OOD-Barrier: Build a Middle-Barrier for Open-Set Single-Image Test Time Adaptation via Vision Language Models

Boyang Peng[1,2] **Sanqing Qu**[1]* **Tianpei Zou**[1] **Fan Lu**[1] **Ya Wu**[3]
Kai Chen[1] **Siheng Chen**[4] **Yong Wu**[1] **Guang Chen**[1,2]*

[1] Tongji University [2] Shanghai Innovation Institute
[3] CNNC Equipment Technology Research (Shanghai) Co., Ltd. [4] Shanghai Jiao Tong University

Abstract

In real-world environments, a well-designed model must be capable of handling dynamically evolving distributions, where both in-distribution (ID) and out-of-distribution (OOD) samples appear unpredictably and individually, making real-time adaptation particularly challenging. While open-set test-time adaptation has demonstrated effectiveness in adjusting to distribution shifts, existing methods often rely on batch processing and struggle to manage single-sample data stream in open-set environments. To address this limitation, we propose Open-IRT, a novel open-set Intermediate-Representation-based Test-time adaptation framework tailored for single-image test-time adaptation with vision-language models. Open-IRT comprises two key modules designed for dynamic, single-sample adaptation in open-set scenarios. The first is Polarity-aware Prompt-based OOD Filter module, which fully constructs the ID-OOD distribution, considering both the absolute semantic alignment and relative semantic polarity. The second module, Intermediate Domain-based Test-time Adaptation module, constructs an intermediate domain and indirectly decomposes the ID-OOD distributional discrepancy to refine the separation boundary during the test-time. Extensive experiments on a range of domain adaptation benchmarks demonstrate the superiority of Open-IRT. Compared to previous state-of-the-art methods, it achieves significant improvements on representative benchmarks, such as CIFAR-100C and SVHN — with gains of +8.45% in accuracy, -10.80% in FPR95, and +11.04% in AUROC.

1 Introduction

In real-world environments, the model's ability to perform real-time adaptation is particularly crucial for handling the emergence of an unknown category or distributional shifts. This capability is essential in safety-critical applications such as autonomous driving, where failure to adapt can have serious consequences. Despite these demands, most approaches in computer vision [1–3] assume a closed-set paradigm, where training and testing data come from the same distribution. In contrast, real-world scenarios frequently encounter open-set environments [4–6], where models must handle unknown distributions and unseen categories, as shown in Fig. 1. This fundamental deviation from the closed-set assumption introduces fundamental challenges in maintaining model performance.

This transition from closed-world to open-set learning necessitates requiring innovative strategies to effectively filter, classify, and adapt to both distributional and semantic shifts without explicit supervision [7]. However, traditional approaches based on predefined categories are not equipped

[*] Corresponding Author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

to handle the emergence of unknown categories in open-set environments, particularly in real-time inference and single-sample adaptation constraints. As a result, accurate and efficient OOD detection has emerged as a critical research focus [7–9] in dynamic, open-set scenarios.

With the advent of vision-language models (VLMs), recent studies [10, 11] have leveraged their strong generalization for OOD detection tasks. Models such as CLIP [10] facilitate adaptation to unseen categories in open-set situation by learning rich cross-modal representations. Furthermore, recent research [12] have demonstrated that VLMs can perform zero-shot reasoning based on image-text associations and further adapt to single image inference, which enhances both their zero-shot generalization and OOD detection performance.

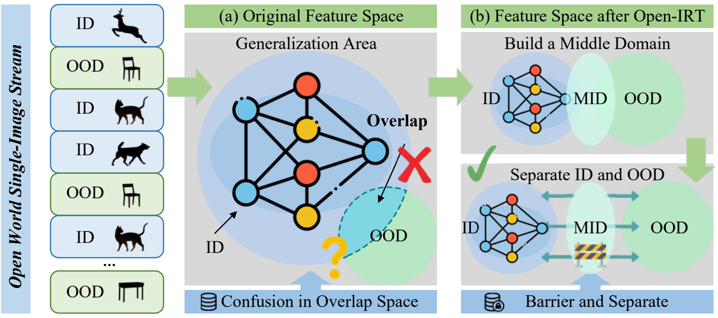


Figure 1: The objective of Open-IRT. In open-set environments, models encounter individual sample streams, which may include unknown categories. Existing test-time adaptation methods mainly focus on closed-set settings or batch image processing, neglecting single-sample data streams. (a) In such cases, the boundary between ID and OOD data can become unclear. (b) Our goal is to improve domain separation via a middle domain (MID) strategy , as shown in Fig. 4, and use test-time adaptation with VLMs to reduce ambiguity.

Test-time adaptation [13–16] offers a promising approach for adapting models to incoming data streams during inference, without relying on labeled data. Recent developments have extended this paradigm to single-sample test-time adaptation [17, 18], which is particularly valuable in domains such as security surveillance and industrial inspection, where real-time adaptation to dynamic environments is critical. However, most existing single-image test-time adaptation methods mainly focus on closed-set assumptions, whereas open-set batch processing strategies struggle to effectively adapt to changes in individual samples, particularly in real-time scenarios where only a single image is available for adaptation, as illustrated in Fig. 1.

To overcome this limitation, we focus on Open-Set Single-Image Test-time adaptation setting, where the model dynamically adapts to each input sample during inference without relying on batch data. Specifically, we propose a novel ***Open**-set **I**ntermediate-**R**epresentation-based **T**est-time adaptation* (**Open-IRT**) framework for single-sample test-time adaptation in OOD detection. The overall motivation is illustrated in Fig. 1. As illustrated in Fig. 1b, Open-IRT establishes a structured separation between ID and OOD samples in feature space. The framework consists of two key modules. First, we introduce a ***P**olarity-aware **P**rompt-based OOD **F**ilter* (**PPF**) module in Fig. 2a. Here, the term "polarity" refers to our utilization of the disparity between positive and negative prompts. PPF leverages the rich cross-modal information in vision language models to fully construct the ID-OOD distribution from both positive and negative prompts. It is guided by Semantic Contrast Hypothesis 1, which considers the absolute semantic alignment and relative semantic polarity. Next, we introduce *Intermediate **D**omain-based **T**est-time adaptation* (**IDT**) module in Fig. 2b based on Intermediate-Domain Hypothesis 2, which indirectly decomposes the distributional discrepancy between ID and OOD representations by modeling an intermediate domain that bridges the gap between these two. As shown in Fig.4, the IDT explicitly models this middle domain, and the learning objective enforces divergence of both ID and OOD samples from the intermediate domain. Therefore, this strategy effectively enlarging the distance between ID and OOD distributions indirectly. In addition, IDT uses a dynamic threshold strategy to generate bidirectional pseudo-labels, encouraging the model to reinforce positive feature representations and suppress intra-class noise.

The main contributions are summarized as follows: (i) We propose the PPF, an effective OOD filtering mechanism based on Semantic Contrast Hypothesis 1. The PPF captures both absolute semantic alignment and relative semantic polarity between an input and its paired prompts, enabling effective ID-OOD separation. (ii) We introduce the Intermediate-Domain Hypothesis 2, which leads to the development of IDT module. This module construct a intermediate domain strategy to establish a real-time two-way repulsion constraint between ID and OOD feature distributions, enhancing ID-OOD separation indirectly. (iii) Open-IRT consistently outperforms prior state-of-the-art methods on standard benchmarks, including ImageNet-C [19], Tiny-ImageNet [20], VisDA [21], CIFAR-10C/100C [19], and digit datasets [22–24]. For instance, on CIFAR-100C and SVHN, it achieves +8.45% accuracy, -10.80% FPR95, and +11.04% AUROC, supporting our motivation and hypothesis.

2 Related Works

2.1 Out-of-Distribution Detection

Out-of-distribution (OOD) detection aims to determine whether a given sample originates from the training distribution or from an unseen distribution. The OOD detection approaches can be broadly categorized into classification-based methods [7, 25, 26] and density-based methods [8, 27]. Classification-based OOD methods relied on a maximum softmax probability to distinguish between ID and OOD samples. These include post hoc techniques such as ODIN [28], which utilizes temperature scaling and input perturbation. In addition, there are classic approaches such as JointEnergy scores [26], Mahalanobis distance [29], and activation space-based techniques [30, 31] that enhance the separability of ID and OOD samples without altering the training process. On the other hand, density-based methods use probabilistic models, such as class-conditional Gaussian distributions [29] and flow-based models [32, 9], to identify OOD samples based on their likelihood. To address high likelihoods challenges of OOD, techniques such as likelihood ratio [33], likelihood regret [34], and SEM scores [35] have been proposed. However, traditional methods are predominantly designed for the training phase, limiting their ability to adapt to distribution and semantic shifts in real-time.

2.2 OOD Detection with Vision Language Models

Vision-language models have gained significant attention in recent years for their ability to integrate visual and textual information. Renowned vision-language models, such as CLIP [10] and MaPLe [11], achieve impressive results by training on large-scale image-text pairs. To adapt vision-language models for downstream tasks (e.g., OOD detection), additional lightweight modules have been introduced, including prompt learners [36, 37], vision adapters [38, 39], and LoRA [40, 41].

Recent research in OOD detection has begun to utilize vision-language models as auxiliary tools, starting with CLIP [42], which aims to distinguish samples that do not belong to any ID class text provided by the user [43]. These approaches often employ techniques such as OOD label retrieval [44], generation [45], or alignment [46]. Training-free methods such as MCM [47] detect OOD using only ID labels, while auxiliary training-based methods, such as CLIPN [48] leverage additional pre-training to enhance OOD detection. From the perspective of specific training methodologies, some approaches implement a specialized handling of prompt words [49, 50], while NegPrompt [51] further explores the use of negative prompt techniques. Moreover, ROSITA [52] consider this task in test-time, yet it remains constrained by its reliance on direct feature alignment. Unlike previous methods, Open-IRT is fundamentally guided by an intermediate domain located near the boundary of the feature space, which indirectly decomposes the distributional discrepancy between ID and OOD representations. It introduces a bidirectional repulsion constraint, combining semantic alignment with relative semantic polarity, to increase the distributional separation.

2.3 Test-Time Adaptation

Test-time adaptation, originating from domain adaptation [53–55], adapts pre-trained models to test data with distribution shifts, without requiring training data. Test-time adaptation is essential for real-world applications, such as autonomous driving in diverse weather conditions. Several methods have been proposed, such as adjusting partial model weights [13, 56] or normalization statistics [14, 15]. Specifically, TENT [15] adapts batch normalization layers by entropy minimization, TTT [56] updates classification layers during testing, and T3A [13] introduces an optimization-free classifier adjustment. As a technology for changing environments, test-time adaptation can be integrated with OOD detection (e.g., RTL [57], UniEnt [58]). However, they typically address the task using batch processing methods, which limits their ability to handle dynamic scenarios effectively.

To reduce reliance on multiple test samples, some approaches prioritize single-sample adaptation. MEMO [59] enforces consistency through augmentations of the same test sample, while TPT [12] fine-tunes prompts for vision-language models during testing. DiffTPT [60] enhances this by using pre-trained diffusion models to augment test data. TDA [61] addresses computational efficiency

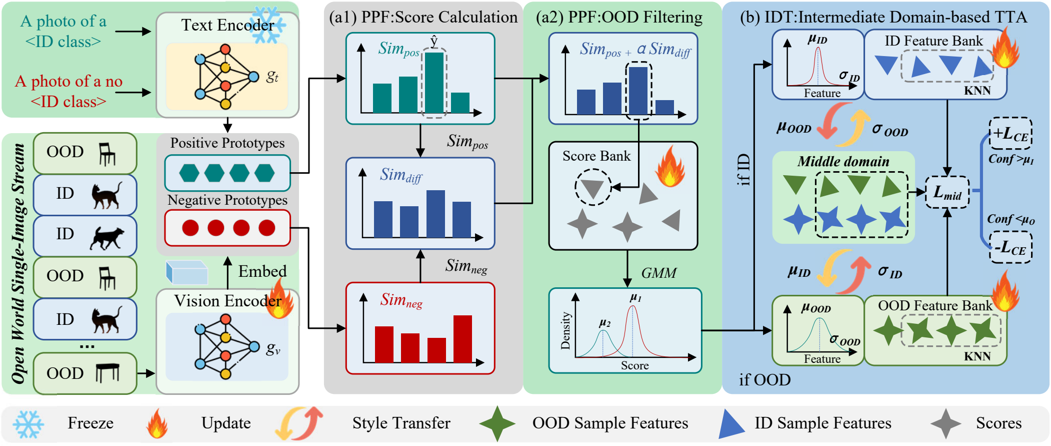


Figure 2: The architecture of Open-IRT comprises two key modules: (a) Polarity-aware Prompt-based OOD Filter and (b) Intermediate Domain-based Test-time Adaptation. (a) filters ID and OOD samples using positive and negative prompts, computing scores in Eq. 3 and utilizing a GMM strategy for OOD filtering. (b) generates an intermediate domain in Fig. 4 and Eq. 7 by leveraging mean and variance from feature banks, enhancing the model's adaptability with contrastive learning loss $\mathcal{L}_{mid}$ in Eq. 10 and pesudo-label loss $\mathcal{L}_{psd}$ in Eq. 13 during the test phrase.

with a cache-based model. While these methods assume closed-set settings, Open-IRT enables single-image Test-time adaptation in open-set environments. Unlike continual test-time adaptation, it further considers semantic shift in addition to distribution shift and refines the ID-OOD boundary in terms of the intermediate representation.

3 Methodology

In open-world environments with real-time perception demands, models frequently encounter single-image data streams that contain OOD samples. To address this challenge, we propose *Open-set Intermediate-Representation-based Test-time Adaptation* (Open-IRT) in Fig. 2. Open-IRT is an open-set, single-image test-time adaptation strategy composed of two key modules: Polarity-aware Prompt based OOD Filter module and Intermediate Domain-based Test-time adaptation module.

3.1 PPF: Polarity-aware Prompt-based OOD Filtering Mechanism

As shown in Fig. 2a, we introduce the *Polarity-aware Prompt-based OOD Filter* (PPF) module, which improves ID–OOD separability by exploiting the dual-polarity semantics of vision-language models. Specifically, we design positive prompts p_c (e.g., "a photo of a [CLS]") to represent ID prototypes, and negative prompts p'_c (e.g., "a photo of no [CLS]") to encode inverse semantics [48]. Given a vision feature $f = g_v(x)$ after L2 normalization, we compute its cosine similarity with both the positive prompt p_c and the negative prompt p'_c for class c. $sim_{pos} = \mathrm{sim}(f, p_c) = \frac{f^\top p_c}{\|f\|\|p_c\|}$, $sim_{neg} = \mathrm{sim}(f, p'_c) = \frac{f^\top p'_c}{\|f\|\|p'_c\|}$, $sim_{diff} = |sim_{pos} - sim_{neg}|$.

Hypothesis 1 (Semantic Contrast Hypothesis)**.** *For ID samples* $\forall c \in [C]$, $f \sim \mathcal{P}_I$ *and OOD* $f \sim \mathcal{P}_O$*:*

$$\mathbb{E}_{f\sim\mathcal{P}_I}[\mathrm{sim}(f, p_c)] \geq \tau_I, \quad \mathbb{E}_{f\sim\mathcal{P}_O}[\mathrm{sim}(f, p_c)] \leq \tau_O \tag{1}$$

$$\mathbb{E}_{f\sim\mathcal{P}_I}[\mathrm{sim}(f, p_c) - \mathrm{sim}(f, p'_c)] \geq \Delta_I, \quad \mathbb{E}_{f\sim\mathcal{P}_O}[\mathrm{sim}(f, p_c) - \mathrm{sim}(f, p'_c)] \leq \Delta_O \tag{2}$$

where thresholds satisfy $\tau_I > \tau_O$ *and* $\Delta_I > \Delta_O$*.*

First, we propose the Semantic Contrast Hypothesis. For ID samples, the alignment with positive prompts is strong, i.e., $\mathrm{sim}(f, p_c)$ is high, and the polarity gap $|\mathrm{sim}(f, p_c) - \mathrm{sim}(f, p'_c)|$ is also large, leading to a confidently larger overall sum. For OOD samples, the alignment with positive prompts is weaker, and the polarity gap is less marked due to semantic ambiguity, leading to suppressed overall sum. The proposed scoring function $\mathcal{S}(f)$ jointly captures *absolute semantic alignment* and *relative semantic polarity* between an input and its paired prompts. This promotes effective ID–OOD separability by maximizing the distributional contrast between them.

$$\mathcal{S}(f) = \phi\left(\sup_{c\in[C]}\left[\mathrm{sim}(f,p_c) + \alpha\left|\mathrm{sim}(f,p_c) - \mathrm{sim}(f,p_c')\right|\right]\right), \tag{3}$$

where ϕ is a min-max normalization operator to ensure the score lies within the range $[0,1]$, and α controls the contrast intensity. As shown in Fig. 3, while individual sample have unique semantics, the resulting aggregate score distribution exhibits a clear bimodal pattern, demonstrating consistent statistical regularity across large-scale data. Furthermore, the introduction of the polarity gap term significantly enhances ID-OOD separation, thereby validating the Hypothesis 1. This dual-polarity design—leveraging alignment with positive prompts and contrast with negative prompts—offers an effective mechanism for ID-OOD distinction. Detailed theoretical analyses are provided in appendix.

Then, we introduce the score bank $\mathcal{B}^s$ and feature bank $\mathcal{B}^f$, both updated using a sliding window strategy, where it functions equivalently to a FIFO queue. The score bank $\mathcal{B}^s$ stores the scores of individual samples and is divided into two components: ID ($\mathcal{B}^s_I$) and OOD ($\mathcal{B}^s_O$) score banks, based on the Gaussian Mixture Model as described in Eq. 4.

$$\mathcal{P}(x) = \pi(x)\mathcal{N}(x \mid \mu^s_I, {\sigma^s_I}^2) + (1-\pi(x))\mathcal{N}(x \mid \mu^s_O, {\sigma^s_O}^2) \tag{4}$$

Here, $\pi(x)$ denotes the probability that $S(x)$ belongs to the ID class, and $\mu^s_{I/O}$, ${\sigma^s_{I/O}}^2$ are the mean and variance of the ID/OOD components. The probability $\pi(x)$ is computed using the Expectation-Maximization algorithm.

Upon the arrival of a new sample x, its score $S(x)$ is appended to both $\mathcal{B}^s_I$ and $\mathcal{B}^s_O$, and Eq. 4 is used to classify it as either ID ($\hat{b} = 1$) or OOD ($\hat{b} = 0$). To mitigate the impact of limited samples during the early stages of learning, we employ bootstrapped resampling, which helps prevent overfitting to sparse observations. The classified features are then inserted into the corresponding feature banks $\mathcal{B}^f_{I/O}$ for use in the subsequent module.

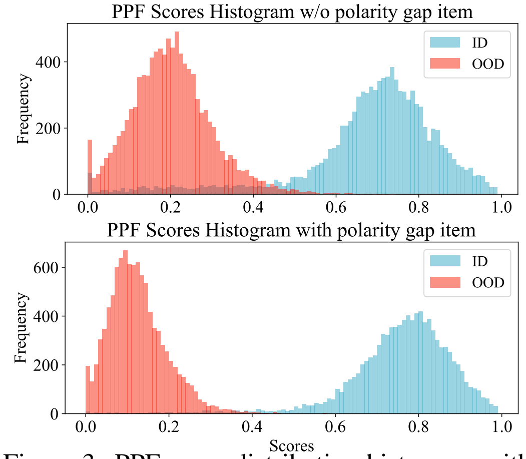


Figure 3: PPF score distribution histogram with CIFAR-10C [19] as ID and MNIST [22] as OOD.

3.2 IDT: Intermediate Domain-based Test-time adaptation

We now introduce the *Intermediate Domain-based Test-time adaptation* (IDT) module, which leverages the feature banks $\mathcal{B}^f_I$ and $\mathcal{B}^f_O$ to further refine adaptation. To address the challenge in Fig. 1, we first propose Hypothesis 2 that models the intermediate domain ($\mathcal{F}_M$).

Hypothesis 2 (Intermediate Domain Characterization)**.** *Let $\mathcal{F}_I \triangleq \{f|f \sim \mathcal{P}_I\}$ and $\mathcal{F}_O \triangleq \{f|f \sim \mathcal{P}_O\}$ denote the ID/OOD feature distributions respectively. There exists a measurable transformation $\mathcal{T} : \mathcal{F}_I \cup \mathcal{F}_O \rightarrow \mathcal{F}_M$ such that for the induced intermediate feature space $\mathcal{F}_M$ that satisfies the following approximate equality:*

$$d_{\mathcal{H}}(\mathcal{F}_I, \mathcal{F}_M) + d_{\mathcal{H}}(\mathcal{F}_O, \mathcal{F}_M) \approx d_{\mathcal{H}}(\mathcal{F}_I, \mathcal{F}_O), \tag{5}$$

where $d_{\mathcal{H}}$ represents the Hilbert-Schmidt independence criterion-based dissimilarity.

Rather than directly maximizing the ID-OOD distance $(\mathcal{F}_I, \mathcal{F}_O)$, which is challenging due to limited knowledge of OOD distribution, we instead construct an intermediate domain $\mathcal{F}_M$ to act as a bridge. By encouraging both ID and OOD samples to move away from this intermediate domain (i.e., increasing $d_{\mathcal{H}}(\mathcal{F}_I, \mathcal{F}_M)$ and $d_{\mathcal{H}}(\mathcal{F}_O, \mathcal{F}_M)$), we indirectly enhance the overall discrepancy $d_{\mathcal{H}}(\mathcal{F}_I, \mathcal{F}_O)$. This aligns with the intuition of margin-based separation in representation space.

Contrastive Learning-based Middle Domain Loss. We first model the intermediate domain feature f_m as per Hypothesis 2. To normalize the sample features f_s, we compute their mean and variance based on augmented features, which is derived the f_s as Eq. 6. The method for calculating f_s has been validated for its rationality in [62], where it is computed across spatial dimensions independently for each feature channel. The feasibility of this approach lies in preserving channel-specific style information. The variance for each channel emphasizes fine-grained structural characteristics in that feature dimension, such as color, texture intensity, etc., without mixing with other channels, thereby achieving refined style representation and alignment.

$$f_s = \frac{f - \mu(f)}{\sigma(f)} \tag{6}$$

Then, we extract the style features stored in the feature memory banks $\mathcal{B}^f_{I/O}$ to compute $\mu^f_{I/O}$ and $\sigma^f_{I/O}$. The design of intermediate style feature f_m is theoretically motivated by domain adaptation principles [62], which suggest that constructing an intermediate feature space through style-transfer-like transformations can effectively position the new domain between source and target domains, enabling more controlled feature interpolation. f_m is then re-assigned as follows.

$$f_m = \begin{cases} f_s \cdot \sigma^f_O + \mu^f_O & \text{if } \hat{b} = 1; \text{Conf}(x) > \mu_I \\ f_s \cdot \sigma^f_I + \mu^f_I & \text{if } \hat{b} = 0; \text{Conf}(x) < \mu_O \end{cases} \tag{7}$$

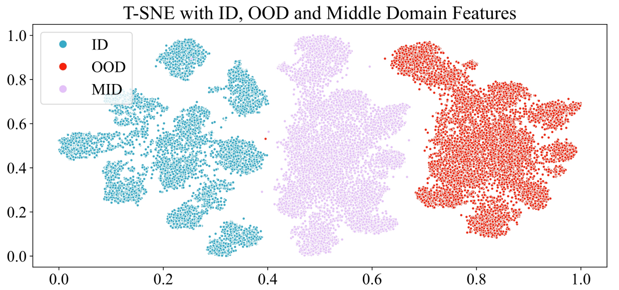


Figure 4: T-SNE visualization of MNIST [22] (ID) and CIFAR-10C [19] (OOD) and their Middle-Domain (MID)

The confidence Conf(x) is defined in Eq. 11. To further enhance inter-class discriminability and suppress intra-class noise, we introduce two thresholds, μ_I and μ_O. They represent the mean confidence scores of the accumulated 512 ID/OOD Conf(x), respectively, updated by a sliding window mechanism. If Conf$(x) > \mu_I$, we treat the input as a confident ID sample. If Conf(x) $< \mu_O$, it is more likely to be a noisy sample or an OOD sample6. The effectiveness of f_m construction is visually confirmed in Fig. 4, where the T-SNE visualization shows a clear separation between ID and OOD features, with a distinct intermediate cluster. This indicates that the constructed intermediate representations indeed satisfy the desired relational constraints in the feature space, thereby validating the theoretical motivation of Hypothesis 2. Then, we select the K nearest neighbors $\mathcal{N}_I$ and $\mathcal{N}_O$, for the sample feature f from the feature banks $\mathcal{B}^f_I$ and $\mathcal{B}^f_O$, respectively. When the sample is confidently distinguished as ID ($\hat{b} = 1$ and Conf$(x) > \mu_I$), the loss $\mathcal{L}_I$ based on InfoNCE loss [63] is constructed as follows.

$$\begin{aligned} \mathcal{L}_I = -\frac{1}{K^+} \sum_{z^+ \in \mathcal{N}_I} (\log \frac{\exp(\text{sim}(f, z^+)/\tau)}{\sum_{z^- \in \mathcal{N}_O} \exp(\text{sim}(f, z^-)/\tau)} \\ -\lambda \log \frac{\exp(\text{sim}(f_m, z^+)/\tau)}{\sum_{z^- \in \mathcal{N}_O} \exp(\text{sim}(f_m, z^-)/\tau)}) \mathbb{1}(y^+ = \hat{y}_p) \end{aligned} \tag{8}$$

where $K^+ = \sum_{z^+ \in \mathcal{N}_I} \mathbb{1}(y^+ = \hat{y}_p)$ denotes the number of positively matched neighbors with the pseudo label $\hat{y}_p$. For confident ID samples, the objective of $\mathcal{L}_I$ is to align with ID while minimizing similarity to OOD and distancing from the intermediate feature f_m.

$$\mathcal{L}_O = -\frac{1}{K} \sum_{z^+ \in \mathcal{N}_O} \left(\log \frac{\exp(\text{sim}(f, z^+)/\tau)}{\sum_{z^- \in \mathcal{N}_I} \exp(\text{sim}(f, z^-)/\tau)} - \lambda \log \frac{\exp(\text{sim}(f_m, z^+)/\tau)}{\sum_{z^- \in \mathcal{N}_I} \exp(\text{sim}(f_m, z^-)/\tau)} \right) \tag{9}$$

For confident OOD samples ($\hat{b} = 0$ and Conf$(x) < \mu_O$), the objective of $\mathcal{L}_O$ is to improve sensitivity to OOD samples by pulling OOD features away from both ID and the intermediate domain feature.

The differences between Eq. 8 and Eq. 9 arise from the distinction between ID and OOD categories in the standard OOD detection task. For ID data, the model needs to learn fine-grained semantic consistency and achieve precision for each class, which is why we set $\mathbb{1}(y^+ = \hat{y}_p)$ in Eq. 8.

$$\mathcal{L}_{mid} = \begin{cases} \mathcal{L}_I & \text{if } \hat{b} = 1;\ \text{Conf}(x) > \mu_I \\ \mathcal{L}_O & \text{if } \hat{b} = 0;\ \text{Conf}(x) < \mu_O \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Threshold-based Bidirectional Pseudo Loss. To facilitate positive knowledge while mitigating intra-class noise, we design an effective pseudo-label strategy. First, we calculate the confidence $\text{Conf}(x)$ using the maximum cosine similarity between feature $g_v(x)$ and classifier weights $\mathbf{C}$:

$$\text{Conf}(x) = \max_i (g_v(x) \cdot \mathbf{C}^T), \quad \hat{y}_p = \arg\max_i (g_v(x) \cdot \mathbf{C}^T) \tag{11}$$

We analyze $\text{Conf}(x)$ from two perspectives: For high-confidence scenarios, we align predictions with $\hat{y}_p$. For low-confidence scenarios, $\hat{y}_p$ may introduce noise, so we stop relying on it and use reverse optimization to push predictions away from it, mitigating the risk of misclassification.

$$\mathcal{L}_{CE}(x, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i \log(p_i) + (1 - \hat{y}_i) \log(1 - p_i)) \tag{12}$$

If $\text{Conf}(x) > \mu_I$, we treat the input as an confident sample and optimize the model to align predictions with $\hat{y}_p$ using $\mathcal{L}_{CE}$, enhancing the inter-class discriminability. If $\text{Conf}(x) < \mu_O$, we deviate the prediction results from the low-quality pseudo-label $\hat{y}_p$ by $-\mathcal{L}_{CE}$, suppressing the intra-class noise.

$$\mathcal{L}_{psd} = \begin{cases} \mathcal{L}_{CE}(x, \hat{y}_p) & \text{if } \hat{b} = 1;\ \text{Conf}(x) > \mu_I \\ -\mathcal{L}_{CE}(x, \hat{y}_p) & \text{if } \hat{b} = 0;\ \text{Conf}(x) < \mu_O \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

Total Test-time Adaptation Loss. The total loss function $\mathcal{L}_{TTA}$ combines the pseudo-label loss $\mathcal{L}_{psd}$ and the intermediate domain loss $\mathcal{L}_{mid}$. $\mathcal{L}_{psd}$ aims to enhance inter-class discriminability and suppress intra-class noise, while $\mathcal{L}_{mid}$ indirectly enlarge the ID-OOD distance by increasing both $d_{\mathcal{H}}(\mathcal{F}_I, \mathcal{F}_M)$ and $d_{\mathcal{H}}(\mathcal{F}_O, \mathcal{F}_M)$ in Hypothesis 2. This combination enables dynamic adjustment of the model's adaptation strategy.

$$\mathcal{L}_{TTA} = \mathcal{L}_{psd} + \mathcal{L}_{mid} \tag{14}$$

During test time, each incoming sample is used to update the model parameters via backpropagation with the test-time adaptation loss $\mathcal{L}_{TTA}$. Subsequently, the same sample is used to compute the evaluation metrics. This online update-then-evaluate approach simulates a realistic scenario where the model continuously adapts to distributional shifts without access to ground-truth labels.

4 Experiments

4.1 Implementation Details

Datasets. We utilize representative ID and OOD datasets to ensure a complete evaluation. For ID datasets, we leverage CIFAR-10C/100C [19], ImageNet-C [19], and VisDA [21]. The OOD datasets include MNIST [22], MNIST-M [23], SVHN [24], Tiny-ImageNet [20], and CIFAR-10C/100C [19].

Metrics. We utilize the Area Under the Receiver Operating Characteristic Curve (AUROC), the False Positive Rate at 95% True Positive Rate (FPR95) and Acc_{HM} as main metrics. Here, Acc_{HM} denotes the harmonic mean of Acc_I and Acc_O, where Acc_O is the precision of ID-OOD binary classification, and Acc_I is the general accuracy to correctly identify ID categoriess.

Baselines. We utilize the CLIP [10] and MaPLe [11] as models, which are based on the ViT-B16/ViT-B32 [64] architectures. We adopt ZS-Eval [52], TPT/TPT-C [12], PAlign/PAlign-C [46], TDA [61], DPE [65], UniEnt [58], OWTTT [66] and ROSITA [52] as baselines.

Details. All experiments are reproduced based on publicly available code, with ImageNet-C experiments conducted on an NVIDIA A6000 GPU and all other experiments on an NVIDIA 3090 GPU. In main experiments, the test size for both ID and OOD datasets is 10,000, except for the VisDA in

Table 3, which is 50,000. In Table 5, the OOD test size is $ratio \times 10,000$. The text encoder is kept fixed, while the vision encoder is updated with a SGD optimizer with learning rate of 1.5e-3, and batch size is set to 1 for all experiments. The size B of both score bank $\mathcal{B}^s$ and feature bank $\mathcal{B}^f$ are set to 128, the number of nearest neighbors K in Eq. 8, 9 configured to 5. The α in Eq. 3 is set to 0.2, and λ in Eq. 8, 9 are set to 0.1, respectively. Details on the baselines and analysis of hyper-parameters are provided in appendix.

Method			MNIST			SVHN			Tiny-ImageNet			CIFAR-100C		
			AUC ↑	FPR ↓	HM ↑	AUC ↑	FPR ↓	HM ↑	AUC ↑	FPR ↓	HM ↑	AUC ↑	FPR ↓	HM ↑
CIFAR10-C	CLIP	ZS-Eval [52]	91.91	85.22	75.60	89.94	64.25	74.11	91.33	27.13	74.24	82.57	67.96	68.92
		TPT [12]	91.90	85.70	75.78	89.93	64.54	74.30	91.31	27.26	74.98	82.57	68.09	69.13
		TPT-C [12]	83.21	67.03	75.05	60.83	69.47	50.63	74.12	57.34	48.88	63.76	93.05	51.98
		ROSITA [52]	99.43	3.25	83.95	94.94	31.22	79.12	96.37	12.69	80.07	83.01	64.54	69.64
		OWTTT [66]	98.05	12.50	83.27	80.74	50.33	70.10	87.09	52.29	73.98	62.55	91.68	56.46
		TDA [61]	92.94	71.11	77.06	92.02	52.68	76.64	91.68	25.37	75.94	**83.54**	66.06	**70.13**
		UniEnt [58]	91.98	85.20	75.62	89.97	64.38	74.18	91.40	26.96	74.73	82.59	68.14	68.98
		DPE [65]	46.97	99.10	27.60	84.15	85.24	68.52	89.92	31.30	69.90	79.18	75.06	62.34
		Open-IRT	**99.73**	**1.28**	**84.55**	**96.52**	**18.34**	**80.62**	**97.07**	**10.09**	**80.95**	82.65	**61.69**	69.20
	MAPLE	ZS-Eval [52]	98.16	5.50	82.43	98.35	**7.82**	83.58	90.86	27.53	76.01	86.15	52.00	71.68
		TPT [12]	98.16	69.35	81.74	98.34	7.88	82.67	90.86	27.55	75.40	86.15	52.10	70.84
		TPT-C [12]	98.22	5.15	83.34	98.35	7.85	83.55	90.91	27.44	75.84	86.20	51.96	71.60
		PAlign [46]	98.16	5.62	82.57	98.34	7.88	83.44	90.86	27.55	76.03	86.15	52.10	71.50
		PAlign-C [46]	98.61	3.45	83.91	**98.35**	8.13	83.45	91.17	26.95	76.12	86.53	50.64	71.11
		ROSITA [52]	99.45	3.84	87.71	98.02	11.45	84.56	91.76	25.23	77.60	86.92	48.12	72.79
		OWTTT [66]	98.34	9.63	86.52	71.01	78.78	68.70	71.20	85.81	68.29	62.35	88.44	61.89
		TDA [61]	98.42	4.13	81.97	98.60	6.20	83.95	91.27	27.00	76.84	86.72	51.40	72.61
		UniEnt [58]	98.17	5.49	82.64	98.35	7.85	83.65	90.90	27.41	76.08	86.16	51.91	71.72
		DPE [65]	83.82	92.73	55.52	97.42	12.95	79.41	89.10	31.13	74.32	73.57	73.67	53.64
		Open-IRT	**99.51**	**2.85**	**88.11**	97.62	15.92	**85.01**	**91.83**	**24.38**	**77.80**	**87.42**	**46.40**	**73.20**

Table 1: Open-set Single-Image Test-time adaptation results with CIFAR-10C as ID, MNIST, SVHN, Tiny-ImageNet, and CIFAR-100C as OOD. The metrics include AUROC (AUC), FPR95 (FPR), and Acc_{HM} (HM) as defined in Section 4.1. Results in bold represent the best performance, while underlined results indicate the second-best ones.

4.2 Main Result

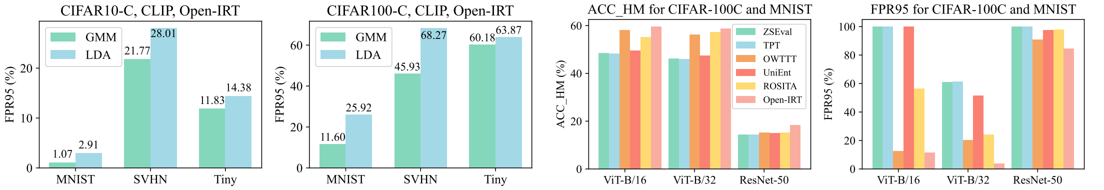
Cifar Benchmark. The cifar benchmark leverages CIFAR-10C/100C [19] as ID datasets, and MNIST [22], SVHN [24], and Tiny-ImageNet [20] as OOD datasets. CIFAR-100C/10C is also treated as an OOD case with small distribution shifts, applying label offsets to distinguish from ID data. As shown in Table 1, Open-IRT has demonstrated significant performance advantages. For example, in the CIFAR-10C → SVHN case with CLIP, Open-IRT achieved notable improvements by +1.58% AUROC, -12.88% FPR95, and +1.50% Acc_{HM}. Additionally, for the CIFAR-100C → SVHN case with CLIP, Open-IRT achieves an +8.45% in Acc_{HM}, -10.80% in FPR95, and +11.04% in AUROC, as detailed in appendix. OpenIRT also achieves better results in MaPLe, a multi-modal fine-tuning framework for CLIP, due to its adaptability across fine-tuning methods through direct feature space operation. Furthermore, Open-IRT exhibits obvious improvements in the FPR95 (up to -12.88%), highlighting its effectiveness in reducing the misclassification of OOD samples. This can be attributed to Open-IRT's ability to model the ID-OOD distribution effectively (see Fig. 3).

Method	CLIP						MAPLE					
	IN-C→MNIST			IN-C→SVHN			IN-C→MNIST			IN-C→SVHN		
	AUC ↑	FPR ↓	HM ↑	AUC ↑	FPR ↓	HM ↑	AUC ↑	FPR ↓	HM ↑	AUC ↑	FPR ↓	HM ↑
ZS-Eval [52]	93.34	57.34	41.41	85.72	74.34	40.84	81.24	93.97	41.29	83.05	73.63	42.44
TPT [12]	91.89	59.54	41.02	85.03	48.98	40.16	80.31	93.54	39.13	82.67	73.57	39.90
TPT-C [12]	57.84	98.92	6.37	10.31	99.59	7.29	82.88	87.95	41.13	82.17	72.10	41.37
OWTTT [66]	95.76	10.43	42.95	87.75	26.23	38.50	**98.58**	**3.35**	48.69	77.17	39.74	38.10
TDA [61]	90.54	76.23	43.66	86.76	75.45	43.07	76.76	99.02	42.98	82.46	91.75	44.63
UniEnt [58]	94.19	46.98	41.53	87.56	67.03	41.10	81.53	93.45	41.50	83.41	70.84	42.78
DPE [65]	87.92	91.94	42.87	82.96	77.90	41.93	73.97	99.59	41.39	80.06	87.10	44.05
ROSITA [52]	98.97	8.55	45.74	91.90	45.66	38.86	97.19	9.56	48.28	91.86	29.21	44.47
Open-IRT	**99.44**	**1.06**	**49.49**	**98.45**	**9.37**	**48.19**	97.54	10.97	**50.07**	**95.54**	**23.11**	**48.77**

Table 2: Open-set Single-Image Test-time adaptation. The ID data is ImageNet-C, while the OOD data comprises MNIST and SVHN.

	$\mathcal{L}_{psd}$	$\mathcal{L}_{mid}$	CIFAR-10C→MNIST			CIFAR-10C→SVHN			VisDA→MNIST			VisDA→SVHN		
			AUC ↑	FPR ↓	HM ↑	AUC ↑	FPR ↓	HM ↑	AUC ↑	FPR ↓	HM ↑	AUC ↑	FPR ↓	HM ↑
CLIP	✗	✗	91.91	84.91	78.07	89.94	64.03	75.14	93.55	65.83	84.34	90.45	64.98	79.02
	✗	✓	99.60	1.90	78.91	94.58	32.42	75.46	99.62	2.51	89.73	99.10	5.35	89.29
	✓	✗	98.15	1.74	**85.52**	93.52	37.04	78.57	95.66	35.40	87.74	95.04	33.70	85.67
	✓	$\lambda = 0$	99.70	1.90	83.52	95.92	24.29	79.24	99.79	1.53	89.83	97.58	4.26	89.81
	✓	✓	**99.71**	**1.28**	84.52	**96.52**	**18.35**	**80.61**	**99.84**	**1.27**	**90.85**	**99.20**	**3.05**	**90.39**

Table 4: Ablation Experiments. The ID data is a combination of CIFAR-10C and VisDA, while the OOD data comprises MNIST and SVHN.



(a) Analyze of Open-IRT with GMM and LDA.

(b) Backbone ViT-B/16, ViT-B/32 and ResNet50.

Figure 5: Comparison of GMM and LDA in PPF and different backbone architectures.

VisDA and ImageNet Benchmarks. The VisDA [21] and ImageNet-C [19] serve as ID datasets, while MNIST [22], MNIST-M [23], and SVHN [24] serve as OOD datasets. The complexity of VisDA and ImageNet-C introduces additional challenges for OOD detection. In Table 2, Open-IRT achieves notable improvements over existing baselines. For example, in ImageNet-C→MNIST experiment with CLIP, Open-IRT achieves gains of +0.47% AUROC, -7.49% FPR95, and +3.75% Acc_{HM}. Similarly, in Table 3, Open-IRT achieves gains of +1.58% AUROC, -7.46% FPR95, and +1.09% Acc_{HM} in VisDA → MNIST-M with CLIP. Moreover, compared to Table 1, when the ID dataset becomes more complex (e.g., ImageNet-C), the model's performance in OOD detection tasks does not fully saturate, resulting in a relatively lower Acc_{HM}. In such cases, Open-IRT's improvement in Acc_{HM} is more obvious, such as a +5.12% increase in the ImageNet-C → SVHN experiment with CLIP. Additionally, Open-IRT achieves gains of -16.86% FPR95 and +6.55% AUROC. Since Open-IRT operates primarily in the feature space and enhances the separation of the ID-OOD boundary (see Fig. 4), its effectiveness becomes increasingly apparent as the feature distribution grows more complex.

Method		VisDA→MNIST			VisDA→MNIST-M		
		AUC ↑	FPR ↓	HM ↑	AUC ↑	FPR ↓	HM ↑
CLIP	ZS-Eval [52]	93.55	65.86	78.30	87.25	67.10	74.84
	TPT [12]	93.55	66.11	78.44	87.25	67.19	75.05
	TPT-C [12]	81.81	85.12	75.09	87.44	62.31	77.32
	ROSITA [52]	99.63	2.99	90.59	97.10	15.14	86.88
	Open-IRT	**99.85**	**1.27**	**90.88**	**98.68**	**7.68**	**87.97**
MAPLE	ZS-Eval [52]	93.07	66.13	80.29	92.31	45.66	78.83
	TPT [12]	93.07	66.03	80.35	92.30	45.70	78.87
	TPT-C [12]	93.40	59.35	80.35	92.48	44.17	78.93
	PAlign [46]	93.07	66.03	80.62	92.29	45.70	79.17
	PAlign-C [46]	95.61	27.65	81.93	94.13	32.97	81.48
	ROSITA [52]	99.80	1.40	**90.84**	98.90	5.79	89.40
	Open-IRT	**99.87**	**1.01**	90.82	**99.15**	**4.99**	**89.56**

Table 3: VisDA (ID), MNIST/MNIST-M (OOD) Results.

4.3 Experiment Analysis

Score Ablation in PPF. We assess the effectiveness of our PPF mechanism in Fig. 5a, comparing PPF with GMM and LDA [67]. The results indicate that the PPF with GMM performs better results, such as -2.55% FPR95 improvement in CIFAR-10C → Tiny-ImageNet task. This performance gain can be attributed to GMM's ability to model complex and non-linearly separable feature distributions, which more accurately capture the characteristics of real-world domain shifts. In contrast, LDA relies on the assumption of linear decision boundaries, which limits its capacity to handle such complexities.

Loss Ablation in IDT. We analyze the effectiveness of each component of IDT in Eq.14, as shown in Table 4, with the first row representing the zero-shot evaluation scenario. The results demonstrate that $\mathcal{L}_{mid}$ significantly improves FPR95 and AUROC, enhancing ID-OOD binary classification by increasing the separation between ID and OOD samples in feature space. On complex datasets like VisDA, $\mathcal{L}_{mid}$ outperforms $\mathcal{L}_{psd}$ by +1.99% and +3.62% in Acc_{HM}, respectively. However, on simpler datasets like CIFAR-10C, $\mathcal{L}psd$ performs better in Acc_{HM} as it helps reduce intra-class noise, which is more beneficial for simpler features. Furthermore, setting $\lambda = 0$ to evaluate the intermediate-domain strategy shows that introducing the intermediate-domain improves performance by up to -5.94% in FPR95, supporting our Hypothesis 2.

Analyze Backbones. In Fig. 5b, we utilize ViT-B/16, ViT-B/32 and ResNet50 as the backbones in CIFAR-100C → MNIST task with CLIP. The results demonstrate that Open-IRT achieves the highest Acc_{HM} (e.g., 59.56% in ViT-B/16, 58.62% in ViT-B/32) and the lowest FPR95 (e.g., 11.29% in ViT-B/16, 3.74% in ViT-B/32). These results highlight the robustness of Open-IRT, as it directly optimizes the feature space, independent of the backbone type, demonstrating its broad applicability.

Varying OOD Ratios. To examine the robustness of Open-IRT, we conduct CIFAR-10C → MNIST experiments on CLIP under different OOD ratios with additional experiments are in appendix. As shown in Table 5, Open-IRT consistently achieves better results across all OOD ratios with Acc_{HM} fluctuate by only 1.01%. This stability suggests that Open-IRT is better suited to handle the uncertainties associated with fluctuating OOD proportions.

	ratio	0.2	0.4	0.6	0.8	1.0
CLIP	ZS-Eval [52]	75.55	75.60	75.59	75.59	75.60
	TPT [12]	75.77	75.78	75.81	75.76	75.78
	TPT-C [12]	73.05	74.29	74.75	75.05	75.05
	DPE [65]	65.67	66.12	56.38	29.98	27.60
	OWTTT [66]	62.31	68.85	81.70	82.90	83.27
	TDA [61]	72.45	75.04	77.54	77.91	77.06
	ROSITA [52]	83.21	84.68	83.90	83.89	83.95
	Open-IRT	**85.10**	**85.53**	**85.08**	**84.63**	**84.52**

Table 5: Experiments (Acc_{HM}) on CLIP with different OOD ratios with CIFAR-10C (ID), and MNIST (OOD).

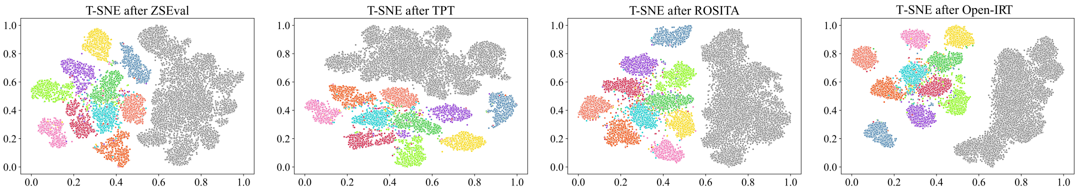


Figure 6: T-SNE Visualization with gray points as OOD.

T-SNE Visualization. We conduct T-SNE experiments of CIFAR-10C → MNIST. The T-SNE visualization in Fig. 6 reveals that Open-IRT achieves better separation and more compact clustering within the ID classes, while also establishing a clearer boundary between ID and OOD classes.

5 Conclusion

In this paper, we address the challenges of real-time adaptation in open-set environments with single-sample stream and propose the *Open-set Intermediate-Representation-based Test-time adaptation* (Open-IRT) framework. Its Polarity-aware Prompt-based OOD Filter module leverages the rich cross-modal information of vision language models, corporating both absolute semantic alignment and relative semantic polarity. The Intermediate Domain-based Test-time adaptation module constructs an intermediate domain and indirectly decomposes and enlarge the ID-OOD distributional discrepancy in real-time. Experiments across various benchmarks underscores the Open-IRT's potential to enhance the robustness and adaptability in dynamic real-world. Future work can focus on extending its application to real-world object detection or semantic segmentation. Moreover, a tension exists between the method's requirement for a sizable memory bank and its premise of single-image test-time adaptation. Future work should aim to reconcile this, achieving robust performance under a strictly single-image setting.

Acknowledgements: This work was supported by the National Key Research and Development Program of China (No. 2024YFE0211000), in part by the National Natural Science Foundation of China (No. 62372329), in part by the Shanghai Scientific Innovation Foundation (No. 23DZ1203400), in part by the China Postdoctoral Science Foundation (No. BX20250383, GZB20250385, 2025M771530, 2025M771539), in part by Tongji-Qomolo Autonomous Driving Commercial Vehicle Joint Lab Project, and in part by Xiaomi Young Talents Program.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[2] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 1

[4] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *CVPR*, pages 5830–5840, 2021. 1

[5] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE TPAMI*, 43(10):3614–3631, 2020.

[6] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE TPAMI*, 35(7):1757–1772, 2012. 1

[7] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *IJCV*, pages 1–28, 2024. 1, 2, 3

[8] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021. 3

[9] Ev Zisselman and Aviv Tamar. Deep residual flow for out of distribution detection. In *CVPR*, pages 13994–14003, 2020. 2, 3

[10] 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. 2, 3, 7

[11] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023. 2, 3, 7

[12] 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. *NeurIPS*, 35:14274–14289, 2022. 2, 3, 7, 8, 9, 10

[13] Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. *NeurIPS*, 34:2427–2440, 2021. 2, 3

[14] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *NeurIPS*, 33:11539–11551, 2020. 3

[15] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 3

[16] Chentao Cao, Zhun Zhong, Zhanke Zhou, Tongliang Liu, Yang Liu, Kun Zhang, and Bo Han. Noisy test-time adaptation in vision-language models. *ICLR*, 2025. 2

[17] Haoyu Dong, Nicholas Konz, Hanxue Gu, and Maciej A Mazurowski. Medical image segmentation with intent: Integrated entropy weighting for single image test-time adaptation. In *CVPR*, pages 5046–5055, 2024. 2

[18] Klara Janouskova, Tamir Shor, Chaim Baskin, and Jiri Matas. Single image test-time adaptation for segmentation. *arXiv preprint arXiv:2309.14052*, 2023. 2

[19] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 3, 5, 6, 7, 8, 9

[20] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 3, 7, 8

[21] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. 3, 7, 9

[22] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 3, 5, 6, 7, 8, 9

[23] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(59):1–35, 2016. 7, 9

[24] 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*, volume 2011, page 4, 2011. 3, 7, 8, 9

[25] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *NeurIPS*, 33:21464–21475, 2020. 3

[26] Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don't know? *NeurIPS*, 34:29074–29087, 2021. 3

[27] Alireza Zaeemzadeh, Niccolo Bisagno, Zeno Sambugaro, Nicola Conci, Nazanin Rahnavard, and Mubarak Shah. Out-of-distribution detection using union of 1-dimensional subspaces. In *CVPR*, pages 9452–9461, 2021. 3

[28] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 3

[29] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *NeurIPS*, 31, 2018. 3

[30] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *NeurIPS*, 34:144–157, 2021. 3

[31] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *ECCV*, pages 691–708, 2022. 3

[32] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE TPAMI*, 43(11):3964–3979, 2020. 3

[33] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *NeurIPS*, 32, 2019. 3

[34] Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. *NeurIPS*, 33:20685–20696, 2020. 3

[35] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *IJCV*, 131(10): 2607–2622, 2023. 3

[36] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 3

[37] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 3

[38] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *IJCV*, 132(2):581–595, 2024. 3

[39] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *ECCV*, pages 493–510, 2022. 3

[40] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[41] Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *CVPR*, pages 1593–1603, 2024. 3

[42] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *NeurIPS*, 34:7068–7081, 2021. 3

[43] Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Yueqian Lin, Qing Yu, Go Irie, Shafiq Joty, Yixuan Li, Hai Li, et al. Generalized out-of-distribution detection and beyond in vision language model era: A survey. *arXiv preprint arXiv:2407.21794*, 2024. 3

[44] Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided ood detection with pretrained vision-language models. *arXiv preprint arXiv:2403.20078*, 2024. 3

[45] Chentao Cao, Zhun Zhong, Zhanke Zhou, Yang Liu, Tongliang Liu, and Bo Han. Envisioning outlier exposure by large language models for out-of-distribution detection. *arXiv preprint arXiv:2406.00806*, 2024. 3

[46] Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. *NeurIPS*, 36, 2024. 3, 7, 8, 9

[47] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *NeurIPS*, 35:35087–35102, 2022. 3

[48] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *ICCV*, pages 1802–1812, 2023. 3, 4

[49] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. *NeurIPS*, 36, 2024. 3

[50] Yichen Bai, Zongbo Han, Bing Cao, Xiaoheng Jiang, Qinghua Hu, and Changqing Zhang. Id-like prompt learning for few-shot out-of-distribution detection. In *CVPR*, pages 17480–17489, 2024. 3

[51] Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. Learning transferable negative prompts for out-of-distribution detection. In *CVPR*, pages 17584–17594, 2024. 3

[52] Manogna Sreenivas and Soma Biswas. Effectiveness of vision language models for open-world single image test time adaptation. *arXiv preprint arXiv:2406.00481*, 2024. 3, 7, 8, 9, 10

[53] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015. 3

[54] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[55] Boyang Peng, Sanqing Qu, Yong Wu, Tianpei Zou, Lianghua He, Alois Knoll, Guang Chen, and Changjun Jiang. Map: Mask-pruning for source-free model intellectual property protection. In *CVPR*, pages 23585–23594, 2024. 3

[56] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, pages 9229–9248, 2020. 3

[57] Ke Fan, Tong Liu, Xingyu Qiu, Yikai Wang, Lian Huai, Zeyu Shangguan, Shuang Gou, Fengjian Liu, Yuqian Fu, Yanwei Fu, et al. Test-time linear out-of-distribution detection. In *CVPR*, pages 23752–23761, 2024. 3

[58] Zhengqing Gao, Xu-Yao Zhang, and Cheng-Lin Liu. Unified entropy optimization for open-set test-time adaptation. In *CVPR*, pages 23975–23984, 2024. 3, 7, 8

[59] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *NeurIPS*, 35:38629–38642, 2022. 3

[60] 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. 3

[61] 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. 3, 7, 8, 10

[62] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. *ICCV*, 2017. 6

[63] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 6

[64] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7

[65] Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Dual prototype evolving for test-time generalization of vision-language models. *NeurIPS*, 37:32111–32136, 2025. 7, 8, 10

[66] 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. 7, 8, 10

[67] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2): 179–188, 1936. 9

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In both our abstract and introduction, we explicitly define the scope of our study as addressing open-world single-sample test-time domain adaptation. Our contribution lies in proposing a state-of-the-art strategy that effectively addresses the aforementioned challenges, supported by a solid theoretical foundation.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the appendix, future work will focus on extending our approach to real-world tasks such as object detection and semantic segmentation. These tasks present more significant distribution estimation challenges due to the shift from image-level understanding to fine-grained spatial localization and semantic parsing.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The theoretical framework and corresponding proofs are comprehensively elaborated in the Methodology section (Section 3) of the main text, with supplementary derivations and extended demonstrations provided in the Appendix section.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper elaborates on the technical details in the experimental section (Section 4), while supplementary materials provide additional experimental investigations along with fully reproducible code implementation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The source code has been archived in the supplementary materials, with all experimental datasets being publicly available benchmark datasets that are comprehensively documented in the main manuscript and supplementary materials, and are publicly accessible through their original repositories.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper meticulously documents experimental details in the Section 4, including hardware specifications (e.g., GPU models, memory capacity) and hyperparameter settings (learning rate, batch size, etc.). Additional analyses, such as robustness checks, are provided in the Supplementary Materials to ensure reproducibility and thorough validation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Our experimental results strictly follow the standardized benchmarking protocols to ensure fair and reproducible performance assessment. Furthermore, all reported results are averaged over multiple runs to account for randomness and enhance the statistical reliability of our findings.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: In the experimental section, we specify that our computational resources are based on NVIDIA A6000 GPUs and NVIDIA 3090 GPUs. For detailed configurations, please refer to Section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: Our submitted code fully conforms to the official conference standards, including strict adherence to the anonymity policy.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In the introduction (Section 1) and Methodology (Section 3) and appendix, we provide an in-depth discussion of the societal impact of our method, particularly emphasizing that the model's ability to perform real-time adaptation in open environments is crucial for addressing the emergence of unknown categories or distributional shifts. Such capability is essential for ensuring safety in high-stakes applications like autonomous driving.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: As stated in the experimental section (Section 4), all datasets used in our study are publicly available and legally obtained, and all models are based on open-source implementations. Therefore, there is no foreseeable risk of misuse associated with our work.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: As indicated in the experimental section (Section 4) and supplementary materials, our method is properly cited and fully complies with the authors' licensing terms.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The assets associated with our paper, including the code, are submitted collectively in a compressed archive.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Only used for improve writing.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.