Dual Distribution Estimation for Zero-shot Noisy Test-Time Adaptation with VLMs

Wenjie Zhu[1], Yabin Zhang[3], Liang Xu[4], Xin Jin[2], Wenjun Zeng[2⋆], and Lei Zhang[1⋆]

[1] The Hong Kong Polytechnic University
[2] Eastern Institute of Technology, Ningbo
[3] Harbin Institute of Technology (Shenzhen)
[4] Shanghai Jiao Tong University
22040319r@connect.polyu.hk

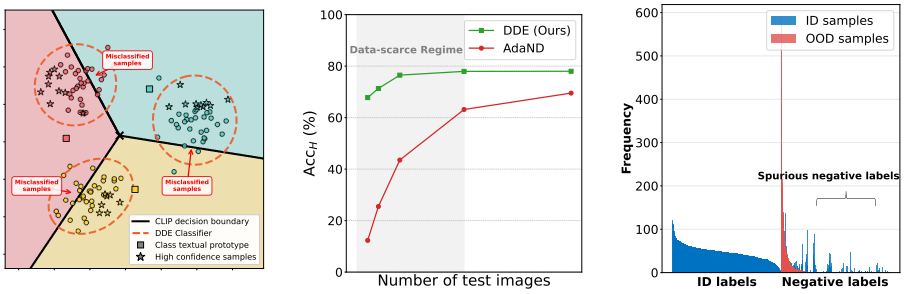
Abstract. While test-time adaptation (TTA) empowers vision-language models to adapt without costly retraining, it remains highly vulnerable to out-of-distribution (OOD) outliers prevalent in real-world applications. This discrepancy motivates Noisy TTA (NTTA), an online task to filter noisy OOD samples on the fly while maximizing in-distribution (ID) classification accuracy. Existing zero-shot NTTA approaches typically rely on test-time discriminative training, leading to overconfident misclassifications and significantly degraded inference efficiency. To address these limitations, we propose a novel framework named Dual Distribution Estimation (DDE), shifting the zero-shot NTTA paradigm from instance-level learning to training-free Gaussian distribution modeling. DDE incorporates two novel modules: Positive Feature Distribution Estimation (PFDE) and Negative Label Distribution Estimation (NLDE). PFDE explicitly models class-wise inclusion and exclusion Gaussian distributions to formulate a calibrated contrastive score, robustly enhancing ID accuracy. In parallel, NLDE improves OOD identification by explicitly modeling the negative label distribution to mine highly discriminative labels, effectively mitigating spurious correlations. Extensive experiments show that on the large-scale ImageNet benchmark, DDE achieves an improvement of 3.70% in harmonic mean accuracy and reduces the FPR95 for OOD detection by 6.20%, while ensuring highly scalable and efficient online inference. Furthermore, DDE is zero-shot and training-free, demonstrating remarkable robustness in data-scarce scenarios. Codes are available at `https://github.com/PolyU-VCLab/OpenOOD-VLM`.

Keywords: Noisy Test-time Adaptation · Dual Distribution Estimation · Out of Distribution · Vision Language Model

1 Introduction

Test-time adaptation (TTA) for vision-language models (VLMs) has garnered increasing attention, as it enhances pre-trained models directly during deployment without requiring costly human annotation or retraining. Most existing

⋆ Corresponding author



(a) CLIP vs. DDE Classifiers **(b)** Data scarce setting **(c)** Label distribution

Fig. 1: (a) CLIP vs. DDE Classifiers: AdaND [3] uses the CLIP classifier frequently misclassifying samples that deviate significantly from textual prototypes. Other TTA methods [60] (e.g. DMN) cache high-confidence samples, which still struggle to accurately represent complete visual class distribution. In contrast, our DDE achieves a more precise visual representation by effectively modeling the test distribution. **(b) Data scarce setting:** Unlike AdaND, which relies on extensive ID and noisy images for training, DDE dynamically models discriminative negative labels. This allows DDE to maintain higher robustness in data-scarce environments. **(c) Label distribution:** Some negative labels exhibit spurious correlations with clean ID data, whereas OOD outliers strongly align with only a narrow subset of negative labels.

TTA methods operate under a closed-world assumption that all test data strictly fall within a specific in-distribution (ID) label space [4, 19, 55, 60]. However, real-world test streams inevitably contain out-of-distribution (OOD) outliers. This discrepancy motivates the emerging task of Noisy TTA (NTTA) [3], which operates under a strict online setting with dual objectives: simultaneously filtering out noisy OOD samples on the fly and maximizing classification accuracy on clean ID data. Crucially, this dual-objective, online nature distinguishes NTTA from TTA for OOD detection [53, 58], which typically focuses solely on offline outlier evaluation with metrics like AUROC.

NTTA methodologies generally fall into two categories: those requiring task-specific training [24], and zero-shot NTTA, which leverages frozen VLMs for task-agnostic inference [3]. In this work, we focus on the more challenging and versatile zero-shot NTTA setting. A representative approach in this domain is AdaND [3], which trains a noise detector online using selected test samples. However, several limitations persist in this approach: (1) AdaND does not model test distribution historical test image features, relying instead on the CLIP text-image alignment for classification (Fig. 1a). This leads to frequent misclassifications when samples deviate significantly from the class textual prototypes. (2) AdaND necessitates a substantial number of samples to achieve satisfactory performance, which limits the robustness of the system in data-scarce scenarios (Fig. 1b). (3) AdaND relies on online training, which significantly degrades inference efficiency (Tab. 7). These limitations are summarized in Tab. 1.

Table 1: Comparison between methods of TTA, TTOOD (Test time OOD Detection), NTTA and our DDE. DDE estimates dual-distributions from historical image features, uniquely combining the ability to handle noisy data with a robust, training-free, high-efficiency architecture.

Methods	**DMN** [60] (TTA)	**AdaNeg** [58] (TTOOD)	**AdaND** [3] (NTTA)	**DDE** (NTTA)
OOD generalization	✓	✗	✓	✓
OOD(Noisy) detection	✗	✓	✓	✓
Metrics	Accuracy	AUROC, FPR95	Acc_S, Acc_N, Acc_H	Acc_S, Acc_N, Acc_H
Distribution Modeling	✗	✗	✗	✓
Robustness to data-scarce scenarios	✓	✓	✗	✓
Training-free	✓	✓	✗	✓
Fast inference speed	✓	✓	✗	✓

To address these limitations, we propose Dual Distribution Estimation (DDE), which is empowered by two novel modules: Positive Feature Distribution Estimation (PFDE) and Negative Label Distribution Estimation (NLDE). Specifically, PFDE enhances ID sample recognition by modeling mined positive image features as class-conditional Gaussian distributions and employing Bayes' theorem for robust posterior inference. To capture a more precise decision boundary, we introduce a dual Gaussian modeling approach that explicitly characterizes both the inclusion and exclusion distributions for each class. The final prediction is formulated as a calibrated contrastive score by subtracting the exclusion probability from the inclusion one.

In parallel, NLDE enhances noisy OOD perception by mining highly discriminative negative labels. This module is motivated by a key observation in Fig. 1c that some negative labels exhibit spurious correlations with clean ID data and true OOD outliers strongly align with only a narrow subset of specific negative labels. To exploit this, we model the image-to-label distribution and mine the most discriminative negative labels from the corpora dataset by quantifying their similarity discrepancy across positive and negative images. Finally, we integrate an adaptive thresholding mechanism [24] to dynamically separate ID and OOD samples during the online inference process.

We conduct extensive experiments to validate the advantages of DDE. On the large-scale ImageNet dataset, our DDE achieves a significant 3.70% improvement in harmonic mean accuracy while reducing FPR95 for OOD detection by 6.20%. Moreover, our method operates in a zero-shot and training-free manner, demonstrating strong scalability. We summarize our contributions as follows:

- We first identify several critical limitations of the existing AdaND [3] method in the zero-shot Noisy TTA setting: (1) it neglects historical test image feature distribution, thereby failing to capture class-specific diversity; (2) it lacks robustness in data-scarce scenarios, as it relies on extensive samples to train its noise detector; and (3) it requires online training, which significantly degrades inference efficiency.

- To address these pitfalls, we propose a novel framework named Dual Distribution Estimation, which consists of two novel, complementary modules. The Positive Feature Distribution Estimation formulates a calibrated contrastive score via class-wise inclusion and exclusion Gaussian distributions to enhance ID accuracy. In parallel, the Negative Label Distribution Estimation improves OOD identification by mining highly discriminative negative labels, effectively isolating true OOD outliers and mitigating spurious similarities among generic negative labels.
- Extensive experiments on large-scale benchmarks validate the superiority of our approach. On ImageNet, DDE achieves a remarkable 3.70% improvement in harmonic mean accuracy and reduces the FPR95 for OOD detection by 6.20%. Crucially, as a fully zero-shot and training-free method, DDE ensures highly scalable and efficient online inference.

2 Related Work

TTA [34, 45] allows pre-trained models to adjust to unlabeled data during deployment. By updating parameters or statistics on-the-fly, TTA addresses distribution shifts between training and real-world environments without requiring access to the source data. Existing approaches often rely on self-supervised objectives [27, 29, 38], such as entropy minimization [34, 45], or recalibrate batch normalization statistics to align feature distributions [31, 62]. Recent studies [4, 13, 19, 21, 26, 40, 55, 60, 61, 64, 69] have extended these techniques to VLMs. Many of these VLM-based methods [40, 55] employ test-time prompt tuning or residual learning to derive adaptive representations for individual samples. Alternatively, some frameworks [19, 60] investigate training-free mechanisms, such as dynamic adapters, to improve efficiency. Other strategies [13, 69] involve learning prompt distributions to account for diverse visual features. Despite these advancements, current methods often fail to account for noisy samples in the data stream, which can lead to substantial performance degradation.

Robustness under Distribution Shift. Recent research has studied model robustness under noisy and distribution-shifted. Recent works [7–9, 11, 15, 25, 32, 33, 42, 47, 50–52, 65, 66] enhance model interpretation by refining feature robustness. OWTTT [24] introduced the Open-World Test-Time Adaptation benchmark and utilized adaptive thresholding to exclude OOD samples. However, it relies on source data and is therefore unsuitable for source-free VLMs such as CLIP. AdaND [3] addresses zero-shot noisy TTA but depends on pseudo-labels to train a noise detector, which is ineffective in data-scarce settings.

Test-time OOD Detection on Vision Language Models. The cross-modal alignment of VLMs has driven progress in OOD detection, with recent work [12, 22, 23, 28, 41, 48, 53, 54, 56, 57, 57–59, 67, 67, 68] shifting toward test-time paradigms that adjust to the inference stream. AdaNeg [58] and OODD [53] employ memory banks and dynamic dictionaries to maintain adaptive negative proxies, thereby addressing the semantic gaps found in static labels. Nevertheless, these strategies frequently rely on high-confidence samples, which can prevent them from fully capturing the underlying data distribution.

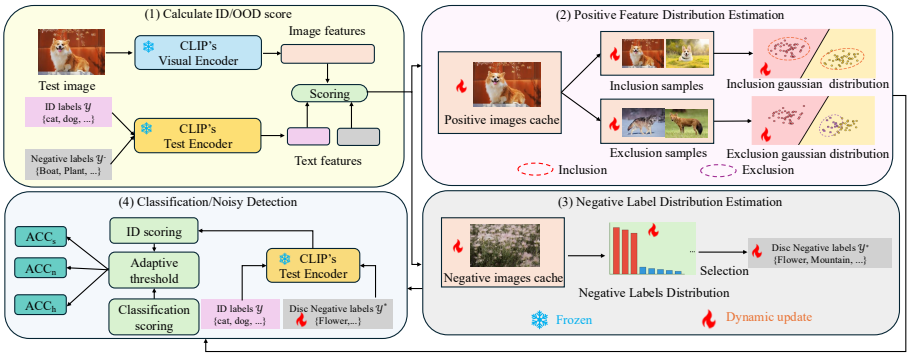


Fig. 2: Overview of our DDE framework. (1) Calculate the ID/OOD score for test images, dynamically caches positive and negative samples. (2) Model positive feature Gaussian distributions via inclusion and exclusion GDA estimating, (3) Filter the discriminative negative labels via negative labels distribution estimation, and (4) Employ an adaptive threshold for simultaneous ID classification and noise detection.

3 Method

3.1 Preliminaries

Gaussian Discriminant Analysis. Inspired by [13, 14], we assume that the embedding distribution of each class k follows a Gaussian distribution as $P(x \mid y = k) \sim \mathcal{N}(\mu_k, \Sigma_k)$, where μ_k and Σ_k are the mean vector and covariance matrix of class k, respectively. For the k-th class with N_k training samples, the mean and covariance matrix of different classes can be estimated as follows:

$$\mu_k = \frac{\sum_{n=1}^{N_k} x_n}{N_k}, \qquad \Sigma_k = \frac{\sum_{n=1}^{N_k} (x_n - \mu_k)(x_n - \mu_k)^T}{N_k}, \tag{1}$$

where μ_k and Σ_k represent the estimated mean and covariance matrix of the k-th class, respectively. x_n denotes the input embeddings.

OOD Detection with Negative Labels. Enhancing OOD detection with textual knowledge has recently gained traction [30, 58], particularly through the use of negative labels [18]. In addition to standard ID labels $\mathcal{Y}$, these methods typically introduce a disjoint set of negative labels $\mathcal{Y}^-$ and classify test samples as OOD when they exhibit higher similarity to negative labels than to ID classes. Since the performance of this approach relies on the relevance of the negative set, NegLabel [18] constructs the set $\mathcal{Y}^-$ by selecting the top M labels from a large corpus $\mathcal{Y}^{cor} = \{\widetilde{y}_1, \widetilde{y}_2, \ldots, \widetilde{y}_{N_c}\}$ that exhibit the most significant cosine distance from the ID labels. Formally, this selection is defined as $\mathcal{Y}^- = \text{Top}\left(\{d_i\}_{i=1}^{N_c}, \mathcal{Y}^{cor}, M\right)$, where N_c and M denote the number of candidate labels in the corpus and the number of selected negative labels, respectively.

The scoring function for OOD detection is defined as:

$$S_{\text{nl}}(x) = \sum_{i=1}^{K} \frac{\exp(x^\top \mathbf{w}_i/\tau)}{\sum_{j=1}^{K} \exp(x^\top \mathbf{w}_j/\tau) + \sum_{j=1}^{M} \exp(x^\top \tilde{\mathbf{w}}_j/\tau)}. \tag{2}$$

In this formulation, $x \in \mathcal{R}^D$ denotes the test image feature, while $\mathbf{w}_i \in \mathcal{R}^D$ and $\tilde{\mathbf{w}}_j \in \mathcal{R}^D$ are the textual prototypes for the ID labels $y_i \in \mathcal{Y}$ and negative labels $y_j^- \in \mathcal{Y}^-$, respectively, with D being the feature dimension. Meanwhile, $\tilde{\mathbf{w}}_j = f_{\text{txt}}(\rho(\tilde{y}_j)) \in \mathcal{R}^D$ represents the text prototype of the mined negative label $\tilde{y}_j$, and $\tau > 0$ is the temperature parameter.

3.2 Motivation

Current zero-shot NTTA methods, such as AdaND [3], often rely on simple linear classifiers for inference and noise detection. However, these classifiers fail to account for the underlying test distribution, thereby neglecting the diverse image features within each category. Furthermore, the absence of OOD-aware information aggregation results in redundant negative labels, which introduces interference and undermines classification accuracy. These challenges motivate the following research question:

Can we design a method to accurately estimate both the positive feature distribution and the negative label distribution?

In this work, we develop a dual distribution estimation framework, named DDE, as illustrated in Fig. 2. These two distribution estimation components facilitate ID classification and OOD detection, respectively, ultimately enhancing the performance of NTTA.

3.3 Dual Distribution Estimation

Positive and Negative Images Selection. Following eq. (2), we identify positive and negative samples for each batch by applying predefined thresholds λ_{pos} and λ_{neg}. Specifically, we follow AdaNeg [58] defining the positive set as $\mathcal{X}_{pos}^{batch} = \{v \in \mathcal{R}^D \mid S_{nl}(v) \geq \lambda_{pos}\}$ and the negative set as $\mathcal{X}_{neg}^{batch} = \{v \in \mathcal{R}^D \mid S_{nl}(v) < \lambda_{neg}\}$. where λ_{pos} and λ_{neg} represent the thresholds for isolating high-confidence pseudo-positive and pseudo-negative samples. Once the positive images $\mathcal{X}_{pos}^{batch}$ and negative images $\mathcal{X}_{neg}^{batch}$ are identified, they are stored in two separate caches, $\mathcal{C}_{pos}$ and $\mathcal{C}_{neg}$ as:

$$\begin{aligned} \mathcal{C}_{pos} &\leftarrow \text{Concat}(\mathcal{C}_{pos}, \mathcal{X}_{pos}^{batch}) \quad \text{s.t.} \quad |\mathcal{C}_{pos}| \leq Q, \\ \mathcal{C}_{neg} &\leftarrow \text{Concat}(\mathcal{C}_{neg}, \mathcal{X}_{neg}^{batch}) \quad \text{s.t.} \quad |\mathcal{C}_{neg}| \leq Q. \end{aligned} \tag{3}$$

Q indicates the maximum capacity constraints for each cache. Then we use them to estimate the positive features distribution and the negative labels distribution.

Positive Feature Distribution Estimation

(1) Gaussian Discriminant Analysis. Following the generative framework of Gaussian Discriminant Analysis (GDA) [1], we adopt a dual Gaussian modeling approach. In this setting, the classifier is derived from specific assumptions about the data distribution for each class. Following the standard GDA framework, the features are assumed to follow Gaussian distributions with a class-conditional covariance matrix, i.e., $P(x|y=k) \sim \mathcal{N}(\mu_k, \Sigma_k)$. Then we use the logit function defined in GDA [1], this function measures how well a sample x fits the distribution of class k as:

$$f_k(x) = \mu_k^T \Sigma_k^{-1} x - \frac{1}{2} \mu_k^T \Sigma_k^{-1} \mu_k + \log p_k. \tag{4}$$

Then we use Bayes' theorem, the posterior probability $P(y=k \mid x)$ for a given sample $x \in \mathcal{X}_{pos}^{batch}$ is determined by combining the class-conditional density $P(x \mid y=k)$ with a uniform prior $P(y=k) = 1/K$:

$$P(y=k \mid x) = \frac{P(x \mid y=k)P(y=k)}{\sum_{j=1}^{K} P(x \mid y=j)P(y=j)} = \frac{\exp(f_k(x))}{\sum_{j=1}^{K} \exp(f_j(x))}. \tag{5}$$

(2) Partition the Positive Images. Unlike prior works [13,69] that aggregate all soft assignments into a single distribution, we explicitly disentangle the positive feature space into inclusion and exclusion distributions. For each ID class k, we maintain two Gaussian distributions, $\mathcal{N}(\boldsymbol{\mu}_k^{\text{in}}, \boldsymbol{\Sigma}_k^{\text{in}})$ and $\mathcal{N}(\boldsymbol{\mu}_k^{\text{ex}}, \boldsymbol{\Sigma}_k^{\text{ex}})$, modeled from inclusion and exclusion features, respectively. Both distributions are updated via test-time streaming. At each test-time batch, PFDE is updated in a class-wise manner using the positive samples $\mathcal{X}_{pos}^{batch}$ selected from the current batch. Specifically, for a given class k, the inclusion distribution $\mathcal{N}(\boldsymbol{\mu}_k^{\text{in}}, \boldsymbol{\Sigma}_k^{\text{in}})$ is updated using features of samples in $\mathcal{X}_{pos}^{batch}$ whose top-1 CLIP prediction is class k. The exclusion distribution $\mathcal{N}(\boldsymbol{\mu}_k^{\text{ex}}, \boldsymbol{\Sigma}_k^{\text{ex}})$ is updated using features of samples in $\mathcal{X}_{pos}^{batch}$ whose top-1 prediction is not k but whose second-highest CLIP prediction corresponds to class k.These samples represent near-boundary cases that are easily confused with class k, allowing the exclusion distribution to characterize ambiguous regions in the feature space.

(3) Estimating Gaussian Distribution Parameters. Given a batch of positive test samples $\mathcal{X}_{pos}^{batch} = \{x_n\}_{n=1}^{N^{batch}}$ at time step t, we estimate the class-specific parameters for each category k, namely the means $\{\boldsymbol{\mu}_k^{\text{in}}, \boldsymbol{\mu}_k^{\text{ex}}\}_{k=1}^{K}$ and covariances $\{\boldsymbol{\Sigma}_k^{\text{in}}, \boldsymbol{\Sigma}_k^{\text{ex}}\}_{k=1}^{K}$. We estimate the parameters via the Expectation-Maximization (EM) algorithm. Specifically, the expectation step derives zero-shot predictions, while the maximization step refines the model parameters by maximizing the log-likelihood of the observed data as:

$$\begin{aligned}
\boldsymbol{\mu}_{k,t}^{s} &= \frac{N_{k,t-1}^{s} \boldsymbol{\mu}_{k,t-1}^{s} + \sum_{i=1}^{N^{batch}} P_k(y=k \mid x_i) x_i}{N_{k,t}^{s} + \sum_{i=1}^{N^{batch}} P_k(y=k \mid x_i)} \\
\boldsymbol{\Sigma}_{k,t}^{s} &= \frac{N_{k,t-1}^{s} \boldsymbol{\Sigma}_{k,t-1}^{s} + \sum_{i=1}^{N^{batch}} P_k(y=k \mid x_i)(x_i - \boldsymbol{\mu}_{k,t}^{s})(x_i - \boldsymbol{\mu}_{k,t}^{s})^{\top}}{N_{k,t}^{s} + \sum_{i=1}^{N^{batch}} P_k(y=k \mid x_i)}
\end{aligned} \tag{6}$$

$N^s_{k,t-1}$ represents the effective sample size, which we define as the cumulative number of observed samples for class k in branch s up to step $t-1$. where $s \in \{in, ex\}$ denotes the specific branch. x_i denotes the i-th image feature of $\mathcal{X}^{batch}_{in}$ or $\mathcal{X}^{batch}_{ex}$. To avoid introducing additional computations, We follow [10] by employing a shrinkage-based approach: $\hat{\mathbf{\Lambda}} = [(1-\epsilon)\hat{\mathbf{\Sigma}} + \epsilon\mathbf{I}]^{-1}$, with $\epsilon = 10^{-4}$. We effectively prevent the matrix from becoming singular during inversion.

(4) Test Time Classification with CLIP and GDA. The fused GDA logits of the inclusion and exclusion classifier are defined as: $f_k(x) = f^{in}_k(x) - \beta f^{ex}_k(x)$, where $\beta \in [0, 1]$ controls the weights of exclusion distributions. As the number of test samples increases and the estimated GDA distributions become more reliable, we integrate the logits from both the CLIP and GDA classifiers. Formally, the final fusion probability is defined as:

$$P_t(y = k \mid x) = \frac{\exp\big(\cos(x, \mathbf{w}_k)/\tau + \alpha_t \left(f^{in}_k(x) - \beta f^{ex}_k(x)\right)\big)}{\sum_{j=1}^{K} \exp\big(\cos(x, \mathbf{w}_j)/\tau + \alpha_t \left(f^{in}_j(x) - \beta f^{ex}_j(x)\right)\big)}. \tag{7}$$

where τ is the temperature parameter and $\mathbf{w}_k$ is the k-th zero-shot text prototype. We adopt a dynamic weighting strategy where the parameter α_t balances the contribution of the GDA classifier. Specifically, α_t remains low when test samples are scarce and increases as more data is accumulated. This is defined as $\alpha_t = \min\left(\rho \cdot (B \cdot t), \alpha_{max}\right)$, where ρ is a scaling factor, B is the batch size, t is the current iteration step, and α_{max} is a predefined upper bound (typically 1.0). This constraint prevents the GDA classifier from overly dominating the zero-shot prior, ensuring the model relies on the zero-shot classifier when the estimated test distribution remains unreliable.

Negative Label Distribution Estimation. Since negative labels contain a large number of irrelevant negative labels(Fig. 1c), we leverage these pseudo samples to identify the negative labels that can accurately characterize the OOD distribution and effectively distinguish ID from OOD samples. Specifically, we estimate the negative labels distribution corresponding to negative images by computing the similarity scores, the similarity scores are defined as follows:

$$\mathrm{Sim}(\mathcal{X}, y^-_i) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \frac{\exp(x^\top \tilde{\mathbf{w}}_i/\tau)}{\sum_{j=1}^{K} \exp(x^\top \mathbf{w}_j/\tau) + \sum_{j=1}^{N} \exp(x^\top \tilde{\mathbf{w}}_j/\tau)}. \tag{8}$$

An effective negative label will have a high similarity score on $\mathcal{X}_{neg}$ and a low similarity score on $\mathcal{X}_{pos}$. After obtaining the similarity scores, we compute a contrastive score to evaluate the ability to distinguish negative images from positive images. This discriminative score is defined as follows:

$$\Delta\mathrm{Sim}(y^-_i) = \mathrm{Sim}(\mathcal{X}_{neg}, y^-_i) - \mathrm{Sim}(\mathcal{X}_{pos}, y^-_i). \tag{9}$$

Then we select the most discriminative $\hat{M}$ negative labels as follow:

$$\mathcal{Y}^* = \mathrm{Top}\left(\{\Delta\mathrm{Sim}(y^-_i)\}^M_{i=1},\ \mathcal{Y}^-,\ \hat{M}\right). \tag{10}$$

We use the discriminative negative labels $\mathcal{Y}^*$ to calculate the confidence score:

$$S_{\text{nl}}(x) = \sum_{i=1}^{K} \left(\frac{\exp(x^\top \mathbf{w}_i/\tau)}{\sum_{j=1}^{K} \exp(x^\top \mathbf{w}_j/\tau) + \sum_{j=1}^{\hat{M}} \exp(x^\top \mathbf{w}_j^*/\tau)} \right). \tag{11}$$

Here, $\mathbf{w}_j^* = f_{\text{txt}}(\rho(y_j^*)) \in \mathbb{R}^D$ denotes the text feature of the $y_j^* \in \mathcal{Y}^*$.

Adaptive Threshold. The OOD detection process utilizes a scoring function $S(\cdot)$ to distinguish between ID and OOD inputs. Specifically, the OOD detector $G_\lambda(\cdot)$ is defined by a threshold $\lambda \in \mathbb{R}$, such that an input x_i is classified as follows: $G_\lambda(x_i) =$ Clean if $S(x_i) \geq \lambda$, and $G_\lambda(x_i) =$ Noise otherwise. A test sample x_i is classified as ID if $S(x_i) \geq \lambda$. A fixed threshold λ generalizes poorly across diverse ID datasets in NTTA. To address this, [24] proposes an adaptive threshold that dynamically calibrates λ by minimizing intra-class variance, leveraging the bimodal nature of OOD scores.

$$\min_{\lambda} \frac{1}{Q_{id}} \sum_{i} \left[S(x_i) - \frac{1}{Q_{id}} \sum_{j} \mathbb{1}(S(x_j) > \lambda) S(x_j) \right]^2 + \frac{1}{Q_{ood}} \sum_{i} \left[S(x_i) - \frac{1}{Q_{ood}} \sum_{j} \mathbb{1}(S(x_j) \leq \lambda) S(x_j) \right]^2, \tag{12}$$

where $Q_{id} = \sum_i^Q \mathbb{1}(S(x_i) > \lambda)$, $Q_{ood} = \sum_i^Q \mathbb{1}(S(x_i) \leq \lambda)$ where Q_{id} and Q_{ood} are the length of the test time pseudo ID and OOD samples.
The overall procedure of our DDE method is detailed in algorithm 1.

4 Experiments

4.1 Experimental Setup

Datasets and Benchmarks. We evaluate our method under the NTTA setting, using ImageNet-1K [6] as the primary ID dataset along with its distribution-shifted variants, namely ImageNet-S [46], ImageNet-A [17], ImageNet-V2 [37], and ImageNet-R [16]. To assess generalization beyond ImageNet-style categories, we also incorporate fine-grained ID datasets including CUB-200-2011 [44], Stanford Cars [20], Food-101 [2], and Oxford-IIIT Pet [35]. For OOD evaluation, we utilize iNaturalist [43], SUN [49], Texture [5], and Places [63] . Following the NTTA protocol, the test stream is constructed by sequentially mixing ID and noisy samples, which include both OOD. All experiments are conducted in a zero-shot and source-free manner.
Implementation Details. We use the visual encoder of ViT-B/16 pretrained by CLIP [36] for all experiments, while DDE performs fully training-free online adaptation. For the Positive and Negative Image Selection stage, we follow the configuration of AdaNeg [58], setting $\lambda_{pos} = 0.75$ and $\lambda_{neg} = 0.25$ for all the experiments, the detailed setting analysis can be found in the Supplementary Material. The queue length is maintained at $Q = 1000$. For the PFDE stage, we set the exclusion GDA weight $\beta = 0.5$ and the scaling factor $\rho = 0.005$. In the NLDE stage, we employ $\hat{M} = 500$ for the ImageNet benchmark [6] and $\hat{M} = 100$ for fine-grained dataset.

Algorithm 1 Dual Distribution Estimation (DDE)

Require: ID label space $\mathcal{Y}$, candidate negative labels $\mathcal{Y}^-$, batch data $\mathcal{X}^{batch}$.
Initialize dual Gaussian parameters $\mathcal{N}(\boldsymbol{\mu}_k^{\text{in}}, \boldsymbol{\Sigma}_k^{\text{in}})$ and $\mathcal{N}(\boldsymbol{\mu}_k^{\text{ex}}, \boldsymbol{\Sigma}_k^{\text{ex}})$;
Initialize two caches $\mathcal{C}_{pos}$ and $\mathcal{C}_{neg}$;
for each incoming batch $\mathcal{X}^{batch}$ **do**
  {**Positive and Negative Images Selection**}
  Select positive/negative features and update caches $\mathcal{C}_{\text{pos}}$ and $\mathcal{C}_{\text{neg}}$ via eq. (3);
  {**Positive Features Distribution Estimation**}
  Calculate class probability of GDA via eq. (4) and eq. (5)
  Partition the Positive Images.
  Estimate inclusive and exclusive Gaussian Distribution Parameters $(\boldsymbol{\mu}_k^{\text{in}}, \boldsymbol{\Sigma}_k^{\text{in}})$ and $(\boldsymbol{\mu}_k^{\text{ex}}, \boldsymbol{\Sigma}_k^{\text{ex}})$ via eq. (6);
  Fuse classifier of CLIP and GDA to predict labels via eq. (7);
  {**Negative Labels Distribution Estimation**}
  Calculate similarity score $\text{Sim}(\mathcal{X}, y_i^-)$ and discriminative score $\Delta\text{Sim}(y_i^-)$ via eq. (8) and eq. (9);
  Select top-$\hat{M}$ discriminative negative labels via eq. (10);
  Calculate the NegLabel OOD scores with discriminative negative labels via eq. (11);
  {**Adaptive Threshold**}
  Calculate the adaptive threshold and separate ID and OOD via eq. (12).
end for
Output: Predicted labels and OOD scores.

Evaluation Metrics. In the noisy TTA setting, we follow OWTTT [24] to report ID accuracy (Acc_S), noisy detection accuracy (Acc_N), and their harmonic mean (Acc_H). The definition details are available in the Supplementary Material.

4.2 Main Results

Zero-Shot Noisy TTA on ImageNet and its Variants. Tab. 2 presents the results for ImageNet and its distribution-shifted variants. DDE consistently achieves the highest harmonic mean accuracy (Acc_H) across all five ID datasets. Specifically, on ImageNet, DDE improves the average Acc_H to 76.79%, outperforming the strongest baseline, AdaND (73.09%), by 3.70%. DDE simultaneously improves ID accuracy and noisy detection performance, reaching 67.73% Acc_S and 99.42% Acc_N on iNaturalist. Similar improvements are observed on ImageNet-A (+6.57% over AdaND), ImageNet-V2 (+3.57%), and ImageNet-R (+3.01%). These findings suggest that dual distribution modeling effectively balances classification and noise detection under distribution shifts.

Zero-Shot Noisy TTA across Various ID Datasets. Tab. 3 evaluates generalization to fine-grained ID datasets. DDE consistently delivers the best overall performance across CUB-200-2011, Stanford Cars, Food-101, and Oxford-IIIT Pet. For example, on Food-101, DDE achieves an average Acc_H of 93.48%, outperforming AdaND (92.23%). On Stanford Cars, DDE reaches 79.14% average Acc_H, improving over AdaND (77.05%). The gains are particularly significant

Table 2: Zero-shot noisy TTA results for ImageNet-1k, ImageNet-S, ImageNet-A, ImageNet-V2, ImageNet-R as the ID datasets. **Bold** indicates the best performance.

ID	Method	iNaturalist			SUN			Texture			Places			Avg		
		Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H
ImageNet	ZS-CLIP [36]	54.01	86.53	66.51	53.43	83.96	65.30	52.71	78.52	63.08	53.35	80.50	64.17	53.38	82.38	64.77
	Tent [45]	48.56	35.74	41.18	55.44	75.54	63.95	54.94	70.93	61.92	55.76	73.98	63.59	53.67	64.05	57.66
	TPT [39]	52.58	88.93	66.09	51.91	86.09	64.77	51.11	80.01	62.38	51.80	82.89	63.76	51.85	84.48	64.25
	DMN [60]	63.41	98.74	77.22	61.77	88.53	72.77	60.32	46.26	52.36	61.38	77.96	68.68	61.73	77.87	67.76
	AdaNeg [58]	62.00	99.21	76.31	61.14	92.95	73.76	61.40	86.65	71.87	61.36	82.23	70.28	61.48	90.26	73.06
	OODD [53]	59.40	99.50	74.39	60.01	90.28	72.10	59.99	77.84	67.76	59.92	79.51	68.34	59.83	86.78	70.65
	AdaND [3]	63.26	96.87	76.54	61.34	89.44	72.77	62.45	83.54	71.47	61.92	84.82	71.58	62.24	88.67	73.09
	DDE	**67.72**	**99.45**	**80.57**	**65.57**	**96.25**	**78.00**	**63.60**	**94.04**	**75.88**	**65.31**	82.02	**72.71**	**65.55**	**92.94**	**76.79**
ImageNet-S	ZS-CLIP [36]	34.17	83.46	48.49	33.46	81.20	47.39	32.61	75.57	45.56	33.40	77.10	46.61	33.41	79.33	47.01
	Tent [45]	30.46	26.86	28.55	36.57	71.82	48.46	36.63	66.63	47.06	36.87	70.32	48.38	35.07	58.91	43.11
	TPT [39]	32.16	86.52	46.89	31.55	83.86	45.85	30.74	77.39	44.00	31.56	80.05	45.27	31.50	81.95	45.50
	DMN [60]	42.69	98.36	59.54	41.54	87.94	56.42	40.80	75.12	52.88	41.32	78.53	54.14	41.59	84.99	55.75
	AdaNeg [58]	39.90	99.82	57.01	39.90	96.23	56.41	39.21	89.01	54.44	39.90	89.11	55.12	39.73	93.54	55.75
	OODD [53]	42.41	98.31	59.26	42.81	69.54	53.00	32.83	89.75	48.08	32.67	89.17	47.82	37.68	86.69	52.04
	AdaND [3]	40.97	93.54	56.98	40.25	85.06	54.64	38.31	74.43	50.58	39.60	79.57	52.88	39.78	83.15	53.77
	DDE	**43.23**	**99.47**	**60.27**	41.42	**96.55**	**57.97**	**43.61**	**89.15**	**58.57**	40.42	83.59	54.49	**42.17**	**92.19**	**57.83**
ImageNet-A	ZS-CLIP [36]	34.73	80.69	48.56	34.20	78.83	47.70	33.97	76.60	47.07	33.96	75.11	46.77	34.22	77.81	47.53
	Tent [45]	34.99	77.19	48.15	34.83	77.05	47.97	34.36	75.19	47.17	34.60	73.83	47.12	34.70	75.81	47.60
	TPT [39]	34.12	81.17	48.04	33.20	80.23	46.97	33.12	79.92	46.83	33.05	77.00	46.25	33.37	79.58	47.02
	DMN [60]	42.59	97.17	59.22	41.90	84.57	56.04	40.62	58.74	48.03	40.95	74.45	52.84	41.52	78.73	54.03
	AdaNeg [58]	42.55	98.58	59.45	42.11	86.94	56.74	42.15	92.30	57.88	41.46	75.23	53.46	42.07	88.26	56.88
	OODD [53]	41.50	99.22	58.52	36.23	92.73	52.10	36.94	91.22	52.58	36.18	72.01	48.16	37.71	88.80	52.84
	AdaND [3]	43.59	91.19	58.98	41.96	80.93	55.27	45.04	79.97	57.62	42.85	72.13	53.76	43.36	81.06	56.41
	DDE	**48.85**	**99.20**	**65.46**	**47.37**	**94.38**	**63.08**	**51.14**	**90.21**	**65.28**	**46.70**	**76.83**	**58.09**	**48.51**	**90.16**	**62.98**
ImageNet-V2	ZS-CLIP [36]	48.01	85.72	61.55	47.37	83.23	60.38	46.81	77.54	58.38	47.39	79.41	59.36	47.39	81.47	59.92
	Tent [45]	47.94	76.98	59.08	48.28	80.50	60.36	47.56	74.47	58.05	48.34	77.37	59.50	48.03	77.33	59.25
	TPT [39]	46.63	88.37	61.05	46.12	85.58	59.94	45.21	79.14	57.55	46.02	81.95	58.94	46.00	83.76	59.37
	DMN [60]	55.22	98.02	70.64	54.33	86.09	66.62	53.84	70.50	61.05	54.18	75.53	63.10	54.39	82.53	65.35
	AdaNeg [58]	52.36	99.43	66.27	50.73	95.53	66.27	51.21	88.88	64.98	50.75	86.70	64.02	51.26	92.64	65.97
	OODD [53]	53.01	99.45	69.16	53.60	90.48	67.32	55.53	77.80	63.42	53.57	79.52	64.02	53.43	86.81	65.98
	AdaND [3]	56.32	97.06	71.28	54.78	86.64	67.12	57.28	80.61	66.97	55.81	79.24	65.49	56.05	85.89	67.72
	DDE	**59.84**	**99.52**	**74.74**	**56.64**	**97.51**	**71.66**	**55.41**	**94.77**	**69.93**	**56.72**	**87.52**	**68.83**	**57.15**	**94.83**	**71.29**
ImageNet-R	ZS-CLIP [36]	61.99	94.39	74.83	61.82	88.95	72.94	60.91	77.05	68.04	61.68	84.86	71.44	61.60	86.31	71.81
	Tent [45]	65.22	91.45	76.14	65.06	85.61	73.93	63.33	69.99	66.49	64.93	82.38	72.62	64.64	82.36	72.30
	TPT [39]	60.95	94.80	74.20	60.85	89.98	72.60	59.98	77.79	67.73	60.67	85.79	71.08	60.61	87.09	71.40
	DMN [60]	69.44	98.43	81.43	69.23	92.78	79.29	68.94	88.12	77.36	69.00	83.93	75.73	69.15	90.81	78.46
	AdaNeg [58]	71.55	99.51	83.24	71.36	96.65	82.10	70.54	87.30	78.03	71.28	85.01	77.54	71.18	92.12	80.23
	OODD [53]	68.52	99.85	81.27	69.76	95.41	80.59	69.58	91.42	79.02	69.62	85.20	76.63	69.37	92.97	79.38
	AdaND [3]	72.21	99.59	83.72	71.02	95.94	81.62	70.44	81.43	75.54	70.85	92.14	80.10	71.13	92.28	80.25
	DDE	**75.84**	**98.88**	**85.84**	**76.05**	**97.26**	**85.35**	69.06	**93.65**	**79.50**	**75.61**	90.18	**82.25**	**74.14**	**94.99**	**83.24**

on challenging fine-grained datasets such as CUB, where accurately modeling intra-class variance is essential. These results demonstrate that our positive feature distribution estimation generalizes beyond ImageNet-style categories and remains robust in fine-grained recognition settings.

Traditional OOD Detection. Tab. 4 reports zero-shot OOD detection results on ImageNet. DDE achieves the best overall performance, reaching 97.89% AUROC and reducing the average FPR95 to 9.80%, substantially outperforming AdaND (95.58% AUROC, 16.00% FPR95). These results demonstrate the effectiveness of dynamically estimating discriminative negative semantic distributions for OOD separation.

4.3 Analyses and Discussions

Ablation Study. Tab. 5 illustrates the contribution of each module in DDE. Specifically, replacing PFDE (in) with PFDE (in+ex) improves the harmonic mean from 72.77% to 73.52% in the noisy stream, confirming the benefit of

Table 3: Zero-shot noisy TTA results for CUB-200-2011, STANFORD-CARS, Food-101, and Oxford-IIIT Pet as the ID datasets. **Bold** indicates the best performance.

Dataset	Method	iNaturalist			SUN			Texture			Places			Avg		
		Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H
CUB-200-2011	ZS-CLIP [36]	38.13	88.06	53.22	38.10	87.86	53.15	37.56	79.11	50.94	38.00	87.81	53.04	37.95	85.71	52.59
	Tent [45]	37.02	46.95	41.40	38.61	55.55	45.56	34.98	41.77	38.07	40.41	74.83	52.48	37.75	54.78	44.38
	TPT [39]	37.41	89.57	52.78	37.49	89.67	52.87	36.88	81.67	50.81	37.44	89.45	52.79	37.30	87.59	52.31
	DMN [60]	55.96	98.66	71.41	56.02	98.72	71.48	56.19	99.36	71.78	56.11	96.52	70.96	56.07	98.32	71.41
	AdaNeg [58]	56.82	99.65	72.38	56.79	99.14	72.21	56.96	93.37	70.75	56.92	99.31	72.37	56.87	97.87	71.93
	OODD [53]	56.06	99.73	71.77	56.02	99.84	71.77	56.06	99.91	71.82	56.06	98.27	71.39	56.05	99.44	71.69
	AdaND [3]	52.34	96.40	67.84	52.41	93.91	67.27	51.82	81.24	63.28	51.82	91.51	66.17	52.10	90.77	66.14
	DDE	**58.56**	**99.58**	**73.77**	**58.74**	**99.55**	**73.91**	**59.89**	**99.80**	**74.88**	**58.47**	**96.28**	**72.71**	**56.92**	**98.80**	**73.82**
Stanford-CARS	ZS-CLIP [36]	50.18	96.62	66.05	53.48	98.81	69.40	53.59	99.05	69.55	53.36	98.05	69.11	52.65	98.13	68.53
	Tent [45]	44.12	52.33	47.88	54.27	94.51	68.95	54.60	97.37	69.97	54.33	96.65	69.56	51.83	85.22	64.09
	TPT [39]	49.24	96.97	65.31	52.40	98.83	68.49	52.75	99.27	68.89	52.42	98.39	68.40	51.70	98.36	67.77
	DMN [60]	64.33	99.99	78.29	64.35	99.63	78.19	64.35	99.70	78.21	64.53	98.29	77.91	64.39	99.40	78.15
	AdaNeg [58]	63.66	99.98	77.80	63.51	99.89	77.65	63.71	99.87	77.79	63.54	99.66	77.60	63.61	99.85	77.71
	OODD [53]	63.55	99.99	77.71	63.55	99.88	77.68	63.56	99.95	77.71	63.55	99.04	77.42	63.55	99.72	77.63
	AdaND [3]	62.80	99.79	77.09	62.73	99.82	77.04	62.91	99.75	77.16	62.76	99.29	76.91	62.80	99.66	77.05
	DDE	**65.51**	**99.99**	**79.16**	**65.30**	**99.89**	**78.98**	**66.35**	**99.93**	**79.75**	**65.43**	98.66	**78.68**	**65.65**	99.62	**79.14**
Food-101	ZS-CLIP [36]	80.60	94.76	87.11	80.75	96.08	87.75	80.51	93.12	86.36	80.62	94.62	87.06	80.62	94.65	87.07
	Tent [45]	75.83	25.09	37.70	82.86	85.10	83.97	82.54	87.03	84.73	82.26	80.13	81.18	80.87	69.34	71.90
	TPT [39]	79.70	94.93	86.65	79.92	96.19	87.30	79.70	93.86	86.20	79.76	95.14	86.77	79.77	95.03	86.73
	DMN [60]	84.17	99.92	91.37	84.13	99.87	91.33	84.14	95.71	89.55	84.21	99.49	91.21	84.16	98.75	90.87
	AdaNeg [58]	84.17	99.74	91.29	84.16	98.86	90.92	84.23	96.82	90.09	84.16	93.65	88.65	84.18	97.27	90.24
	OODD [53]	83.63	99.99	91.09	83.77	99.99	91.17	83.75	97.22	89.98	83.76	99.91	91.13	83.73	99.28	90.84
	AdaND [3]	86.50	99.87	92.71	86.40	99.64	92.55	86.44	96.51	91.20	86.42	99.40	92.46	86.44	98.85	92.23
	DDE	**88.35**	**99.93**	**93.78**	**88.23**	**99.78**	**93.65**	**88.30**	**97.94**	**92.87**	**88.24**	**99.65**	**93.61**	**88.28**	**99.33**	**93.48**
Oxford-IIIT Pet	ZS-CLIP [36]	78.58	88.30	83.16	79.75	87.30	83.35	80.20	91.16	85.33	79.59	84.17	81.82	79.53	87.73	83.41
	Tent [45]	80.07	78.09	79.07	81.19	68.30	74.19	81.48	74.72	77.95	80.64	62.51	70.43	80.84	70.91	75.41
	TPT [39]	77.56	89.71	83.19	78.87	89.82	83.99	79.17	92.26	85.22	78.62	87.32	82.74	78.56	89.78	83.78
	DMN [60]	87.98	97.62	92.55	87.90	97.55	92.47	88.17	96.77	92.27	88.12	96.91	92.31	88.04	97.21	92.40
	AdaNeg [58]	88.06	99.49	93.43	87.98	91.72	89.81	88.25	94.56	91.30	88.23	91.29	89.73	88.13	94.26	91.07
	OODD [53]	85.04	99.99	91.91	85.58	99.96	92.21	85.69	99.82	92.22	85.58	99.74	92.12	85.47	99.88	92.12
	AdaND [3]	85.81	98.78	91.84	85.82	98.19	91.59	85.86	98.68	91.82	85.88	96.58	90.92	85.84	98.06	91.54
	DDE	**88.28**	**99.97**	**93.76**	**88.72**	99.14	**93.64**	**88.74**	**99.63**	**93.87**	**88.58**	**96.88**	**92.54**	**88.58**	98.90	**93.45**

modeling exclusion distributions. Furthermore, NLDE significantly boosts noisy sample detection, increasing Acc_N from 83.96% to 96.25%. The full DDE (incorporating Ada λ) achieves the best balance, reaching the highest harmonic mean of **78.00%**. This demonstrates that the three modules work synergistically to handle noise effectively.

Robustness to Noise. As shown in Fig. 3a, the model exhibits remarkable robustness, with Acc_H decreasing only marginally from 78.00% to 76.87% as the noise rate increases from 0.1 to 0.7. This negligible fluctuation confirms that DDE effectively filters out corrupted samples, ensuring reliable adaptation even in heavily noisy environments.

Analysis of Different Backbones. As shown in Fig. 3b, our method scales effectively across various architectures, with performance improving consistently as model capacity increases. ViT-L achieves the highest Acc_H of 81.20%, outperforming the RN50 baseline by 10.12%.

Performance Stability. Fig. 3c shows that DDE maintains consistent performance across all adaptation stages. The stable Acc_H across segments confirms that our method keeps performance stable over time.

Comparison with Fixed Threshold. As shown in Fig. 3d, the model performance is highly sensitive to the choice of fixed thresholds. In contrast, our adaptive mechanism achieves competitive results without manual searching, nearly matching the optimal fixed parameter and demonstrating its effectiveness.

Table 4: Zero-shot OOD detection results for ImageNet as the ID dataset. **Bold** indicates the best performance.

Method	iNaturalist		SUN		Texture		Places		Avg	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Max-Logit	89.31	61.66	87.43	64.39	71.68	86.61	85.95	63.67	83.59	69.08
Energy	85.09	81.08	84.24	79.02	65.56	93.65	83.38	75.08	79.57	82.21
MCM	94.61	30.91	92.57	37.59	86.11	57.77	89.77	44.69	90.77	42.74
NegLabel	99.49	1.91	95.49	20.53	90.22	43.56	91.64	35.59	94.21	25.40
NegRefine	99.57	1.51	94.63	22.93	94.68	21.15	90.41	39.10	94.82	21.16
OODD	99.74	0.51	96.84	16.30	93.19	34.09	93.77	28.68	95.89	19.90
AdaNeg	98.71	0.59	97.44	9.50	94.55	34.34	94.93	31.27	96.66	18.92
AdaND	98.91	4.19	95.86	17.08	93.01	21.76	94.55	20.95	95.58	16.00
DDE	**99.83**	**0.47**	**98.97**	**3.45**	**95.98**	**20.85**	**96.78**	**14.42**	**97.89**	**9.80**

Table 5: Ablation of DDE components on ImageNet under noisy and clean streams. For PFDE, "(in)" denotes using only the inclusion distribution, while "(in+ex)" uses both inclusion and exclusion distributions.

Components				Clean Stream			Noisy Stream		
PFDE (in)	PFDE (in+ex)	NLDE	Adaλ	Acc_S ↑	Acc_N ↑	Acc_H ↑	Acc_S ↑	Acc_N ↑	Acc_H ↑
				66.62	–	–	53.43	83.96	65.30
✓				69.62	–	–	61.87	88.34	72.77
	✓			70.34	–	–	62.69	88.90	73.52
		✓		66.83	–	–	55.72	95.99	70.51
	✓	✓		70.81	–	–	64.14	95.76	76.82
	✓	✓	✓	71.14	–	–	**65.57**	**96.25**	**78.00**

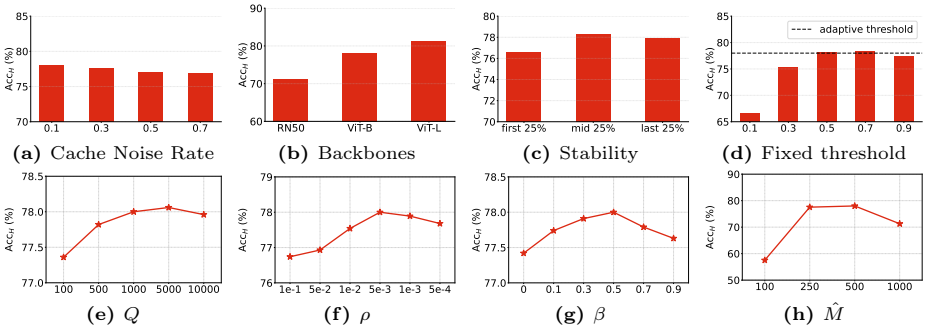
Cache Size Q. We analyze Q in Fig. 3e. While performance peaks at $Q = 5000$ due to the increased diversity of candidate samples, maintaining such a large cache incurs substantial memory overhead. To strike an optimal balance between adaptation accuracy and memory efficiency, we set $Q = 1000$ by default.

The Scaling Factor ρ. We analyze ρ in Fig. 3f. Acc_H remains relatively stable across different ρ values, reaching its maximum of 78.00% at $\rho = 0.005$. The performance remains robust even at a very small ρ of 0.0005 (77.68%).

The Weights of Exclusion GDA Model β. We analyze β in Fig. 3g. Acc_H gradually improves from 77.42% to its peak of 78.00% as β increases to 0.5. The performance stays consistently high across the tested range, showing the method's insensitivity to β.

The Number of Discriminative Negative Labels $\hat{M}$. We analyze $\hat{M}$ in Fig. 3h. The best Acc_H of 78.00% is achieved at $\hat{M} = 500$. Either too small or too large values of $\hat{M}$ significantly degrade the performance, with Acc_H dropping to 71.28% at $\hat{M} = 1000$.

Robustness to Diverse OOD Settings. We further evaluate DDE under four challenging OOD settings, including heterogeneous, noisy, medical, and industrial scenarios. As shown in Table 6, DDE consistently achieves the best performance across all settings, demonstrating strong robustness under diverse real-world OOD streams.



(a) Cache Noise Rate **(b)** Backbones **(c)** Stability **(d)** Fixed threshold

(e) Q **(f)** ρ **(g)** β **(h)** $\hat{M}$

Fig. 3: Comprehensive analysis of model components and hyperparameters. We evaluate the harmonic mean (Acc_H) across various settings: (a) noise ratios of positive and negative images, (b) backbones, (c) stability, and (d) fixed threshold. The bottom row shows sensitivity analysis for (e) cache size for positive and negative images, (f) scaling factor ρ, and (g) exclusion GDA model weight β, (h) number of discriminative negative labels $\hat{M}$. All experiments are conducted on ImageNet.

Table 6: Performance (ACC_H) under diverse OOD scenarios.

Methods	Heter	Noise	Medical	Industrial
CLIP	63.23	65.92	33.45	60.63
AdaND	66.25	67.03	36.82	64.25
DDE	**74.89**	**75.90**	**45.96**	**71.15**

Table 7: Complexity analyses. Results are obtained using a GeForce RTX 3090.

Methods	**Testing (min)**	**Memory (GiB)**	**Param (k)**
TENT [45]	3.72	14.99	40
TPT [39]	321.90	21.23	8.2
AdaNeg [58]	2.10	5.90	–
AdaND [3]	6.59	4.57	1.0
DDE	1.84	5.41	–

Complexity Analyses. Table 7 summarizes the efficiency of DDE. Compared with optimization-based methods such as TPT [39], DDE achieves the lowest test-time latency (1.84 min) while introducing no additional learnable parameters and maintaining low memory consumption, demonstrating excellent efficiency and scalability.

5 Conclusion

We present Dual Distribution Estimation (DDE), a novel, training-free, zero-shot framework tailored for NTTA. We first identify several limitations inherent in existing NTTA methods. To address these pitfalls, DDE incorporates two key components: (1) Positive Feature Distribution Estimation, which utilizes dual Gaussian distributions (inclusion and exclusion) to refine ID accuracy; and (2) Negative Label Distribution Estimation, which selects discriminative negative labels to filter out noise from generic ones. Extensive experiments on large-scale benchmarks demonstrate that DDE achieves state-of-the-art performance while maintaining both robustness and efficiency.

References

1. Bishop, C.M., Nasrabadi, N.M.: Pattern recognition and machine learning, vol. 4. Springer (2006)
2. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: European conference on computer vision. pp. 446–461. Springer (2014)
3. Cao, C., Zhong, Z., Zhou, Z., Liu, T., Liu, Y., Zhang, K., Han, B.: Noisy test-time adaptation in vision-language models. arXiv preprint arXiv:2502.14604 (2025)
4. Chen, X., Zhai, H., Zhang, C., Shi, X., Li, R.: 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)
5. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3606–3613 (2014)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Feng, W., Ge, Z.: Generalized category discovery under domain shift: A frequency domain perspective. Advances in Neural Information Processing Systems **38**, 111721–111749 (2026)
8. Feng, W., Jiang, Y., Zhou, S., Ge, Z.: Beyond the static world: Continual category discovery under visual drift. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25032–25042 (2026)
9. Feng, W., Jiang, Y., Zhou, S., Qi, Z., Xu, Z., Wang, Z., Tang, F., Ge, Z.: Seeing through the shift: Causality-inspired robust generalized category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17766–17775 (2026)
10. Friedman, J.H.: Regularized discriminant analysis. Journal of the American statistical association **84**(405), 165–175 (1989)
11. Fu, R., Tang, W., Wang, Z., Tan, J.Y., Zhang, Z., Kang, Z., Qi, M., Zhang, S., Fong, S.: Modalimmune: Immunity driven unlearning via self destructive training. arXiv preprint arXiv:2602.16197 (2026)
12. Guo, J., Luo, X., Zheng, J., Wang, Y., Chang, K.W., Wang, W., Liu, J.: Quantized-tinyllava: a new multimodal foundation model enables efficient split learning. arXiv preprint arXiv:2511.23402 (2025)
13. Han, Z., Yang, J., Wang, G., Li, J., Xu, Q., Shou, M.Z., Zhang, C.: Dota: Distributional test-time adaptation of vision-language models. arXiv preprint arXiv:2409.19375 (2024)
14. Hastie, T., Tibshirani, R.: Discriminant analysis by gaussian mixtures. Journal of the Royal Statistical Society Series B: Statistical Methodology **58**(1), 155–176 (1996)
15. He, Y., Zhang, C., Chen, F., Cao, J.: Cinematte: Background matting for virtual production and beyond. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8725–8735 (2026)
16. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8340–8349 (2021)

17. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15262–15271 (2021)
18. Jiang, X., Liu, F., Fang, Z., Chen, H., Liu, T., Zheng, F., Han, B.: Negative label guided ood detection with pretrained vision-language models. arXiv preprint arXiv:2403.20078 (2024)
19. Karmanov, A., Guan, D., Lu, S., El Saddik, A., Xing, E.: Efficient test-time adaptation of vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14162–14171 (2024)
20. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013)
21. Lafon, M., Hakim, G.A.V., Rambour, C., Desrosier, C., Thome, N.: Cliptta: Robust contrastive vision-language test-time adaptation. arXiv preprint arXiv:2507.14312 (2025)
22. Li, Y., Dong, J., Yang, C., Wen, S., Koniusz, P., Huang, T., Tian, Y., Ong, Y.S.: Mmt-ard: Multimodal multi-teacher adversarial distillation for robust vision-language models. arXiv preprint arXiv:2511.17448 (2025)
23. Li, Y., Yang, C., Dong, J., Yao, Z., Xu, H., Dong, Z., Zeng, H., An, Z., Tian, Y.: Ammkd: Adaptive multimodal multi-teacher distillation for lightweight vision-language models. arXiv preprint arXiv:2509.00039 (2025)
24. Li, Y., Xu, X., Su, Y., Jia, K.: On the robustness of open-world test-time training: Self-training with dynamic prototype expansion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11836–11846 (2023)
25. Lu, S., Wang, Y., Sheng, L., He, L., Zheng, A., Liang, J.: Out-of-distribution detection: A task-oriented survey of recent advances. ACM Computing Surveys **58**(2), 1–39 (2025)
26. Luo, W., Ou, Y., Deng, J., Deng, Z., Yan, X., Wen, Z., Tan, M.: Protodcs: Towards robust and efficient open-set test-time adaptation for vision-language models. arXiv preprint arXiv:2602.23653 (2026)
27. Ma, W., Chen, R., Zhang, K., Wu, S., Ding, S.: Instruct where the model fails: Generative data augmentation via guided self-contrastive fine-tuning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 5991–5999 (2025)
28. Ma, W., Sun, S., Yu, T., Wang, R., Chua, T.S., Bian, J.: Thinking with blueprints: Assisting vision-language models in spatial reasoning via structured object representation. arXiv preprint arXiv:2601.01984 (2026)
29. Meng, G., Gu, P., Liang, P., Lalor, J.P., Chambers, E.W., Chen, D.Z.: Topocl: Topological contrastive learning for medical imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 42681–42690 (2026)
30. Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., Li, Y.: Delving into out-of-distribution detection with vision-language representations. Advances in neural information processing systems **35**, 35087–35102 (2022)
31. Mirza, M.J., Micorek, J., Possegger, H., Bischof, H.: The norm must go on: Dynamic unsupervised domain adaptation by normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14765–14775 (2022)
32. Ning, J., Huang, W., Ding, C., Wang, J., Zhu, Z.: Physics-informed unsupervised domain adaptation framework for cross-machine bearing fault diagnosis. Advanced Engineering Informatics **62**, 102774 (2024)

33. Ning, J., Huang, W., Guo, P., Ding, C., Huangfu, Y., Shen, C., Zhu, Z.: A physics-guided memory enhancement and causality-inspired generalization framework for continual fault diagnosis. Knowledge-Based Systems **325**, 114044 (2025), corresponding author: Weiguo Huang
34. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. arXiv preprint arXiv:2302.12400 (2023)
35. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
37. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International conference on machine learning. pp. 5389–5400. PMLR (2019)
38. Shi, H., Zhang, Y., Tang, S., Zhu, W., Li, Y., Guo, Y., Zhuang, Y.: On the efficacy of small self-supervised contrastive models without distillation signals. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 2225–2234 (2022)
39. Shu, M., Nie, W., Huang, D.A., Yu, Z., Goldstein, T., Anandkumar, A., Xiao, C.: Test-time prompt tuning for zero-shot generalization in vision-language models. Advances in Neural Information Processing Systems **35**, 14274–14289 (2022)
40. Sui, E., Wang, X., Yeung-Levy, S.: Just shift it: Test-time prototype shifting for zero-shot generalization with vision-language models. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 825–835. IEEE (2025)
41. Tang, H., Liu, Y., Yan, S., Shen, F., He, S., Qin, J.: Cross-modal proxy evolving for ood detection with vision-language models. arXiv preprint arXiv:2601.08476 (2026)
42. Tao, Y., Wang, Y., Song, X., Luo, X., Liu, K., Liu, J.: Grasp: Plan-guided graph retrieval with adaptive fusion and reranking on semi-structured knowledge bases. arXiv preprint arXiv:2605.30237 (2026)
43. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018)
44. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S., et al.: The caltech-ucsd birds-200-2011 dataset. Tech. rep.
45. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
46. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. Advances in neural information processing systems **32** (2019)
47. Wang, T., Li, Y., Li, L., Chen, Y., Huang, S., Chen, Y., Li, P., Liu, Y., Chen, G.: Sppo: Sequence-level ppo for long-horizon reasoning tasks (2026), `https://arxiv.org/abs/2604.08865`
48. Xiao, C., Xu, T., Ma, S., Jiang, Y., Gao, H., Wu, Y.: Reversible primitive–composition alignment for continual vision–language learning. In: The Fourteenth International Conference on Learning Representations (2026)

49. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: 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)
50. Yang, P., Akhtar, N., Jiang, J., Mian, A.: Backdoor-based explainable ai benchmark for high fidelity evaluation of attribution methods. arXiv preprint arXiv:2405.02344 (2024)
51. Yang, P., Akhtar, N., Shah, M., Mian, A.: Regulating model reliance on non-robust features by smoothing input marginal density. In: European Conference on Computer Vision. pp. 329–347. Springer (2024)
52. Yang, P., Akhtar, N., Wen, Z., Shah, M., Mian, A.S.: Re-calibrating feature attributions for model interpretation. In: International Conference on Learning Representations (2023)
53. Yang, Y., Zhu, L., Sun, Z., Liu, H., Gu, Q., Ye, N.: Oodd: Test-time out-of-distribution detection with dynamic dictionary. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30630–30639 (2025)
54. Zeng, S., Chang, X., Xie, M., Liu, X., Bai, Y., Pan, Z., Xu, M., Wei, X.: Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. arXiv preprint arXiv:2505.17685 (2025)
55. Zhang, C., Stepputtis, S., Sycara, K., Xie, Y.: Dual prototype evolving for test-time generalization of vision-language models. Advances in Neural Information Processing Systems **37**, 32111–32136 (2024)
56. Zhang, J., Zhao, C., Xiao, C., Duan, R., Mo, W., Gao, H., Wang, W.: Pi-cca: Prompt-invariant cca certificates for replay-free continual multimodal learning. In: The Fourteenth International Conference on Learning Representations (2026)
57. Zhang, Y., Varma, M., Gao, Y., Delbrouck, J.B., Liu, J., Wang, C., Langlotz, C.: Activation matters: Test-time activated negative labels for ood detection with vision-language models. arXiv preprint arXiv:2603.25250 (2026)
58. Zhang, Y., Zhang, L.: Adaneg: Adaptive negative proxy guided ood detection with vision-language models. Advances in Neural Information Processing Systems **37**, 38744–38768 (2024)
59. Zhang, Y., Zhu, W., He, C., Zhang, L.: Lapt: Label-driven automated prompt tuning for ood detection with vision-language models. In: European conference on computer vision. pp. 271–288. Springer (2024)
60. Zhang, Y., Zhu, W., Tang, H., Ma, Z., Zhou, K., Zhang, L.: 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 (2024)
61. Zhang, Y., Wang, X., Zhou, T., Yuan, K., Zhang, Z., Wang, L., Jin, R.: Model-free test time adaptation for out-of-distribution detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
62. Zhao, B., Chen, C., Xia, S.T.: Delta: degradation-free fully test-time adaptation. arXiv preprint arXiv:2301.13018 (2023)
63. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017)
64. Zhou, L., Ye, M., Li, S., Li, N., Zhu, X., Deng, L., Liu, H., Lei, Z.: Bayesian test-time adaptation for vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29999–30009 (2025)
65. Zhu, C., Lin, J., Tan, G., Zhu, N., Li, K., Wang, C., Li, S.: Advancing ultrasound medical continuous learning with task-specific generalization and adaptability. In:

2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 3019–3025. IEEE (2024)
66. Zhu, C., Lin, Y., Chen, S., Wang, Y., Lin, J.: Medeyes: Learning dynamic visual focus for medical progressive diagnosis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 13916–13924 (2026)
67. Zhu, W., Zhang, Y., Jin, X., Zeng, W., Zhang, L.: Knowledge regularized negative feature tuning of vision-language models for out-of-distribution detection. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3565–3574 (2025)
68. Zhu, W., Zhang, Y., Jin, X., Zeng, W., Zhang, L.: Ants: Adaptive negative textual space shaping for ood detection via test-time mllm understanding and reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20–30 (2026)
69. Zhu, X., Zhu, B., Tan, Y., Wang, S., Hao, Y., Zhang, H.: Enhancing zero-shot vision models by label-free prompt distribution learning and bias correcting. Advances in Neural Information Processing Systems **37**, 2001–2025 (2024)