BEYOND HOLISTIC REPRESENTATIONS: DUAL DISTRIBUTION MODELING FOR TEST-TIME ADAPTATION OF VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Although Vision-Language Models (VLMs) exhibit strong zero-shot generalization, their performance degrades substantially under domain shifts and out-of-distribution (OOD) scenarios. Test-Time Adaptation (TTA) addresses this issue using only unlabeled test samples, among which training-free TTA is particularly appealing for its efficiency. Existing training-free methods refine predictions using cached samples or representation distributions, but their reliance on holistic image representations limits the extraction of more semantically meaningful features. To this end, we decompose visual features into in-distribution features and class-related OOD features based on their alignment with textual semantics, and empirically show that both exhibit clear class-wise clustering in the embedding space. Based on this finding, we propose Dual Distribution Modeling for Adaptation (DDMA), a novel training-free TTA framework that separates these two feature types to better exploit semantic information while reducing interference from semantically irrelevant features. DDMA independently models their class-wise distributions during test time and jointly refines the original predictions by their distribution statistics. Extensive experiments across cross-domain and OOD benchmarks show that DDMA outperforms state-of-the-art TTA methods, achieving average accuracy gains of 3.37% and 4.02% over the pre-adaptation performance. Moreover, DDMA maintains competitive efficiency with an average throughput of 29.37 items/s.

1 INTRODUCTION

Vision-Language Models (VLMs), largely built upon the Contrastive Language-Image Pretraining (CLIP) paradigm (Radford et al., 2021; Jia et al., 2021; Yao et al., 2022), exhibit strong cross-modal understanding and zero-shot generalization, and have developed modern large VLMs such as LLaVA (Liu et al., 2023), GPT-4V (Hurst et al., 2024), and MiniGPT-4 (Zhu et al., 2024). However, their performance degrades significantly under domain shifts and out-of-distribution (OOD) scenarios, particularly in domain-specific deployments (Dong et al., 2025). Existing solutions rely primarily on fine-tuning or post-training with domain-specific labeled data, incurring substantial computation and manual annotation costs (Sung et al., 2022; Zhao et al., 2025; Li et al., 2024).

To this end, recent studies have explored Test-Time Adaptation (TTA), which adapts models using only unlabeled test samples. Early VLM TTA methods primarily rely on prompt tuning, such as CoOp (Zhou et al., 2022b), CoCoOp (Zhou et al., 2022a), and TPT (Shu et al., 2022), which optimize textual prompts to improve test-time predictions. These methods require backpropagation and parameter updates, leading to additional computational costs and limiting their applicability to online TTA (Wang et al., 2025b). Recent studies have therefore explored training-free TTA methods, which avoid backpropagation process and can be broadly categorized into cache- and distribution-based approaches. Cache-based methods (*e.g.*, TDA (Karmanov et al., 2024) and DMN (Zhang et al., 2024)) store high-quality samples and refine predictions based on their similarity to incoming inputs. But limited cache capacity restricts their ability to capture the evolving test distribution, potentially leading to unstable adaptation and catastrophic forgetting over time (Han et al., 2024). In contrast, distribution-based methods (*e.g.*, ZERO (Farina et al., 2024) and TT-RAA (Fan et al., 2025)) construct global distributions from reliable test samples and refine predictions by their statistics. By aggregating information from all historical samples, they enable more stable and consistent adaptation.

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.

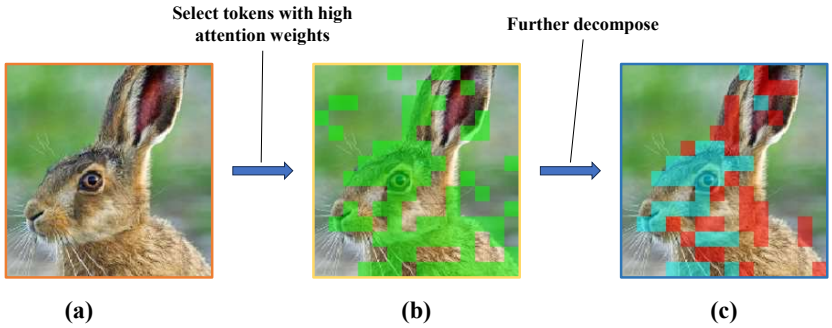


Figure 1: **Illustration of token selection and feature decomposition.** (a) Original image. (b) High-attention tokens (green), which primarily capture class-related semantics. (c) Further distinguishing the selected tokens based on textual alignment. Blue tokens exhibit strong alignment and correspond to ID features, while red tokens show weaker alignment but retain some class-related semantics (e.g., rabbit ears and fur texture), corresponding to class-related OOD features.

However, existing training-free TTA methods share a common limitation: they only leverage entire image representations, without further distinguishing image features according to their semantics and how well they are learned by the VLM. When test samples contain substantial insufficiently learned and poorly aligned features during the VLM pre-training stage (referred to as OOD features), these features interfere with class-related semantics, making holistic representations less effective for exploiting discriminative information during adaptation. To address this challenge, we move beyond holistic image representations and perform a token-level feature decomposition. As shown in Fig. 1 (b), high-attention tokens predominantly capture class-related features. We further distinguish these features based on their alignment with textual semantics, separating them into well-aligned in-distribution (ID) features and less aligned yet semantically meaningful features, namely class-related OOD features (shown in Fig. 1 (c)). Interestingly, we then reveal that both feature types exhibit clear class-wise clustering in the VLM visual embedding space.

Based on these findings, we propose Dual Distribution Modeling for Adaptation (DDMA), a novel training-free TTA method for VLMs. DDMA decomposes test samples into ID and class-related OOD features and maintains separate class-wise distributions for each type with high-quality samples. During inference, the original predictions are then jointly refined using the similarities between the two feature types of inputs and the statistical centers of corresponding distributions. Extensive experiments across a wide range of cross-domain and OOD benchmarks demonstrate that DDMA outperforms existing SOTA TTA methods, achieving average accuracy gains of 3.37% and 4.02% over the pre-adaptation performance. Moreover, DDMA achieves an average throughput of 29.37 items/s with 15.46 GFLOPs, also demonstrating competitive efficiency with existing training-free methods.

Overall, our main contributions are summarized as follows: (1) We introduce a new perspective for analyzing VLM representations by decomposing visual features into ID and class-related OOD features, and reveal their class-wise clustering properties in the embedding space. (2) Based on our findings, we propose **DDMA**, a training-free TTA method that separately models the distributions of ID and class-related OOD features and jointly leverages them to refine predictions. (3) Extensive experiments on diverse cross-domain and OOD benchmarks demonstrate that DDMA outperforms existing SOTA TTA methods while maintaining competitive computational efficiency.

2 Related Work

2.1 Vision-Language Models

Vision-Language Models (VLMs) can be traced back to VisualBERT (Li et al., 2019) and UNITER (Chen et al., 2020), which integrate visual and textual representations to address fundamental multimodal tasks like image captioning and visual question answering. CLIP (Radford et al., 2021) marks a major milestone by aligning image and text embeddings through contrastive learning on large-scale image–text pairs, enabling strong zero-shot capability. Subsequent works, including ALIGN

(Jia et al., 2021) and FILIP (Yao et al., 2022), further improve noise robustness and fine-grained cross-modal representation alignment, respectively. Recent works have combined CLIP-like VLMs with large language models (LLMs) to develop large vision–language models (LVLMs), significantly enhancing multimodal understanding and reasoning. For example, GPT-4V (Hurst et al., 2024) incorporates visual inputs into GPT-4, while LLaVA (Liu et al., 2023) uses a CLIP-based vision encoder with a linear projection to align visual and textual representations. Since CLIP underlies many existing VLMs and LVLMs, we adopt CLIP as a representative for studying VLM adaptation.

2.2 Test-Time Adaptation for VLMs

VLMs often suffer substantial performance degradation under domain shifts and out-of-distribution scenarios (Dong et al., 2025). To mitigate this issue under the test-time setting, recent studies have proposed various test-time adaptation (TTA) methods for VLMs. Early TTA methods primarily rely on prompt tuning, including TPT (Shu et al., 2022), C-TPT (Yoon et al., 2024), and Diff-TPT (Feng et al., 2023), which optimize or enhance textual prompt embeddings to improve test-time predictions. However, these methods require iterative prompt updates via a backpropagation process, incurring considerable computational cost and latency and limiting their applicability to real-time and online TTA scenarios (Wang et al., 2025b; Zanella et al., 2025).

Recent studies have explored training-free TTA methods, which can be broadly categorized into cache-based and distribution-based approaches. Cache-based methods, such as TDA (Karmanov et al., 2024), DMN (Zhang et al., 2024), and TCA (Wang et al., 2025a), store high-quality test features and refine subsequent predictions based on their similarities to cached features. However, limited cache capacity restricts the use of historical samples, potentially limiting adaptation over long test streams (Han et al., 2024). In contrast, distribution-based methods, including MTA (Zanella & Ben Ayed, 2024), DOTA (Han et al., 2024), ADAPT (Zhang et al., 2025), and TT-RAA (Fan et al., 2025), maintain global feature distributions and use their statistics for prediction refinement. By aggregating information across test samples, these methods are more memory-efficient and provide a holistic view of the test data (Fan et al., 2025; Farina et al., 2024).

3 Methodology

We aim to develop a distribution-based method to improve both the performance and efficiency of TTA. Following prior works, we study VLM adaptation in the CLIP-based image classification task, which directly reflects CLIP's multimodal representation and alignment capabilities.

3.1 CLIP-based Image Classification

CLIP comprises an image encoder $E_v(\cdot)$ and a text encoder $E_t(\cdot)$ that map an image I and text T into a shared D-dimensional embedding space as $z_i = E_v(I) \in \mathbb{R}^D$ and $z_t = E_t(T) \in \mathbb{R}^D$, respectively. CLIP aligns image and text representations in a shared embedding space, where semantically similar inputs have higher similarity. For classification task, we construct class textual prototypes (*e.g.*, "an image of <*class name*>") and compute class probabilities from their cosine similarities with the image representation, as shown in Eq. (1).

$$p_{i,c} = \text{softmax}(l_{i,c}) = \frac{\exp(\cos(z_i, z_{t,c})/\tau)}{\sum_{j=1}^{C} \exp(\cos(z_i, z_{t,j})/\tau)} \tag{1}$$

where C is the number of classes, $z_{t,c}$ represents the embedding of the textual prototype for the c-th class, $l_{i,c} = \cos(z_i, z_{t,c})/\tau$ is the logits before softmax for class c, $p_{i,c}$ denotes the predicted probability, and τ is the temperature parameter. Then the final prediction is $y_i = \arg\max_c p_{i,c}$.

3.2 ID Features and Class-related OOD Features

Existing training-free TTA methods for VLMs typically rely on holistic image representations, limiting their ability to exploit fine-grained information in test samples. To this end, we perform a token-level analysis of image representations encoded by CLIP, inspired by Wang et al. (Wang et al., 2025a). Specifically, the token embeddings at the l-th attention layer are denoted as $E^l = e_{cls}, e_1, e_2, \ldots, e_m \in \mathbb{R}^{(m+1)\times d}$, where d is the embedding dimension, m is the number of image

tokens, and e_{cls} denotes the global <cls> token embedding. The corresponding attention weights are denoted as $A^l = A^l_1, A^l_2, \ldots, A^l_H \in \mathbb{R}^{H\times(m+1)}$, where H is the number of attention heads.

Following (Cai et al., 2025), we select token embeddings from the $(L-1)$-th attention layer, which encodes richer semantic information, and apply a set of masks to investigate the semantics of patch tokens and their impact on prediction results. The masked embeddings are then fed into the remaining part of CLIP to obtain the representation z'_i, as formulated in Eq. (2), allowing us to assess whether the selected token semantics can be recognized and utilized by CLIP.

$$z'_i = f^L(\mathbf{M} \odot E^{(L-1)}) \tag{2}$$

where $f^L(\cdot)$ denotes the remaining forward computation after the $(L-1)$-th attention layer, and $\mathbf{M} \in \{0,1\}^m$ is the designed mask for each patch token.

First, we construct a mask $\mathbf{M}1$ to separate class-related features from irrelevant ones. Since tokens with higher attention weights typically contribute more to the prediction, we compute the mean attention weights across all heads in the $(L-1)$-th layer, $\bar{A} = \frac{1}{H}\sum A^{L-1}_h$, and retain the top 80% tokens with the largest values in $\bar{A}$ by $\mathbf{M}1$. We apply $\mathbf{M}_1$ to $E^{(L-1)}$ and evaluate the resulting predictions on the ImageNet ILSVRC 2012 validation set (Russakovsky et al., 2015), measuring both accuracy and prediction entropy. The computation of entropy is given in Eq. (3).

$$\mathbf{H}(p_{i,c}) = -\sum_{c=1}^{C} p_{i,c} \log p_{i,c} \tag{3}$$

where $p_{i,c}$ is the probability of being predicted as class c, calculated according to Eq. (1). A smaller entropy value indicates more stable predictions.

The results are reported in Tab. 1. Compared with the original input, applying $\mathbf{M}_1$ causes slight accuracy degradation while reducing prediction entropy. In contrast, selecting low-attention tokens with $\bar{\mathbf{M}}_1$ leads to collapse in prediction performance. These results indicate that $\mathbf{M}_1$ preserves most class-related features while filtering out irrelevant ones, reducing their interference and yielding more confident predictions. We summarize this finding as **Observation I**.

Observation I: *Visual patch tokens with higher attention weights contain most class-related features, while tokens with lower attention weights mainly contain class-irrelated semantics.*

We further examine whether the class-related features in high-attention tokens can be stably aligned with textual semantics and used for prediction. Specifically, for each token selected by $\mathbf{M}_1$, we compute the entropy of its class logits based on its cosine similarities with class textual prototypes, following Eq. (1). We then construct $\mathbf{M}_2$ to retain the 50% lowest-entropy tokens among those selected by $\mathbf{M}_1$. The classification accuracy and prediction entropy are reported in Tab. 1.

Tab. 1 shows that although $\mathbf{M}_2$ causes accuracy drop to some extent, $\bar{\mathbf{M}}_2$ results in a more substantial accuracy decrease and much higher prediction entropy. These results indicate that $\mathbf{M}_2$ preserves features that are class-related and aligned with textual semantics, while $\bar{\mathbf{M}}_2$ primarily captures class-related OOD features that are poorly aligned and insufficiently learned during pre-training stage. Besides, although the two groups have similar mean attention weights, higher-entropy tokens exhibit substantially larger variance across attention heads (0.9526 vs. 0.5861), also indicating less stable perception. We summarize this finding as **Observation II**.

Observation II: *Class-related features in high-attention patch tokens can be divided into **in-distribution (ID) features**, which are stably captured by CLIP and well aligned with textual semantics, and **class-related out-of-distribution (OOD) features**, which retain class-related semantics but exhibit unstable perception and weak text alignment.*

To fully leverage all class-related features during TTA, we further investigate whether class-related OOD features exhibit class-wise clustering behavior. Specifically, we obtain the final representations from features selected by $\bar{\mathbf{M}}_2 \odot \mathbf{M}_1$ and group them by their ground-truth labels. We then report the average intra- and inter-class cosine similarities in Tab. 2, and visualize the representation distributions of ID and class-related OOD features for randomly selected ImageNet classes in Fig. 2.

Both quantitative and qualitative results show that class-related OOD features exhibit high intra-class similarity and clear inter-class separation in CLIP's visual embedding space, consistent with the

Table 1: Prediction accuracy and entropy under different mask settings.

Settings	Origin	$\mathbf{M}_1$	$\bar{\mathbf{M}}_1$	$\mathbf{M}_2 \odot \mathbf{M}_1$	$\bar{\mathbf{M}}_2 \odot \mathbf{M}_1$
Acc(%)	65.75	62.48	12.76	42.21	23.45
Entropy	0.2705	0.2556	1.057	0.3276	0.6973

Table 2: Average intra-class and inter-class similarities under different mask settings.

Settings	Origin	$\mathbf{M}_2 \odot \mathbf{M}_1$	$\bar{\mathbf{M}}_2 \odot \mathbf{M}_1$
Intra-class	0.7262	0.7637	0.7545
Inter-class	0.5776	0.6011	0.6142

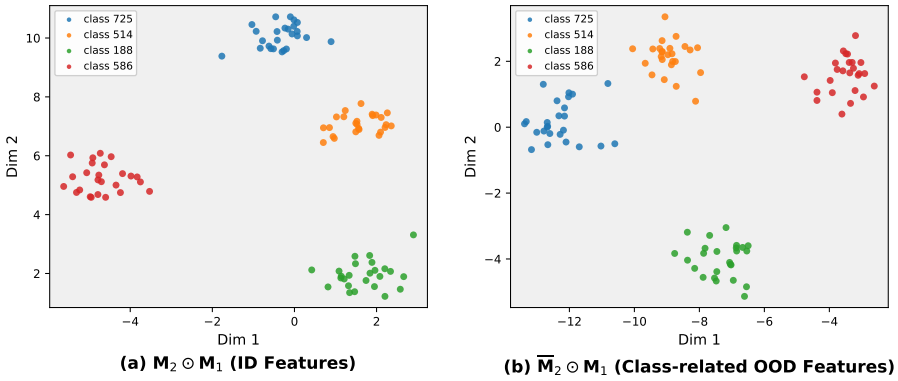


Figure 2: Visualization results of feature representations by t-SNE. Both ID features and class-related OOD features exhibit clear class-wise clustering in the embedding space.

class-wise clustering observed for ID features and holistic image representations. We summarize this finding as Observation III, providing empirical support for separately modeling ID and class-related OOD feature distributions to refine predictions during TTA.

Observation III: *Representations derived from both ID and class-related OOD features exhibit clear class-wise clustering behavior in the visual embedding space, characterized by higher intra-class similarity and lower inter-class similarity.*

3.3 Dual Distribution Modeling for Adaptation

Based on these observations, we propose Dual Distribution Modeling for Adaptation (DDMA), which separately models the distributions of ID and class-related OOD features to refine the original logits. An overview is shown in Fig. 3.

Adaptive Entropy-Guided Sample Selection. Following prior works (Wang et al., 2025a; Fan et al., 2025), we select high-confidence test samples for each class based on the entropy H_i of their original predictions, which tend to contain more class-related features. Specifically, we select samples with pseudo label in class c and its entropy $H_i \leq \theta_c$, where the pseudo label is obtained from the original prediction and θ_c is progressively updated. The update rule is given in Eq. (4).

$$\theta_c = \frac{n_c}{n_c + 1}\theta_c + \eta * \frac{1}{n_c + 1}H_i, \tag{4}$$

where n_c denotes the number of already selected samples for class c, η is the update step. For each class, the first encountered test sample is directly selected to initialize θ_c with H_i, without applying the threshold criterion. To retain potentially useful class-related OOD features from moderately uncertain samples, we introduce a relaxation term R and select samples satisfying $H_i \leq \theta_c + R$.

Token-level Feature Decomposition. We then construct masks to separately extract ID and class-related OOD features from each input image. Following Sec. 3.2, we first design $\mathbf{M}_1$ to select class-related tokens based on their attention weights. Since tokens containing class-related OOD features exhibit larger variance across attention heads, we use the maximum attention weight across heads to better preserve these tokens. The computation is given in Eq. (5), and then the mask $\mathbf{M}_1$ can be obtained by Eq. (6).

$$\hat{a}_i = \max_{1 \leq h \leq H} a_{h,i}^{L-1} \tag{5}$$

$$\mathbf{M}_{1,i} = \begin{cases} 1, \text{if } \hat{a}_i \geq T(r_1, m), \\ 0, \text{otherwise.} \end{cases} \tag{6}$$

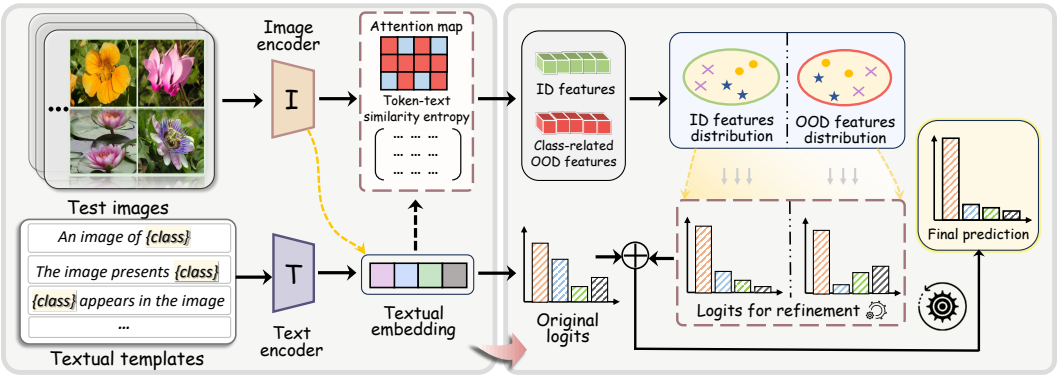


Figure 3: **Overview of DDMA.** High-quality test samples are decomposed into ID and class-related OOD features based on attention maps and textual alignment entropy. Their representations are used to construct and update separate feature distributions, which are leveraged to refine the original logits.

where $a_{h,i}^{L-1}$ denotes the attention weight of the i-th token in A_h^{L-1}. $\mathbf{M}_{1,i}$ denotes the i-th entry of $\mathbf{M}_1$, and $T(r_1, m)$ is the threshold for retaining the top $r_1 \cdot m$ tokens with the largest $\hat{a}$ values.

We design $\mathbf{M}_2$ to further distinguish ID features from class-related OOD features among the selected tokens. Following Observation 2, we select tokens based on their alignment (measured by cosine similarity) entropy with class textual prototypes, as shown in Eqs. (7) and (8).

$$s_i = -\mathbf{H}(p_{t_i}), \quad p_{t_i,c} = \frac{\exp(\cos(t_i, z_{t,c})/\tau)}{\sum_{j=1}^{C} \exp(\cos(t_i, z_{t,j})/\tau)} \tag{7}$$

$$\mathbf{M}_{2,i} = \begin{cases} 1, \text{if } s_i \geq T(r_2, r_1 m), \\ 0, \text{otherwise.} \end{cases} \tag{8}$$

where t_i is the representation of i-th patch token selected by $\mathbf{M}_1$, and $T(r_2, r_1 m)$ is the threshold for retaining r_2 fraction of tokens with the lowest entropy among those selected by $\mathbf{M}_1$.

Based on $\mathbf{M}_1$ and $\mathbf{M}_2$, we obtain the representations of ID and class-related OOD features for distribution modeling and prediction refinement. Specifically, we apply the two masks to the $(L-1)$-th layer embeddings and feed the masked embeddings through the remaining layers to complete the forward pass, as formulated in Eqs. (9). We further design a projection for the representations to enhance their class-wise clustering. A detailed discussion is provided in Appendix C.

$$z_{\text{id}} = f^L(\mathbf{M}_2 \odot \mathbf{M}_1 \odot E^{(L-1)}), \quad z_{\text{c-ood}} = f^L(\bar{\mathbf{M}}_2 \odot \mathbf{M}_1 \odot E^{(L-1)}) \tag{9}$$

Class-wise Distribution Modeling. For each class, we maintain two distributions for ID and class-related OOD features, respectively. When a high-quality test sample with pseudo label c is selected, the corresponding distribution centers are updated accordingly, as shown in Eq. (10), where $\mu_{c,id}$ and $\mu_{c,c\text{-}ood}$ are both initialized to 0. .

$$\mu_{c,id} = \frac{n_c}{n_c+1}\mu_{c,id} + \eta * \frac{1}{n_c+1} z_{id}, \quad \mu_{c,c\text{-}ood} = \frac{n_c}{n_c+1}\mu_{c,c\text{-}ood} + \eta * \frac{1}{n_c+1} z_{c\text{-}ood} \tag{10}$$

Distribution-Guided Logit Correction. During test time, we refine each sample's logits using the two learned distributions. Specifically, we compute the cosine similarities between its ID and class-related OOD representations and the corresponding class-wise distribution centers, yielding two additional logits, as shown in Eq. (11).

$$\begin{aligned} \delta_{id} &= \sum_{c=1}^{C} \alpha \exp(-\beta(1 - \cos(z_{id}, \mu_{c,id})))\mathbb{I}_c \\ \delta_{c\text{-}ood} &= \sum_{c=1}^{C} \alpha \exp(-\beta(1 - \cos(z_{c\text{-}ood}, \mu_{c,c\text{-}ood})))\mathbb{I}_c \end{aligned} \tag{11}$$

Table 3: Results on the cross-domain benchmarks. The metric is classification Acc(%).

Method \ Dataset	Aircraft	Caltech101	Cars	DTD	EuroSAT	Flower102	Food101	Pets	SUN397	UCF101	Average
CLIP	23.22	93.55	66.11	45.04	50.42	66.99	82.86	86.92	65.63	65.16	64.59
CoOP	18.47	93.70	64.51	41.92	46.93	68.71	85.30	89.14	64.15	66.55	63.94
CoCoOP	22.29	93.79	64.90	45.45	39.23	70.85	83.97	**90.46**	66.89	68.44	64.63
TPT	24.78	94.16	66.87	**47.75**	42.44	68.98	84.67	87.79	65.50	68.04	65.10
DiffTPT	**25.60**	92.94	67.01	47.00	43.13	70.10	87.23	88.22	65.74	62.67	64.96
TDA	23.52	94.18	67.24	47.26	52.67	71.89	86.10	88.47	67.76	**70.03**	66.91
DMN	24.65	94.03	65.84	44.13	48.26	71.40	84.79	88.85	66.54	66.39	65.49
TCA	24.33	93.34	65.07	46.15	**62.01**	69.49	85.24	89.17	65.69	68.74	66.92
MTA	24.97	94.06	67.72	45.59	36.48	68.37	85.10	87.91	64.78	68.28	64.33
ZERO	25.05	93.73	**67.88**	45.60	35.07	67.64	86.71	87.56	65.11	67.74	64.21
TT-RAA	24.89	94.16	66.18	46.45	58.80	70.37	86.10	88.63	67.35	69.15	67.21
DDMA (Ours)	25.29	**94.25**	66.32	45.86	60.39	**72.53**	**87.48**	89.57	**68.03**	69.83	**67.96**

where $\mathbb{I}c \in \{0,1\}^{1\times C}$ is the one-hot vector for class c, and α and β are scaling factors to align δ_{id} and $\delta_{c\text{-}ood}$ with the original logits. The final prediction are obtained as $p_{final} = \text{softmax}(l_i + w_1\delta_{id} + w_2\delta_{c\text{-}ood})$, where l_i is the logits before softmax from Eq. (1), and w_1 and w_2 are weighting factors.

4 Experiment

In this section, we evaluate our method under both cross-domain and out-of-distribution benchmarks. Additional ablation studies and analyses of hyperparameter are in Appendix.

4.1 Setup

Models and datasets. We conduct experiments using two typical attention-based CLIP models, namely CLIP ViT-B/16 and CLIP ViT-B/32. For cross-domain scenarios, we evaluate on Aircraft (Maji et al., 2013), Caltech101 (Fei-Fei et al., 2004), Stanford Cars (Cars) (Krause et al., 2013), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), Flower102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), Oxford Pets (Pets) (Parkhi et al., 2012), SUN397 (Xiao et al., 2010), and UCF101 (Soomro et al., 2012). For OOD scenarios, we adopt several ImageNet variants, including 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).

Baselines. We compare DDMA with representative SOTA VLM TTA methods, including prompt tuning-based methods (CoOp (Zhou et al., 2022b), CoCoOp (Zhou et al., 2022a), TPT (Shu et al., 2022), and Diff-TPT (Feng et al., 2023)), cache-based methods (TDA (Karmanov et al., 2024), DMN (Zhang et al., 2024), and TCA (Wang et al., 2025a)), and distribution-based methods (MTA (Zanella & Ben Ayed, 2024), ZERO (Farina et al., 2024), and TT-RAA (Fan et al., 2025)). All baselines are implemented using their official open-source code and evaluated with a batch size of 1, following the challenging single-image TTA setting.

Metrics. We use classification accuracy (Acc) to evaluate prediction performance and adaptation effectiveness. For efficiency, we report GFLOPs and throughput, measured as images processed per second (IPS), where lower GFLOPs and higher throughput indicate greater computational efficiency.

Settings of our method. We set the relaxation term R to 0.5, which is used for selecting high-confidence test samples. For the token selection ratios in Eqs. (6) and (8), we set r_1 and r_2 to 0.8 and 0.8 for cross-domain benchmarks, and to 0.8 and 0.7 for OOD scenarios. The remaining hyper-parameters are tuned separately for each dataset. Detailed configurations can be found in our source code, and the experimental environment and hardware setup is provided in Appendix B.

4.2 Evaluation of Adaptation Effectiveness

We conduct experiments on benchmark datasets under cross-domain scenarios and report the accuracy of our method and the selected baselines. The results for CLIP ViT-B/16 are shown in Tab. 3 and 4, while those for CLIP ViT-B/32 are provided in Appendix E.

As shown in Tab. 3, prompt tuning-based methods achieve limited gains on cross-domain benchmarks, with an average improvement of less than 1%. Cache-based methods perform better, with TDA and TCA achieving gains of around 2%. Among distribution-based methods, TT-RAA achieves the

Table 4: Results on the out-of-distribution benchmarks. The metric is classification Acc(%).

Method \ Dataset	ImageNet	ImageNet-A	ImageNet-V2	ImageNet-S	ImageNet-R	Average
CLIP	68.34	49.89	61.88	48.24	77.65	61.20
CoOP	69.18	49.71	64.20	47.99	75.21	61.26
CoCoOP	69.85	50.63	64.07	48.75	76.18	61.90
TPT	68.98	54.77	63.45	47.94	77.06	62.44
DiffTPT	**70.30**	55.68	**65.10**	46.80	75.00	62.58
TDA	68.51	59.76	64.15	50.67	79.82	64.58
DMN	68.40	57.63	64.94	**53.16**	78.39	64.50
TCA	68.27	52.85	62.96	48.53	77.05	61.93
MTA	69.43	55.60	63.47	48.77	76.05	62.66
ZERO	68.46	58.29	63.58	48.61	77.25	63.24
TT-RAA	68.59	60.18	64.22	49.45	80.34	64.56
DDMA (Ours)	69.72	**61.46**	64.50	49.86	**80.53**	**65.22**

Table 5: The average efficiency metrics across cross-domain and OOD benchmarks. "V" denotes throughput, and "L-Para." denotes the number of learnable parameters. GFLOPs and learnable parameters are manually estimated from the adaptation algorithm and model architecture and are independent of the evaluation datasets.

Method \ Metrics	Cross-Domain		OOD		GFLOPS	L-Para.
	Time (s) ↓	V (items/s) ↑	Time (s) ↓	V (items/s) ↑		
CLIP	183.79	42.94	683.35	44.29	15.05	0
CoOP	569.05	14.73	2281.60	13.34	17.74	2048
CoCoOP	1162.49	6.89	4156.72	7.16	17.74	34816
TPT	575.68	14.06	2264.83	13.45	1094.20	2048
DiffTPT	503.36	16.59	1961.46	15.88	-	-
TDA	308.53	24.38	1328.61	24.06	15.87	0
DMN	325.42	22.61	1435.38	21.24	15.87	0
TCA	236.67	34.17	906.70	33.93	14.93	0
MTA	289.06	27.48	1274.58	25.67	-	0
ZERO	245.22	32.30	940.92	32.44	15.29	0
TT-RAA	272.14	29.63	1067.09	30.15	15.29	0
DDMA (Ours)	261.80	30.11	1149.94	28.62	15.46	0

best performance, with an average accuracy of 67.21%. In contrast, DDMA achieves an average accuracy gain of over 3%, outperforming existing SOTA methods. Moreover, DDMA achieves the best performance on four benchmarks, namely Caltech101, Flower102, Food101, and SUN397.

In OOD scenarios, all methods show obvious improvements of accuracy, with prompt tuning-based methods achieving gains of approximately 4% to 5%. Among training-free methods, TDA, DMN, and TT-RAA perform competitively, all exceeding 64.50% accuracy, while TCA is less effective, likely due to the abundance of class-ambiguous tokens in OOD data. In contrast, DDMA outperforms all baselines, achieving an average accuracy gain of over 4% and the best performance on ImageNet-A and ImageNet-R.

4.3 Evaluation of Efficiency

To evaluate efficiency, we report the average inference time and throughput across datasets in both CD and OOD scenarios, along with the number of learnable parameters and manually estimated GFLOPs. Results are presented in Tab. 5.

As shown in Tab. 5, prompt tuning-based methods result in substantially higher GFLOPs and lower throughput due to gradient computation and backpropagation. Among training-free methods, TDA and DMN introduce additional overhead from computing similarities with cached samples, making them less efficient than distribution-based methods. In contrast, TCA reduces the number of tokens processed through token selection and merging, achieving the highest throughput among the baselines. DDMA incurs slightly higher computational cost than existing distribution-based methods due to maintaining two distributions. Nevertheless, its inference efficiency remains competitive: DDMA achieves throughput comparable to TT-RAA and is only slightly slower than MTA and ZERO.

Table 6: Ablation study on feature decomposition. The first row corresponds to modeling the distribution using the holistic image representations.

ID Feat.	C-OOD Feat.	Cars	Pets	ImageNet-A	ImageNet-V2
-	-	66.10	88.71	58.45	63.69
✓		66.27	88.93	58.07	63.74
	✓	65.49	87.54	59.38	64.18
✓	✓	**66.32**	**89.57**	**61.46**	**64.50**

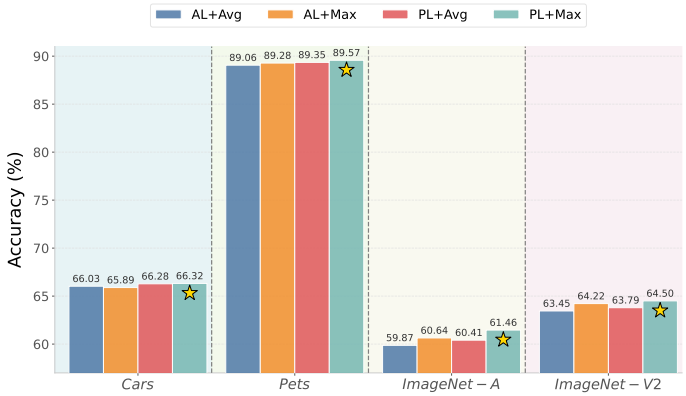


Figure 4: Ablation study on the token selection strategy. "AL" and "PL" denote attention weights from all layers and the penultimate layer, while "Avg" and "Max" denote mean and maximum aggregation across attention heads.

4.4 Ablation Study

To assess the contribution of each DDMA component, we conduct ablation studies on Cars, Pets, ImageNet-A and ImageNet-V2 for both cross-domain and OOD scenarios, with all other settings unchanged. Additional ablations on more modules and hyperparameter are provided in Appendix F.

Ablation on feature decomposition. To evaluate the contributions of ID and class-related OOD features, we conduct an ablation study on the feature decomposition strategy in Eqs. (10) and (11). Specifically, we consider four settings: (1) whole-image representations, (2) ID features only, (3) class-related OOD features only, and (4) both feature types (our method). The results are reported in Tab. 6. In CD scenarios, ID features provide more class-discriminative information than class-related OOD features, yielding better performance than Setting (3). In contrast, class-related OOD features contribute more to adaptation under OOD scenarios. Jointly leveraging both feature types allows DDMA to better exploit semantic information in test samples. Moreover, compared with Setting (1), separating the two feature types avoids their interference within a single distribution, achieving the best overall adaptation performance.

Ablation on token selection strategy. For class-related token selection based on attention weights (Eqs. (6) and (7)), DDMA uses only the penultimate-layer attention weights and the maximum weight across heads. We compare this design with two alternatives commonly-used in prior works: using attention weights from all layers and averaging across heads. As shown in Fig. 4, our design consistently achieves the best performance, with gains exceeding 1% on OOD datasets. This indicates that our strategy better preserves class-related tokens for subsequent adaptation.

5 Conclusion

In this paper, we first introduce ID features and class-related OOD features in VLM visual representations and then systematically investigate their properties through empirical studies. Based on these findings, we propose DDMA, a training-free TTA method that separately models their distributions to refine original predictions more effectively. Extensive experiments across cross-domain and OOD benchmarks show that DDMA outperforms existing cache- and distribution-based TTA methods while maintaining competitive efficiency.

AI USE STATEMENT

In this work, we used generative AI tools only for designing and providing feedback on experiments, and assisting with translation. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, proposing or refining hypotheses, implementing methods, or interpreting results. Generating synthetic data sets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, cleaning and reformating dataset, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for suggesting experimental parameters and proposing a title for this paper.

We have reviewed all AI-assisted work. Specifically, we have carefully reviewed and proofread all text and codes generated with the assistance of AI tools, and verified the suggested experimental settings to ensure the accuracy and reliability of the resulting content. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6229–6242. PMLR, 13–19 Jul 2025.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020.

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.

Hao Dong, Lijun Sheng, Jian Liang, Ran He, Eleni Chatzi, and Olga Fink. Adapting vision-language models without labels: A comprehensive survey. *arXiv preprint arXiv:2508.05547*, 2025.

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

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

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*, pp. 178–178. IEEE, 2004.

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 (ICCV)*, pp. 2704–2714, October 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. *arXiv preprint arXiv:2409.19375*, 2024.

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

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, 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 (ICCV)*, pp. 8340–8349, October 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15262–15271, June 2021b.

Xuefeng Hu, Ke Zhang, Min Sun, Albert Chen, Cheng-Hao Kuo, and Ram Nevatia. Bafta: Backprop-free test-time adaptation for zero-shot vision-language models. *arXiv preprint arXiv:2406.11309*, 2024a.

Xuefeng Hu, Ke Zhang, Lu Xia, Albert Chen, Jiajia Luo, Yuyin Sun, Ken Wang, Nan Qiao, Xiao Zeng, Min Sun, Cheng-Hao Kuo, and Ram Nevatia. Reclip: Refine contrastive language image pre-training with source free domain adaptation. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2982–2991, 2024b.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14162–14171, June 2024.

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

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

Ming Li, Jike Zhong, Chenxin Li, Liuzhuozheng Li, Nie Lin, and Masashi Sugiyama. Vision-language model fine-tuning via simple parameter-efficient modification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14394–14410, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the 36th Advances in Neural Information Processing Systems (NeurIPS'23)*, 2023.

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

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

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

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

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

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

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*, volume 35, pp. 14274–14289, 2022.

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

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5227–5237, June 2022.

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

Zixin Wang, Dong Gong, Sen Wang, Zi Huang, and Yadan Luo. Is less more? exploring token condensation as training-free test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 144–154, October 2025a.

Zixin Wang, Yadan Luo, Liang Zheng, Zhuoxiao Chen, Sen Wang, and Zi Huang. In search of lost online test-time adaptation: A survey. *International Journal of Computer Vision*, 133(3): 1106–1139, 2025b.

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.

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

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

Maxime Zanella and Ismail Ben Ayed. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23783–23793, June 2024.

Maxime Zanella, Clément Fuchs, Christophe De Vleeschouwer, and Ismail Ben Ayed. Realistic test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25103–25112, June 2025.

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 (CVPR)*, pp. 28718–28728, June 2024.

Youjia Zhang, Youngeun Kim, Young-Geun Choi, Hongyeob Kim, Huiling Liu, and Sungeun Hong. Backpropagation-free test-time adaptation via probabilistic gaussian alignment. *arXiv preprint arXiv:2508.15568*, 2025.

Yumiao Zhao, Bo Jiang, Yuhe Ding, Xiao Wang, Jin Tang, and Bin Luo. Fine-grained vlm fine-tuning via latent hierarchical adapter learning. *arXiv preprint arXiv:2508.11176*, 2025.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16816–16825, June 2022a.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

A ALGORITHM OF DDMA

Algorithm 1 Detailed algorithm of DDMA

Require: Test sample sequence T, Pretrained VLM, hyper-parameters η, w_1, w_2, r_1, r_2, R, α and β.
Ensure: Adapted prediction p_{final}

```
For each class, initialize two distribution centers and set the number of selected samples to 0.
for I_i ∈ T do
    Get the holistic representation of image I_i.
    Get the original logits l_i, prediction p_i, pseudo label c and its entropy H_i by Eq. (1) and (3).
    Generate two masks M_{1,i} and M_{2,i} based on Eq. (6) and (8).
    Obtain the representations z_id and z_{c-ood} of ID and class-related OOD features by Eq. (9).
    (Optional: Project the representations for improved performance by Eq. (13)).
    if n_c = 0 or H_i ≤ θ_c + R then
        Update two distributions according to Eq. (10).
        if n_c = 0 then
            Initialize threshold θ_c to H_i and set n_c = 1.
        else
            Update threshold θ_c by Eq. (4) and set n_c = n_c + 1.
        end if
    end if
    Get the correction logits δ_id and δ_{c-ood} using two distributions by Eq. (11).
    Produce adapted prediction p_final = softmax(l_i + w_1 δ_id + w_2 δ_{c-ood}) in real time.
end for
```

B EXPERIMENT ENVIRONMENT

All experiments are conducted on NVIDIA A100-SXM4 GPUs, each equipped with 80GB of GPU memory. All models and adaptation methods are implemented and evaluated under the same hardware environment to ensure a fair and consistent comparison. The reported computational costs and throughput are measured under a single-GPU setting. Unless otherwise specified, the reported results are obtained using this hardware configuration.

C FEATURE REPRESENTATION PROJECTION

After obtaining the ID and class-related OOD feature representations using Eq. (11), we further introduce a projection to improve their class-wise separability. Specifically, the projection maps both feature types into a new space, where representations from the same class are pulled closer while those from different classes are pushed apart. Following prior works (Hu et al., 2024a), we first perform singular value decomposition on the textual embeddings of all classes, as formulated in Eq. (12).

$$U, S, V = svd(Z_t) \tag{12}$$

where $Z_t = [z_{t,1}, z_{t,2}, ..., z_{t,c}] \in \mathbb{R}^{d \times C}$ denotes the textual embeddings of all classes. Then we obtain $U = [u_1, u_2, ..., u_c]$ and construct a transformed matrix UU^T that filters out most class-irrelevant semantic information. Moreover, following (Hu et al., 2024b), the first component u_1 in U captures the most class-overlapped semantic information. We thus remove u_1 to enhance the separability of different class centers in the projection space. The projection is formulated as Eq. (13).

$$z' = \frac{U'U'^T}{\|U'U'^T\|} z, \quad U' = [u_2, u_3, ..., u_c] \tag{13}$$

We further conduct ablation studies to demonstrate the effectiveness of the projection. The settings are the same as those in Sec. 5.4. The results are reported in Tab. 7. With the proposed projection module, DDMA achieves a significant performance improvement, particularly on Stanford Cars and ImageNet-A, where the accuracy increases by approximately 10%. This demonstrates that the projection to enhance the class-wise separability of the representations of ID and class-related OOD features contributes to more effective adaptation.

Table 7: Ablation study on the projection module.

Settings	Cars	Pets	ImageNet-A	ImageNet-V2
w/o proj.	61.69	80.31	50.27	59.98
baseline	**66.32**	**89.57**	**61.46**	**64.50**

Table 8: Accuracy Results (%) of adaptation for CLIP ViT-B/32 on cross domain benchmarks.

Method	Aircraft	Caltech101	Cars	DTD	EuroSAT	Flower102	Food101	Pets	SUN397	UCF101	Average
CLIP	18.52	91.07	59.38	44.83	45.06	64.29	79.75	85.68	60.51	62.14	**61.12**
DDMA (Ours)	20.12	91.81	60.87	45.15	48.39	67.44	81.49	88.33	65.62	64.27	**63.35**

D VISUALIZATION OF TWO TYPES OF FEATURES

To provide a direct understanding of the token-level feature decomposition mechanism in DDMA, we visualize the tokens identified as containing ID features and class-related OOD features on representative images from different datasets. The results are shown in Fig. 5.

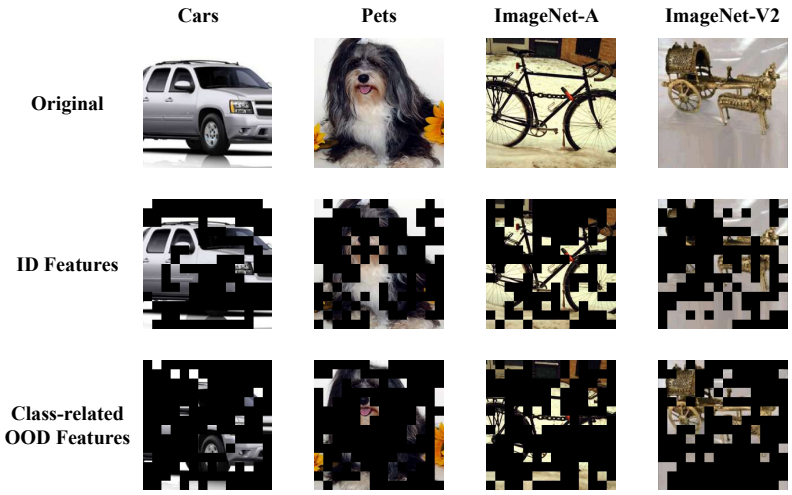


Figure 5: Visualization results of selected tokens containing ID and class-related OOD features.

According to the visualization results, tokens containing ID features are typically concentrated around the central regions of the object, such as a dog's face and eyes, which encode the majority of class-related semantics. In contrast, some tokens containing class-related OOD features are around the object boundaries, yet still capture semantically meaningful cues, particularly in OOD datasets. These observations further demonstrate that our feature decomposition strategy in DDMA is able to effectively separate two complementary types of class-related information, both of which contribute to improving adaptation performance.

E GENERALIZATION TO DIFFERENT MODELS

To evaluate the generalization of our method across different model backbones, we further conduct experiments with CLIP ViT-B/32. We follow the same experimental procedure and settings as those in Sec. 5.2. The results are presented in Tab. 8 and Tab. 9. On cross-domain benchmarks, DDMA improves the average classification accuracy of the original CLIP by approximately 2%. In particular, obvious gains are observed on EuroSAT, Flower102, and SUN397 by over 3%. Similarly, DDMA also achieves substantial improvements over CLIP ViT-B/32 in OOD scenarios, reaching an average accuracy of 57.88%. Overall, these results demonstrate that DDMA has generalization across different CLIP backbones and consistently provides effective adaptation.

Table 9: Accuracy Results (%) of adaptation for CLIP ViT-B/32 on OOD benchmarks.

Method	ImageNet	ImageNet-A	ImageNet-V2	ImageNet-S	ImageNet-R	Average
CLIP	62.36	43.28	54.19	43.85	70.57	**54.85**
DDMA (Ours)	63.94	47.59	57.42	46.37	74.06	**57.88**

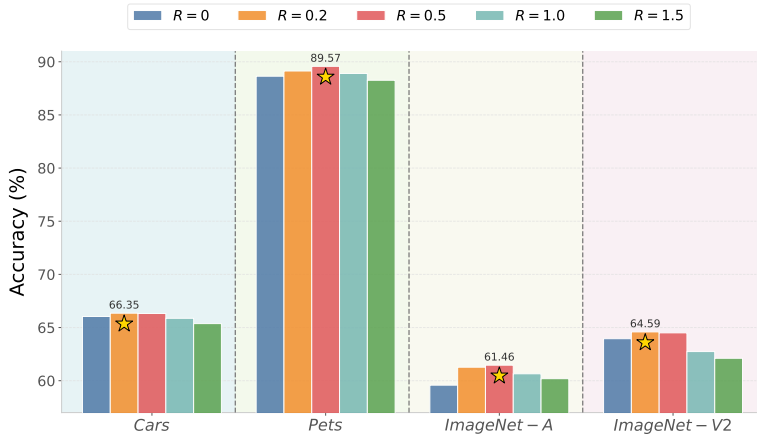


Figure 6: The results under different values of relaxation term R.

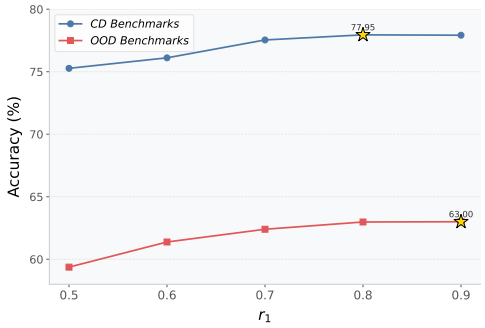


Figure 7: The average results across two types of benchmarks under different values of r_1.

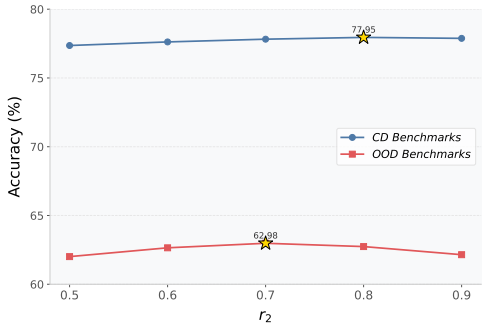


Figure 8: The average results across two types of benchmarks under different values of r_2.

F Impact of Hyper-parameters

In this section, we further investigate the impact of different hyper-parameter settings in DDMA on adaptation performance. Unless otherwise specified, all experimental procedures and settings remain the same as those described in Sec. 5.2 and Sec. 5.4, with only the values of the investigated hyper-parameters varied.

Impact of relaxation term R. When selecting high-confidence test samples, DDMA introduces a relaxation term R to appropriately loosen the selection threshold, allowing more class-related features to be incorporated. We therefore evaluate a range of relaxation term values to investigate their impact on adaptation performance. The results are reported in Fig. 6. The adaptation accuracy first increases and then decreases as R grows. When R is small, some moderately confident samples containing abundant class-related OOD features are excluded, resulting in insufficient feature information for distribution modeling and thus weaker adaptation effectiveness. In contrast, a large R overly relaxes the selection process, allowing low-quality samples to contaminate the feature distributions and consequently degrade performance. Based on these results, we set $R = 0.5$ as the default value, achieving a favorable trade-off in adaptation performance.

Impact of r_1 and r_2. The hyper-parameters r_1 and r_2 in Eq. (6) and (8) are key to constructing the masks that filter out class-irrelevant features and separate ID features from class-related OOD

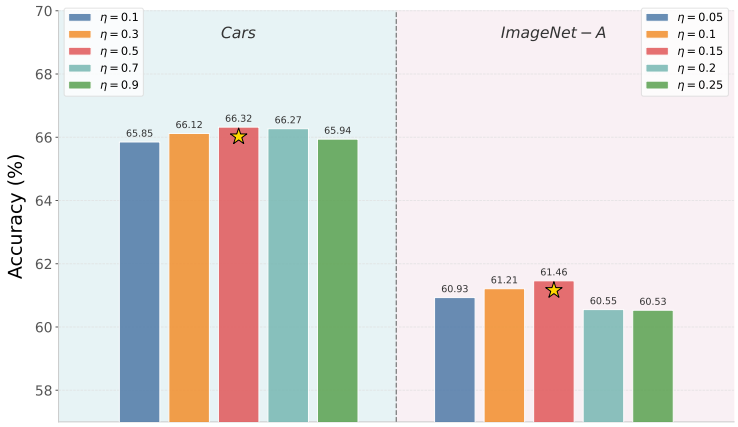


Figure 9: The results on StanfordCars and ImageNet-A under a series of η values.

Table 10: The results on StanfordCars and ImageNet-A under a series of weights w_1.

w1	Cars	w1	ImageNet-A
1.5	65.76	0.4	59.21
2	66.27	0.8	60.98
2.5	**66.32**	1.2	**61.46**
3	66.14	1.6	60.53

Table 11: The results on StanfordCars and ImageNet-A under a series of weights w_2.

w2	Cars	w2	ImageNet-A
0.05	66.15	0.05	58.87
0.2	66.29	0.2	60.08
0.6	**66.32**	0.4	**61.46**
1.2	65.57	0.8	61.13

features. We therefore conduct ablation studies to evaluate their effects, with the results reported in Fig. 7 and 8, respectively.

For r_1, a larger value allows $\mathbf{M}_1$ to retain more class-related features, resulting in richer semantic information for distribution modeling and better adaptation performance. Notably, increasing r_1 from 0.8 to 0.9 yields only a marginal accuracy gain, suggesting that most class-related features have already been retained. Since processing fewer tokens reduces the computational overhead, we set $r_1 = 0.8$ as the default value.

When r_2 is small (0.5 and 0.6), the extracted ID features are incomplete and cannot fully contribute to prediction refinement during adaptation. In contrast, a large setting $r_2 = 0.9$ introduces some class-related OOD features into the ID features, thus reducing their purity and weakening the performance. Based on these results, we set $r_2 = 0.8$ for cross-domain benchmarks and $r_2 = 0.7$ for OOD benchmarks that include more OOD features, to achieve a better balance.

Impact of update step η. After selecting a high-confidence test sample, the hyper-parameter η in Eq. (4) and (10) is used to update the sample selection threshold and the centers of the two feature distributions. We therefore evaluate the adaptation performance under different values of η. Since the optimal η value varies substantially across benchmarks, we select StanfordCars and ImageNet-A as representative, with the results reported in Fig. 9. The results show that both excessively large and small η values degrade adaptation performance. Detailed experimental settings of other datasets are provided in our source code.

Impact of correction weight w_1 and w_2. These two hyper-parameters control the weights of contribution from the ID and class-related OOD feature distributions during prediction refinement.

Their optimal values also vary across different benchmarks, so we report the results on StanfordCars and ImageNet-A as examples. As shown in Tab. 10 and 11, the optimal w_2 is generally higher on OOD benchmarks than on cross-domain datasets, as samples in OOD scenarios tend to contain more class-related OOD features that are insufficiently learned by CLIP. Details of other datasets are provided in our source code.