate more data-efficient adaptation strategies (Luo et al., 2025b).

Test-time Adaptation. Test-time adaptation adjusts the model on the fly at inference time based solely on unlabeled test data, allowing VLMs to better cope with distribution shifts (Farina et al., 2024; Nguyen et al., 2025). Early TTA for VLMs focused on prompt optimization at test time. For instance, TPT (Shu et al., 2022) adapts soft prompts via entropy minimization over augmented views, while DiffTPT (Feng et al., 2023) extends this by incorporating more diverse, diffusion-generated augmentations. Despite their simplicity, these approaches require per-sample backpropagation and frequent model resets, leading to substantial computational and latency costs. To improve efficiency, recent work has increasingly turned to cache-based methods that exploit memories constructed from test data. Representative methods (Karmanov et al., 2024; Zhang et al., 2024a) maintain caches of test features and pseudo-labels, querying them during inference to refine model outputs. A more recent cache-based method, DPE (Zhang et al., 2024a), treats per-class cached features as visual prototypes and learns residual updates to the textual prototypes to edit decision boundaries. Despite promising efficiency, these methods still struggle to maintain representative memories under evolving streams and ensure reliable visual-space retrieval.

3 The Proposed DANCE

3.1 Problem Definition

Consider a Vision-Language Model (e.g., CLIP) consisting of an image encoder E_{img} and a text encoder E_{txt} that map inputs into a shared d-dimensional space. For a downstream task with classes $\mathcal{C} = \{1, \ldots, C\}$, each class c is represented by a text embedding $f_c^{\text{txt}} = E_{\text{txt}}(T_c)$. Given a test image x, the image encoder produces a visual feature $f = E_{\text{img}}(x)$. The zero-shot probability distribution p_0 is computed as follows:

$$p_0(y = c \mid x) = \frac{\exp(s(f, f_c^{\text{txt}})/t)}{\sum_{c'=1}^{C} \exp(s(f, f_{c'}^{\text{txt}})/t)}, \quad (1)$$

where $s(\cdot, \cdot)$ denotes cosine similarity and t represents the temperature parameter. The initial prediction is $\hat{y}_0 = \arg\max_c p_0(y = c \mid x)$. In the on-the-fly TTA setting, the model encounters an unlabeled target stream $\{x_t\}_{t=1}^{T}$ under an evolving distribution shift. The goal is to leverage such streaming target information to boost test-time adaptation performance in the absence of ground-truth labels.

3.2 Framework Overview

We present DANCE, a novel method for test-time adaptation in Vision-Language Models, as illustrated in the overview in Figure 3. For each incoming image x_t, we (1) form a proxy representation by selecting reliable augmented views, (2) update the predicted-class cache using a reliability–novelty admission mechanism, and (3) retrieve cached prototypes to produce a cache-based logit correction and apply a lightweight prototype alignment. Notably, we employ asymmetric quantization (AQ) similarity as a unified retrieval primitive in both cache updating and prototype retrieval.

3.3 Stable Augmented-view Selection

In test-time adaptation, a single test image x_t is typically transformed into multiple augmented views $\{x_a\}_{a=1}^{A}$, each passed through VLM to produce an image feature f_a and logits z_a. Due to the stochastic nature of augmentation, a common heuristic is to select the views with low prediction entropy e_a and average their features and logits. However, entropy is a purely per-view confidence measure and does not enforce consistency in the representation space, which can introduce semantic noise into the proxy representation of the sample.

To obtain a faithful proxy, we propose a stable view selection mechanism that filters noisy aug-

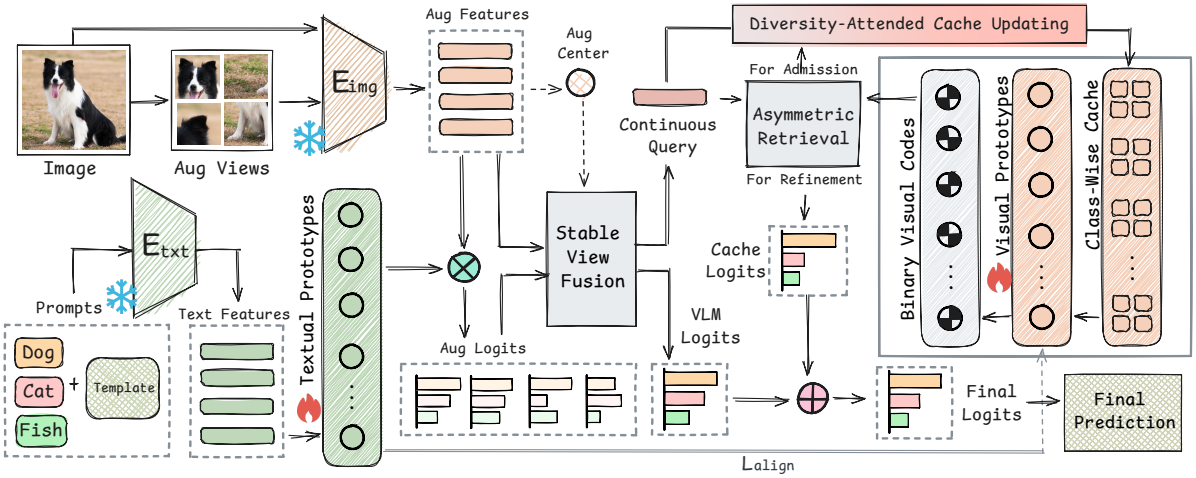

Figure 3: Overview of our DANCE. Stable view fusion constructs a proxy feature from augmented views, which updates a class-wise cache with a diversity-attended rule. Retrieved visual prototypes provide cache logits that refine the predictions of VLMs, with AQ serving as the shared similarity for cache updating and retrieval. This design helps maintain informative prototypes over the test stream and improves robustness under distribution shift.

mentations by enforcing feature-space consistency alongside prediction confidence. We first compute the consistency between each augmented feature f_a and augmentations center $\bar{f}$, where $\bar{f}$ denotes the mean feature over all augmentations:

$$c_a = \frac{1}{2}\left(f_a^\top \bar{f} + 1\right). \tag{2}$$

Then, we combine reliability and consistency into a scalar view selection score:

$$u_a = -e_a + \alpha_{\text{cons}}\, c_a, \tag{3}$$

where α_{cons} controls the influence of semantic consistency. We retain the top-M augmentations with the largest view scores $\{u_a\}$ and denote their indices by $\mathcal{A}^\star$. We then compute $w_a = \text{softmax}_{a\in\mathcal{A}^\star}(-e_a)$ to fuse features and logits:

$$f_t = \sum_{a\in\mathcal{A}^\star} w_a\, f_a, \qquad z_t = \sum_{a\in\mathcal{A}^\star} w_a\, z_a. \tag{4}$$

The resulting pseudo-label $\hat{y}_t = \arg\max_c z_{t,(c)}$ and mean entropy e_t jointly form the basis for subsequent cache updates and logit refinement.

3.4 Diversity-attended Dynamic Cache for Shift Tracking

A typical entropy-only cache update rule may repeatedly admit near-duplicate high-confidence samples, progressively narrowing cache coverage within each class. To keep the cache maximally useful for downstream refinement, DANCE uses a reliability-gated, diversity-aware update. Specifically, reliability serves as a guardrail against noisy pseudo-labels, while diversity is used to select complementary exemplars that expand intra-class coverage rather than repeatedly storing a few easy samples. An intuitive overview of this cache update mechanism is provided in Figure 2 for clarity.

Let $\mathcal{C}_{\hat{y}_t} = \{f_i\}_{i=1}^{K}$ denote the per-class cache associated with the predicted class $\hat{y}_t$. To maintain a diverse exemplar set, we quantify the novelty of each new sample f_t by its incremental contribution to the existing feature distribution. Specifically, we retrieve the nearest neighbor of f_t from $\mathcal{C}_{\hat{y}_t}$ and define novelty as the corresponding feature dissimilarity (Zhou et al., 2025b):

$$\mathcal{N}(f_t, \mathcal{C}_{\hat{y}_t}) = 1 - \max_{i\in[K]} \text{sim}(f_t, f_i). \tag{5}$$

Here $\text{sim}(\cdot,\cdot)$ is a similarity operator for retrieval in the visual feature space, whose concrete instantiation will be specified in Sec. 3.5. A high novelty score indicates that f_t resides in a previously underrepresented region of the feature space, thereby expanding the diversity of the exemplar bank. We then assign each candidate an admission score for cache admission, coupling entropy-based reliability with diversity-attended incrementality:

$$\begin{aligned} S_t &= -e_t + \lambda_{\text{eff}} \cdot \mathcal{N}(f_t, \mathcal{C}_{\hat{y}_t}), \\ \lambda_{\text{eff}} &= \lambda_{\text{nov}} \exp(-\alpha_{\text{nov}} e_t), \end{aligned} \tag{6}$$

where λ_{nov} and α_{nov} are hyperparameters controlling the novelty scale and its sensitivity to uncertainty. In this formulation, λ_{eff} acts as an adaptive coefficient that automatically suppresses the novelty term when prediction entropy is high, ensuring that cache diversity expands only when the

pseudo-label is likely reliable, thereby preventing boundary drift from noisy outliers. We maintain each per-class cache as a priority queue of size K. Each cache entry stores (f, ℓ_p, S), i.e., the feature, pseudo-label, and its priority key. Once the capacity K is reached, the sample f_t replaces the entry with the minimum score $S_{\min}$ if $S_t > S_{\min}$. This dynamic maintenance ensures the cache remains a compact yet representative summary of the target stream, which in turn enables more stable adaptation under evolving distribution drift.

3.5 Prototypical Logit Adjustment with Asymmetric Quantization

Given the proxy representation (f_t, z_t, e_t), a natural way to improve zero-shot predictions is to aggregate target-stream evidence from a class-wise cache and use it to adjust decision boundaries. However, cache-based refinement is only as reliable as the underlying *visual-space retrieval* that identifies which cached exemplars should contribute to the correction. In CLIP-style cross-modal embedding spaces, image–image similarities can exhibit substantial overlap, leading to frequent *near-ties* among top candidates (as discussed in Sec. 1). Consequently, small feature perturbations may flip the top-k ordering and produce high-magnitude but incorrect cache corrections. This issue is especially critical in our framework because retrieval appears in two coupled places: (i) cache maintenance, where novelty is estimated via nearest-neighbor search in Eq. 5, and (ii) logit refinement, where prototypes are retrieved to form cache correction. We therefore treat *retrieval stability* as a first-class design goal and introduce *asymmetric quantization (AQ)* as the core retrieval primitive in DANCE.

Asymmetric Quantization. Robust nearest-neighbor retrieval has been widely studied in the visual retrieval community (Malkov and Yashunin, 2018; Neyshabur et al., 2013; Liu et al., 2025), where *asymmetric binary coding* improves ranking robustness by binarizing database features while keeping queries continuous (Neyshabur et al., 2013; Shen et al., 2017; Da et al., 2017). Motivated by this principle, we binarize only the cached keys and keep the query feature continuous. Given a query feature $f_t \in \mathbb{R}^d$ and a cache key $f_i \in \mathbb{R}^d$, we compute a binary code for the key:

$$b_i = \operatorname{sign}(f_i) \in \{-1, +1\}^d. \tag{7}$$

We ℓ_2-normalize the query as $\bar{f}_t = \operatorname{Norm}(f_t)$ and define the asymmetric similarity:

$$\operatorname{sim}_{\mathrm{AQ}}(f_t, f_i) = \frac{1}{d}\, \bar{f}_t^{\top} b_i \in [-1, 1]. \tag{8}$$

Beyond efficiency, the appeal of AQ lies in its asymmetric design, which stabilizes neighbor ranking by reducing sensitivity to small perturbations in the stored keys, while continuous queries preserve expressive power. Based on this, we adopt $\operatorname{sim}_{\mathrm{AQ}}$ as the unified retrieval primitive throughout DANCE, and apply it consistently to both prototypical logit correction and dynamic cache maintenance.

Prototypical Logit Correction. We summarize each class in the cache by its averaged feature, yielding a prototype bank $F_{\text{cache}} \in \mathbb{R}^{N \times d}$, where N is the number of available class prototypes in the cache. Following the residual adaptation scheme in Zhang et al. (2024a), we use a lightweight residual parameterization to adjust both the text prototypes (text embeddings W) and the visual prototypes:

$$\begin{aligned} \tilde{W} &= \operatorname{Norm}(W + \Delta W), \\ \tilde{F}_{\text{cache}} &= \operatorname{Norm}(F_{\text{cache}} + \Delta F), \end{aligned} \tag{9}$$

where ΔW and ΔF are learnable residuals optimized via entropy minimization and prototype alignment on the cached data. The final prediction for the test query f_t is then derived from a fusion of the original zero-shot VLM logit with a cache-based correction computed via AQ retrieval:

$$z_t^{\text{final}} = \underbrace{f_t^{\top} \tilde{W}}_{\textit{VLM}} + \underbrace{g\Big(\operatorname{sim}_{\mathrm{AQ}}\big(f_t, \tilde{F}_{\text{cache}}\big)\Big)^{\top} L_{\text{cache}}}_{\textit{Cache Correction}}, \tag{10}$$

where $L_{\text{cache}} \in \{0, 1\}^{N \times C}$ is a one-hot row matrix mapping each prototype to its class, and $g(x) = \alpha \exp(-\beta(1 - x))$ is a modulation function (Karmanov et al., 2024). This formulation allows the model to rectify zero-shot misclassifications by retrieving relevant target-domain prototypes.

Finally, we use the same AQ similarity in cache maintenance by instantiating Eq. 5 with $\operatorname{sim}_{\mathrm{AQ}}$:

$$\mathcal{N}(f_t, \mathcal{C}_{\hat{y}_t}) = 1 - \max_{f_i \in \mathcal{C}_{\hat{y}_t}} \operatorname{sim}_{\mathrm{AQ}}(f_t, f_i), \tag{11}$$

with $\mathcal{N}(f_t, \mathcal{C}_{\hat{y}_t}) = 1$ when $|\mathcal{C}_{\hat{y}_t}| = 0$. This unifies cache admission and prototypical logit adjustment under a single retrieval primitive, improving robustness to similarity near-ties and yielding more reliable cache corrections under distribution shift.

Method	Aircraft	Caltech	Cars	DTD	EuroSAT	Flower	Food101	Pets	SUN397	UCF101	AVG
CLIP-ResNet-50	15.66	85.88	55.70	40.37	23.69	61.75	73.97	83.57	58.80	58.84	55.82
Ensemble	16.11	87.26	55.89	40.37	25.79	62.77	74.82	82.97	60.85	59.48	56.60
CoOp (Zhou et al., 2022b)	15.12	86.53	55.32	37.29	26.20	61.55	75.59	87.00	58.15	59.05	56.18
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
DiffTPT (Feng et al., 2023)	17.60	86.89	60.71	40.72	41.04	63.53	**79.21**	83.40	62.72	62.67	59.85
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
DPE (Zhang et al., 2024a)	19.80	**90.83**	59.26	50.18	41.67	67.60	77.83	85.97	64.23	61.98	61.93
BCA (Zhou et al., 2025a)	19.89	89.70	58.13	48.58	**42.12**	66.30	77.19	85.58	63.38	63.51	61.44
DANCE	**22.65**	**90.83**	**61.46**	**55.12**	42.03	**69.71**	78.40	**86.26**	**65.83**	63.60	**63.59**
CLIP-ViT-B/16	23.67	93.35	65.48	44.27	42.01	67.44	83.65	88.25	62.59	65.13	63.58
Ensemble	23.22	93.55	66.11	45.04	50.42	66.99	82.86	86.92	65.63	65.16	64.59
CoOp (Zhou et al., 2022b)	18.47	93.70	64.51	41.92	46.39	68.71	85.30	89.14	64.15	66.55	63.88
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
DiffTPT (Feng et al., 2023)	25.60	92.49	67.01	47.00	43.13	70.10	**87.23**	88.22	65.74	62.67	65.47
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
DPE (Zhang et al., 2024a)	28.95	94.81	67.31	54.20	55.79	75.07	86.17	91.14	70.07	70.44	69.40
BCA (Zhou et al., 2025a)	28.59	94.69	66.86	53.49	56.63	73.12	85.97	90.43	68.41	67.59	68.59
GS-Bias + E (Huang et al., 2025)	26.49	94.60	67.33	46.10	52.42	71.94	86.09	90.38	67.40	67.59	67.03
DANCE	**31.29**	**95.82**	**69.01**	**57.45**	**58.30**	**75.80**	86.61	**91.74**	**70.77**	**70.90**	**70.74**

Table 1: Compared performance of different methods on cross-dataset generalization across ten benchmark datasets.

3.6 Theoretical analysis

In this part, we present a theoretical analysis of the DANCE framework, focusing on how the diversity-attended dynamic cache offers superior adaptability to distribution shifts compared to conventional entropy-based caches. To facilitate the analysis, we consider a binary classification setting ($C = 2$) with balanced classes and two environments: the source environment $\mathcal{E}_s$ and the target environment $\mathcal{E}_t$, representing a distribution shift.

We model the distribution of the feature f in each environment as the multivariate normal distribution $\mathcal{N}(\mu_s, \sigma_{s,n}^2 I)$ and $\mathcal{N}(\mu_t, \sigma_t^2 I)$, respectively, where μ_s and μ_t denote the class means, and $\sigma_{s,n}^2$, $\sigma_{t,n}^2$ are the variances. We assume that both $\sigma_{s,n}$ and $\sigma_{t,n}$ converge to zero as the number of samples n increases, implying that features become concentrated around their respective class means and that within-environment variation diminishes.

Correspondingly, we model the prediction probability of the first class in each environment as following a Beta distribution: $\mathrm{Beta}(\alpha_{s,n}, \beta_{s,n})$ and $\mathrm{Beta}(\alpha_t, \beta_t)$, where $\alpha_{s,n} \to 1$, $\beta_{s,n} \to 0$, $\alpha_t > 0$, and $\beta_t > 0$ as $n \to \infty$. This captures the intuition that, as more samples are observed, the model becomes more confident in its predictions, leading to lower entropy values that are sharply concentrated near zero in source environment. With the prediction probability, we can compute the prediction entropy e under each environment.

Theorem 3.1. *Assume we have a fixed cache size K. As the number of samples $n \to \infty$,*

1. *Under an entropy-only cache update rule, the probability that all top-K samples originate from the source environment $\mathcal{E}_s$ converges to 1.*
2. *Under the diversity-attended dynamic cache update rule proposed in* DANCE*, there exists a sufficiently large novelty weight λ_{nov} such that the probability of selecting at least one sample from the target environment $\mathcal{E}_t$ converges to 1.*

Theorem 3.1 indicates that, under an entropy-only cache update rule, the cache may become dominated by samples from the source environment, which can lead to poor adaptability to distribution shifts. In contrast, the diversity-attended dynamic cache in DANCE encourages the selection of samples from the target environment, thereby enhancing the representativeness of the cache and improving adaptability under distribution shift.

4 Experiment

4.1 Experiment Settings

Datasets. Following prior TTA work, we report results under two main benchmarking scenarios: (1) *Cross-dataset generalization*. We evaluate on 10 widely used target datasets: Aircraft (Maji et al., 2013), Caltech101 (Fei-Fei et al., 2004), Cars (Krause et al., 2013), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), Flowers (Nilsback and Zisserman, 2008), Food101 (Bossard et al., 2014), Pets (Parkhi et al., 2012), SUN397 (Xiao et al., 2010), and UCF101 (Soomro et al., 2012) (Table 1). This setting measures adaptability and generalization under distribution shift in real-world transfer. (2) *Robustness under natural distribution*

Method	ImageNet	-A	-V2	-R	-Sketch	Average	OOD Avg.
CLIP-ResNet-50	58.16 $_{0.00}$	21.83 $_{0.00}$	51.41 $_{0.00}$	56.15 $_{0.00}$	33.37 $_{0.00}$	44.18 $_{0.00}$	40.69 $_{0.00}$
Ensemble	59.81 $_{\uparrow 1.65}$	23.24 $_{\uparrow 1.41}$	52.91 $_{\uparrow 1.50}$	60.72 $_{\uparrow 4.57}$	35.48 $_{\uparrow 2.11}$	46.43 $_{\uparrow 2.25}$	43.09 $_{\uparrow 2.40}$
CoOp (Zhou et al., 2022b)	63.33 $_{\uparrow 5.17}$	23.06 $_{\uparrow 1.23}$	55.40 $_{\uparrow 3.99}$	56.60 $_{\uparrow 0.45}$	34.67 $_{\uparrow 1.30}$	46.61 $_{\uparrow 2.43}$	42.43 $_{\uparrow 1.74}$
TPT (Shu et al., 2022)	60.74 $_{\uparrow 2.58}$	26.67 $_{\uparrow 4.84}$	54.70 $_{\uparrow 3.29}$	59.11 $_{\uparrow 2.96}$	35.09 $_{\uparrow 1.72}$	47.26 $_{\uparrow 3.08}$	43.89 $_{\uparrow 3.20}$
DiffTPT (Feng et al., 2023)	60.80 $_{\uparrow 2.64}$	**31.06** $_{\uparrow 9.23}$	55.80 $_{\uparrow 4.39}$	58.80 $_{\uparrow 2.65}$	37.10 $_{\uparrow 3.73}$	48.71 $_{\uparrow 4.53}$	45.69 $_{\uparrow 5.00}$
TDA (Karmanov et al., 2024)	61.35 $_{\uparrow 3.19}$	30.29 $_{\uparrow 8.46}$	55.54 $_{\uparrow 4.13}$	62.58 $_{\uparrow 6.43}$	38.12 $_{\uparrow 4.75}$	49.58 $_{\uparrow 5.40}$	46.63 $_{\uparrow 5.94}$
DPE (Zhang et al., 2024a)	63.41 $_{\uparrow 5.25}$	30.15 $_{\uparrow 8.32}$	**56.72** $_{\uparrow 5.31}$	**63.72** $_{\uparrow 7.57}$	40.03 $_{\uparrow 6.66}$	50.81 $_{\uparrow 6.63}$	47.66 $_{\uparrow 6.97}$
BCA (Zhou et al., 2025a)	61.81 $_{\uparrow 3.65}$	30.35 $_{\uparrow 8.52}$	56.58 $_{\uparrow 5.17}$	62.89 $_{\uparrow 6.74}$	38.08 $_{\uparrow 4.71}$	49.94 $_{\uparrow 5.76}$	46.98 $_{\uparrow 6.29}$
GS-Bias + E (Huang et al., 2025)	62.05 $_{\uparrow 3.89}$	27.83 $_{\uparrow 6.00}$	55.91 $_{\uparrow 4.50}$	63.01 $_{\uparrow 6.86}$	36.98 $_{\uparrow 3.61}$	49.16 $_{\uparrow 4.98}$	45.93 $_{\uparrow 5.24}$
DANCE	**64.03** $_{\uparrow 5.87}$	30.96 $_{\uparrow 9.13}$	56.63 $_{\uparrow 5.22}$	63.66 $_{\uparrow 7.51}$	**40.49** $_{\uparrow 7.12}$	**51.15** $_{\uparrow 6.97}$	**47.94** $_{\uparrow 7.25}$
CLIP-ViT-B/16	66.73 $_{0.00}$	47.87 $_{0.00}$	60.86 $_{0.00}$	73.98 $_{0.00}$	46.09 $_{0.00}$	59.11 $_{0.00}$	57.20 $_{0.00}$
Ensemble	68.34 $_{\uparrow 1.61}$	49.89 $_{\uparrow 2.02}$	61.88 $_{\uparrow 1.02}$	77.65 $_{\uparrow 3.67}$	48.24 $_{\uparrow 2.15}$	61.20 $_{\uparrow 2.09}$	59.42 $_{\uparrow 2.22}$
CoOp (Zhou et al., 2022b)	71.51 $_{\uparrow 4.78}$	49.71 $_{\uparrow 1.84}$	64.20 $_{\uparrow 3.34}$	75.21 $_{\uparrow 1.23}$	47.99 $_{\uparrow 1.90}$	61.72 $_{\uparrow 2.61}$	59.28 $_{\uparrow 2.08}$
TPT (Shu et al., 2022)	68.98 $_{\uparrow 2.25}$	54.77 $_{\uparrow 6.90}$	63.45 $_{\uparrow 2.59}$	77.06 $_{\uparrow 3.08}$	47.94 $_{\uparrow 1.85}$	62.44 $_{\uparrow 3.33}$	60.81 $_{\uparrow 3.61}$
DiffTPT (Feng et al., 2023)	70.30 $_{\uparrow 3.57}$	55.68 $_{\uparrow 7.81}$	65.10 $_{\uparrow 4.24}$	75.00 $_{\uparrow 1.02}$	46.80 $_{\uparrow 0.71}$	62.28 $_{\uparrow 3.17}$	60.52 $_{\uparrow 3.32}$
TDA (Karmanov et al., 2024)	69.51 $_{\uparrow 2.78}$	60.11 $_{\uparrow 12.24}$	64.67 $_{\uparrow 3.81}$	80.24 $_{\uparrow 6.26}$	50.54 $_{\uparrow 4.45}$	65.01 $_{\uparrow 5.90}$	63.89 $_{\uparrow 6.69}$
DPE (Zhang et al., 2024a)	71.91 $_{\uparrow 5.18}$	59.63 $_{\uparrow 11.76}$	**65.44** $_{\uparrow 4.58}$	80.40 $_{\uparrow 6.42}$	52.26 $_{\uparrow 6.17}$	65.93 $_{\uparrow 6.82}$	64.43 $_{\uparrow 7.23}$
BCA (Zhou et al., 2025a)	70.22 $_{\uparrow 3.49}$	**61.14** $_{\uparrow 13.27}$	64.90 $_{\uparrow 4.04}$	**80.72** $_{\uparrow 6.74}$	50.87 $_{\uparrow 4.78}$	65.37 $_{\uparrow 6.26}$	64.16 $_{\uparrow 6.96}$
GS-Bias + E (Huang et al., 2025)	70.57 $_{\uparrow 3.84}$	56.61 $_{\uparrow 8.74}$	64.62 $_{\uparrow 3.76}$	80.49 $_{\uparrow 6.51}$	50.33 $_{\uparrow 4.24}$	64.52 $_{\uparrow 5.41}$	63.01 $_{\uparrow 5.81}$
DANCE	**72.21** $_{\uparrow 5.48}$	60.06 $_{\uparrow 12.19}$	65.41 $_{\uparrow 4.55}$	80.46 $_{\uparrow 6.48}$	**52.91** $_{\uparrow 6.82}$	**66.21** $_{\uparrow 7.10}$	**64.71** $_{\uparrow 7.51}$

Table 2: Performance comparisons on ImageNet and its distribution shifts. Each entry is annotated with the absolute change (percentage points) relative to CLIP zero-shot under the same backbone.

shifts. We evaluate on ImageNet (Deng et al., 2009) and four established shifted variants: ImageNet-A (Hendrycks et al., 2021b), ImageNet-V2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021a), and ImageNet-S (Wang et al., 2019) (Table 2). This benchmark tests the model's ability to generalize reliably to diverse unseen conditions.

Baselines. We compare DANCE with CLIP and a set of representative TTA baselines. As non-adaptive references, we report zero-shot CLIP and an augmentation ensemble. As label-dependent offline adaptation baselines, we include CoOp (Zhou et al., 2022b). For prompt-based test-time adaptation, we compare with TPT (Shu et al., 2022) and DiffTPT (Feng et al., 2023). For cache-based adaptation, we compare with TDA (Karmanov et al., 2024), DPE (Zhang et al., 2024a), and BCA (Zhou et al., 2025a). In addition, we also include GS-Bias+E (Huang et al., 2025), a parameter-efficient baseline that adapts via explicit bias terms.

Implementation Details. Following prior work, we adopt a pretrained CLIP model with standard backbone choices. We follow the standard image augmentation pipeline from Shu et al. (2022). In line with Zhang et al. (2024a), prototypes undergo a single AdamW update step, while the cache modulation is configured to match their setup. We use a per-class cache capacity of $K=10$ by default, which provides adequate candidates for prototype construction with manageable overhead. We further study the effect of cache capacity in Appendix B. For simplicity and reproducibility, we use the same hyperparameter setting across all datasets, using top-1 accuracy as the primary evaluation metric. Results are averaged over multiple random seeds and reported as the mean values.

4.2 Main Results

Cross-Datasets Generalization. We report cross-dataset results in Table 1, comparing DANCE with representative TTA baselines. DANCE consistently outperforms strong baselines, with clear gains over DPE in particular, which the average accuracy improves from 61.93% to 63.54% on ResNet-50 and from 69.40% to 70.75% on ViT-B/16. Notably, DANCE yields larger gains on challenging targets, e.g., DTD improves from 50.18% to 55.12%, while remaining competitive across the other datasets. We attribute these gains to more reliable cache evolution and retrieval under shift, yielding a more representative cache and reliable prediction.

Robustness under Natural Distribution Shifts. We further report results on ImageNet and four established natural distribution shifts in Table 2. DANCE achieves strong robustness in aggregate, obtaining the best overall performance on both backbones (Average: 51.15% on ResNet-50 and 66.21% on ViT-B/16; OOD Average: 47.94% and 64.71%). The gains are most pronounced on challenging shifts such as ImageNet-S, where DANCE attains the top accuracy for both backbones. We attribute this to the improved quality of cached prototypes and the more reliable retrieval process. Overall, these results highlight DANCE's strong ability to maintain robust test-time performance across diverse datasets under natural distribution shifts.

Components			RN50	ViT-B/16
m1	m2	m3	avg Δ	avg Δ
✓	✓	✓	**63.59** 0.00	**70.74** 0.00
×	✓	✓	63.49 ↓0.10	70.54 ↓0.20
✓	×	✓	63.50 ↓0.09	70.61 ↓0.13
✓	✓	×	62.59 ↓1.00	69.98 ↓0.76
✓	×	×	62.75 ↓0.84	70.01 ↓0.73
×	✓	×	62.60 ↓0.99	69.96 ↓0.78
×	×	✓	63.47 ↓0.12	70.44 ↓0.30
×	×	×	61.93 ↓1.66	69.40 ↓1.34

Table 3: Ablation study on cross-dataset generalization, reported as the average top-1 accuracy over 10 benchmark datasets. Each row removes one or more key components of DANCE to assess their individual and combined contributions to overall performance.

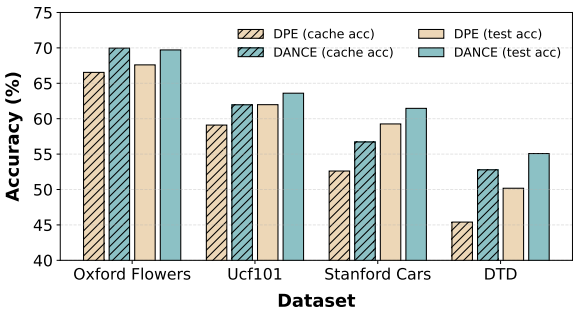

Figure 4: Cache accuracy and final accuracy on four cross-dataset targets, comparing DPE and DANCE.

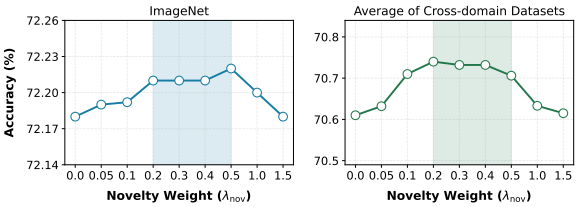

Figure 5: Sensitivity analysis of the novelty weight λ_{nov} on ImageNet and the cross-domain datasets.

Dataset	retr@1 (%)	hurthit (%)	top-1 acc (%)
DTD	47.64→52.66	1.07→0.62	50.16→55.12
	↑5.02	↓0.45	↑4.96
Flowers	66.75→67.76	1.51→0.86	69.14→69.71
	↑1.01	↓0.65	↑0.57
Caltech101	88.80→89.41	0.41→0.12	90.51→90.83
	↑0.61	↓0.29	↑0.32
Cars	54.83→56.82	2.21→1.03	61.26→61.46
	↑1.99	↓1.18	↑0.20

Table 4: Diagnostic comparison between cosine retrieval and AQ for cache-based refinement. Results are consistently reported as cosine → AQ, with absolute changes (percentage points) shown in parentheses.

4.3 Further Analysis

Ablation Study. To verify the contribution of each component in DANCE, Table 3 reports an ablation study on cross-dataset generalization under two backbones. Reverting our key designs to their baseline counterparts results in consistent performance drops, validating the necessity of each component: ❶ *Stable View Selection (m1):* Replacing it with entropy-based selection leads to a performance decay, suggesting that enforcing feature-space consistency yields more trustworthy anchors and prevents noise accumulation in early adaptation stages. ❷ *Diversity-attended Admission (m2):* Replacing it with entropy-only filtering degrades results across all backbones. This highlights that maintaining cache diversity is crucial for covering evolving test distributions and strengthening generalization. ❸ *AQ-based Retrieval (m3):* Replacing AQ-based retrieval with cosine similarity causes clear drop, indicating that AQ provides a more reliable retrieval mechanism. ❹ *Component Synergy:* Beyond single-module ablations, removing arbitrary combinations leads to significantly larger degradations. This underscores that these components provide complementary benefits and collectively form a robust framework for TTA.

Sensitivity Analysis. We study the sensitivity to the novelty weight λ_{nov} used in cache admission. We vary λ_{nov} and report accuracy on ImageNet and the average over 10 cross-domain datasets in Figure 5 and observe a stable region where performance varies slightly. This parameter insensitivity suggests that the proposed novelty gating is robust to the choice of λ_{nov}. To maintain a consistent test-time protocol, we use a single setting of $\lambda_{\text{nov}} = 0.2$ in all experiments. Additional hyperparameter sensitivity results are provided in the Appendix C.

Cache Reliability. We evaluate cache reliability using the top-1 accuracy of the cache-branch prediction (Eq. 10), denoted as cache acc, and report it alongside the final model accuracy (test acc). Across the four datasets, DANCE yields higher cache accuracy than DPE, and the same trend is reflected in the final prediction accuracy. Notably, the substantial improvement in cache acc validates that our diversity-attended admission and AQ-based retrieval effectively curate a more representative and high-fidelity exemplar bank. This superior cache quality provides a more dependable foundation for logit correction, ultimately translating into higher final prediction accuracy under distribution shift.

Retrieval Reliability. To further understand why AQ out-performs standard cosine similarity, we introduce two diagnostic metrics to evaluate the reliability of prototype retrieval: (i) Retr@1, the hit rate of top-1 retrieved prototypes matching the query's ground-truth class; and (ii) HurtHit, the rate of correct base predictions being miscorrected by faulty visual retrieval. As shown in Table 4, AQ consistently improves Retr@1 while suppressing HurtHit rate. These results demonstrate that AQ effectively retrieves more class-faithful prototypes and prevents detrimental updates, directly underpinning the observed gains in overall accuracy.

5 Conclusion

We propose DANCE, a novel test-time adaptation framework for VLMs under distribution shift. DANCE enhances adaptation by integrating stable augmented-view selection, diversity-attended cache updates, and robust visual-space retrieval. Extensive experiments support the effectiveness of this design for test-time adaptation. More broadly, DANCE shows how coherently selecting, retaining, and exploiting historical target evidence can improve shift tracking and test-time recognition.

Limitations

While DANCE demonstrates promising performance in test-time adaptation, an important limitation remains to be considered. Although our diversity-attended cache update strategy effectively expands the coverage of visual prototypes by incorporating features that are both confident and novel, the framework does not explicitly account for cases where the backbone model exhibits overconfidence in out-of-distribution noise. This may introduce subtle drift in class prototypes in scenarios characterized by high-intensity adversarial perturbations.

Acknowledgements

The work of Ziyue Qiao is partially supported by the National Natural Science Foundation of China (No. 62406056), the Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515140114).

References

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

Xinyu Chen, Haotian Zhai, Can Zhang, Xiupeng Shi, and Ruirui Li. 2025. Multi-cache enhanced prototype learning for test-time generalization of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2281–2291.

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*, pages 3606–3613.

Cheng Da, Shibiao Xu, Kun Ding, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. 2017. Amvh: Asymmetric multi-valued hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 736–744.

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*, pages 248–255. Ieee.

Xinqi Fan, Xueli Chen, Luoxiao Yang, Chuin Hong Yap, Rizwan Qureshi, Qi Dou, Moi Hoon Yap, and Mubarak Shah. 2025. Test-time retrieval-augmented adaptation for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8810–8819.

Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. 2024. Frustratingly easy test-time adaptation of vision-language models. *Advances in Neural Information Processing Systems*, 37:129062–129093.

Li Fei-Fei, Rob Fergus, and Pietro 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*, pages 178–178. IEEE.

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*, pages 2704–2714.

Zenghao Guan, Zhou Yucan, Wu Liu, and Xiaoyan Gu. Statistics caching test-time adaptation for vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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):2217–2226.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai,

Tyler Zhu, Samyak Parajuli, Mike Guo, and 1 others. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349.

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

Zhaohong Huang, Yuxin Zhang, Jingjing Xie, Fei Chao, and Rongrong Ji. 2025. Gs-bias: Global-spatial bias learner for single-image test-time adaptation of vision-language models. *arXiv preprint arXiv:2507.11969*.

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 *International conference on machine learning*, pages 4904–4916. PMLR.

Siddharth Joshi, Arnav Jain, Ali Payani, and Baharan Mirzasoleiman. 2024. Data-efficient contrastive language-image pretraining: Prioritizing data quality over quantity. In *International Conference on Artificial Intelligence and Statistics*, pages 1000–1008. PMLR.

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*, pages 14162–14171.

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 workshops*, pages 554–561.

Shuaifeng Li, Mao Ye, Lihua Zhou, Nianxin Li, Siying Xiao, Song Tang, and Xiatian Zhu. 2024. Cloud object detector adaptation by integrating different source knowledge. *Advances in Neural Information Processing Systems*, 37:25251–25283.

Shuaifeng Li, Mao Ye, Xiatian Zhu, Lihua Zhou, and Lin Xiong. 2022. Source-free object detection by learning to overlook domain style. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8014–8023.

Jian Liang, Ran He, and Tieniu Tan. 2025a. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64.

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. *Advances in Neural Information Processing Systems*, 35:17612–17625.

Yiwen Liang, Hui Chen, Yizhe Xiong, Zihan Zhou, Mengyao Lyu, Zijia Lin, Shuaicheng Niu, Sicheng Zhao, Jungong Han, and Guiguang Ding. 2025b. Advancing reliable test-time adaptation of vision-language models under visual variations. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4788–4797.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2025. Robust neural information retrieval: An adversarial and out-of-distribution perspective. *ACM Transactions on Information Systems*, 44(1):1–48.

Junyu Luo, Xiao Luo, Xiusi Chen, Zhiping Xiao, Wei Ju, and Ming Zhang. 2025a. Semi-supervised fine-tuning for large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2795–2808.

Junyu Luo, Bohan Wu, Xiao Luo, Zhiping Xiao, Yiqiao Jin, Rong-Cheng Tu, Nan Yin, Yifan Wang, Jingyang Yuan, Wei Ju, and 1 others. 2025b. A survey on efficient large language model training: From data-centric perspectives. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30904–30920.

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

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

Behnam Neyshabur, Nati Srebro, Russ R Salakhutdinov, Yury Makarychev, and Payman Yadollahpour. 2013. The power of asymmetry in binary hashing. *Advances in neural information processing systems*, 26.

Khanh-Binh Nguyen, Phuoc-Nguyen Bui, Hyunseung Choo, and Duc Thanh Nguyen. 2025. Adaptive cache enhancement for test-time adaptation of vision-language models. *arXiv preprint arXiv:2508.07570*.

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*, pages 722–729. IEEE.

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

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15691–15701.

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

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

Fumin Shen, Yang Yang, Li Liu, Wei Liu, Dacheng Tao, and Heng Tao Shen. 2017. Asymmetric binary coding for image search. *IEEE Transactions on Multimedia*, 19(9):2022–2032.

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. *Advances in Neural Information Processing Systems*, 35:14274–14289.

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):1–7.

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

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*, pages 3485–3492. IEEE.

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*, pages 11621–11631.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

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*, pages 10899–10909.

Haotian Zhai, Xinyu Chen, Can Zhang, Tianming Sha, and Ruirui Li. 2025. Mitigating cache noise in test-time adaptation for large vision-language models. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18123–18133.

Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. 2024a. Dual prototype evolving for test-time generalization of vision-language models. *Advances in Neural Information Processing Systems*, 37:32111–32136.

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 *European conference on computer vision*, pages 493–510. Springer.

Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. 2024b. Boostadapter: Improving vision-language test-time adaptation via regional bootstrapping. *Advances in Neural Information Processing Systems*, 37:67795–67825.

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

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

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

Yujun Zhou, Zhenwen Liang, Haolin Liu, Wenhao Yu, Kishan Panaganti, Linfeng Song, Dian Yu, Xiangliang Zhang, Haitao Mi, and Dong Yu. 2025b. Evolving language models without labels: Majority drives selection, novelty promotes variation. *arXiv preprint arXiv:2509.15194*.

A Proof of Theorem 3.1

Proof. We fix the cache size to K and, without loss of generality, focus on samples from the first class. Let e_s and e_t denote the prediction entropy of a sample drawn from the source environment $\mathcal{E}_s$ and target environment $\mathcal{E}_t$, respectively.

We begin by analyzing prediction entropy in each environment. For a Beta distribution $\mathrm{Beta}(\alpha, \beta)$, the mean and variance are given by:

$$\mu = \frac{\alpha}{\alpha+\beta}, \qquad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Under the source environment $\mathcal{E}_s$, as $n \to \infty$, we have

$$\mu_{s,n} = \frac{\alpha_{s,n}}{\alpha_{s,n}+\beta_{s,n}} \to 1,$$

and

$$\sigma^2_{s,n} = \frac{\alpha_{s,n}\beta_{s,n}}{(\alpha_{s,n}+\beta_{s,n})^2(\alpha_{s,n}+\beta_{s,n}+1)} \to 0.$$

This implies that the predicted class probability concentrates around 1, and consequently, the prediction entropy e_s converges to 0 in probability. Moreover, since the maximum operator is continuous, the continuous mapping theorem implies that the largest K entropy values among samples from $\mathcal{E}_s$ also converge to 0 in probability.

In contrast, under the target environment $\mathcal{E}_t$, we assume that α_t and β_t are positive constants. As a result, the predicted class probability follows a non-degenerate distribution with mean

$$\mu_t = \frac{\alpha_t}{\alpha_t+\beta_t}, \qquad \sigma^2_t = \frac{\alpha_t\beta_t}{(\alpha_t+\beta_t)^2(\alpha_t+\beta_t+1)}.$$

Accordingly, the prediction entropy e_t also has a non-degenerate distribution. Since entropy is non-negative and strictly positive with probability one under this setting, we have $\mathbb{P}(e_t > 0) = 1$.

Therefore, if samples are selected solely based on entropy, the probability that all top-K selected samples originate from the source environment $\mathcal{E}_s$ converges to 1, potentially leading to a cache that is unrepresentative of the target environment $\mathcal{E}_t$.

Next, we analyze the effect of incorporating a novelty term into the sample selection process. The novelty term encourages the selection of samples that are dissimilar to those already present in the cache, thereby promoting diversity. Mathematically, we define the novelty of a sample f_t with respect to the cache $\mathcal{C}$ as:

$$\mathcal{N}(f_t, \mathcal{C}) = 1 - \max_{f_i \in \mathcal{C}} \mathrm{sim}(f_t, f_i),$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes the cosine similarity.

We now examine the behavior of the initial samples under the diversity-attended dynamic cache mechanism. Since the cache is initially empty, the first sample f_1 is admitted solely based on its entropy e_1. From the earlier analysis, as $n \to \infty$, f_1 is drawn from the source environment $\mathcal{E}_s$ with probability approaching 1.

Now consider a second sample f_2 also drawn from $\mathcal{E}_s$. The similarity $\mathrm{sim}(f_2, f_1)$ converges to 1 in probability as $n \to \infty$, causing the novelty term $\mathcal{N}(f_2, \{f_1\})$ to converge to 0. The resulting admission score for f_2 is given by:

$$S_2 = -e_2 + \lambda_{\text{eff}} \cdot \mathcal{N}(f_2, \{f_1\}),$$

which converges to $-e_2$ in probability. Since $e_2 \to 0$ in probability, we have $S_2 \to 0$ in probability.

Now suppose the third sample f_3 is drawn from the target environment $\mathcal{E}_t$. In this case, the similarity $\mathrm{sim}(f_3, f_1)$ converges to a value less than 1 in probability, resulting in a positive novelty score $\mathcal{N}(f_3, \{f_1\})$ with high probability. Since the entropy e_3 is bounded above by $2\log 2$ in binary classification, it remains finite. Thus, for a suitably chosen λ_{eff}, the admission score

$$S_3 = -e_3 + \lambda_{\text{eff}} \cdot \mathcal{N}(f_3, \{f_1\})$$

can be positive with probability approaching 1. This implies that f_3 has a high chance of being admitted to the cache, despite its entropy being potentially higher than that of samples from $\mathcal{E}_s$. Therefore, the second sample in the cache comes from the target environment $\mathcal{E}_t$ with probability approaching 1, improving the representativeness of the cache across environments. □

B Effect of Cache Size (Shot-Capacity)

We analyze the effect of the per-class cache size K (shot-capacity) on cross-dataset performance. Table 5 shows that increasing the cache capacity consistently improves over the conventional small-cache regime: compared with K=3, both K=10 and K=16 yield higher accuracy. Notably, K=3 corresponds to the setting that is commonly adopted in prior cache-based adaptation methods (and is typically their best-performing choice), yet such a small capacity is more prone to redundancy and coverage shrinkage, which can lead to cache collapse and less representative class-wise prototypes under streaming shift. In contrast, DANCE remains effective when scaling the cache to larger

Cache Size	DTD	Pets	Caltech	UCF101
DPE (K=3)	54.20	91.14	94.81	70.44
DANCE (K=3)	57.09	91.52	95.42	70.61
DANCE (K=10)	**57.45**	**91.74**	**95.82**	**70.90**
DANCE (K=16)	<u>57.39</u>	<u>91.71</u>	<u>95.78</u>	<u>70.76</u>

Table 5: Effect of cache size K (shot-capacity) on cross-dataset performance (Top-1 accuracy, %). Best in bold; second-best underlined (among DANCE variants).

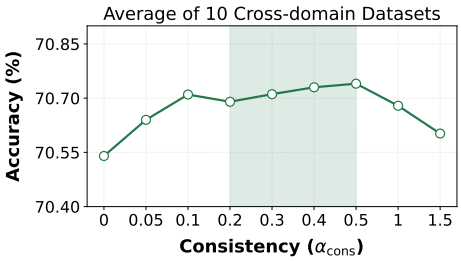

Figure 6: Sensitivity to the consistency weight α_{cons}. Average Top-1 accuracy over 10 cross-domain datasets as α_{cons} varies in stable augmented-view selection.

capacities, indicating that it can exploit additional target-stream evidence without being overly susceptible to accumulating noise. The comparison with DPE at its default setting K=3 suggests that the gains are not merely from enlarging the cache, but from our reliability–diversity admission strategy, which maintains reliable prototypes as K increases.

C Additional Hyperparameter Studies

C.1 Sensitivity of Stable View Selection

We evaluate the sensitivity of the consistency weight α_{cons} used in stable augmented-view selection, where α_{cons} trades off feature-space consistency with entropy-based reliability when selecting augmented views for fusion. Figure 6 reports the average accuracy across 10 cross-domain datasets under different α_{cons} values. We observe a relatively stable operating region across moderate α_{cons} choices, and use a single shared $\alpha_{\text{cons}} = 0.5$ in all experiments for consistency and simplicity.

C.2 Sensitivity to the Entropy-Adaptation Coefficient

We examine the effect of the entropy-adaptation coefficient α_{nov} in the entropy-adaptive novelty weighting (Fig. 7). We vary α_{nov} and report the average accuracy over a representative subset of datasets, including FGVC, Caltech101, DTD, Flowers102, OxfordPets, and SUN397. The coefficient α_{nov} controls how strongly the effective novelty weight $\lambda_{\text{eff}} = \lambda_{\text{nov}} \exp(-\alpha_{\text{nov}} e_t)$ is attenuated by prediction entropy, i.e., the sensitivity of novelty-driven admission to uncertainty. As shown in Fig. 7, moderate values balance diversity-seeking updates under confident predictions while suppressing novelty contributions when predictions are uncertain, leading to stable behavior across the sweep. Accordingly, we use $\alpha_{\text{nov}} = 0.2$ in all experiments.

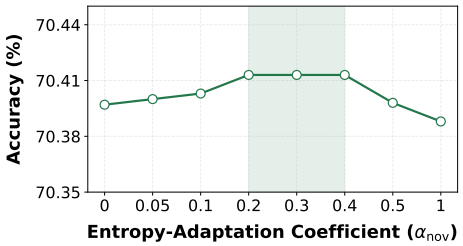

Figure 7: Sensitivity to the entropy-adaptation coefficient α_{nov}. We sweep α_{nov} and report the average accuracy over six representative datasets.

D Language-side Extension of DANCE

We further examine whether the core reliability-and-novelty principle in DANCE also benefits the text side of VLMs, in line with broader evidence that unlabeled data can benefit language-side adaptation (Luo et al., 2025a). Specifically, instead of uniformly averaging all prompts to construct the class text prototype, we apply a reliability-and-novelty-driven prompt selection strategy to obtain a more representative prompt subset for each class, and then fuse the selected prompts with score-based weights. Here, reliability measures how confidently a prompt supports its class on test images predicted as that class, while novelty encourages selecting prompts that are complementary rather than redundant. For simplicity, the selection score is defined as the sum of the two terms, and the subset size is fixed to approximately 30% of the full prompt pool per class. Table 6 reports the gains over the standard all-prompt averaging baseline. We observe consistent improvements across datasets and backbones, suggesting that the reliability-and-novelty principle may also provide useful guidance for text-side prototype construction in VLMs.

E Additional Implementation Details

E.1 Dataset Statistics

Table 7 summarizes the datasets used in our evaluation, including the number of categories, the sizes of the official splits (train/val/test when available),

Backbone	DTD	FGVC	Flowers	Caltech	UCF101	Pets	Cars	SUN397	Food101	EuroSAT	Avg.
ViT-B/16	↑2.08	↑0.60	↑1.58	↑0.73	↑1.32	↑1.17	↑0.24	↑1.36	↑0.09	↑0.25	↑0.94
RN50	↑3.07	↑1.08	↑1.99	↑0.37	↑1.85	↑3.03	↑0.18	↑1.94	↑0.35	↑0.33	↑1.42

Table 6: Language-side extension of DANCE. Top-1 accuracy gain (%) over standard all-prompt averaging.

Dataset	Classes	Train	Val	Test	Task
Caltech101 (Fei-Fei et al., 2004)	101	4,128	1,649	2,465	Object recognition
DTD (Cimpoi et al., 2014)	47	2,820	1,128	1,692	Texture recognition
EuroSAT (Helber et al., 2019)	10	13,500	5,400	8,100	Remote sensing recognition
FGVCAircraft (Maji et al., 2013)	100	3,334	3,333	3,333	Fine-grained aircraft recognition
Flowers102 (Nilsback and Zisserman, 2008)	102	4,093	1,633	2,463	Fine-grained flower recognition
Food101 (Bossard et al., 2014)	101	50,500	20,200	30,300	Fine-grained food recognition
OxfordPets (Parkhi et al., 2012)	37	2,944	736	3,669	Fine-grained pet recognition
StanfordCars (Krause et al., 2013)	196	6,509	1,635	8,041	Fine-grained car recognition
SUN397 (Xiao et al., 2010)	397	15,880	3,970	19,850	Scene recognition
UCF101 (Soomro et al., 2012)	101	7,639	1,898	3,783	Action recognition
ImageNet (Deng et al., 2009)	1,000	1,281,167	50,000	50,000	Large-scale object recognition
ImageNet-V2 (Recht et al., 2019)	1,000	–	–	10,000	Distribution shift benchmark
ImageNet-Sketch (Wang et al., 2019)	1,000	–	–	50,889	Sketch-domain shift benchmark
ImageNet-A (Hendrycks et al., 2021b)	200	–	–	7,500	Natural adversarial shift benchmark
ImageNet-R (Hendrycks et al., 2021a)	200	–	–	30,000	Artistic renditions benchmark

Table 7: Dataset statistics used in our experiments. Evaluation-only ImageNet variants provide test splits only.

Dataset	Prompt templates
ImageNet / V2 / Sketch / A / R	"a photo of a {CLASS}." "itap of a {CLASS}." "a bad photo of a {CLASS}." "a photo of the large {CLASS}." "a photo of the small {CLASS}." "an origami {CLASS}." "art of a {CLASS}."
Caltech101 (Fei-Fei et al., 2004)	"a photo of a {CLASS}."
DTD (Cimpoi et al., 2014)	"a {CLASS} texture."
EuroSAT (Helber et al., 2019)	"a centered satellite photo of {CLASS}."
FGVCAircraft (Maji et al., 2013)	"a photo of a {CLASS}, a type of aircraft."
Flowers102 (Nilsback and Zisserman, 2008)	"a photo of a {CLASS}, a type of flower."
Food101 (Bossard et al., 2014)	"a photo of {CLASS}, a type of food."
OxfordPets (Parkhi et al., 2012)	"a photo of a {CLASS}, a type of pet."
StanfordCars (Krause et al., 2013)	"a photo of a {CLASS}."
SUN397 (Xiao et al., 2010)	"a photo of a {CLASS}."
UCF101 (Soomro et al., 2012)	"a photo of a person doing {CLASS}."

Table 8: Prompt templates used in our experiments. {CLASS} indicates the dataset-specific class name.

and the primary recognition task of each benchmark. For evaluation-only ImageNet variants, only the test split is provided by default.

E.2 Prompt Templates

We use a small set of hand-crafted prompt templates to instantiate class names for each dataset. Besides these prompts, we utilize CuPL (Pratt et al., 2023) prompts as in DPE (Zhang et al., 2024a).

F Discussion of Recent Related Methods

Recent works also approach test-time adaptation from a cache-based perspective, but improve cache quality from different directions.

- **ReTA.** (Liang et al., 2025b) ReTA improves cache reliability mainly from the text side, by enforcing consistency through multi-prompt voting and modeling text prototypes with Gaussian distributions. In contrast, DANCE operates primarily in the visual feature space: it first stabilizes

augmented-view selection through feature-space consistency, and then explicitly controls how the cache evolves through reliability-and-novelty-based admission. Therefore, ReTA mainly improves *sample trustworthiness*, whereas DANCE additionally emphasizes *distributional coverage* of the evolving target stream. These two directions are largely complementary.

- **MCP.** (Chen et al., 2025) MCP adopts a multi-cache design with different caches serving distinct functional roles. DANCE maintains a single unified class-wise cache, which simplifies storage and coordination. Moreover, MCP promotes compactness through prototype proximity, whereas DANCE explicitly reduces redundancy by favoring reliable yet novel samples during admission. As a result, MCP and DANCE differ not only in design (multi-cache versus unified cache), but also in cache curation (compactness-oriented versus diversity-attended).
- **SCA.** (Guan et al.) SCA follows a statistics-caching paradigm, where target information is accumulated implicitly through running feature statistics. DANCE instead keeps explicit representative exemplars and forms class-wise visual prototypes from them. This explicit exemplar-based design provides two advantages: it enables direct cache diagnostics and interpretability, and it naturally supports asymmetric quantization as a unified retrieval primitive for both cache updating and logit refinement. Such capabilities are less natural under statistics-based aggregation.
- **CRG.** (Zhai et al., 2025) CRG mitigates cache noise mainly at inference time, while retaining an entropy-only caching strategy for memory construction. DANCE instead introduces a diversity-attended admission mechanism to reduce redundancy and improve the quality of cached samples during cache evolution. In addition, DANCE incorporates asymmetric quantization to enhance retrieval robustness in the visual feature space. Overall, CRG focuses on improving cache usage at inference time, whereas DANCE provides a complementary perspective by emphasizing cache construction and retrieval.