LET CLASSES COMPETE: DISCRIMINATIVE CACHE MATCHING FOR CLIP TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cache-based test-time adaptation (TTA) improves CLIP under distribution shifts by retaining previous test images and reusing them as visual references. Orthogonal to how such caches are constructed, we study a complementary question: once an observation has been stored, how should a newly arriving test image retrieve useful evidence from it? We identify two limitations of conventional cache matching. First, cache membership is determined by CLIP's image–text predictions, whereas subsequent retrieval relies on full-space image–image similarity, so even correctly retained references can be misranked. Second, correcting the global comparison geometry still represents each image with a single pooled feature, which can overlook localized class evidence. We propose **Discriminative Cache Matching (DCM)**, a training-free plug-in that preserves the pseudo-labeling and image-cache update rules of the underlying adapter while improving retrieval at both levels. Globally, DCM uses a fixed text-induced subspace to guide image–image cache matching. Locally, it reuses ViT patch tokens from the same retained images to perform fine-grained, class-competitive patch–patch retrieval. We further provide a conditional margin analysis characterizing when projection and re-normalization can correct misleading cache-reference rankings. Experiments on cross-dataset and out-of-distribution benchmarks demonstrate the effectiveness of DCM.

1 INTRODUCTION

Pretrained vision–language models such as CLIP (Radford et al., 2021) enable zero-shot classification by comparing image representations with textual class descriptions. Their performance can nevertheless degrade under distribution shifts, motivating test-time adaptation (TTA) using unlabeled target data. While optimization-based methods adapt prompts or selected model parameters at test time (Shu et al., 2022b; Yoon et al., 2024; Xiao et al., 2025), training-free cache-based approaches reuse observations accumulated from the test stream. TDA (Karmanov et al., 2024) maintains dynamic caches of image features and pseudo-labels, while MCP (Chen et al., 2025) refines class prototypes through multiple caches. Such methods separate two related decisions: which test observations are retained, and how their evidence is reused for subsequent queries. Rather than revisiting cache construction, we study how retained visual evidence should be matched to each incoming query.

Cache construction and cache retrieval rely on different relations. Retained observations are organized by CLIP's image–text predictions, whereas their contribution to a future query is determined by image–image similarity. These relations need not preserve the same class ordering. Even when both the query and cached references are correctly classified, shared context, background, style, or appearance can make a different-class reference more similar than a same-class one. Thus, improving which observations enter the cache does not by itself determine how effectively those observations will support future queries. This motivates a complementary question: *how should retained visual evidence be matched to a newly arriving test image?*

We propose **Discriminative Cache Matching (DCM)**, a training-free method that improves cache retrieval while preserving the pseudo-labeling and image-cache update rules of the underlying adapter. We first revisit the relation used for image-level retrieval. Since the cache is organized through CLIP's category-aware image–text predictions, discarding this categorical reference during retrieval can make image–image comparison sensitive to visual factors unrelated to the class decision. DCM therefore reintroduces textual category structure into image–image matching: text determines which visual

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

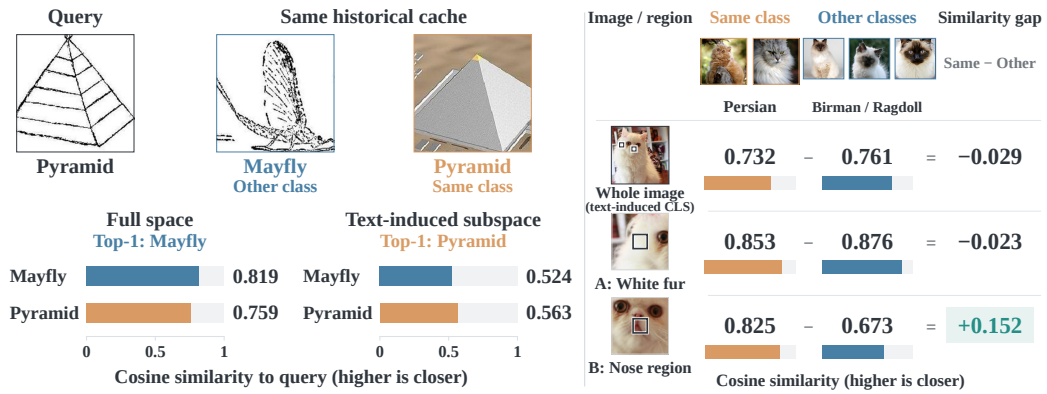

Figure 1: **Why cache retrieval benefits from text guidance and finer granularity. Left:** With the historical cache fixed, full-space image similarity ranks a different-class reference above an available same-class reference, while text-induced matching corrects the ordering. **Right:** Text-induced whole-image matching and shared white fur favor competing classes. In this example, the nose region provides discriminative evidence for distinguishing Persian from Birman and Ragdoll, yielding a positive same-class similarity margin.

directions influence retrieval, while the cached images remain the actual references. We instantiate this principle with a fixed comparison geometry derived from the leading principal directions of a broad reference-text vocabulary. Figure 1 (left) illustrates the resulting change in cache retrieval: full-space similarity favors a different-class reference, whereas text-induced matching reverses the ordering within the same historical cache.

However, improving the image-level relation does not resolve the granularity of retrieval. Both the query and each cached image are still summarized by a single pooled representation. Fine-grained evidence confined to a small region may therefore contribute only weakly to the global affinity even when a strong local correspondence exists between the query and a same-class reference. In Figure 1 (right), the nose region favors the same class even though both the projected whole-image match and the white-fur region favor competing classes. Changing the global projection does not explicitly expose such spatial correspondences because it changes the comparison geometry rather than the spatial granularity of the comparison. DCM therefore extends cache retrieval from image–image matching to patch–patch matching over the same retained images. Rather than introducing additional region encoders, DCM reuses the ViT patch tokens available from the frozen image encoder and treats the retained images as a fine-grained visual memory. An image-level pseudo-label does not identify which patches provide class-specific evidence. Inspired by discriminative patch mining, we select patch representatives from the online cache by contrasting their support within the assigned pseudo-class against that from confusing pseudo-classes. Patch-level retrieval allows different candidate classes to draw evidence from different regions of the query, while competitive matching prevents a strong local correspondence from being evaluated independently of alternative classes.

Together, DCM improves cache retrieval along two complementary axes: *what relation* should determine similarity between images, and *at what granularity* should their evidence be compared. The resulting global and local cache scores are combined with the frozen CLIP prediction, and no model parameters are updated at test time.

Our contributions are threefold:

- We introduce DCM, a training-free method that combines text-induced global matching with class-discriminative patch memory and class-conditioned, competitive patch matching over retained test images.
- We provide a conditional analysis of how projection and re-normalization affect pairwise cache margins, identifying sufficient conditions for correcting misleading reference rankings.
- We evaluate DCM with TDA and MCP on ten cross-domain recognition benchmarks and ImageNet distribution shifts, obtaining consistent improvements while preserving the underlying image-cache construction and update rules.

2 Related Work

2.1 Test-Time Adaptation of Vision-Language Models

Pre-trained vision–language models, exemplified by CLIP (Radford et al., 2021), exhibit strong zero-shot transfer but can suffer from test-time distribution shifts (Shu et al., 2022a; Karmanov et al., 2024). Optimization-based approaches adapt prompts or selected model parameters using unlabeled test samples. TPT (Shu et al., 2022a) minimizes marginal entropy over confidence-selected augmented views, while DiffTPT (Feng et al., 2023) and PromptAlign (Samadh et al., 2023) improve adaptation through diffusion-based augmentation and feature-statistics alignment, respectively. Subsequent methods improve prompt adaptation through calibration (Yoon et al., 2024), or update lightweight model components using objectives based on bimodal alignment (Maharana et al., 2025), contrastive learning (Lafon et al., 2025), or margin-aware adaptation (Huang et al., 2026a). A complementary line explicitly accumulates historical test observations to refine subsequent predictions. TDA (Karmanov et al., 2024) maintains dynamic positive and negative caches without backpropagation, while DMN (Zhang et al., 2024c) organizes historical test features through a dynamic memory. COSMIC (Huang et al., 2025) further models multi-granular cross-modal semantic relations in test-time memory. Related methods represent accumulated test-time information through evolving visual and textual prototypes (Zhang et al., 2024a; Chen et al., 2025; Zhu et al., 2025a), distributional summaries of the incoming test stream (Han et al., 2025), or attributive and associative memory representations (Zhang et al., 2026). Our method follows the training-free cache-based setting, retaining the base pseudo-labeling and image-cache update protocol while focusing on how historical visual evidence is matched to the current query.

2.2 Text-guided geometry for visual matching.

Textual structure has been exploited to refine visual representations and their comparison geometry. APE (Zhu et al., 2023) analyzes downstream inter-class disparity to select informative feature channels and refine cache-based priors for few-shot CLIP adaptation. COLA (Zhu et al., 2025b) projects adversarial image embeddings into a text-induced subspace and further restores cross-modal alignment through optimal transport. SPACE (Manesco et al., 2026) constructs an SVD-based semantic subspace from downstream class texts for domain adaptation, while SubTTA (Zeng et al., 2026) aligns textual and online visual principal subspaces and performs test-time adaptation in the projected space. Herzog and Wang (Herzog & Wang, 2026) study image–image retrieval using principal directions derived from a fixed ImageNet text vocabulary, attributing the resulting gains to reduced task ambiguity rather than intrinsic intra-modal misalignment. In contrast, our method applies text-induced geometry specifically to image–image cache retrieval in TTA, addressing the mismatch between the image–text predictions used to construct the cache and the full-space image–image similarity used to retrieve it.

2.3 Fine-grained visual evidence in CLIP.

Recent vision–language methods exploit multiple views and localized visual evidence beyond a single global representation. BoostAdapter (Zhang et al., 2024b) incorporates regional views of the current image into a lightweight test-time memory, while RITA (Zhu et al., 2026) models augmented visual views as a distribution and performs optimal-transport alignment with textual prototypes together with a dynamic cache for online refinement. At a finer granularity, CDPS (Ye et al., 2026) mines discriminative patches from few-shot support using intra-class consistency and inter-class ambiguity, whereas ABS (Cai et al., 2025) associates local visual evidence with fine-grained textual descriptions through soft matching. ERICT (Dong et al., 2025) further exploits class-related concept tokens to modify the visual attention of the current image.

3 Preliminaries

3.1 Zero-shot classification with CLIP

We consider the standard test-time adaptation setup for CLIP on a C-way classification task. Let $\mathcal{C} = \{1, \ldots, C\}$ denote the set of target classes, for each class $c \in \mathcal{C}$, we construct a class-specific

textual prompt $\mathcal{T}_c$. Given a test image $x \in \mathcal{X}_{\text{test}}$, the pre-trained CLIP visual encoder E_v and text encoder E_t map the visual and textual inputs into a shared D-dimensional embedding space:

$$f(x) = \frac{E_v(x)}{\|E_v(x)\|_2}, \qquad t_c = \frac{E_t(\mathcal{T}_c)}{\|E_t(\mathcal{T}_c)\|_2}, \qquad f(x), t_c \in \mathbb{R}^D. \tag{1}$$

The corresponding zero-shot logits and posterior probability for class c are then given by

$$z_c^{\text{zs}} = \frac{\cos(f(x), t_c)}{\tau}, \qquad P(y = c \mid x) = \frac{\exp(z_c^{\text{zs}})}{\sum_{c'=1}^{C} \exp(z_{c'}^{\text{zs}})}, \tag{2}$$

where $\tau > 0$ denotes the fixed CLIP logit scale.

3.2 Cache-Based Test-Time Adaptation

Cache-based test-time adaptation improves zero-shot CLIP predictions by exploiting information accumulated from previously observed test samples. Let $\mathcal{H}_t = \{(f_i, \hat{y}_i)\}_{i=1}^{N_t}$ denote the cache available at time t, where $f_i \in \mathbb{R}^D$ is a normalized global image feature and $\hat{y}_i \in \{0, 1\}^C$ is its pseudo-label. Given a query feature f_t, conventional cache retrieval measures the full-space similarity

$$a_{t,i}^{\text{cache}} = f_t^\top f_i, \qquad w_{t,i}^{\text{cache}} = \exp\left[-\beta \left(1 - a_{t,i}^{\text{cache}}\right)\right], \tag{3}$$

where $\beta > 0$ controls the sharpness of the similarity kernel. The cache logits are obtained by aggregating the cached pseudo-labels,

$$z_{t,c}^{\text{cache}} = \sum_{i=1}^{N_t} w_{t,i}^{\text{cache}} \hat{y}_i, \qquad z_{t,c}^{\text{TTA}} = z_{t,c}^{\text{zs}} + \alpha z_{t,c}^{\text{cache}}, \tag{4}$$

where $z_{t,c}^{\text{zs}}$ denotes the zero-shot CLIP logits defined in equation 2, and α controls the contribution of the cache. Our method builds on this cache-based paradigm while refining how the current query matches and retrieves global and local evidence from the existing image cache.

4 Method

Figure 2 provides an overview of DCM. Given the image cache maintained by the underlying adapter, DCM improves cache retrieval along two complementary axes: text-induced image–image matching refines the global comparison relation, while patch–patch matching operates at a finer spatial granularity.

4.1 Text-Induced Global Cache Matching

Cache construction is guided by CLIP's image–text predictions, whereas cache retrieval compares retained samples through image–image similarity. A test image can be confidently assigned to a class by the former, yet this confidence does not imply that it is close to other samples of the same class in the full visual embedding space. Same-class images can differ substantially in context, background, style, or appearance, while different-class images may share these factors. Consequently, full-space image–image similarity can distort the relative ranking of cached references. As a result, the geometry used to retrieve the cache need not preserve the class relation that was used to construct it.

We therefore use a separate fixed reference vocabulary to determine which directions should contribute to global image–image matching. Its class text embeddings provide a category-level reference in the shared CLIP space, but some directions vary much more across reference classes than others. Inspired by the text-based PCA analysis of CLIP (Herzog & Wang, 2026), we retain the dominant directions of this cross-class variation and use the resulting truncated subspace to constrain global cache matching.

Specifically, we encode a fixed reference vocabulary of C_r classes with the frozen CLIP text encoder and obtain normalized text representations $\{r_j\}_{j=1}^{C_r}$, $r_j \in \mathbb{R}^D$. After centering the reference embeddings, we retain their top d principal directions,

$$U = [u_1, \ldots, u_d] \in \mathbb{R}^{D \times d}, \qquad U^\top U = I. \tag{5}$$

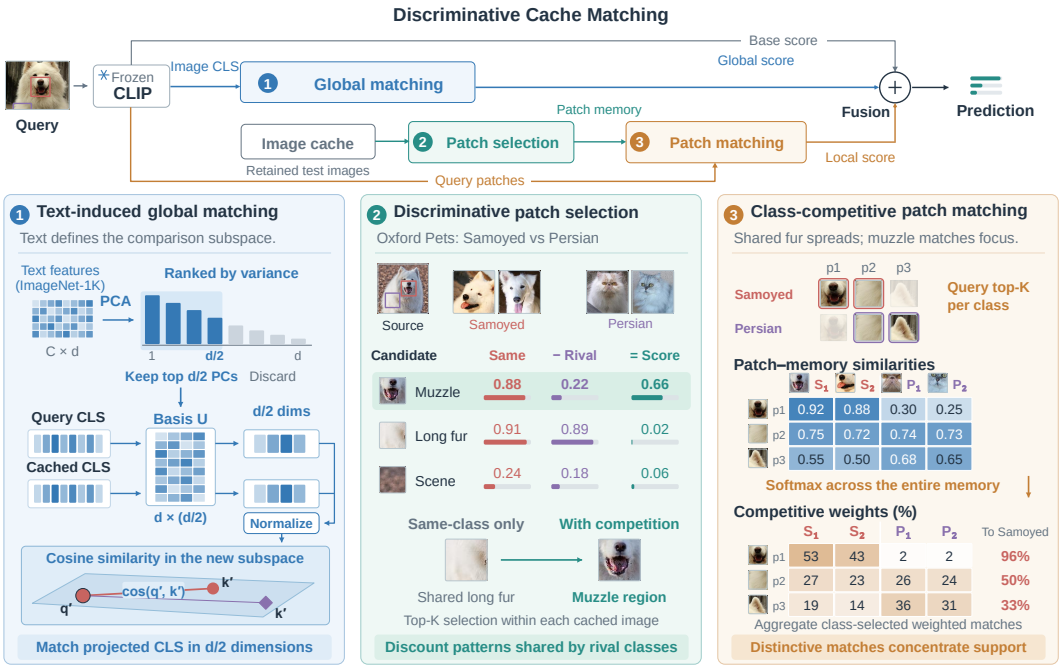

Figure 2: **Overview of Discriminative Cache Matching.** (1) Text-induced global matching retains the top (d/2) text principal components and computes cosine similarity between projected, re-normalized query and cached CLS features. (2) Discriminative patch selection contrasts same-class support with support from confusing classes, favoring discriminative regions over shared patterns, as illustrated by muzzle regions versus long fur. (3) Class-competitive patch matching selects query patches per class and aggregates similarities weighted by a softmax over the entire patch memory.

The columns of U capture the dominant directions of cross-class variation among the reference text prototypes and define the subspace used for global cache matching. We instantiate U using ImageNet-1K reference texts. Using a large reference vocabulary also avoids the rank limitation of target-text PCA: for a target vocabulary of C classes, the centered text covariance has rank at most $C - 1$. Given an image representation f, we project it onto the retained text-induced subspace and re-normalize:

$$\phi(f) = \frac{U^\top f}{\|U^\top f\|_2}. \tag{6}$$

Global cache affinity remains image–image, but is now computed in this restricted space:

$$a^g_{t,i} = \phi(f_t)^\top \phi(f_i), \tag{7}$$

replacing the full-space cosine similarity used in equation 3. A conditional analysis of when projection and re-normalization can correct pairwise cache-reference rankings is provided in Appendix A.

We replace $a^{\text{cache}}_{t,i}$ in equation 3 with the projected affinity $a^{\text{g}}_{t,i}$ for positive and negative cache retrieval, yielding $z^{\text{global}}_{t,c}$. Cache values, kernel parameters, and aggregation weights remain unchanged.

4.2 Class-discriminative Patch Cache Matching

The global branch regulates which feature directions contribute to image cache matching, but each cached image is still represented by a global feature, which can dilute fine-grained evidence confined to small object regions. One option is to explicitly encode selected regions or crops, as in region-based adaptation methods (Zhang et al., 2024b), but this requires additional visual encoding and is sensitive to localization errors, inaccurate regions may retain excessive context or miss discriminative details. We instead reuse the patch tokens already produced by the frozen ViT in the forwards required by the base protocol, avoiding separate region encoding.

Using these tokens requires identifying which local patterns provide useful class evidence, since patches may also capture backgrounds, generic object parts, and other patterns shared across classes.

Moreover, CLIP is trained with image-level image–text supervision rather than explicit class-specific patch–text correspondence (Zhong et al., 2022), making patch–text similarity an indirect signal for discriminative patch selection. We therefore use cached test samples themselves as visual references. A useful local pattern should recur across images assigned to the same pseudo-class, while receiving weaker support from confusing pseudo-classes. We use this criterion to select class-discriminative local patch evidence from the current image cache. These discriminative patches are then matched to the current image through class-conditioned, class-competitive patch matching.

Class-Discriminative Patch Cache Selection. We derive the local patch cache directly from the current positive image cache. Let $\mathcal{H}_c$ denote the cached images currently assigned to pseudo-class c, and let $\mathcal{F}_i$ denote the candidate patches associated with cached image i; we omit the test-step index below for clarity.

We first define a patch-to-cache affinity that measures the support of a candidate patch across a set of cached images. For a candidate patch q and a reference image set $\mathcal{R}$,

$$A(q, \mathcal{R}) = \frac{1}{|\mathcal{R}|} \sum_{j \in \mathcal{R}} \max_{u \in \mathcal{F}_j} q^\top u. \tag{8}$$

Each reference image contributes through its best-matching local pattern, so $A(q, \mathcal{R})$ is large when similar local evidence for q is repeatedly observed across $\mathcal{R}$. When q is itself drawn from the cache, its source image is omitted from the reference set to avoid trivial self-matching.

For a candidate patch $q \in \mathcal{F}_i$ from cached image i, let c denote its pseudo-label. For a candidate patch q assigned to pseudo-class c, $A(q, \mathcal{H}_c)$ measures its support from the same-class cache. However, a large value alone does not imply class specificity, since background, context, and generic object parts may also receive strong support from competing class caches. We therefore contrast the same-class support with the average support from confusing classes:

$$D(q, c) = A(q, \mathcal{H}_c) - \eta \frac{1}{|\mathcal{N}_c|} \sum_{c' \in \mathcal{N}_c} A(q, \mathcal{H}_{c'}). \tag{9}$$

where $\mathcal{N}_c \subseteq \mathcal{C} \setminus \{c\}$ denotes the top-N confusing classes considered for class c. Thus, $D(q, c)$ favors local patterns that recur within the assigned pseudo-class but are less prevalent among its competitors.

For each cached image $i \in \mathcal{H}_c$, we retain the K_{mem} patches in $\mathcal{F}_i$ with the largest $D(q, c)$. We pool the selected patches across all images in $\mathcal{H}_c$ and apply greedy coverage selection to reduce redundancy. The resulting class-wise patch cache $\mathcal{P}_c$ contains at most M representatives. We rebuild $\mathcal{P}_c$ whenever the membership of $\mathcal{H}_c$ changes.

Class-Competitive Patch Matching. Given the class-wise patch caches $\{\mathcal{P}_c\}_{c=1}^{C}$, we estimate local class evidence for image i from its patch set $\mathcal{F}_i$. Class-specific evidence may be confined to only a few patches, while aggregating the full patch set can dilute informative local matches. For each class c, we therefore select the K_{qry} patches in $\mathcal{F}_i$ with the strongest match to $\mathcal{P}_c$:

$$\mathcal{Q}_{i,c} = \text{TopK}_{q \in \mathcal{F}_i}^{K_{\text{qry}}} \left[\max_{\pi \in \mathcal{P}_c} q^\top \pi \right]. \tag{10}$$

Hence, different classes can retrieve evidence from different regions of the image.

A selected query patch may still match patterns stored by several classes. Inspired by the soft-matching formulation of ABS (Cai et al., 2025), for each $q \in \mathcal{Q}_{i,c}$ we compute a soft assignment over all patches currently stored in the local cache. The weight assigned to a cached patch π is

$$w_q(\pi) = \frac{\exp\big(\beta q^\top \pi\big)}{\sum_{d=1}^{C} \sum_{\pi' \in \mathcal{P}_d} \exp\big(\beta q^\top \pi'\big)}. \tag{11}$$

Since the normalization spans all class-wise patch caches, patches from different classes compete for the same query patch. We then aggregate the assignment-weighted similarities over the patches stored for class c:

$$z_{i,c}^{\text{loc}} = \frac{\beta}{|\mathcal{Q}_{i,c}|\,|\mathcal{P}_c|} \sum_{q \in \mathcal{Q}_{i,c}} \sum_{\pi \in \mathcal{P}_c} w_q(\pi)\, q^\top \pi. \tag{12}$$

4.3 Overall Prediction

We combine the base zero-shot score with global and local cache evidence:

$$z_{t,c}^{\text{final}} = z_{t,c}^{zs} + \alpha_{\text{g}} z_{t,c}^{\text{global}} + \alpha_{\text{loc}} z_{t,c}^{\text{loc}}, \qquad \widehat{y}_t = \arg\max_{c \in \mathcal{C}} z_{t,c}^{\text{final}}, \tag{13}$$

where $\alpha_{\text{g}} \geq 0$ and $\alpha_{\text{loc}} \geq 0$ control the global and local cache contributions, respectively.

5 Experiment

5.1 Experimental Setup

Datasets. Following prior CLIP test-time adaptation studies (Shu et al., 2022a; Karmanov et al., 2024), we evaluate the proposed method on fifteen datasets under two benchmark settings: cross-dataset classification and natural distribution shifts. The cross-dataset benchmark comprises Caltech101 (Fei-Fei et al., 2004), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), FGVC-Aircraft (Maji et al., 2013), Food101 (Bossard et al., 2014), Flowers102 (Nilsback & Zisserman, 2008), Oxford-IIIT Pets (Parkhi et al., 2012), Stanford Cars (Krause et al., 2013), SUN397 (Xiao et al., 2010), and UCF101 (Soomro et al., 2012). For natural distribution shifts, we evaluate on ImageNet (Deng et al., 2009) and four shifted variants: ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a), ImageNet-Sketch (Wang et al., 2019), and ImageNet-V2 (Recht et al., 2019). We report top-1 classification accuracy (%) on each dataset.

Implementation Details. We conduct experiments using pretrained CLIP models with ResNet-50 and ViT-B/16 visual backbones, keeping both the visual and text encoders frozen. Following recent practice in cache-based TTA (Chen et al., 2025; Wu et al., 2026), we construct text classifiers using hand-crafted prompts and CuPL (Pratt et al., 2023). DCM serves as a plug-in for cache-based TTA, and we evaluate its integration with TDA and MCP. For global matching, the projection basis is computed from ImageNet-1K reference text embeddings and remains fixed across datasets. We use a projection rank of $d = 256$. For local matching, we set $K_{\text{mem}} = K_{\text{qry}} = 15$, $N = 5$, $M = 12$ and set the inter-class penalty coefficient to $\eta = 0.1$. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

Baselines. We compare DCM with zero-shot CLIP (Radford et al., 2021) and representative TTA methods covering two main adaptation paradigms. The first optimizes prompts or learnable residuals at test time, including TPT (Shu et al., 2022a), DiffTPT (Feng et al., 2023), DPE (Zhang et al., 2024a), and DLAE (Wu et al., 2026). The second adapts predictions through updates to memories, class prototypes, or distributional statistics, including TDA (Karmanov et al., 2024), DMN-ZS (Zhang et al., 2024c), BoostAdapter (Zhang et al., 2024b), BCA (Zhou et al., 2025), MCP (Chen et al., 2025), ADAPT (Zhang et al., 2025), PTA (Huang et al., 2026b), and A^{2}Memory (Zhang et al., 2026). Following recent practice in cache-base TTA (Chen et al., 2025; Wu et al., 2026), we supplement TDA with CuPL prompts (Pratt et al., 2023) and adopt the corresponding augmentation operators whe multi-view inference is used. We denote this configuration as TDA†, while retaining TDA's original number of views.

5.2 Main Results

Results on Cross-Domain Datasets. Tables 1 and 5 report results on ten cross-domain recognition datasets. Additional results with RN-50 are provided in Appendix D.1. DCM+MCP achieves the highest average accuracy among all compared methods, reaching 73.71% with ViT-B/16 and 66.31% with ResNet-50. It outperforms DLAE, the strongest backpropagation-based baseline, by 1.83 and 1.50 pts., respectively, while requiring no test-time optimization. Compared with A^{2}Memory, DCM+MCP improves average accuracy by 2.30 and 2.75 pts.. MCP provides the strongest baseline performance on both backbones, and integrating DCM further improves its average accuracy by 1.73 and 1.23 pts., with gains on all ten datasets. The benefits also extend to TDA†: with ViT-B/16, DCM improves accuracy on every dataset and raises the average from 69.63% to 72.55%. These results demonstrate both the competitive performance of DCM and its effectiveness when integrated with different cache-based adapters.

Table 1: Results on cross-domain recognition datasets using CLIP ViT-B/16. The best result in each column is highlighted in bold.

Method	Caltech	DTD	EuroSAT	Flowers	Pets	UCF101	Food101	SUN397	Aircraft	Cars	Average
CLIP-ViT-B/16	93.35	44.27	42.01	67.44	88.25	65.13	83.65	62.59	23.67	65.48	63.58
Backpropagation-Based											
TPT (Shu et al., 2022a)	94.16	47.75	42.44	68.98	87.79	68.04	84.67	65.50	24.78	66.87	65.10
DiffTPT (Feng et al., 2023)	92.49	47.00	43.13	70.10	88.22	62.67	87.23	65.74	25.60	67.01	65.47
DPE (Zhang et al., 2024a)	94.81	54.20	55.79	75.07	91.14	70.44	86.17	70.07	28.95	67.31	69.40
DLAE (Wu et al., 2026)	95.33	**58.86**	61.30	77.50	91.88	73.46	87.05	70.83	32.58	69.99	71.88
Backpropagation-Free											
TDA (Karmanov et al., 2024)	94.24	47.40	58.00	71.42	88.63	70.66	86.14	67.62	23.91	67.28	67.53
TDA† (Karmanov et al., 2024)	94.16	53.61	61.79	75.36	90.71	70.90	86.19	68.91	28.32	66.4	69.63
DMN-ZS (Zhang et al., 2024c)	95.38	55.85	59.43	74.49	92.04	72.51	85.08	70.18	30.03	67.96	70.30
BoostAdapter (Zhang et al., 2024b)	94.77	45.69	61.22	71.66	89.51	71.93	87.17	68.09	27.45	69.30	68.68
BCA (Zhou et al., 2025)	94.69	53.49	56.63	73.12	90.43	67.59	85.97	68.41	28.59	66.86	68.59
MCP (Chen et al., 2025)	95.33	55.91	68.42	75.88	91.88	74.36	86.85	71.04	30.18	69.99	71.98
ADAPT (Zhang et al., 2025)	94.48	55.20	68.19	75.56	92.01	70.66	83.81	70.57	28.95	68.19	70.76
PTA (Huang et al., 2026b)	94.81	47.70	61.57	75.23	91.06	73.12	86.44	69.21	26.10	68.55	69.38
A^{2}Memory (Zhang et al., 2026)	95.01	57.28	68.19	76.11	92.14	70.66	86.19	70.43	29.92	68.20	71.41
Ours+TDA†	94.81	56.21	67.33	76.78	92.72	**76.02**	87.12	70.64	32.28	71.63	72.55
Ours+MCP	**95.54**	58.81	**70.79**	**78.12**	**93.13**	75.92	**87.62**	**71.82**	**33.00**	**72.33**	**73.71**

Table 2: Results on ImageNet and out-of-distribution benchmarks using CLIP ViT-B/16. The best result in each column is highlighted in bold.

Method	ImageNet-A	ImageNet-R	ImageNet-S	ImageNet-V2	ImageNet	OOD Avg.	Average
CLIP-ViT-B/16	47.87	73.98	46.09	60.86	66.73	57.20	59.11
Backpropagation-Based							
TPT (Shu et al., 2022a)	54.77	77.06	47.94	63.45	68.98	60.81	62.44
DiffTPT (Feng et al., 2023)	55.68	75.00	46.80	65.10	70.30	60.52	62.28
DPE (Zhang et al., 2024a)	59.63	80.40	52.26	65.44	71.91	64.43	65.93
DLAE (Wu et al., 2026)	64.06	80.43	53.06	65.50	72.41	65.76	67.09
Backpropagation-Free							
TDA (Karmanov et al., 2024)	60.11	80.24	50.54	64.67	69.51	63.89	65.01
TDA† (Karmanov et al., 2024)	63.56	80.94	52.20	65.23	71.77	65.48	66.74
DMN-ZS (Zhang et al., 2024c)	58.28	78.55	53.20	65.17	72.25	63.80	65.49
BCA (Zhou et al., 2025)	61.14	80.72	50.87	64.90	70.22	64.16	65.37
MCP (Chen et al., 2025)	59.71	81.59	53.89	65.16	72.37	65.09	66.54
PTA (Huang et al., 2026b)	61.15	80.79	51.00	64.85	70.28	64.45	65.61
ADAPT (Zhang et al., 2025)	63.32	80.66	53.13	64.64	70.91	65.44	66.53
A^{2}Memory (Zhang et al., 2026)	64.38	80.98	53.21	65.38	72.21	65.99	67.23
Ours+TDA†	**65.27**	81.78	54.16	66.00	72.69	**66.80**	**67.98**
Ours+MCP	60.47	**82.85**	**55.52**	**66.11**	**73.37**	66.24	67.66

Results on OOD Datsets. Tables 2 and 6 report results on ImageNet and its four shifted variants. Additional results with ResNet-50 are provided in Appendix D.1. DCM+TDA† achieves the highest average accuracy across the four OOD benchmarks, reaching 66.80% with ViT-B/16 and 49.31% with ResNet-50. It outperforms A^{2}Memory by 0.81 and 0.68 pts., respectively, and exceeds DLAE by 1.04 and 0.25 pts. without test-time optimization. DCM+MCP also improves upon MCP on every OOD benchmark with both backbones. Across individual datasets, DCM+TDA† achieves the highest accuracy on ImageNet-A and ImageNet-V2, while DCM+MCP performs best on ImageNet-R and ImageNet-Sketch; this pattern holds for both backbones. The improvements extend to ImageNet, where DCM+MCP achieves 73.37% and 64.90% accuracy with ViT-B/16 and ResNet-50, surpassing MCP by 1.00 and 0.71 pts., respectively.

5.3 Ablation Studies

Effect of individual Branches. Table 3 evaluates the contributions of the global and local branches using TDA† with ViT-B/16. Global matching and the local branch individually improve average accuracy by 1.34 and 2.06 pts., respectively. Combining both achieves 72.55% accuracy, a gain

Table 3: Ablation of text-induced global matching and local patch matching using TDA† with CLIP ViT-B/16. We report mean accuracy (%) across ten cross-domain datasets.

Text-induced global matching	Local Patch matching	Avg. acc. (%)
–	–	69.63
✓	–	70.97
–	✓	71.69
✓	✓	**72.55**

Table 4: Effect of the reference vocabulary. IN subsets are fixed random subsets of IN-1K.

Reference vocabulary	No. of classes	Avg. acc. (%)
No projection	–	71.69
Target classes	C	71.70
ImageNet subset	250	72.33
ImageNet subset	500	72.50
ImageNet subset	750	72.51
ImageNet-1K	1,000	**72.55**

of 2.92 pts. over the baseline. Adding global matching to the local branch provides a further improvement of 0.86 pts., while incorporating the local branch into global matching yields an additional 1.58 pts. These gains demonstrate the complementary contributions of text-induced global matching and local patch evidence.

Effect of Projection Rank and Component Selection. Figure 3 compares the leading d text principal components (Top-d) with d components randomly selected from the same PCA basis (Random-d), keeping the local branch enabled throughout. Top-d outperforms Random-d at every reduced rank. At our default rank of $d = 256$, Top-d achieves 72.55% accuracy, exceeding Random-d and full-dimensional matching by 0.89 and 0.86 pts., respectively. Accuracy remains between 72.33% and 72.55% across the tested ranks from 128 to 384, while retaining only 32 components lowers accuracy to 71.25%. These results demonstrate the benefit of prioritizing leading text principal components for cache matching, with stable performance across intermediate ranks.

Effect of the reference vocabulary. As shown in Table 4, constructing the text-induced basis from the target classes alone provides virtually no improvement over full-space matching (71.70% vs. 71.69%). The performance increases to 72.33% with 250 ImageNet reference classes and to 72.50% with 500 classes, after which the gain largely saturates. Using the complete ImageNet-1K vocabulary achieves the best average accuracy of 72.55%. This result indicates that effective text-induced cache matching requires a sufficiently broad categorical reference rather than the target label set alone. Accordingly, we use ImageNet-1K as the fixed reference vocabulary.

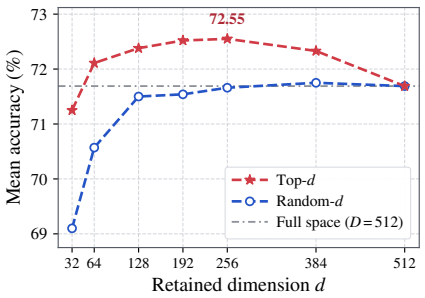

Figure 3: Cross-domain accuracy at different projection ranks.

6 Conclusion

This paper separates cache construction from cache retrieval in training-free test-time adaptation of CLIP and focuses on how retained visual evidence should be matched to future queries. We propose Discriminative Cache Matching (DCM), which improves cache retrieval along two complementary axes. At the image level, DCM introduces a text-induced comparison geometry to better align image–image retrieval with the category structure used to organize the cache. At the local level, it increases the retrieval granularity from images to patches, allowing localized evidence to be matched and evaluated relative to competing class memories. We further provide a conditional margin analysis that characterizes when projection and re-normalization can correct misleading cache-reference rankings. Experiments on cross-domain recognition and natural distribution shifts show consistent improvements when DCM is integrated with both TDA and MCP, while ablations confirm the complementary benefits of global and local retrieval. DCM remains affected by cache errors and limited coverage, motivating future work on jointly improving cache construction and retrieval.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *Computer Vision – ECCV 2014*, volume 8694 of *Lecture Notes in Computer Science*, pp. 446–461. Springer, 2014. doi: 10.1007/978-3-319-10599-4_29. URL https://doi.org/10.1007/978-3-319-10599-4_29.

Lincan Cai, Jingxuan Kang, Shuang Li, Wenxuan Ma, Binhui Xie, Zhida Qin, and Jian Liang. From local details to global context: Advancing vision-language models with attention-based selection. In *International Conference on Machine Learning*, 2025.

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

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. doi: 10.1109/CVPR.2014.461. URL https://doi.org/10.1109/CVPR.2014.461.

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. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Xinpeng Dong, Min Zhang, Didi Zhu, Ye Jun Jian, Zhang Keli, Aimin Zhou, Fei Wu, and Kun Kuang. Erict: Enhancing robustness by identifying concept tokens in zero-shot vision language models. In *Forty-second International Conference on Machine Learning*, 2025.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*. IEEE, 2004. doi: 10.1109/CVPR.2004.383. URL https://doi.org/10.1109/CVPR.2004.383.

Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2704–2714, 2023.

Zongbo Han, Jialong Yang, Guangyu Wang, Junfan Li, Qianli Xu, Mike Zheng Shou, and Changqing Zhang. Dota: Distributional test-time adaptation of vision-language models. In *Advances in Neural Information Processing Systems*, 2025.

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. doi: 10.1109/JSTARS.2019.2918242. URL https://doi.org/10.1109/JSTARS.2019.2918242.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. 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. doi: 10.1109/ICCV48922.2021.00823. URL https://doi.org/10.1109/ICCV48922.2021.00823.

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. doi: 10.1109/CVPR46437.2021.01501. URL https://doi.org/10.1109/CVPR46437.2021.01501.

Jonas Herzog and Yue Wang. Reevaluating the intra-modal misalignment hypothesis in CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Fanding Huang, Jingyan Jiang, Qinting Jiang, Hebei Li, Faisal Nadeem Khan, and Zhi Wang. Cosmic: Clique-oriented semantic multi-space integration for robust clip test-time adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9772–9781, 2025.

Yan Huang, Guowei Wang, Xu Wang, Kangjun Liu, and Xin Lin. Local margin restoration for test-time adaptation of vision-language models. In *arXiv preprint arXiv:2608.02216*, 2026a.

Zhaohong Huang, Yuxin Zhang, Wenjing Liu, Fei Chao, and Rongrong Ji. Prototype-based test-time adaptation of vision-language models, 2026b.

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

Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. In *Second Workshop on Fine-Grained Visual Categorization*, 2013. URL https://ai.stanford.edu/~jkrause/papers/fgvc13.pdf.

Marc Lafon, Gustavo Vargas Hakim, Clément Rambour, Christian Desrosiers, and Nicolas Thome. Cliptta: Robust contrastive vision-language test-time adaptation. In *Advances in Neural Information Processing Systems*, 2025.

Sarthak Maharana, Baoming Zhang, Leonid Karlinsky, Rogerio Feris, and Yunhui Guo. Batclip: Bimodal online test-time adaptation for clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft, 2013. URL https://arxiv.org/abs/1306.5151.

João Renato Ribeiro Manesco, Danilo Samuel Jodas, Douglas Rodrigues, Leandro Aparecido Passos, and João Paulo Papa. Space: Semantic projection and alignment of clip embeddings for domain adaptation. In *2026 IEEE International Conference on Image Processing (ICIP)*, 2026.

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. doi: 10.1109/ICVGIP.2008.47. URL https://doi.org/10.1109/ICVGIP.2008.47.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505. IEEE, 2012. doi: 10.1109/CVPR.2012.6248092. URL https://doi.org/10.1109/CVPR.2012.6248092.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. 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*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Jameel Hassan Abdul Samadh, Hanan Gani, Noor Hazim Hussein, Muhammad Uzair Khattak, Muzammal Naseer, Fahad Khan, and Salman Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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, 2022a.

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. In *Advances in Neural Information Processing Systems*, 2022b.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. Technical Report CRCV-TR-12-01, Center for Research in Computer Vision, University of Central Florida, November 2012. URL https://www.crcv.ucf.edu/wp-content/uploads/2019/03/UCF101_CRCV-TR-12-01.pdf.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, volume 32, pp. 10506–10518, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3eefceb8087e964f89c2d59e8a249915-Abstract.html.

Haoyan Wu, Yahao Liu, Yinjie Lei, Lixin Duan, and Wen Li. Dynamic logits adjustment and exploration for test-time adaptation in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

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. doi: 10.1109/CVPR.2010.5539970. URL https://doi.org/10.1109/CVPR.2010.5539970.

Zehao Xiao, Shilin Yan, Jack Hong, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiayi Shen, Qi Wang, and Cees G. M. Snoek. DynaPrompt: Dynamic test-time prompt tuning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhenzhang Ye, Duolikun Danier, Bo Zhao, and Hakan Bilen. Training-free discriminative patch mining for robust few-shot recognition with clip. In *European Conference on Computer Vision*, 2026.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang Yoo. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *International Conference on Learning Representations*, 2024.

Zhichen Zeng, Wenxuan Bao, Xiao Lin, Ruizhong Qiu, Tianxin Wei, Xuying Ning, Yuchen Yan, Chen Luo, Monica Xiao Cheng, Jingrui He, et al. Subspace alignment for vision-language model test-time adaptation. *arXiv preprint arXiv:2601.08139*, 2026.

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

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

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

Youjia Zhang, Youngeun Kim, Young-Geun Choi, Hongyeob Kim, Huiling Liu, and Sungeun Hong. Backpropagation-free test-time adaptation via probabilistic gaussian alignment. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yuchao Zhang, Hao Wang, Fan Zhang, Qirui Mi, Mengyue Yang, Yisen Wang, Jun Wang, Haoxuan Li, and Zhouchen Lin. Learning to memorize with attributive and associative memory for online test-time adaptation of vision-language models. In *Forty-third International Conference on Machine Learning*, 2026.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

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*, 2025.

Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2605–2615, October 2023.

Xingyu Zhu, Shuo Wang, Beier Zhu, Miaoge Li, Yunfan Li, Junfeng Fang, Zhicai Wang, Dongsheng Wang, and Hanwang Zhang. Dynamic multimodal prototype learning in vision-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2501–2511, 2025a.

Xingyu Zhu, Beier Zhu, Shuo Wang, Kesen Zhao, and Hanwang Zhang. Enhancing clip robustness via cross-modality alignment. In *Advances in Neural Information Processing Systems*, 2025b.

Xingyu Zhu, Huanshen Wu, Shuo Wang, Beier Zhu, Jiannan Ge, Jiaheng Zhang, and Long Chen. Robustifying vision-language models via test-time prompt adaptation. In *Forty-third International Conference on Machine Learning*, 2026.

USE OF LARGE LANGUAGE MODELS

In this work, we used generative AI tools, including ChatGPT and Kimi, to polish the language and improve the readability of the manuscript. We have reviewed all AI-assisted work. Specifically, we checked that the language edits preserve the intended technical meaning and do not alter the scientific claims. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

A CONDITIONAL ANALYSIS OF PAIRWISE CACHE RETRIEVAL

Retaining the dominant text directions does not guarantee that the resulting subspace coincides with the task-relevant geometry of a downstream dataset. We therefore analyze the projection conditionally, asking when the text-induced geometry preserves the category relation required by cache retrieval. Let $P^\star$ denote an ideal orthogonal projector that preserves only task-relevant variation. For a unit image representation x, we write

$$x = a_x u_x + r_x, \qquad a_x = \|P^\star x\|_2, \qquad u_x = \frac{P^\star x}{\|P^\star x\|_2}, \qquad r_x = (I - P^\star)x. \tag{14}$$

Here, u_x is the normalized task-relevant direction, a_x is the magnitude retained in this subspace, and r_x contains the remaining visual variation. For a query q, a same-class reference k^+, and a different-class reference k^-, we define the ideal task-relevant margin as

$$\Delta^\star = u_q^\top u_+ - u_q^\top u_-. \tag{15}$$

A positive $\Delta^\star$ means that the same-class reference is preferred when only task-relevant variation is considered.

In the original image space, the corresponding similarity margin admits the exact decomposition

$$\Delta_I = q^\top k^+ - q^\top k^- = a_q a_+ \Delta^\star + \underbrace{a_q(a_+ - a_-)u_q^\top u_-}_{B_{\text{amp}}} + \underbrace{r_q^\top (r_+ - r_-)}_{B_{\text{res}}}. \tag{16}$$

Thus, even when $\Delta^\star > 0$, the raw image similarity can rank the different-class reference higher if residual visual similarity B_{res} or unequal projected magnitudes B_{amp} overwhelm the task-relevant margin. This decomposition also separates the roles of the two operations in equation 6: under

the ideal projector, projection removes the direct residual contribution B_{res}, while re-normalization removes the magnitude factors a_x responsible for B_{amp}. Their combination therefore reduces the pairwise comparison to $\Delta^\star$.

The practical text-induced projector $P = UU^\top$ need not recover $P^\star$ exactly. Instead, we measure its distortion only on the samples participating in a cache relation:

$$e_x = \frac{\|(P - P^\star)x\|_2}{\|P^\star x\|_2}. \tag{17}$$

Let $\Delta_P = \phi(q)^\top \phi(k^+) - \phi(q)^\top \phi(k^-)$ denote the corresponding margin in the text-induced space.

Proposition 1 (Sample-wise margin stability)**.** *Assume $P^\star x \neq 0$ and $e_x < 1$ for $x \in \{q, k^+, k^-\}$, and let $\varepsilon_{\mathrm{tri}} = 4e_q + 2e_+ + 2e_-$. Then*

$$|\Delta_P - \Delta^\star| \leq \varepsilon_{\mathrm{tri}}. \tag{18}$$

The proof is provided in Appendix B.

Proposition 1 shows that the projected margin remains close to the ideal task-relevant margin when the text-induced projector has small distortion on the relevant samples. In particular, $\Delta^\star > \varepsilon_{\mathrm{tri}}$ guarantees $\Delta_P > 0$. Combined with equation 16, this characterizes a correctable failure mode in which structured visual variation yields $\Delta_I < 0$, while the projected metric preserves the same-class ordering with $\Delta_P > 0$. Importantly, the condition is sample-local and does not require P to recover $P^\star$ uniformly over the entire embedding space. Because the global branch changes only the query–cache affinity and the cache kernel is monotonic in this affinity, the corrected pairwise ordering is directly inherited by the corresponding relative cache weights.

B Proof of Proposition 1

We first establish a stability result for the normalized projected representation.

Lemma 1 (Normalized projection stability)**.** *Let*

$$u_x = \frac{P^\star x}{\|P^\star x\|_2}, \qquad \hat{u}_x = \frac{Px}{\|Px\|_2},$$

and let

$$e_x = \frac{\|(P - P^\star)x\|_2}{\|P^\star x\|_2}.$$

If $P^\star x \neq 0$ and $e_x < 1$, then $\hat{u}_x$ is well defined and

$$\|\hat{u}_x - u_x\|_2 \leq 2e_x. \tag{19}$$

Proof. By the reverse triangle inequality,

$$\|Px\|_2 \geq \|P^\star x\|_2 - \|(P - P^\star)x\|_2 = (1 - e_x)\|P^\star x\|_2 > 0,$$

so $\hat{u}_x$ is well defined. Moreover,

$$\begin{aligned}
\|\hat{u}_x - u_x\|_2 &= \left\| \frac{Px}{\|Px\|_2} - \frac{P^\star x}{\|P^\star x\|_2} \right\|_2 && (20)\\
&\leq \left\| \frac{Px}{\|Px\|_2} - \frac{Px}{\|P^\star x\|_2} \right\|_2 + \frac{\|(P - P^\star)x\|_2}{\|P^\star x\|_2} && (21)\\
&= \frac{\big|\|P^\star x\|_2 - \|Px\|_2\big|}{\|P^\star x\|_2} + e_x && (22)\\
&\leq 2e_x. && (23)
\end{aligned}$$

□

Algorithm 1 Discriminative Cache Matching (DCM)

Require: Frozen CLIP; base cache adapter $\mathcal{B}$; pre-computed text-induced basis U; $K_{\mathrm{mem}}, K_{\mathrm{qry}}, N, \eta, \beta$; $\alpha_{\mathrm{g}}, \alpha_{\mathrm{loc}}$
Ensure: Online predictions $\{\widehat{y}_t\}_{t=1}^{T}$
Initialize the class-wise patch caches $\mathcal{P}_c \leftarrow \emptyset$ for all $c \in \mathcal{C}$
for each test image x_t **do**
  Obtain the Base zero-shot score z_t^{zs}, global feature f_t, candidate patches $\mathcal{F}_t$, and current image-cache state from $\mathcal{B}$
  Text-induced global matching
  Project and re-normalize the query and cached global features using equation 6
  Compute projected image–image affinities using equation 7
  Replace the original cache affinities in the base cache rule with $a_{t,i}^{\mathrm{g}}$ to obtain z_t^{global}
  Class-discriminative patch cache selection
  for each class c whose positive image cache $\mathcal{H}_c$ has changed **do**
    Compute patch-to-cache affinities $A(q, \cdot)$ using equation 8
    Compute class-discriminative scores $D(q, c)$ using equation 9
    For each $i \in \mathcal{H}_c$, retain the K_{mem} patches in $\mathcal{F}_i$ with the largest $D(q, c)$
    Rebuild $\mathcal{P}_c$ by pooling the retained patches from all $i \in \mathcal{H}_c$
  end for
  Class-competitive patch matching
  for each class c with $\mathcal{P}_c \neq \emptyset$ **do**
    Select class-conditioned query patches $\mathcal{Q}_{t,c}$ using equation 10
    Compute competitive patch weights $w_q(\pi)$ using equation 11
    Compute $z_{t,c}^{\mathrm{loc}}$ using equation 12
  end for
  Compute z_t^{final} and $\widehat{y}_t$ using equation 13
  Update the underlying image cache according to the original rule of $\mathcal{B}$
end for

We now prove Proposition 1. Let $\hat{u}_q$, $\hat{u}_+$, and $\hat{u}_-$ denote the normalized projections under P. Using Lemma 1 and the fact that all normalized representations have unit norm,

$$\left|\hat{u}_q^\top \hat{u}_+ - u_q^\top u_+\right| \leq \|\hat{u}_q - u_q\|_2 + \|\hat{u}_+ - u_+\|_2 \tag{24}$$

$$\leq 2e_q + 2e_+, \tag{25}$$

and similarly,

$$\left|\hat{u}_q^\top \hat{u}_- - u_q^\top u_-\right| \leq 2e_q + 2e_-. \tag{26}$$

Therefore,

$$|\Delta_P - \Delta^\star| \leq \left|\hat{u}_q^\top \hat{u}_+ - u_q^\top u_+\right| + \left|\hat{u}_q^\top \hat{u}_- - u_q^\top u_-\right| \tag{27}$$

$$\leq 4e_q + 2e_+ + 2e_- \tag{28}$$

$$= \varepsilon_{\mathrm{tri}}. \tag{29}$$

C Algorithm

We present the overall algorithm for our proposed method DCM in Algorithm.1.

D Additional Experiments

D.1 Main results on CLIP RN50

Local Feature Extraction for ResNet-50. For CLIP ResNet-50, we obtain local representations from the 49 spatial locations of the final 7×7 convolutional feature map. The feature map is flattened into spatial tokens, whose representations are extracted at the frozen attention-pooling stage for local matching. We retain the same proportion of spatial tokens as in the ViT-B/16 configuration and apply the same class-discriminative patch cache construction and competitive matching procedure.

We complement the CLIP ViT-B/16 experiments in the main text with results using CLIP ResNet-50. We evaluate DCM integrated with TDA† and MCP on cross-dataset benchmarks and natural distribution shifts, reporting top-1 accuracy (%) for each dataset.

Table 5: Results on cross-domain recognition datasets using CLIP-RN50. The best result in each column is highlighted in bold.

Method	Caltech	DTD	EuroSAT	Flowers	Pets	UCF101	Food101	SUN397	Aircraft	Cars	Average
CLIP-RN50	85.88	40.37	23.69	61.75	83.57	58.84	73.97	58.80	15.66	55.70	55.82
Backpropagation-Based											
TPT (Shu et al., 2022a)	87.02	40.84	28.33	62.69	84.49	60.82	74.88	61.46	17.58	58.46	57.66
DiffTPT (Feng et al., 2023)	86.89	40.72	41.04	63.53	83.40	62.67	79.21	62.72	17.60	60.71	59.85
DPE (Zhang et al., 2024a)	90.83	50.18	41.67	67.60	85.97	61.98	77.83	64.23	19.80	59.26	61.93
DLAE (Wu et al., 2026)	90.91	**55.56**	50.04	69.63	87.68	64.76	78.76	65.43	23.34	61.98	64.81
Backpropagation-Free											
TDA (Karmanov et al., 2024)	89.70	43.74	42.11	68.74	86.18	64.18	77.75	62.53	17.61	57.78	61.03
TDA† (Karmanov et al., 2024)	89.94	52.42	36.65	68.01	86.21	65.29	77.85	63.96	20.49	57.74	61.86
DMN-ZS (Zhang et al., 2024c)	90.14	50.41	48.72	67.93	86.78	65.34	76.70	64.39	22.77	60.02	63.71
BoostAdapter (Zhang et al., 2024b)	91.24	41.13	43.10	64.40	84.39	64.38	**80.12**	63.14	20.35	62.06	61.43
BCA (Zhou et al., 2025)	89.70	48.58	42.12	66.30	85.58	63.51	77.19	63.38	19.89	58.13	61.44
MCP (Chen et al., 2025)	90.99	53.19	55.60	68.49	87.35	67.14	78.37	65.39	23.04	61.27	65.08
ADAPT (Zhang et al., 2025)	89.37	51.89	50.47	70.04	86.43	64.29	75.57	63.12	18.00	58.38	62.82
PTA (Huang et al., 2026b)	88.64	44.98	51.63	68.49	86.67	66.43	76.83	62.86	17.44	59.35	62.33
A^{2}Memory (Zhang et al., 2026)	90.36	51.89	51.22	69.91	86.92	64.29	76.26	65.22	19.91	59.66	63.56
Ours+TDA†	90.91	54.02	50.40	70.12	**88.17**	67.49	78.53	65.16	21.75	62.17	64.87
Ours+MCP	**91.48**	55.38	**56.47**	**71.25**	87.84	**67.88**	79.04	**66.25**	**23.79**	**63.70**	**66.31**

Table 6: Results on ImageNet and out-of-distribution benchmarks using CLIP-RN50. The best result in each column is highlighted in bold.

Method	ImageNet-A	ImageNet-R	ImageNet-S	ImageNet-V2	ImageNet	OOD Avg.	Average
CLIP-RN50	21.83	56.15	33.37	51.41	58.16	40.69	44.18
Backpropagation-Based							
TPT (Shu et al., 2022a)	26.67	59.11	35.09	54.70	60.74	43.89	47.26
DiffTPT (Feng et al., 2023)	31.06	58.80	37.10	55.80	60.80	45.69	48.71
DPE (Zhang et al., 2024a)	30.15	63.72	40.03	56.72	63.41	47.66	50.81
DLAE (Wu et al., 2026)	34.80	63.63	40.68	57.13	64.10	49.06	52.06
Backpropagation-Free							
TDA (Karmanov et al., 2024)	30.29	62.58	38.12	55.54	61.35	46.63	49.58
TDA† (Karmanov et al., 2024)	34.53	63.47	40.22	57.00	63.57	48.81	51.76
DMN-ZS (Zhang et al., 2024c)	28.57	61.44	39.84	56.12	63.87	46.49	49.97
BCA (Zhou et al., 2025)	30.35	62.89	38.08	56.58	61.81	46.98	49.94
MCP (Chen et al., 2025)	29.30	64.40	41.16	56.93	64.19	48.02	51.20
ADAPT (Zhang et al., 2025)	33.08	62.69	40.21	55.97	62.16	47.99	50.82
A^{2}Memory (Zhang et al., 2026)	34.51	63.86	40.16	55.99	63.96	48.63	51.70
Ours+TDA†	**35.01**	63.80	41.15	**57.29**	64.08	**49.31**	**52.27**
Ours+MCP	29.95	**65.17**	**42.18**	57.27	**64.90**	48.64	51.89

D.2 Additional Ablation Study

Effect of patch selection strategies. In Table 7, we replace our class-discriminative patch selection criterion with several alternative strategies while keeping the remaining pipeline unchanged. Our method achieves the best performance on all six datasets. The improvements are particularly notable on EuroSAT, Pets, and Aircraft, while remaining consistent across all evaluated datasets.

Effect of the Patch Selection Budget. Figure 4 examines how the patch selection budget K_{mem} affects local matching. A small budget restricts the local evidence available for comparison: increasing K_{mem} from 1 to 15 raises mean accuracy from 62.14% to 72.55%. Further expanding the selected set provides no additional gain, with accuracy decreasing by only 0.13 and 0.38 pts. at $K_{\text{mem}} = 20$ and $K_{\text{mem}} = 25$, respectively. These results support a moderate patch selection budget, with stable performance across the tested settings from 15 to 25.

Table 7: Comparison of different selection strategies across six datasets. Best results are shown in bold.

Method	Caltech101	EuroSAT	Flowers	Pets	Food101	Aircraft
Random	94.20	66.98	75.88	91.20	86.72	29.64
Low Entropy	93.83	64.73	75.68	89.81	86.43	25.05
Attention	94.48	65.47	75.84	91.11	86.17	28.95
Low Sim	94.56	65.60	76.13	92.10	86.83	30.39
Ours	**94.69**	**67.31**	**76.53**	**92.67**	**87.10**	**31.98**

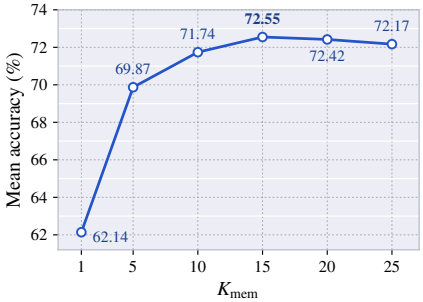

Figure 4: Effect of the patch selection budget K_{mem}. We report mean top-1 accuracy (%) across ten cross-domain datasets.