Open-Set Test-Time Adaptation in VLMs via Noise-Immune Self-Purification

Mingxu Feng, **Fengqiang Wan**, **Yang Yang***
Nanjing University of Science and Technology
{mingxuf, fqwan, yyang}@njust.edu.cn

Abstract

Unlike test-time adaptation (TTA), open-set TTA (OSTTA) aims to robustly adapt to in-distribution (ID) domain shifts while suppressing the detrimental impact of out-of-distribution (OOD) samples encountered at test time. To this end, we propose a novel OSTTA approach named Noise-Immune Self-Purification (NISP). By exploiting the zero-shot priors of pre-trained vision-language models (VLMs), NISP advances from coarse pseudo-labeling to fine-grained, noise-resilient adaptation. Technically, we first introduce a dual-cluster Bayesian Gaussian mixture model to fit VLMs-derived scores, achieving coarse pseudo-ID/OOD separation via posterior-risk thresholding. Subsequently, NISP constructs a noise-immune fine-grained adaptation where the adapter enforces consensus-discrepancy constraints to refine coarse pseudo-labels and suppress noise propagation. We then devise the Jaccard Consistency Score for OOD discrimination. Overall the coarse-to-fine pipeline enables rigorous self-purification and robust online adaptation. Theoretically we show that consensus-discrepancy losses mitigate the deleterious effects of noise. Empirically, NISP achieves state-of-the-art results across multiple OSTTA benchmarks, validating its efficacy. The code is available at https://github.com/njustkmg/IJCAI26-NISP.

1 Introduction

Despite the remarkable generalization of deep learning models, they still suffer performance degradation when encountering distribution shifts during testing. Test-time adaptation (TTA) has been introduced to mitigate this issues [Shu *et al.*, 2022; Ma *et al.*, 2023; Jia *et al.*, 2024]. Nevertheless, TTA is vulnerable in open-world scenarios, where out-of-distribution (OOD) samples bias the optimization direction, thereby undermining the discriminability of known classes [Li *et al.*, 2023b; Yang *et al.*, 2024]. This motivates the open-set TTA (OSTTA) challenge, which aims to robustly adapt to in-distribution (ID) shifts while mitigating the detrimental impact of OOD samples [Gao *et al.*, 2024; Yang and Xu, 2025; Dong *et al.*, 2025; Zhao *et al.*, 2025].

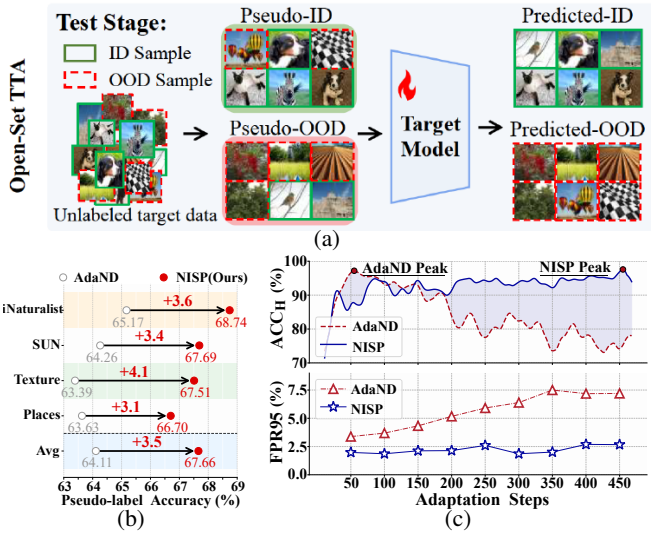


Figure 1: (a) Open-Set TTA pipeline for pseudo-labeling and adaptation. (b) Pseudo-label accuracy evaluation on ImageNet under varying data distributions. (c) Classification (ACC_H) and OOD detection (FPR95) performance in different steps on ImageNet.

To tackle the OSTTA challenge, existing methods typically adopt a filtering-then-adaptation paradigm, i.e., Figure 1(a): thresholding OOD scores to generate pseudo-labels, and then supervising the model via discriminative losses. State-of-the-art (SOTA) approaches like AdaND [Cao *et al.*, 2025] and Open-IRT [Peng *et al.*, 2025] enhance this pipeline by leveraging vision-language models' (VLMs) zero-shot generalization [Chao *et al.*, 2025] to derive OOD scores. However, their thresholding schemes rely on idealized priors (e.g., balanced bimodality) that fail under complex real-world distributions, leading to erroneous pseudo-labeling. Figure 1(b) reveals that AdaND fails to maintain robust OOD filtering across diverse evaluation datasets. Crucially, this filtering failure triggers a cascading error accumulation. Instead of rejecting noise, the model is compelled to overfit to the mislabeled samples during updates, distorting learned decision boundaries [Wang *et al.*, 2019; Zhang *et al.*, 2023]. This results in the catastrophic collapse observed in Figure 1(c): performance de-

*Corresponding author.

grades severely after an initial transient peak as noise dominates the adaptation process, highlighting the inability to sustain robustness.

Motivated by these limitations, we propose a novel OSTTA framework named Noise-Immune Self-Purification (NISP) that implements self-purification of OOD samples to safeguard adaptation quality. First, we propose a dual-cluster Bayesian Gaussian mixture model (BGMM) with posterior-risk thresholding to extract high-fidelity pseudo-labels. As illustrated in Figure 1(b), our strategy yields superior pseudo-label accuracy and maintains exceptional robustness regardless of the dataset variations. Second, we develop a noise-resilient adaptation governed by consensus-discrepancy constraints to refine coarse pseudo-labels. The synergistic regularization effectively mitigates noise interference from pseudo-labels and prevents error accumulation. Capitalizing on this adaptation dynamic, we derive the Jaccard Consistency Score (JCS) to enable precise test-time OOD discrimination. As evidenced in Figure 1(c), NISP exhibits a sustained performance trajectory: ascending and then stabilizing, validating its efficacy in noise-robust adaptation. Furthermore, theoretical analysis substantiates that the collaborative optimization of consensus and discrepancy objectives can significantly reduce the deleterious impact of pseudo-label noise, thereby enhancing the robustness of the model in open-world scenarios.

The contributions of this work are summarized as follows:

- **Technically**, we propose NISP, a coarse-to-fine self-purification framework that leverages VLMs to safeguard adaptation against open-set noise.
- **Theoretically**, we prove that our optimization objective provides an upper-bound on Bayes risk under noisy supervision, ensuring rigorous denoising adaptation.
- **Empirically**, NISP attains state-of-the-art (SOTA) performance across multiple OSTTA benchmarks, validating its efficacy in open-world settings.

2 Related Work

2.1 Open-Set Test-Time Adaptation

Test-Time Adaptation (TTA) aims to address performance degradation caused by the distribution shifts. Traditional TTA methods adapt models to target domains via self-supervised objectives, such as entropy minimization [Niu *et al.*, 2022; Wu *et al.*, 2025], or updating internal normalization statistics [Schneider *et al.*, 2020; Yuan *et al.*, 2023; Lim *et al.*, 2023]. Recent studies have exploited the zero-shot generalization of VLMs for TTA, like TPT[Shu *et al.*, 2022], DiffTPT [Feng *et al.*, 2023], and TDA [Karmanov *et al.*, 2024], which facilitate online calibration via prompt optimization or key-value caching. However, these methods overlook the presence of OOD samples in open-world deployments, which gives rise to the open-set TTA (OSTTA) challenge. While approaches such as SoTTA [Gong *et al.*, 2023], UniEnt [Gao *et al.*, 2024] and AEO [Dong *et al.*, 2025] incorporate OOD detection into TTA, they do not leverage the VLMs for adaptation. Prior work, such as AdaND [Cao *et al.*, 2025], utilizes VLMs to generate pseudo-labels, updating only the OOD detector at test time with BCE loss. Yet, without addressing the inherent noise in pseudo-labels, the model could overfit to erroneous signals during adaptation.

2.2 Vision-Language Models

Pretrained vision-language models (VLMs) typically consist of image and text encoders, which are jointly trained on billions of image-text pairs to align cross-modal representations. For instance, ALIGN [Jia *et al.*, 2021] leverages a trillion-scale noisy web dataset to improve alignment precision and scalability. CLIP [Radford *et al.*, 2021] establishes a standardized contrastive pretraining framework using templated text prompts and four million image-text pairs, setting a new paradigm for zero-shot transfer tasks. BLIP [Li *et al.*, 2023a] introduces cross-modal attention into the contrastive loss and combines generative and discriminative objectives, enhancing fine-grained semantic understanding and generation. EVA-CLIP [Sun *et al.*, 2023] augments the CLIP architecture with a more powerful visual backbone and richer pretraining goals, employing LAMB [You *et al.*, 2020] optimization to reduce computational overhead while boosting downstream task performance. Empowered by large-scale pretraining data and contrastive learning, VLMs exhibit exceptional zero-shot generalization capacities.

3 Proposed Method

3.1 Preliminaries

In OSTTA, we consider an unlabeled test stream $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$, where each sample is from an underlying distribution with an unobserved ground-truth label y_i. The universal label space $\mathcal{Y}$ comprises two disjoint subsets: the ID categories $\mathcal{Y}_{\text{id}}$ and OOD categories $\mathcal{Y}_{\text{ood}}$, such that $\mathcal{Y} = \mathcal{Y}_{\text{id}} \cup \mathcal{Y}_{\text{ood}}$ and $\mathcal{Y}_{\text{id}} \cap \mathcal{Y}_{\text{ood}} = \emptyset$. During the adaptation phase, the model has access solely to the unlabeled images in $\mathcal{X}$ and the ID class name set $\mathcal{Y}_{\text{id}} = \{y_1, y_2, \ldots, y_C\}$. Our objective is to leverage the VLM's zero-shot generalization to assign each $\boldsymbol{x}_i$ an OOD score, thereby facilitating precise ID/OOD discrimination. Standard VLMs utilize a dual-encoder architecture consisting of an image encoder $f(\cdot)$ and a text encoder $g(\cdot)$. For an image $\boldsymbol{x}_i$ and a set of class-specific prompts $\{\boldsymbol{t}_k\}_{k=1}^{C}$ (e.g., "a photo of a y_k"), we obtain the corresponding embeddings as $\boldsymbol{v}_i = f(\boldsymbol{x}_i)$ and $\boldsymbol{u}_k = g(\boldsymbol{t}_k)$.

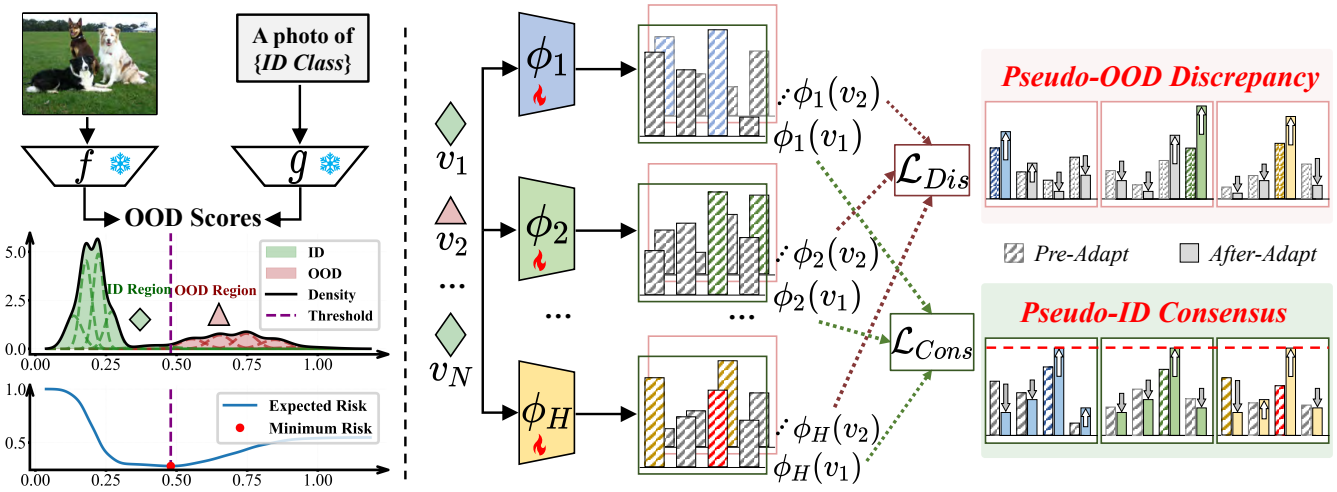
3.2 Unsupervised Coarse Pseudo-Labeling

With the extracted embeddings, we adopt the Maximum Concept Matching (MCM) score [Ming *et al.*, 2022] to assign the coarse OOD score to each test sample:

$$S(\boldsymbol{x}_i) = \max_{1 \le k \le C} \frac{\exp\big(\text{sim}(\boldsymbol{v}_i, \boldsymbol{u}_k)/\tau\big)}{\sum_{j=1}^{C} \exp\big(\text{sim}(\boldsymbol{v}_i, \boldsymbol{u}_j)/\tau\big)}, \qquad (1)$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity and τ is the temperature hyperparameter. Then, we propose an unsupervised strategy to partition the test data into coarse pseudo-ID/OOD subsets. This is achieved by modeling the MCM score distribution as a mixture of Gaussian components, identifying their semantics (ID vs. OOD) and minimizing the decision risk.

To capture the complex data structure without pre-specifying the number of components, we introduce BGMM



(a) Coarse Pseudo-Labeling (b) Noise-Immune Fine-Grained Adaptation

Figure 2: Illustration of the proposed NISP. (a) *Coarse Pseudo-Labeling*: OOD scores derived from VLMs are modeled by Bayesian GMM, enabling coarse pseudo-label assignment through component-level clustering and posterior-risk thresholding. (b) *Noise-Immune Fine-Grained Adaptation*: A multi-head adapter is trained using jointly designed consensus and discrepancy losses to refine the pseudo-labels and suppress label noise in a modality-aware manner.

with a Dirichlet Process prior [Blei and Jordan, 2004] fit the score distribution, defined as an infinite mixture:

$$p(S;\mathbf{\Theta}) = \sum_{m=1}^{\infty} \pi_m \mathcal{N}\big(S \mid \mu_m, \sigma_m^2\big), \tag{2}$$

where $\mathbf{\Theta} = \{\pi_m, \mu_m, \sigma_m\}_{m=1}^{\infty}$ comprises the mixture weights π_m, the component means μ_m, and variances σ_m. The weights π_m follow a stick-breaking process. To ensure computational tractability, we approximate Eq. (2) using a variational truncation bound M. Parameters $\mathbf{\Theta}$ are estimated via variational inference by maximizing the evidence lower bound, where negligible components with weights below ϵ_w (empirically set to 0.001) are pruned.

Once the density is modeled, the next objective is to distinguish ID-dominant components from OOD-dominant ones. Each Gaussian component is characterized by its statistical moments, represented as a bivariate feature vector $\mathbf{f}_m = (\mu_m, \sigma_m) \in \mathbb{R}^2$. We apply K-means clustering [Likas *et al.*, 2003] on these vectors to categorize the component index set $\mathcal{M} = \{1, \ldots, M\}$ into mutually exclusive ID subset $\mathcal{C}_{\text{id}}$ and OOD subset $\mathcal{C}_{\text{ood}}$. The clustering objective minimizes:

$$\mathcal{J}_{\text{cluster}} = \sum_{m \in \mathcal{C}_{\text{id}}} \|\mathbf{f}_m - \mathbf{c}_{\text{id}}\|^2 + \sum_{m \in \mathcal{C}_{\text{ood}}} \|\mathbf{f}_m - \mathbf{c}_{\text{ood}}\|^2, \tag{3}$$

where $\mathbf{c}_{\text{id}}, \mathbf{c}_{\text{ood}} \in \mathbb{R}^2$ denote the cluster centroids.

With the components identified, we determine the optimal decision boundary by leveraging the component-wise posterior probabilities. The posterior responsibility $\gamma_m(S)$ of component m for a score S is calculated via Bayesian inversion:

$$\gamma_m(S) = \frac{\pi_m \mathcal{N}\big(S \mid \mu_m, \sigma_m^2\big)}{\sum_{j=1}^{M} \pi_j \mathcal{N}\big(S \mid \mu_j, \sigma_j^2\big)}. \tag{4}$$

Subsequently, given the score sequence $\{S_{(i)}\}_{i=1}^{N}$ sorted in ascending order, we seek a cut-off index d that minimizes the cumulative misclassification risk $\mathcal{R}(d)$, balancing the error of rejecting ID samples against accepting OOD samples:

$$\mathcal{R}(d) = \sum_{i=1}^{d} \sum_{m \in \mathcal{C}_{\text{id}}} \gamma_m(S_{(i)}) + \sum_{i=d+1}^{N} \sum_{n \in \mathcal{C}_{\text{ood}}} \gamma_n(S_{(i)}). \tag{5}$$

The optimal index $d^* = \arg\min_d \mathcal{R}(d)$ is efficiently computed via prefix sums [Martinez and Raschia, 2024]. Finally, the decision threshold is set to the midpoint $\xi = \frac{1}{2}(S_{(d^*)} + S_{(d^*+1)})$, yielding binary pseudo-labels:

$$\hat{y}_i = \begin{cases} 0 & S_i \le \xi, \\ 1 & S_i > \xi, \end{cases} \tag{6}$$

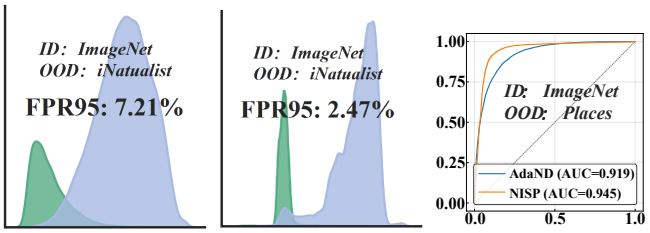
where $\hat{y}_i = 0$ and $\hat{y}_i = 1$ denote pseudo-ID and pseudo-OOD labels, respectively.

3.3 Noise-Immune Fine-Grained Adaptation

At each adaptation step, let $\mathbf{V} = \{\boldsymbol{v}_i\}_{i=1}^{B}$ denote the image embeddings set of current batch with pseudo-labels $\{\hat{y}_i\}_{i=1}^{B}$. A binary mask partitions indices into ID set $\mathcal{I} = \{i : \hat{y}_i = 0\}$ and OOD set $\mathcal{O} = \{i : \hat{y}_i = 1\}$. Subsequently, we introduce a lightweight adapter to perform noise-resilient adaptation, thereby refining the coarse pseudo-labels. The adapter consists of H parallel prediction heads $\{\phi_h\}_{h=1}^{H}$, where each ϕ_h is an independent two-layer MLP with non-linear activation.

The adapter is optimized via a dual-path objective that separately handles pseudo-ID and pseudo-OOD samples. To promote predictive consistency among classifier heads on pseudo-ID samples, we minimize the pairwise mean squared error [Xue *et al.*, 2013] across all unique head pairs:

$$\mathcal{L}_{\text{Cons}} = \frac{1}{|\mathcal{I}|\binom{H}{2}} \sum_{i \in \mathcal{I}} \sum_{\substack{h,h' \in \mathcal{H} \\ h<h'}} \big\|\phi_h(\boldsymbol{v}_i) - \phi_{h'}(\boldsymbol{v}_i)\big\|_2^2, \tag{7}$$



(a) AdaND Density (b) NISP Density (c) AUROC Curve

Figure 3: OOD Detection Performance.

where $\mathcal{H} = \{1, \ldots, H\}$ denotes the set of head indices, $|\mathcal{I}|$ is the number of pseudo-ID samples, and $\binom{H}{2}$ enumerates all unordered head pairs. This objective reduces prediction variance across heads on pseudo-ID samples, yielding more consistent and robust outputs.

In contrast, pseudo-OOD samples are expected to exhibit prediction disagreement. To achieve this, a hinge-activated margin is applied to maximize their pairwise divergence:

$$\mathcal{L}_{\text{Dis}} = \frac{1}{|\mathcal{O}|\binom{H}{2}} \sum_{j \in \mathcal{O}} \sum_{\substack{h,h' \in \mathcal{H} \\ h<h'}} \left[\delta - \left\| \phi_h(\boldsymbol{v}_j) - \phi_{h'}(\boldsymbol{v}_j) \right\|_2^2 \right]_+, \tag{8}$$

where $[\cdot]_+ = \max(0, \cdot)$ is the hinge operator and $\delta > 0$ denotes a predefined divergence margin (set to 0.1 in all experiments). Gradient flow is selectively activated only when head disagreement falls below the margin, allowing the model to preserve uncertainty on OOD samples. The overall adaptation objective is defined to minimize a weighted sum:

$$\mathcal{L} = \mathcal{L}_{\text{Cons}} + \lambda \cdot \mathcal{L}_{\text{Dis}}, \tag{9}$$

where the coefficient λ balances the two terms. As shown in Lemma 1, the discrepancy and consensus losses jointly suppress unreliable gradients and reduce gradient variance, resulting in a tighter bound in Theorem 1.

3.4 Inference Phase

Inspired by this adaptation process, we adopt a novel OOD scoring metric, termed the *Jaccard Consistency Score* (JCS), to enhance the separation between ID and OOD samples. Let $\mathcal{A}$ denote a set of percentile thresholds (typically set to $\{5, 10, \ldots, 55\}$). For any granularity $\alpha \in \mathcal{A}$, the corresponding subset size is defined as $K = \lceil \alpha \cdot C \rceil$. Given an image embedding $\boldsymbol{v}_i$, the set of top-K classes predicted by head h is defined as:

$$Y_h(\boldsymbol{v}_i; \alpha) = \left\{ c_k \;\middle|\; c_k \in \text{argsort}^{(K)}\left(\phi_h(\boldsymbol{v}_i)\right) \right\}, \tag{10}$$

where $\text{argsort}^{(K)}(\cdot)$ returns the indices of the top-K highest confidence classes, and $|Y_h(\boldsymbol{v}_i)| = K$.

We compute the aggregates pairwise Jaccard similarity at this granularity α across all head pairs (p, q):

$$\bar{J}(\boldsymbol{v}_i; \alpha) = \frac{1}{\binom{H}{2}} \sum_{1 \le p < q \le H} \frac{|Y_p(\boldsymbol{v}_i; \alpha) \cap Y_q(\boldsymbol{v}_i; \alpha)|}{|Y_p(\boldsymbol{v}_i; \alpha) \cup Y_q(\boldsymbol{v}_i; \alpha)|}. \tag{11}$$

The final JCS score is obtained by identifying the maximum consensus achieved across the spectrum of granularities:

$$\text{JCS}(\boldsymbol{v}_i) = \max_{\alpha \in \mathcal{A}} \bar{J}(\boldsymbol{v}_i; \alpha). \tag{12}$$

This metric provides a compact measure of cross-head prediction consistency. Higher JCS values indicate strong agreement among heads, which is characteristic of ID samples, while lower values reflect prediction dispersion typical of OOD samples. Empirical effectiveness of this metric is demonstrated in Figure 3.

3.5 Theoretical Understanding

We theoretically justify the effectiveness of our method in suppressing overfitting to noisy pseudo-labels under test-time adaptation. Under the standard assumptions of L-smoothness [Seccia *et al.*, 2025] and the μ-PL condition [Plassier *et al.*, 2025], we analyze the convergence behavior in the presence of pseudo-label noise.

Theorem 1 (Excess Risk Bound)**.** *Let θ_T be the model after T steps of adaptation with learning rate $\eta \le 1/L$, and let $\bar{\sigma}^2_{\text{eff}}$ denote the average gradient variance over noisy samples. Then the expected excess risk satisfies:*

$$\mathbb{E}\left[\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*)\right] \le (1 - \eta\mu)^T \cdot \left(\mathcal{L}(\theta_0) - \mathcal{L}(\theta^*)\right) + \frac{L\eta}{2\mu} \cdot \rho \cdot \bar{\sigma}^2_{\text{eff}}. \tag{13}$$

The key to controlling generalization error lies in reducing $\bar{\sigma}^2_{\text{eff}}$. Our method achieves this through multi-head regularization, which adaptively modulates the gradient contribution of noisy samples.

Lemma 1 (Variance-Controlled Gradient Suppression)**.** *For any noisy sample x_i, the per-sample gradient satisfies:*

$$\left\| \nabla_\theta \mathcal{L}_{\text{total}}(x_i) \right\|^2 \le C \cdot \text{Var}_h\left[\phi_h(x_i)\right], \tag{14}$$

where $C > 0$ is a constant and $\text{Var}_h[\cdot]$ denotes inter-head variance. Consequently,

$$\bar{\sigma}^2_{\text{eff}}(\text{ours}) \ll \bar{\sigma}^2_{\text{eff}}(\text{vanilla}) = \mathcal{O}(1). \tag{15}$$

This behavior arises because the discrepancy loss is inactive for incorrectly assigned pseudo-OOD samples, yielding zero gradients [Li *et al.*, 2025], whereas the consensus loss is activated only when predictions exhibit sufficient disagreement. By contrast, conventional pseudo-labeling applies uniform updates to all samples irrespective of pseudo-label reliability, resulting in elevated gradient variance.

4 Experiments

4.1 Experimental Setup

Benchmark Datasets. Following prior work [Cao *et al.*, 2025], we utilize CIFAR-10/100 [Krizhevsky *et al.*, 2009], CUB-200-2011 [Wah *et al.*, 2011], Food-101 [Bossard *et al.*, 2014], Stanford-Cars [Krause *et al.*, 2013], ImageNet [Deng *et al.*, 2009] and its variants (ImageNet-V2 [Recht *et al.*, 2019], ImageNet-R [Hendrycks *et al.*, 2021], ImageNet-Sketch [Wang *et al.*, 2019]) as ID datasets. The OOD datasets include SVHN [Netzer *et al.*, 2011], LSUN [Yu *et al.*, 2015], iNaturalist [Van Horn *et al.*, 2018], SUN [Xiao *et al.*, 2010], Places [Zhou *et al.*, 2017], Texture [Cimpoi *et al.*, 2014].

ID	Method	Avg			iNaturalist			SUN			Texture			Places		
		AUC	FPR95	ACC_H	AUC	FPR95	ACC_H	AUC	FPR95	ACC_H	AUC	FPR95	ACC_H	AUC	FPR95	ACC_H
ImageNet	ZS-CLIP	83.07	57.97	63.11	86.71	50.39	64.58	84.05	57.43	63.58	79.50	64.20	61.53	82.00	59.84	62.76
	Tent	80.17	63.85	63.49	80.47	64.57	63.35	82.18	60.10	64.55	77.60	67.61	62.52	80.43	63.11	63.55
	SoTTA	79.06	65.73	62.24	77.67	67.19	61.39	80.46	63.48	62.82	78.45	67.30	62.17	79.64	64.93	62.57
	TPT	82.77	57.83	62.25	86.09	49.91	63.69	83.91	57.30	62.75	79.15	64.38	60.56	81.92	59.73	61.99
	TDA	83.02	57.89	63.20	86.72	50.35	64.68	84.06	57.39	63.67	79.30	63.90	61.60	82.00	59.91	62.84
	AdaND	94.09	24.42	70.31	**98.51**	7.21	75.91	94.47	26.79	70.13	91.44	31.49	67.06	91.94	32.20	68.13
	NISP	**95.58**	**12.52**	**73.71**	98.11	**2.47**	**76.41**	**95.37**	**14.04**	**73.22**	**94.38**	**16.19**	**72.76**	**94.48**	**17.40**	**72.46**
ImageNet-R	ZS-CLIP	85.98	58.89	71.18	91.05	49.66	73.95	87.94	56.55	72.27	79.62	69.95	67.65	85.30	59.39	70.83
	Tent	86.30	58.74	73.23	90.64	50.79	76.18	88.20	56.11	74.40	80.60	69.38	69.56	85.76	58.69	72.78
	SoTTA	86.67	56.61	73.73	90.91	50.14	76.59	88.54	53.50	74.82	81.21	65.92	70.17	86.02	56.88	73.34
	TPT	85.33	58.87	70.71	90.64	49.65	73.31	87.43	56.55	71.89	78.65	69.88	67.26	84.59	59.38	70.37
	TDA	85.98	58.80	71.20	91.05	49.50	74.02	87.94	56.43	72.29	79.62	69.89	67.65	85.30	59.37	70.82
	AdaND	96.33	16.02	78.95	**99.58**	1.42	82.23	97.93	10.44	80.49	**91.81**	34.06	**74.40**	96.00	18.17	78.67
	NISP	**96.39**	**9.59**	**79.68**	98.90	**0.43**	**83.24**	**97.94**	**5.09**	**81.08**	91.74	**25.59**	74.09	**97.00**	**7.25**	**80.31**
ImageNet-K	ZS-CLIP	70.60	82.17	45.39	75.22	78.30	46.39	71.66	82.79	45.73	66.29	85.04	44.34	69.21	82.54	45.11
	Tent	70.30	80.73	47.96	70.85	80.83	48.24	72.90	79.21	48.54	66.97	83.07	46.92	70.47	79.79	48.15
	SoTTA	68.79	80.72	47.67	66.63	81.27	47.08	71.22	79.58	48.41	67.60	82.25	47.20	69.69	79.77	47.99
	TPT	70.35	81.94	43.84	74.80	78.19	44.81	71.54	82.32	44.15	65.85	85.00	42.77	69.20	82.23	43.64
	TDA	70.60	82.14	45.62	75.22	78.24	46.66	71.66	82.74	45.93	66.30	84.96	44.59	69.21	82.60	45.30
	AdaND	86.39	53.99	52.10	95.39	27.51	55.71	88.78	52.09	52.91	77.14	70.98	48.49	84.23	65.36	51.27
	NISP	**93.12**	**23.09**	**57.60**	**97.11**	**5.96**	**60.45**	**95.58**	**11.37**	**59.01**	**84.92**	**60.35**	**52.55**	**94.88**	**14.67**	**58.39**
CUB-200-2011	ZS-CLIP	78.61	62.92	52.28	80.82	59.52	53.12	80.37	61.15	53.05	73.38	68.33	50.12	79.88	62.69	52.81
	Tent	60.42	79.18	48.27	51.21	81.52	44.19	61.91	76.58	49.01	56.70	85.04	46.56	71.87	73.58	53.31
	SoTTA	79.96	62.02	55.75	80.23	61.82	56.15	82.21	59.26	56.76	75.72	66.88	53.57	81.69	60.11	56.51
	TPT	78.56	62.92	51.74	80.58	59.53	52.45	80.21	61.21	52.58	73.56	68.32	49.58	79.89	62.60	52.35
	TDA	78.61	62.92	52.45	80.82	59.54	53.32	80.37	61.24	53.22	73.38	68.33	50.27	79.88	62.57	52.97
	AdaND	94.93	20.62	63.90	96.64	14.63	65.30	96.70	14.75	65.27	90.36	36.77	60.40	96.01	16.32	64.63
	NISP	**96.87**	**7.93**	**66.36**	**96.99**	**6.78**	**66.09**	**97.69**	**4.35**	**67.08**	**95.57**	**16.12**	**65.54**	**97.24**	**4.50**	**66.71**

Table 1: OSTTA results for the four benchmarks (ImageNet, ImageNet-R, ImageNet-K and CUB-200-2011) across four OOD datasets. **Bold** indicates best per column. Underline indicates second best.

Method	ImageNet-V2			CIFAR-100			Food-101			CIFAR-10			Stanforrd-Cars		
	AUC	FPR95	ACC_H	AUC	FPR95	ACC_H	AUC	FPR95	ACC_H	AUC	FPR95	ACC_H	AUC	FPR95	ACC_H
ZS-CLIP	79.69	67.07	59.96	72.26	90.52	59.58	97.55	11.88	86.71	94.17	27.35	86.21	97.10	13.30	66.49
Tent	74.14	72.70	57.57	66.29	80.61	57.27	90.27	24.23	81.18	91.86	30.00	82.47	95.85	16.10	66.95
SoTTA	77.35	71.57	59.05	79.34	73.77	66.50	97.35	13.18	86.92	96.16	17.83	88.27	96.73	14.32	67.62
TPT	79.22	66.93	59.24	70.77	90.52	57.81	97.27	12.01	86.28	93.39	27.60	85.46	97.00	13.28	65.47
TDA	80.35	66.35	60.16	72.26	90.48	59.51	97.55	11.90	86.71	94.18	27.36	86.20	97.10	13.30	66.82
AdaND	**93.85**	23.34	67.72	89.48	33.91	69.20	**99.81**	0.52	**92.10**	**99.40**	1.91	92.56	**99.86**	0.32	**77.27**
NISP	93.72	**22.03**	**67.80**	**91.08**	**33.38**	**71.99**	99.54	**0.37**	91.35	98.19	**1.58**	**92.69**	99.44	**0.19**	77.14

Table 2: OSTTA results for per-benchmark average performance over all OOD datasets in ImageNet-V2, CIFAR-100, Food-101, CIFAR-10 and Stanforrd-Cars. **Bold** indicates best per column. Underline indicates second best.

Evaluation Metric. We adopt the harmonic mean accuracy (ACC_H) as the metric for OSTTA evaluation in VLMs, which is computed over the classification accuracies of ID and OOD samples. For OOD detection performance, we employ FPR95 and AUC for evaluation [Cao *et al.*, 2025; Ming *et al.*, 2022]. FPR95 refers to the false positive rate of OOD samples when the true positive rate of ID samples is fixed at 95%, and AUC is a widely used metric that quantifies the overall separability between ID and OOD samples, based on the receiver operating characteristic (ROC) curve.

Compared Methods and Setups. We evaluate ZS-CLIP [Radford *et al.*, 2021], Tent [Wang *et al.*, 2021], SoTTA [Gong *et al.*, 2023], TDA [Karmanov *et al.*, 2024], TPT [Shu *et al.*, 2022] and AdaND [Cao *et al.*, 2025] in the benchmarks. In our main experiments, we utilize CLIP [Radford *et al.*, 2021] with a ViT-B/16 [Dosovitskiy, 2020] backbone. Moreover, we report our method's performance across multiple pretrained VLM backbones in Section 4.3. The batch size is fixed at 128, ensuring each sample is processed exactly once (one-pass adaptation). For the initial 10 steps, we retain predictions from the frozen VLM and derive OOD scores via the MCM detector [Ming *et al.*, 2022]. The model is updated using the Adam [Kingma, 2014] optimizer with a learning rate of 0.005 and no weight decay. While the loss weight λ is tuned via grid search, other hyperparameters remain consistent for datasets with over 50 classes: the truncation bound $M = 6$, adapter heads $H = 5$, and a hidden dimension of 128.

4.2 Main Results

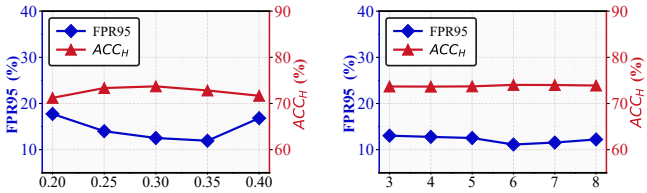
Tables 1 and 2 present a comparative analysis of OSTTA performance across diverse benchmarks, yielding four pivotal insights: (1) Both standard (e.g., Tent) and VLM-specific TTA methods (e.g., TPT, TDA) struggle to consistently surpass the frozen baseline (ZS-CLIP). Notably, under severe distribution

Thresholding	Noise-Resilient	Avg			iNaturalist			SUN			Texture			Places		
		AUC	FPR95	ACC_H	AUC	FPR95	ACC_H	AUC	FPR95	ACC_H	AUC	FPR95	ACC_H	AUC	FPR95	ACC_H
✓	✗	94.06	24.55	70.28	**98.48**	7.28	75.87	94.44	27.07	70.12	91.43	31.52	67.05	91.90	32.34	68.07
✗	✓	93.97	17.21	72.87	97.07	7.03	75.34	93.48	23.41	72.84	92.34	20.53	71.39	92.97	17.88	71.92
✓	✓	**95.58**	**12.52**	**73.71**	98.11	**2.47**	**76.41**	**95.37**	**14.04**	**73.22**	**94.38**	**16.19**	**72.76**	**94.48**	**17.40**	**72.46**

Table 3: Ablation of OOD thresholding and noise-resilient adaptation. Results for the ImageNet benchmark show that each component improves performance, with their combination achieving the best overall performance.

Method	Avg			iNaturalist			SUN			Texture			Places		
	AUC	FPR95	ACC_H	AUC	FPR95	ACC_H	AUC	FPR95	ACC_H	AUC	FPR95	ACC_H	AUC	FPR95	ACC_H
BLIP	83.55	55.42	56.96	83.44	75.36	57.26	80.76	54.45	56.27	81.90	53.83	54.96	88.10	38.04	59.35
BLIP+NISP	89.48	33.66	66.51	93.41	**17.66**	70.55	84.90	41.55	62.87	88.52	**32.71**	65.42	91.11	42.70	67.20
ALIGN	73.36	75.75	58.18	77.39	73.36	60.95	73.97	76.88	58.58	70.83	73.69	56.89	71.26	79.05	56.30
ALIGN+NISP	80.86	65.08	61.10	84.29	78.40	62.30	82.93	48.66	63.52	80.27	57.73	60.88	75.93	75.55	57.69
EVACLIP	88.32	45.99	75.94	92.96	34.48	79.30	89.26	45.24	76.48	84.27	52.06	73.00	86.80	52.18	74.96
EVACLIP+NISP	**93.50**	**31.80**	**84.18**	**95.67**	22.13	**86.02**	**93.98**	**30.14**	**83.94**	**90.62**	40.11	**82.42**	**93.72**	**34.81**	**84.33**

Table 4: OSTTA results in different VLMs for ImageNet benchmark.



(a) Loss weight λ. (b) Adapter heads H.

Figure 4: Sensitivity analysis of key hyperparameters on ImageNet.

Method	Avg	iNaturalist	SUN	Texture	Places
MEAN	63.16	64.38	63.53	61.89	62.82
AdaND	64.11	65.17	64.26	63.39	63.63
NISP	**67.66**	**68.74**	**67.69**	**67.51**	**66.70**

Table 5: Pseudo-label accuracy for ImageNet benchmark.

Resource	ZS-CLIP	Tent	AdaND	NISP
Time (s) ↓	0.0023	0.0059	0.0024	0.0038
Memory (GiB) ↓	4.5889	11.5107	4.6241	4.6494

Table 6: Resource consumption under 128 batchsize.

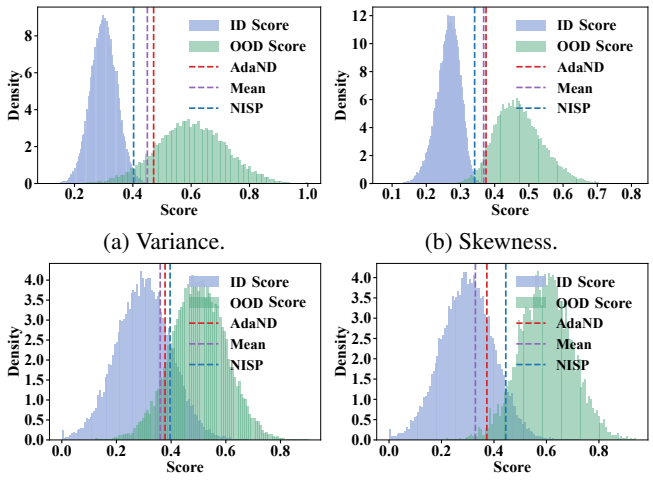
shifts, these approaches often succumb to negative adaptation, yielding lower performance than the non-adapted model. Even specialized OSTTA methods like SoTTA demonstrate only marginal or inconsistent gains across benchmarks. (2) AdaND substantially improves VLMs' performance through decoupling ID classification and OOD detection, establishing a strong baseline in the OSTTA setting. (3) NISP achieves the highest ACC_H and lowest FPR95 across the majority of benchmarks. Specifically on ImageNet, NISP delivers a remarkable ACC_H improvement of 3.40% and a FPR95 reduction of 11.90% relative to AdaND. (4) NISP's advantage is most pronounced in challenging, low-accuracy regimes (e.g., ImageNet-K and CUB-200-2011). Conversely, on saturated benchmarks (e.g., Food-101) where AdaND already performs well, NISP maintains competitive parity. This dichotomy underscores NISP's critical strength in mitigating noise accumulation, a claim further substantiated in Section 4.4.

4.3 Ablation Study

Effectiveness of Each Module. Table 3 presents an ablation study on ImageNet, evaluating the individual contributions of the thresholding strategy and the noise-resilient adaptation module. Excluding the thresholding strategy leads to a noticeable decline in pseudo-label quality, resulting in overall performance degradation. More significantly, ablating the noise-resilient module triggers a 12.03% surge in FPR95 and a 3.43% drop in ACC_H, primarily due to overfitting on noisy pseudo-labels. These findings highlight the critical role of each component in maintaining adaptation robustness. The full model achieves optimal average performance by jointly leveraging both mechanisms in a complementary manner.

Sensitivity Analysis of Key Hyperparameters. Figure 4 visualizes a sensitivity analysisof key hyperparameters on ImageNet. The loss weight λ demonstrates robust adaptation efficacy, striking an optimal trade-off between pseudo-ID consistency and pseudo-OOD discrepancy at approximately 0.3. Furthermore, the model exhibits remarkable insensitivity to the number of adapter heads H. The performance fluctuations are negligible across varying configurations, where our approach consistently outperforms baselines, validating the structural stability of the proposed framework.

Impact of Different VLMs. Table 4 scrutinizes the architectural scalability of NISP across diverse VLM backbones. While EVACLIP establishes a superior baseline due to advanced alignment, incorporating NISP yields universal performance gains regardless of the backbone's capacity. Notably, NISP empowers EVACLIP to achieve peak performance (93.50% Avg AUC and 84.18% Avg ACC_H), significantly surpassing its unadapted counterpart. This confirms that the performance gains are independent of specific architectural inductive biases. These results attest that our framework is architecturally agnostic: it effectively revitalizes weaker models (e.g., BLIP) while further amplifying the capabilities of advanced foundations.



(a) Variance. (b) Skewness.

(c) Number imbalance (7:3). (d) Number imbalance (9:1).

Figure 5: Visualization of threshold placement under different data distributions. The optimal threshold ideally aligns with the intersection of posterior densities.

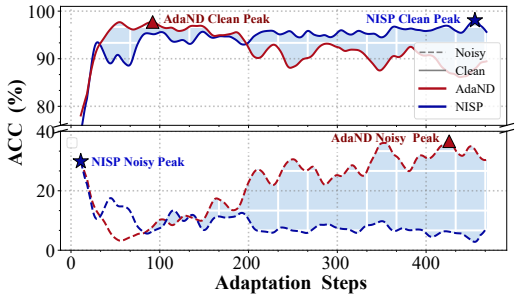


Figure 6: Adaptation dynamics of pseudo-labels. AdaND gradually overfits to noisy samples, degrading clean accuracy, whereas NISP sustains clean supervision and avoids noise-induced collapse.

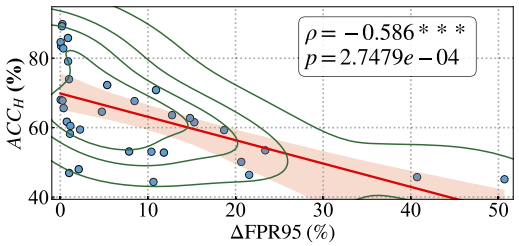


Figure 7: FPR95 gain vs. base pseudo-label accuracy. NISP shows greater improvement on harder benchmarks, with a significant negative correlation.

4.4 Further Analysis

Thresholding under Distribution Shift. Table 5 presents a comparative analysis of pseudo-label accuracy achieved by different thresholding strategies on the ImageNet. MEAN adopts the per-batch average OOD score as the threshold, while AdaND selects the threshold by minimizing intra-cluster score variance. In contrast, NISP achieves the highest pseudo-label accuracy across all datasets. Figure 5 illustrates threshold placement under various data distributions, including settings with variance imbalance, skewness, and number imbalance. NISP consistently locates thresholds near the posterior density intersections to minimize classification risk, demonstrating superior adaptability and reliable pseudo-labeling under complex conditions.

Noisy Overfitting Analysis. To elucidate the performance degradation observed in AdaND during late-stage adaptation, we analyze the dynamics of pseudo-label optimization. As depicted in Figure 6, AdaND initially benefits from clean samples, it progressively succumbs to noise memorization, overfitting incorrect pseudo-labels. This leads to a distinct inverse correlation: rising accuracy on noisy samples at the expense of clean accuracy. In contrast, our method maintains a sustained focus on clean samples throughout adaptation, resulting in stable improvements in clean accuracy while suppressing the influence of noise-induced supervision.

Resource Analysis. Table 6 reports computational efficiency analysis on ImageNet. Even with a multi-head configuration comprising $H = 5$ heads and a hidden dimension of 128, our method maintains inference runtime and memory usage comparable to frozen VLM baseline (ZS-CLIP) and AdaND. This is attributed to the lightweight nature of our adapter architecture (a simple two-layer MLP). Unlike Tent, which requires memory-intensive full-graph backpropagation, NISP isolates gradients within the terminal adapter to prevent storing massive intermediate activations. Consequently, our method allows for efficient online adaptation with significantly lower resource demands than Tent.

Gain Analysis. Experiments reveal heterogeneous performance gains across benchmarks: while reducing FPR95 by 11.90% on ImageNet, our method achieves marginal improvements on other datasets with already high pseudo-label fidelity like Food-101. Figure 7 demonstrates a moderate negative correlation ($\rho = -0.586$, $p < 0.001$) between FPR95 reduction gains and base VLM pseudo-label accuracy across ID/OOD pairs. This stems from AdaND's limited noisy overfitting at high pseudo-label accuracy regimes (e.g., 87.34% on Food-101/SUN), reducing our method's relative advantage. Whereas at low accuracy (e.g., 63.58% on ImageNet/SUN), our method delivers 12.75% FPR95 reduction, validating its noise robustness.

5 Conclusion

In this paper, we propose a novel OSTTA approach named NISP that achieves unsupervised OOD self-purification via a coarse-to-fine noise-resilient adaptation paradigm. Technically, NISP introduces (1) a dual-cluster BGMM posterior-risk thresholding to generate high-fidelity pseudo-labels based on VLM-derived OOD scores, and (2) a fine-grained noise-resilient adapter updates under consensus–discrepancy constraints to refine pseudo-labels and suppress error propagation. Rigorous theoretical analysis demonstrates that NISP upper-bounds the Bayes risk in noisy environments, while extensive experiments across multiple OSTTA benchmarks confirm that NISP outperforms SOTA methods, demonstrating both theoretical soundness and practical efficacy.

Acknowledgments

This work is supported by the NSFC (62276131), Natural Science Foundation of Jiangsu Province of China under Grant (BK20240081).

References

[Blei and Jordan, 2004] David M Blei and Michael I Jordan. Variational methods for the dirichlet process. In *ICML*, page 12, 2004.

[Bossard *et al.*, 2014] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461, 2014.

[Cao *et al.*, 2025] Chentao Cao, Zhun Zhong, Zhanke Zhou, Tongliang Liu, Yang Liu, Kun Zhang, and Bo Han. Noisy test-time adaptation in vision-language models. In *ICLR*, 2025.

[Chao *et al.*, 2025] Dian Chao, Yuxuan Zhang, Luping Zhou, and Yang Yang. Enriching category representations with llms towards robust zero-shot ood detection. In *ECML-PKDD*, pages 20–36, 2025.

[Cimpoi *et al.*, 2014] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[Dong *et al.*, 2025] Hao Dong, Eleni Chatzi, and Olga Fink. Towards robust multimodal open-set test-time adaptation via adaptive entropy-aware optimization. In *ICLR*, 2025.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Feng *et al.*, 2023] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *ICCV*, pages 2704–2714, 2023.

[Gao *et al.*, 2024] Zhengqing Gao, Xu-Yao Zhang, and Cheng-Lin Liu. Unified entropy optimization for open-set test-time adaptation. In *CVPR*, pages 23975–23984, 2024.

[Gong *et al.*, 2023] Taesik Gong, Yewon Kim, Taeckyung Lee, Sorn Chottananurak, and Sung-Ju Lee. Sotta: Robust test-time adaptation on noisy data streams. In *NeurIPS*, pages 14070–14093, 2023.

[Hendrycks *et al.*, 2021] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8340–8349, 2021.

[Jia *et al.*, 2021] 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 *ICML*, pages 4904–4916, 2021.

[Jia *et al.*, 2024] Ziyu Jia, Xihao Yang, Chenyang Zhou, Haoyang Deng, Tianzi Jiang, and Brainnetome Center. Atta: adaptive test-time adaptation for multi-modal sleep stage classification. In *IJCAI*, pages 5882–5890, 2024.

[Karmanov *et al.*, 2024] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *CVPR*, pages 14162–14171, 2024.

[Kingma, 2014] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Krause *et al.*, 2013] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, pages 554–561, 2013.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images., 2009.

[Li *et al.*, 2023a] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742, 2023.

[Li *et al.*, 2023b] Yushu Li, Xun Xu, Yongyi Su, and Kui Jia. On the robustness of open-world test-time training: Self-training with dynamic prototype expansion. In *ICCV*, pages 11836–11846, 2023.

[Li *et al.*, 2025] Zeman Li, Xinwei Zhang, Peilin Zhong, Yuan Deng, Meisam Razaviyayn, and Vahab Mirrokni. Addax: Utilizing zeroth-order gradients to improve memory efficiency and performance of SGD for fine-tuning language models. In *ICLR*, 2025.

[Likas *et al.*, 2003] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461, 2003.

[Lim *et al.*, 2023] Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. Ttn: A domain-shift aware batch normalization in test-time adaptation. In *ICLR*, 2023.

[Ma *et al.*, 2023] Xiaosong Ma, Jie Zhang, Song Guo, and Wenchao Xu. Swapprompt: Test-time prompt adaptation for vision-language models. In *NeurIPS*, volume 36, pages 65252–65264, 2023.

[Martinez and Raschia, 2024] José Martinez and Guillaume Raschia. Revisiting optimal window aggregation in data streams: The prefix-sum approach. In Edoardo Serra and Francesca Spezzano, editors, *CIKM*, pages 1660–1669. ACM, 2024.

[Ming *et al.*, 2022] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *NeurIPS*, volume 35, pages 35087–35102, 2022.

[Netzer *et al.*, 2011] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NeurIPS workshop*, page 7, 2011.

[Niu *et al.*, 2022] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *ICML*, pages 16888–16905, 2022.

[Peng *et al.*, 2025] Boyang Peng, Sanqing Qu, Tianpei Zou, Fan Lu, Ya Wu, Kai Chen, Siheng Chen, Yong Wu, and Guang Chen. Ood-barrier: Build a middle-barrier for open-set single-image test time adaptation via vision language models. In *NeurIPS*, 2025.

[Plassier *et al.*, 2025] Vincent Plassier, Alexander Fishkov, Victor Dheur, Mohsen Guizani, Souhaib Ben Taieb, Maxim Panov, and Eric Moulines. Rectifying conformity scores for better conditional coverage. In *ICML*, volume 267, pages 49459–49492, 2025.

[Radford *et al.*, 2021] 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 *ICML*, pages 8748–8763, 2021.

[Recht *et al.*, 2019] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019.

[Schneider *et al.*, 2020] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *NeurIPS*, volume 33, pages 11539–11551, 2020.

[Seccia *et al.*, 2025] Ruggiero Seccia, Corrado Coppola, Giampaolo Liuzzi, and Laura Palagi. Convergence of ease-controlled random reshuffling gradient algorithms under lipschitz smoothness. *Computational Optimization and Applications*, 57(5):1–37, 2025.

[Shu *et al.*, 2022] 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 *NeurIPS*, volume 35, pages 14274–14289, 2022.

[Sun *et al.*, 2023] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[Van Horn *et al.*, 2018] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, pages 8769–8778, 2018.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*, 2011.

[Wang *et al.*, 2019] Hao Wang, Bing Liu, Chaozhuo Li, Yan Yang, and Tianrui Li. Learning with noisy labels for sentence-level sentiment classification. In *EMNLP*, pages 6286–6292, 2019.

[Wang *et al.*, 2021] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021.

[Wu *et al.*, 2025] Xiangyu Wu, Feng Yu, Qing-Guo Chen, Yang Yang, and Jianfeng Lu. Multi-label test-time adaptation with bound entropy minimization. In *ICLR*, 2025.

[Xiao *et al.*, 2010] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010.

[Xue *et al.*, 2013] Wufeng Xue, Xuanqin Mou, Lei Zhang, and Xiangchu Feng. Perceptual fidelity aware mean squared error. In *ICCV*, pages 705–712, 2013.

[Yang and Xu, 2025] Yang Yang and Haonan Xu. Strengthen out-of-distribution detection capability with progressive self-knowledge distillation. In *ICML*, 2025.

[Yang *et al.*, 2024] Yang Yang, Nan Jiang Jiang, Yi Xu, and De-Chuan Zhan. Robust semi-supervised learning by wisely leveraging open-set data. *IEEE transactions on pattern analysis and machine intelligence*, 46(12):8334–8347, 2024.

[You *et al.*, 2020] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. In *ICLR*, 2020.

[Yu *et al.*, 2015] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[Yuan *et al.*, 2023] Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *CVPR*, pages 15922–15932, 2023.

[Zhang *et al.*, 2023] Xiaobo Zhang, Yutao Liu, Hao Wang, Wei Wang, Panpan Ni, and Ji Zhang. Cosar: Combating label noise using collaborative sample selection and adversarial regularization. In *CIKM*, pages 3184–3194, 2023.

[Zhao *et al.*, 2025] Shiji Zhao, Shao-Yuan Li, Chuanxing Geng, Sheng-Jun Huang, and Songcan Chen. Dm-posa: enhancing open-world test-time adaptation with dual-mode matching and prompt-based open set adaptation. In *IJCAI*, pages 7101–7109, 2025.

[Zhou *et al.*, 2017] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.