

## A Theoretical Analysis

### A.1 Theoretical Analysis of Score Function

#### A.1.1 Formalization of Dual-Polarity Contrast

Let  $\mathcal{P}_I$  and  $\mathcal{P}_O$  denote the ID and OOD distributions respectively. For input  $x$  with visual feature  $f = g_v(x) \in \mathbb{R}^d$  after L2 normalization, we posit:

$$f \sim \begin{cases} \mathcal{P}_I & \text{if } x \in \text{ID data} \\ \mathcal{P}_O & \text{if } x \in \text{OOD data} \end{cases} \quad (1)$$

Define positive prompt embeddings  $p_c$  and negative counterparts  $p'_c$  through prompt engineering. Our core assumption is:

**Assumption 1** (Semantic Contrast Hypothesis). *For ID samples  $\forall c \in [C]$ :*

$$\mathbb{E}_{f \sim \mathcal{P}_I}[\text{sim}(f, p_c)] \geq \tau_I \quad (2)$$

$$\mathbb{E}_{f \sim \mathcal{P}_I}[\delta_c(f)] \geq \Delta_I \quad (3)$$

*For OOD samples:*

$$\mathbb{E}_{f \sim \mathcal{P}_O}[\text{sim}(f, p_c)] \leq \tau_O \quad (4)$$

$$\mathbb{E}_{f \sim \mathcal{P}_O}[\delta_c(f)] \leq \Delta_O \quad (5)$$

where  $\delta_c(f) := |\text{sim}(f, p_c) - \text{sim}(f, p'_c)|$  and thresholds satisfy  $\tau_I > \tau_O$  and  $\Delta_I > \Delta_O$ .

**Lemma 1** (Expectation of Maximum). *For any random variables  $\{X_c\}_{c=1}^C$ :*

$$\mathbb{E}[\max_c X_c] \geq \max_c \mathbb{E}[X_c] \quad (6)$$

#### A.1.2 Scoring Function Derivation

The composite score per class is:

$$s_c(f) = \text{sim}(f, p_c) + \alpha \delta_c(f) \quad (7)$$

**Proposition 1** (Score Separability). *Under Assumption 1, there exists  $\alpha > 0$  such that:*

$$\inf_{f \sim \mathcal{P}_I} \mathbb{E}[\max_c s_c(f)] \geq \tau_I + \alpha \Delta_I \quad (8)$$

$$\sup_{f \sim \mathcal{P}_O} \mathbb{E}[\max_c s_c(f)] \leq \tau_O + \alpha \Delta_O \quad (9)$$

*establishing ID-OOD separability when  $\alpha > \frac{\tau_O - \tau_I}{\Delta_I - \Delta_O}$ .*

*proof* From linearity of expectation and Lemma 1:

$$\begin{aligned} \mathbb{E}[\max_c s_c(f)] &\geq \max_c \mathbb{E}[s_c(f)] \\ &= \max_c (\mathbb{E}[\text{sim}(f, p_c)] + \alpha \mathbb{E}[\delta_c(f)]) \\ &\geq \tau_I + \alpha \Delta_I \quad (\text{ID case}) \\ &\leq \tau_O + \alpha \Delta_O \quad (\text{OOD case}) \end{aligned}$$

The separation condition requires:

$$\tau_I + \alpha \Delta_I > \tau_O + \alpha \Delta_O$$

Solving for  $\alpha$  gives:

$$\alpha > \frac{\tau_O - \tau_I}{\Delta_I - \Delta_O} \quad (10)$$

Since  $\Delta_I > \Delta_O$  by Assumption 1, the denominator is positive. Given  $\tau_I > \tau_O$ , the right-hand side is negative, thus the inequality holds for all  $\alpha > 0$ .

### A.1.3 Gaussian Mixture Modeling

**Assumption 2** (Gaussian Mixture). *The composite scores  $s$  satisfy:*

$$s | \mathcal{P}_k \sim \mathcal{N}(\mu_k, \sigma_k^2), \quad k \in \{I, O\} \quad (11)$$

with  $\mu_I > \mu_O$  and  $\sigma_I^2 \neq \sigma_O^2$ .

#### A.1.4 EM Parameter Estimation

Given score set  $\{s_i\}_{i=1}^N$ , the EM algorithm iterates:

**E-step:** Compute responsibilities

$$\gamma_i^{(t)} = \frac{\pi^{(t)} \phi(s_i; \mu_I^{(t)}, (\sigma_I^{(t)})^2)}{\pi^{(t)} \phi(s_i; \mu_I^{(t)}, (\sigma_I^{(t)})^2) + (1 - \pi^{(t)}) \phi(s_i; \mu_O^{(t)}, (\sigma_O^{(t)})^2)} \quad (12)$$

**M-step:** Update parameters via MLE:

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^N \gamma_{ik}^{(t)} s_i}{\sum_{i=1}^N \gamma_{ik}^{(t)}}, \quad k \in \{I, O\} \quad (13)$$

$$\sigma_k^{2(t+1)} = \frac{\sum_{i=1}^N \gamma_{ik}^{(t)} (s_i - \mu_k^{(t+1)})^2}{\sum_{i=1}^N \gamma_{ik}^{(t)}} \quad (14)$$

$$\pi^{(t+1)} = \frac{1}{N} \sum_{i=1}^N \gamma_{iI}^{(t)} \quad (15)$$

## A.2 On the Efficacy of Polarity-Aware Prompts for OOD Detection

The Polarity-aware Prompt-based OOD Filter (PPF) module, central to our Open-IRT framework, employs a dual-prompting strategy to enhance the discrimination between ID and OOD samples. This strategy utilizes positive prompts, denoted as  $p_c = g_t(w_c^+)$  (e.g., derived from "a photo of a [CLS]"), and negative prompts,  $p'_c = g_t(w_c^-)$  (e.g., derived from "a photo of no [CLS]"), where  $g_t(\cdot)$  is the text encoder mapping prompt sentences  $w_c^+$  and  $w_c^-$  to embeddings in  $\mathcal{T} \subseteq \mathbb{R}^d$ . For a given L2 normalized visual feature  $f = g_v(x)$  derived from an image  $x$  via the visual encoder  $g_v(\cdot)$ , where  $f \in \mathcal{V} \subseteq \mathbb{R}^d$ , we compute cosine similarities  $\text{sim}(f, p_c)$  and  $\text{sim}(f, p'_c)$ . Our Semantic Contrast Hypothesis (Hypothesis ??) posits that these similarities, and particularly the "polarity gap"  $G(f, c) = |\text{sim}(f, p_c) - \text{sim}(f, p'_c)|$ , exhibit distinct statistical properties for ID and OOD samples. This appendix provides a supplementary rationale for this prompting mechanism.

We acknowledge the documented challenges VLMs face in robustly interpreting logical negation in a manner akin to human understanding. However, the efficacy of the negative prompts  $p'_c$  within our proposed framework is not solely contingent upon the VLM achieving perfect logical inference. Instead, our methodology leverages the differential representational capabilities of the VLM's text encoder. The lexical distinction between the positive prompt sentence  $w_c^+$  and its negative counterpart  $w_c^-$  naturally leads to distinct embeddings  $p_c$  and  $p'_c$  within the shared embedding space. Our framework is designed to exploit this induced representational disparity as a discriminative signal for OOD detection.

The core of our approach lies in the scoring function  $\mathcal{S}(f) = \phi \left( \sup_{c \in [C]} [\text{sim}(f, p_c) + \alpha G(f, c)] \right)$ , which captures not only the absolute alignment of  $f$  with  $p_c$  but also the relative contrast expressed by the polarity gap  $G(f, c)$ . For an ID sample  $x_{ID}$  whose visual features  $f_{ID}$  strongly correspond to class  $c$ , we expect a high  $\text{sim}(f_{ID}, p_c)$ . Concurrently,  $p'_c$  is anticipated to occupy a sufficiently different region in the embedding space such that  $\text{sim}(f_{ID}, p'_c)$  is significantly lower. This discrepancy results in a large  $G(f_{ID}, c)$ , and as per Hypothesis ??, we expect  $\mathbb{E}_{f \sim \mathcal{P}_I} [\text{sim}(f, p_c) - \text{sim}(f, p'_c)] \geq \Delta_I$ , where  $\Delta_I$  is substantial. The additive nature of the polarity term in  $\mathcal{S}(f)$  ensures that this large gap contributes positively, thereby amplifying the score for confident ID instances. Conversely, for an OOD sample  $x_{OOD}$  with features  $f_{OOD}$ , its alignment  $\text{sim}(f_{OOD}, p_c)$  with any ID class prototype  $p_c$  is inherently weaker. Furthermore, due to the "semantic ambiguity" of  $f_{OOD}$  relative to the

ID-specific prompts, the differential similarity  $|\text{sim}(f_{OOD}, p_c) - \text{sim}(f_{OOD}, p'_c)|$  is expected to be less pronounced. This aligns with Hypothesis ?? where  $\mathbb{E}_{f \sim \mathcal{P}_O}[\text{sim}(f, p_c) - \text{sim}(f, p'_c)] \leq \Delta_O$ , with  $\Delta_O < \Delta_I$ . A smaller  $G(f_{OOD}, c)$  leads to a comparatively suppressed overall score, aiding in the ID-OOD distinction.

It is important to contextualize the role of  $p'_c$ : it is not designed as a universal logical negator but rather as a task-specific contrastive anchor relative to  $p_c$ . Its utility lies in its ability to generate a differential VLM response that, when compared against the response to  $p_c$ , provides a useful heuristic for our OOD detection task. This use of negative prompts for creating contrast is supported by prior explorations in OOD detection, such as Wang et al. (2023) [16], suggesting the viability of such prompting schemes.

Ultimately, the pragmatic efficacy of this polarity-aware prompting strategy is substantiated by our empirical evaluations. As demonstrated in Fig. ?? and the ablation studies presented in Section ??, the incorporation of the polarity gap term consistently enhances ID-OOD score separation and improves overall OOD detection performance. This indicates that, within the operational context of Open-IRT, the proposed positive and negative prompts, and the way their similarities are combined, furnish an effective compound signal for distinguishing between in-distribution and out-of-distribution data.

### A.3 Theoretical Analysis of Intermediate Domain Hypothesis

We provide a theoretical analysis of Hypothesis ?? from the perspective of measure theory and reproducing kernel Hilbert space (RKHS). The proof consists of four key steps aligned with the original notation.

#### A.3.1 Intermediate Domain Characterization

Let  $\mathcal{F}_I \triangleq \{f | f \sim P_I\}$  and  $\mathcal{F}_O \triangleq \{f | f \sim P_O\}$  denote the ID/OOD feature distributions as per Hypothesis ?. We formally define the intermediate domain  $\mathcal{F}_M$  through a measurable transformation  $\mathcal{T}$  that induces:

$$\mathcal{F}_M = \mathcal{T}(\mathcal{F}_I \cup \mathcal{F}_O), \quad (16)$$

where  $\mathcal{T}$  minimizes the following objective:<sup>1</sup>

$$\min_{\mathcal{T}} [d_{\mathcal{H}}(\mathcal{F}_I, \mathcal{F}_M) + d_{\mathcal{H}}(\mathcal{F}_O, \mathcal{F}_M) - d_{\mathcal{H}}(\mathcal{F}_I, \mathcal{F}_O)]. \quad (17)$$

#### A.3.2 HSIC Distance Decomposition

For the HSIC distance  $d_{\mathcal{H}}$  defined in Hypothesis ??, under characteristic kernels (e.g., Gaussian RBF), we have:

$$d_{\mathcal{H}}(\mathcal{F}_I, \mathcal{F}_O) \leq d_{\mathcal{H}}(\mathcal{F}_I, \mathcal{F}_M) + d_{\mathcal{H}}(\mathcal{F}_M, \mathcal{F}_O) + \epsilon, \quad (18)$$

where  $\epsilon$  is a residual term controlled by  $\mathcal{T}$ 's complexity. This directly corresponds to the approximate equality in (??).

#### A.3.3 Existence of an Approximate Minimizer

We recall that  $\mathcal{T}$  is optimized to minimize the following objective:

$$\min_{\mathcal{T}} \mathcal{J}(\mathcal{T}) \triangleq d_{\mathcal{H}}(\mathcal{F}_I, \mathcal{F}_M) + d_{\mathcal{H}}(\mathcal{F}_O, \mathcal{F}_M) - d_{\mathcal{H}}(\mathcal{F}_I, \mathcal{F}_O).$$

If the function class of  $\mathcal{T}$  is sufficiently expressive, then there exists a transformation  $\mathcal{T}^*$  such that:

$$\mathcal{J}(\mathcal{T}^*) \approx 0,$$

which implies:

$$d_{\mathcal{H}}(\mathcal{F}_I, \mathcal{F}_M^*) + d_{\mathcal{H}}(\mathcal{F}_O, \mathcal{F}_M^*) \approx d_{\mathcal{H}}(\mathcal{F}_I, \mathcal{F}_O).$$

This validates the existence of a transformation yielding the approximate equality in Equation (??).

<sup>1</sup>For the geometric interpretation of this objective,  $d_{\mathcal{H}}$  is used as a direct measure of HSIC-based dissimilarity, not as a squared quantity. This allows for intuitive connections to distance-like properties, such as the triangle inequality, when analyzing the objective's behavior.

#### A.3.4 Empirical Validation

Fig. ?? validates two essential properties from Hypothesis ??:

- **Intermediate Property:**  $\mathcal{F}_M$  lies between  $\mathcal{F}_I$  and  $\mathcal{F}_O$  in feature space
- **Distance Additivity:**  $d_{\mathcal{H}}(\mathcal{F}_I, \mathcal{F}_M) + d_{\mathcal{H}}(\mathcal{F}_O, \mathcal{F}_M) \approx d_{\mathcal{H}}(\mathcal{F}_I, \mathcal{F}_O)$

**Conclusion.** Hypothesis ?? holds when: 1)  $\mathcal{F}_M$  is constructed via measurable transformation  $\mathcal{T}$ , 2) The residual  $\epsilon$  is minimized through proper kernel selection.

## B Experiment Details

### B.1 Datasets

In this section, we introduce the details of the utilized datasets. For ID datasets, we employ CIFAR-10C/100C [5], ImageNet-C [5], and VisDA [12]. For OOD datasets, we exploit MNIST [9], MNIST-M [3], SVHN [11], Tiny-ImageNet [8], and CIFAR-10C/100C [5].

Datasets Components		Images Number		
ID	OOD	ID	OOD	total
CIFAR-10C	MNIST, SVHN, Tiny ImageNet, CIFAR-100C	10,000	10,000	20,000
CIFAR-100C	MNIST, SVHN, Tiny ImageNet, CIFAR-10C	10,000	10,000	20,000
ImageNet-C	MNIST, SVHN	10,000	10,000	20,000
VisDA	MNIST, SVHN, MNISTM	50,000	50,000	100,000

Table 1: Details of ID and OOD dataset combinations.

**CIFAR-10C dataset** [5] is an extension of the CIFAR-10 dataset, designed to test the robustness of image classification models against various corruptions. It consists of 10 categories (airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks), with 15 types of corruptions. It comprises 10,000 images for each type of corruption on average.

**CIFAR-100C dataset** [5] is an extension of the CIFAR-100 dataset, designed to test the robustness of image classification models against various corruptions. It consists of 100 categories, with 15 types of corruptions. It comprises 10,000 images for each type of corruption.

**ImageNet-C dataset** [5] is a large-scale corruption dataset encompassing 1,000 categories, comprising a total of 50,000 images. From these 50,000 images, 15 types of corruption are synthesized.

**VisDA dataset** [12] is a large-scale synthetic-to-real dataset designed to evaluate the performance of image classification models. It comprises 152,397 synthetic training images and 55,388 real testing images, spanning 12 distinct categories. This dataset is instrumental in assessing the robustness and adaptability of models when transitioning from synthetic to real-world data.

**MNIST dataset** [9] is a widely utilized benchmark in the field of machine learning and computer vision. It consists of 70,000 images of handwritten digits, divided into 60,000 training images and 10,000 testing images. This dataset serves as a fundamental resource for evaluating and comparing the performance of various image classification algorithms.

**SVHN dataset** [11], named Street View House Numbers (SVHN) dataset, is a widely used benchmark in the field of machine learning and computer vision. It consists of images of house numbers captured from real-world street scenes. The dataset includes 50,000 training images and 10,000 testing images, making it a valuable resource for evaluating the performance of digit recognition algorithms in realistic settings.

**MNIST-M dataset** [3] is created by blending MNIST digits with patches randomly extracted from color photos of the BSDS500 dataset as their background. It contains 59,001 training images and 90,001 test images. This dataset is designed to evaluate the robustness of image classification models under domain shifts, which makes it a valuable resource for research in the field of domain adaptation and transfer learning.

**Tiny-ImageNet dataset** [8] is a subset of the ImageNet dataset, contains 100,000 images across 200 classes, with each class having 500 images. These images are downsized to 64×64 colored pixels. Each class is divided into 500 training images, 50 validation images, and 50 test images. It is widely used for benchmarking machine learning algorithms, particularly in the fields of image classification due to its manageable size and diverse categories.

### B.2 Metrics

We utilize the AUROC (Area Under the Receiver Operating Characteristic Curve), FPR95 (False Positive Rate at 95% True Positive Rate), and  $Acc_{HM}$  as main metrics. While  $Acc_{HM}$  denotes the harmonic mean of  $Acc_I$  and  $Acc_O$ , in this study,  $Acc_O$  refers to the model’s binary classification accuracy for determining whether a sample is ID ( $\hat{y}_O = 1$  or OOD ( $\hat{y}_O = 0$ ), and  $Acc_I$  indicates the overall accuracy in correctly identifying each category of ID samples.

---

**Algorithm 1** Open-IRT Algorithm

---

**Require:** The target dataset  $\mathcal{X}_t$ , CLIP Text Encoder  $g_t(x; \theta_t)$ , and CLIP Vision Encoder  $g_v(x; \theta_v)$ .

```
1: Initialize  $\theta_v, \theta_t$ , update  $\theta_v$  and fix  $\theta_t$ .
2: Given an input sample  $x \in \mathcal{X}_t$  with a batchsize of 1
3: for  $Step = 1, 2, \dots, M$  do
4:   # Out-of-Distribution Sample Filter-PPF
5:   Generate score  $S(x)$ , update score bank  $\mathcal{B}^s$ 
6:   Classify  $x$  to ID or OOD sample with  $\mathcal{B}^s$ 
7:   Update  $g_v(x)$  to  $\mathcal{B}_{I/O}^f$ , and subdivide  $\mathcal{B}^s$  to  $\mathcal{B}_{I/O}^s$ 
8:   # Test-Time Adaptation-IDT
9:   if  $x$  is ID sample then
10:    Compute loss  $\mathcal{L}_{psd} = \mathcal{L}_{CE}$ 
11:    Gain middle-feature  $f_m$  with  $\sigma_O, \mu_O$ 
12:    Compute loss  $\mathcal{L}_{mid} = \mathcal{L}_I$ 
13:   else
14:    Compute loss  $\mathcal{L}_{psd} = -\mathcal{L}_{CE}$ 
15:    Gain middle-feature  $f_m$  with  $\sigma_I, \mu_I$ 
16:    Compute loss  $\mathcal{L}_{mid} = \mathcal{L}_O$ 
17:   end if
18:   Update  $\theta_v$  with  $\mathcal{L}_{TTA}$ 
19: end for
20: return The finally evaluation metrics and  $\theta_v$ .
```

---

The Area Under the Receiver Operating Characteristic Curve (AUROC) is a fundamental metric used to evaluate the performance of binary classification models. The ROC curve itself is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The TPR, also known as sensitivity or recall, measures the proportion of actual positive samples that are correctly identified by the model, defined as:

$$TPR = \frac{TP}{TP + FN} \quad (19)$$

where  $TP$  represents true positives and  $FN$  denotes false negatives. Conversely, the FPR quantifies the proportion of actual negative samples that are incorrectly classified as positive, defined as:

$$FPR = \frac{FP}{FP + TN} \quad (20)$$

where  $FP$  refers to false positives and  $TN$  indicates true negatives. The AUROC metric summarizes the performance of the model across all possible classification thresholds, yielding a single scalar value between 0 and 1. An AUROC value of 0.5 suggests no discrimination capability, equivalent to random guessing, while a value of 1.0 indicates perfect classification performance. This metric is particularly advantageous as it remains invariant to the class distribution and provides a comprehensive assessment of the model's ability to distinguish between the positive and negative classes across varying conditions.

The False Positive Rate at 95% True Positive Rate (FPR95) is a crucial metric used to assess the performance of binary classifiers, particularly in scenarios where maintaining a high sensitivity is essential. FPR95 specifically quantifies the false positive rate (FPR) corresponding to a threshold that achieves a true positive rate (TPR) of 95%.

To compute FPR95, the model's prediction thresholds are adjusted to find the point at which the TPR reaches 95%. The corresponding FPR at this threshold is then reported as FPR95. This metric is particularly informative in applications where the cost of false positives is significant, as it enables practitioners to evaluate the trade-off between sensitivity and specificity while ensuring a high level of true positive identification. A lower FPR95 value indicates a more effective model in maintaining a low rate of false positives at high sensitivity levels, making it a vital measure in fields such as medical diagnosis and fraud detection.

$Acc_{HM}$  represents the harmonic mean of  $Acc_I$  and  $Acc_O$ , it is important to clarify that  $Acc_O$  specifically denotes the model’s binary classification accuracy in determining whether a given sample is ID ( $\hat{y}_O = 1$ ) or OOD ( $\hat{y}_O = 0$ ). Conversely,  $Acc_I$  indicates the overall accuracy of the model in correctly identifying each category of ID samples.  $Acc_I$  and  $Acc_O$  is defined as:

$$Acc_I = \frac{\sum_{(x_i, y_i) \in \mathcal{P}_t} \mathbb{1}(y_i = \hat{y}_i) \cdot \mathbb{1}(y_i \in \mathcal{Y})}{\sum_{x_i, y_i \in \mathcal{P}_t} \mathbb{1}(y_i \in \mathcal{Y})} \quad (21)$$

$$Acc_O = \frac{\sum_{(x_i, y_i) \in \mathcal{P}_t} \mathbb{1}(\hat{y}_{i, OOD} = 0) \cdot \mathbb{1}(y_i \notin \mathcal{Y})}{\sum_{x_i, y_i \in \mathcal{P}_t} \mathbb{1}(y_i \notin \mathcal{Y})} \quad (22)$$

These definitions provide a comprehensive understanding of the model’s performance metrics in distinguishing between ID and OOD samples, essential for evaluating its efficacy in real-world applications.

### B.3 Baselines

Our experimental results are reproduced based on publicly available code. We leverage CLIP [13] and MaPLe [7] as backbones, and ZS-Eval [15], TPT/TPT-C [14], PAlign/PAlign-C [1], TDA [6], DPE [17], UniEnt [4], OWTTT [10] and ROSITA [15] as baselines. And the samples of ID/OOD datasets are listed in Table 1. While test-time adaptation baselines are based on LDA [2, 10] strategy to separate ID and OOD samples. We use memory banks for methods with batch processing strategy like UniEnt. Details are shown as follows. The parameters selection strategy is based on the conclusion of [15], which analyzed six parameter groups, including Prompts, LayerNorm, Prompts+LayerNorm, First Block, Last Block and Full Networks. The conclusion shows that with lower learning rates (e.g.,  $1e-6$  to  $1e-5$ ), most parameter update methods have relative stable performance. However, updating all network parameters results in the worst outcomes, indicating a significant loss of prior knowledge in the model. So we choose the parameters of the Vision Encoder with a relative lower learning rate is a stable choice.

**CLIP [13]:** Contrastive Language–Image Pre-training (CLIP) is a novel approach developed by OpenAI that leverages large-scale datasets of images and their corresponding textual descriptions to learn visual concepts from natural language supervision. The core idea behind CLIP is to train a model to predict which caption (from a set of randomly sampled captions) is the correct one for a given image, using a contrastive loss function. Mathematically, the contrastive loss can be represented as:

$$L = -\frac{1}{N} \sum_{i=1}^N \log \left( \frac{\exp(\text{sim}(x_i, y_i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(x_i, y_j)/\tau)} \right) \quad (23)$$

where  $\text{sim}(x_i, y_i)$  denotes the cosine similarity between the image embedding  $x_i$  and the text embedding  $y_i$ ,  $\tau$  is a temperature parameter that controls the sharpness of the distribution,  $N$  is the number of image-text pairs.

CLIP models are trained on a diverse range of internet images and their associated text, enabling them to generalize well to various downstream tasks without requiring task-specific fine-tuning. This capability is particularly useful for zero-shot learning, where the model can recognize and classify images it has never seen before based on textual descriptions alone. The architecture of CLIP typically involves two separate neural networks: one for processing images (e.g., a Vision Transformer or ResNet) and another for processing text. These networks are jointly trained to maximize the similarity between the image and text embeddings for matching pairs while minimizing it for non-matching pairs.

**MaPLe [7]:** MaPLe is an innovative multimodal prompt learner model designed to concurrently adapt both vision and text encoders while fine-tuning CLIP for various downstream applications. This can be regarded as another backbone of CLIP with different tuning strategy. This model employs learnable text prompts  $p_T$  and establishes a connection between the two modalities through visual prompts formulated as  $p_V = \text{Proj}(p_T)$ . Additionally, learnable tokens are integrated into the deeper layers of both image and text encoders, facilitating a progressive adaptation of features that enhances the performance. In line with the findings presented in [1], we incorporate MaPLe as an additional

vision-language model backbone to evaluate the efficacy of Open-IRT. In the following sections, we will review the baselines developed based on both CLIP and MaPLe specifically for zero-shot evaluation, highlighting their strengths and contributions to the field. This examination aims to provide a comprehensive understanding of how these models operate and their implications for multimodal learning tasks.

**ZS-Eval [15]:** Given a test image  $x_t$ , the image features  $f_t$  are extracted using the vision encoder, expressed as  $f_t = F_V(x_t)$ . In a classification task involving  $C$  potential categories, the classifier is constructed by concatenating a predefined text prompt  $p_T = \text{"A photo of a"}$  with each class label  $\{c_1, c_2, \dots, c_C\}$ . This results in class-specific text inputs  $\{p_T, c_i\}$  for  $i \in \{1, \dots, C\}$ . These text inputs are subsequently processed by the text encoder  $F_T$  to generate the text embeddings  $t_i = F_T(\{p_T, c_i\})$ , which serve as the class-specific text representations  $\{t_1, t_2, \dots, t_C\}$ . The classification decision is made by selecting the text embedding  $t_i$  that exhibits the highest similarity to the extracted image feature  $f_t$ .

**TPT [14]/TPT-C [15]:** TPT enhances CLIP’s zero-shot generalization by employing learnable text prompts  $p_T$  that are tailored for each image, in contrast to using fixed prompts. These prompts are concatenated with class names, thereby forming flexible text classifiers  $t_i = F_T(\{p_T, c_i\})$ . The learnable prompts are optimized through entropy minimization, represented as  $\arg \min_{p_T} L_{\text{ent}}$ , which effectively reduces uncertainty associated with augmented view scores, leading to improved performance. TPT-C further extends TPT by incorporating continuous updates, where the prompts  $\{p_T\}$  and  $\{p_V, p_T\}$  are iteratively refined to adapt to new test data, thereby ensuring robust and dynamic performance.

**PromptAlign [1]/PromptAlign-C [15]:** PromptAlign (PAlign) leverages the capabilities of the multimodal prompt learning model MaPLe [7] to enhance the adaptation process of both vision and language encoders for each specific test sample. This approach draws on previous advancements in test-time adaptation, focusing on bridging the disparity between the token distributions of the source and target domains. In this context, ImageNet is employed as a representative proxy for the source dataset associated with CLIP. The optimization of MaPLe’s vision and language prompts is achieved by minimizing the combined objective function  $\arg \min_{\{p_V, p_T\}} (\mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{align}})$  for each individual sample  $x_t$ . This methodology not only enhances the alignment of modalities but also fosters a more nuanced understanding of the contextual relationships between visual and textual information, thereby contributing to more robust and effective model performance across diverse applications. Furthermore, PromptAlign-C serves as a continuous learning version of PromptAlign, following the strategy outlined in TPT-C [15].

**OWTTT [10]:** This paper addresses the challenge of test-time adaptation in open-world scenarios where the target domain is contaminated with strong OOD data. It identifies the failure of existing TTT methods in distinguishing strong OOD samples from weak OOD ones, leading to performance degradation. To enhance the robustness of open-world test-time training (OWTTT), the authors propose an adaptive strong OOD pruning method to improve self-training efficacy. Additionally, they introduce a dynamic prototype expansion strategy to better separate weak and strong OOD samples. The model’s self-training is further regularized using distribution alignment to mitigate confirmation bias.

**TDA [6]:** Training-free Dynamic Adapter (TDA) is an efficient test-time adaptation method for vision-language models. Unlike existing test-time prompt tuning (TPT, DiffTPT) approaches that require computationally expensive backpropagation, TDA leverages a lightweight key-value cache that maintains a dynamic queue, where keys store test sample features and values store few-shot pseudo labels. By progressively refining pseudo labels and dynamically adapting, TDA effectively accommodates test data while maintaining high computational efficiency.

**DPE [17]:** Dual Prototype Evolving (DPE) is an advanced test-time adaptation method for vision-language models that effectively accumulates task-specific knowledge from both textual and visual modalities. Unlike prior methods that adapt from a single modality and fail to retain knowledge across test samples, DPE introduces two evolving sets of prototypes—one textual and one visual—to progressively capture more accurate multi-modal representations of target classes. To ensure consistency between these prototypes, DPE optimizes learnable residuals for each test sample, aligning textual and visual embeddings. This approach not only improves adaptation over time but also enhances zero-shot generalization without requiring backpropagation through CLIP’s textual encoder.



Method			MNIST			SVHN			Tiny-ImageNet			CIFAR-100C/10-C		
			AUC ↑	FPR ↓	HM ↑	AUC ↑	FPR ↓	HM ↑	AUC ↑	FPR ↓	HM ↑	AUC ↑	FPR ↓	HM ↑
CIFAR-10C	CLIP	ZS-Eval [15]	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 [14]	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 [14]	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 [15]	<u>99.43</u>	<u>3.25</u>	<u>83.95</u>	<u>94.94</u>	<u>31.22</u>	<u>79.12</u>	<u>96.37</u>	<u>12.69</u>	<u>80.07</u>	<u>83.01</u>	<u>64.54</u>	<u>69.64</u>
		OWTTT [10]	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 [6]	92.94	71.11	77.06	92.02	52.68	76.64	91.68	25.37	75.94	<b>83.54</b>	66.06	<b>70.13</b>
		UniEnt [4]	91.98	85.2	75.62	89.97	64.38	74.18	91.40	26.96	74.73	82.59	68.14	68.98
		DPE [17]	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	<b>99.73</b>	<b>1.28</b>	<b>84.55</b>	<b>96.52</b>	<b>18.34</b>	<b>80.62</b>	<b>97.07</b>	<b>10.09</b>	<b>80.95</b>	82.65	<b>61.69</b>	69.20
	MAPLE	ZS-Eval [15]	98.16	5.50	82.43	98.35	<b>7.82</b>	83.58	90.86	27.53	76.01	86.15	52.00	71.68
		TPT [14]	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 [14]	98.22	5.15	83.34	98.35	<u>7.85</u>	83.55	90.91	27.44	75.84	86.20	51.96	71.60
		PAlign [1]	98.16	5.62	82.57	<u>98.34</u>	7.88	83.44	90.86	27.55	76.03	86.15	52.10	71.50
		PAlign-C [1]	98.61	<u>3.45</u>	83.91	<b>98.35</b>	8.13	83.45	91.17	26.95	76.12	86.53	50.64	71.11
		ROSITA [15]	<u>99.45</u>	3.84	<u>87.71</u>	98.02	11.45	<u>84.56</u>	<u>91.76</u>	<u>25.23</u>	<u>77.60</u>	<u>86.92</u>	<u>48.12</u>	<u>72.79</u>
		OWTTT [10]	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 [6]	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 [4]	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 [17]	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	<b>99.51</b>	<b>2.85</b>	<b>88.11</b>	97.62	15.92	<b>85.01</b>	<b>91.83</b>	<b>24.38</b>	<b>77.80</b>	<b>87.42</b>	<b>46.40</b>	<b>73.20</b>
CLIP	ZS-Eval [15]	77.77	99.93	48.40	64.69	98.64	45.81	67.31	73.83	46.00	63.28	93.22	44.06	
	TPT [14]	77.77	99.93	48.31	64.70	98.62	45.81	67.28	73.80	45.97	63.27	93.19	44.04	
	TPT-C [14]	50.96	99.95	25.69	10.45	99.95	6.16	60.41	82.09	17.70	55.26	<b>86.37</b>	14.19	
	ROSITA [15]	86.41	56.22	55.15	<u>82.24</u>	69.07	<u>47.67</u>	<u>83.97</u>	<u>50.86</u>	<u>55.84</u>	<b>68.22</b>	<u>89.52</u>	<b>47.71</b>	
	OWTTT [10]	<u>96.89</u>	<u>12.15</u>	<u>59.72</u>	75.24	<u>51.64</u>	43.73	41.84	99.61	31.83	54.02	93.93	32.00	
	TDA [6]	80.33	99.57	46.52	71.77	96.11	46.01	70.70	69.63	47.52	<u>66.07</u>	91.90	45.79	
	UniEnt [4]	77.94	99.93	48.32	64.78	98.61	45.84	67.40	73.77	45.83	63.28	93.18	44.04	
	DPE [17]	67.06	99.88	42.54	43.23	99.79	35.69	61.42	80.62	42.80	60.08	92.80	42.21	
	Open-IRT	<b>98.06</b>	<b>12.82</b>	<b>62.01</b>	<b>93.28</b>	<b>40.84</b>	<b>56.12</b>	<b>86.16</b>	<b>42.24</b>	<b>56.78</b>	65.12	92.11	<u>46.70</u>	
MAPLE	ZS-Eval [15]	87.43	64.23	55.23	92.98	40.60	56.41	68.80	74.37	48.26	66.92	88.00	46.30	
	TPT [14]	87.44	64.06	53.33	92.97	40.49	54.61	68.81	74.20	47.02	66.93	87.93	44.83	
	TPT-C [14]	87.66	63.05	55.44	93.10	39.87	56.34	68.98	73.35	48.12	67.04	<u>87.48</u>	44.71	
	PAlign [1]	87.46	63.55	54.80	92.98	40.59	56.32	68.84	73.81	48.23	66.95	87.93	46.14	
	PAlign-C [1]	87.51	63.54	55.47	92.93	41.06	56.24	68.88	73.37	47.96	67.00	87.96	46.02	
	ROSITA [15]	89.26	46.87	61.65	93.33	38.22	<u>60.18</u>	69.44	72.71	47.88	67.55	87.63	45.92	
	OWTTT [10]	<u>96.49</u>	<b>9.42</b>	<b>62.97</b>	65.73	78.63	32.60	42.94	99.95	27.52	53.48	94.26	34.70	
	TDA [6]	89.82	52.24	55.46	<u>95.04</u>	<u>30.76</u>	59.51	<u>72.05</u>	<u>71.83</u>	<u>49.19</u>	<u>69.12</u>	<u>87.36</u>	<b>49.06</b>	
	UniEnt [4]	87.40	64.02	54.86	92.99	40.36	56.42	68.84	74.26	48.41	66.93	87.96	46.09	
	DPE [17]	39.05	98.88	33.66	84.29	76.13	52.20	63.74	82.75	45.74	65.61	90.67	46.36	
	Open-IRT	<b>95.27</b>	<u>16.54</u>	<u>61.82</u>	<b>98.29</b>	<b>10.74</b>	<b>62.66</b>	<b>72.51</b>	<b>72.63</b>	<b>49.90</b>	<b>70.22</b>	<b>81.18</b>	<u>48.44</u>	

Table 2: Open-set Single-Image Test Time Adaptation results with CIFAR-10C and CIFAR-100C as ID datasets. We leverage MNIST, SVHN, Tiny-ImageNet, and CIFAR-100C/CIFAR-10C as OOD datasets. The metrics include AUC, FPR, and HM, which mean AUROC, FPR95, and  $Acc_{HM}$ , respectively, as defined in Section ?? . Notably, while higher AUROC and  $Acc_{HM}$  values indicate better performance, a lower FPR95 is considered better. Results in bold represent the best performance, while underlined results indicate the second-best ones.

**UniEnt [4]:** UniEnt is a unified entropy optimization framework designed for open-set test-time adaptation, aiming to simultaneously adapt to covariate-shifted in-distribution (csID) data and detect covariate-shifted out-of-distribution (csOOD) data. Existing test-time adaptation methods primarily focus on covariate shift while neglecting the impact of semantic shift, leading to performance degradation and inaccurate confidence estimation. UniEnt addresses these issues by introducing a distribution-aware filter to preliminarily distinguish between csID and csOOD samples. It then applies entropy minimization on csID samples to enhance classification performance for known classes and entropy maximization on csOOD samples to improve the detection of unknown classes. Furthermore, UniEnt+ introduces a sample-level confidence-weighted strategy to mitigate errors caused by noisy data partitioning.

## C Additional Experiments

### C.1 Visualization of OOD scores in ablation study

We present the visualization in Fig. 1 from the ablation study conducted in the experiments of CIFAR10C-MNIST with CLIP in section of the main paper. This ablation study analyzes the effectiveness of various loss components. In the methods section, the total loss is defined as  $\mathcal{L}_{TTA} = \mathcal{L}_{psd} + \mathcal{L}_{mid}$ . We can intuitively observe from Fig. 1a to Fig. 1d that as the model components become increasingly complete, the model’s ability to distinguish between ID and OOD data becomes progressively more pronounced. This enables us to better utilize GMM to fit the two distinct distributions.

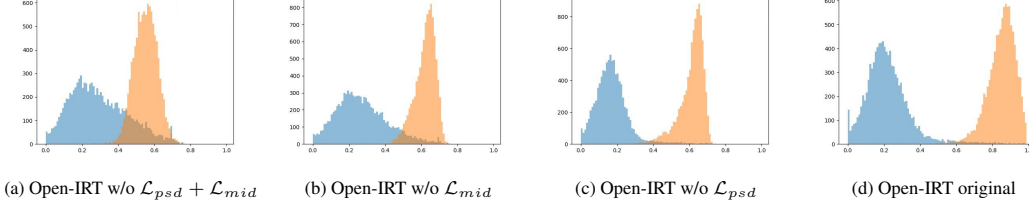


Figure 1: Visualization of OOD Scores with Ablation Loss Components, where (d) represents complete Open-IRT.

### C.2 Analysis on different learning rate.

We analyze Open-IRT in comparison to baselines with varying learning rates. The results presented in Fig. 3 and Fig. 2 indicate that Open-IRT demonstrates notable robustness to changes in the learning rate, regardless of whether it is trained on the CIFAR-10C or CIFAR-100C datasets. Specifically, within a certain range of learning rates (from  $1e-3$  to  $1e-4$ ), the model consistently exhibits substantial performance, suggesting that it maintains stability and effectiveness across different training conditions. This robustness not only highlights the adaptability of Open-IRT but also reinforces its applicability in real-world scenarios where optimal hyper-parameter tuning may be challenging.

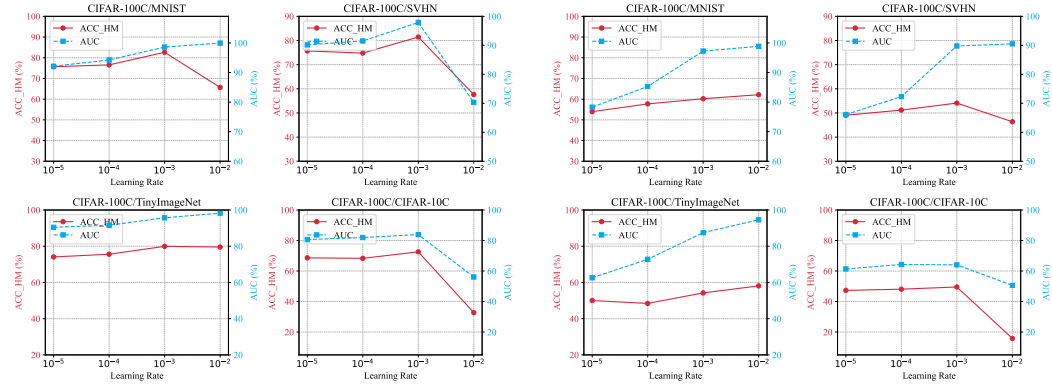


Figure 2: Learning rate with CIFAR-10C (ID). Figure 3: Learning rate with CIFAR-100C (ID).

### C.3 Analysis on different parameter $B$ , $K$ , $\alpha$ and $\lambda$ .

We investigate the impact of varying the hyper-parameter  $K$ , which denotes the number of positive and negative samples selected in the contrastive learning-based test-time adaptation process, and present the corresponding results ( $Acc_{HM}$ ) in Table 3. We set  $K$  to values of 0, 1, 3, and 5. The experimental results indicate that Open-IRT maintains stable performance across different values of  $K$ .

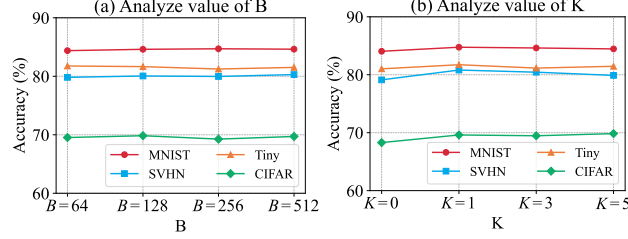


Figure 4: Analyze the  $B$  and  $K$  with CIFAR-10C as ID.

Notably, in the CIFAR-10C and MNIST experiments, the accuracy of Open-IRT fluctuates by no more than 0.70%. This stability suggests that Open-IRT is resilient to variations in sample selection, indicating a robust learning mechanism that can effectively leverage available data without being overly sensitive to hyper-parameter tuning. Such resilience enhances the practical applicability of Open-IRT, as it can perform reliably in diverse scenarios where optimal hyper-parameter settings may not be easily achievable. And we evaluate  $\alpha$  and  $\lambda$  in CIFAR-10C  $\rightarrow$  SVHN, results shows that  $\lambda$  exhibits high robustness within the range of 0.1 to 0.5, with accuracy fluctuations less than 0.57%. In comparison,  $\alpha$  demonstrates even greater robustness, with accuracy variations below 0.33% in the same range. We observe that an excessively large value of  $\lambda$  can degrade model performance, as it biases the optimization towards distinguishing ID and OOD distributions, while relatively diminishing the emphasis on discriminating between classes within the ID distribution. Therefore,  $\lambda$  should be selected within a theoretically sound range.

ID	OOD Dataset	K			
		0	1	3	5
CIFAR-10C	MNIST	84.05	84.75	84.62	84.46
	SVHN	78.12	80.81	80.44	79.89
	Tiny	81.02	81.23	81.15	81.16
	CIFAR-100C	68.09	69.51	69.48	69.24
CIFAR-100C	MNIST	63.08	59.40	60.13	59.58
	SVHN	46.97	59.24	57.24	59.18
	Tiny	53.39	55.58	55.92	56.17
	CIFAR-10C	39.30	47.44	47.55	47.42

Table 3: The  $Acc_{HM}(\%)$  with different  $K$  for Open-IRT.

We conduct experiments of CIFAR-10C with CLIP to analyze the hyper-parameters. We evaluate  $B$ , the size of both the  $B^f$  and  $B^s$  in Fig. 4a, with values of 64, 128, 256, 512. Results show a stable performance, with variations remaining within 0.37% in CIFAR-10C  $\rightarrow$  SVHN. We also investigate the impact of varying the hyper-parameter  $K$  in Eq. ??, ?? with values of 0,1,3,5 in Fig. 4b, with accuracy fluctuations of no more than 0.70% in CIFAR-10C  $\rightarrow$  MNIST. And we evaluate  $\alpha$  and  $\lambda$  in CIFAR-10C  $\rightarrow$  SVHN, results shows that  $\lambda$  exhibits high robustness within the range of 0.1 to 0.5, with accuracy fluctuations less than 0.57%. In comparison,  $\alpha$  demonstrates even greater robustness, with accuracy variations below 0.33% in the same range.

Datasets	Hyper	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
CIFAR-10C $\rightarrow$ USPS	$\alpha$	80.47	80.61	80.73	80.80	80.75	80.73	80.48	80.35	80.22
	$\lambda$	80.61	80.91	80.88	80.64	80.34	79.68	75.87	70.25	68.41

Table 4: Analyze Hyperparameter  $\alpha$  and  $\lambda$ .

#### C.4 Analysis on different backbones.

We also leverage the CLIP ViT-B32 and ViT-B16 architectures to evaluate Open-IRT. The experimental results presented in Table 5 demonstrate that Open-IRT consistently achieves superior performance, even after replacing the original ViT-B16 backbone. This robustness across different backbone architectures suggests that Open-IRT is not only effective but also adaptable, capable of leveraging various model architectures without compromising performance. Such versatility is particularly advantageous in real-world applications, where different deployment scenarios may necessitate the use of diverse model architectures with computational constraints or specific task requirements.

		L											
Method (ViT-B32)		MNIST			SVHN			Tiny-ImageNet			CIFAR-100C/10C		
		AUC $\uparrow$	FPR $\downarrow$	HM $\uparrow$	AUC $\uparrow$	FPR $\downarrow$	HM $\uparrow$	AUC $\uparrow$	FPR $\downarrow$	HM $\uparrow$	AUC $\uparrow$	FPR $\downarrow$	HM $\uparrow$
CIFAR-10C	ZS-Eval [15]	96.58	18.94	73.20	92.01	43.95	71.35	91.55	24.72	72.19	79.27	69.32	64.06
	TPT [14]	96.55	19.44	73.96	91.97	44.31	71.96	91.54	24.81	73.61	79.25	69.48	64.59
	TPT-C [14]	63.79	99.97	50.48	55.96	99.30	40.63	78.71	52.30	43.31	57.83	93.11	42.47
	ROSITA [15]	99.14	3.84	81.65	<b>93.78</b>	<b>33.45</b>	<b>75.18</b>	98.86	4.14	80.91	80.28	64.17	64.34
	Open-IRT	<b>99.47</b>	<b>1.68</b>	<b>82.28</b>	92.93	35.85	73.76	<b>98.99</b>	<b>3.58</b>	<b>81.53</b>	<b>80.45</b>	<b>62.68</b>	<b>65.17</b>
CIFAR-100C	ZS-Eval [15]	89.17	61.01	46.11	78.17	79.92	44.59	72.58	61.21	45.65	64.29	90.53	41.44
	TPT [14]	89.08	61.15	45.99	78.06	80.11	44.78	72.57	61.24	46.25	64.31	90.47	41.65
	TPT-C [14]	61.66	99.96	17.97	30.50	89.96	11.55	83.18	82.01	11.79	53.52	92.74	9.34
	ROSITA [15]	94.34	23.99	57.14	90.26	45.33	51.60	91.22	30.17	56.02	<b>68.33</b>	<b>86.03</b>	<b>44.57</b>
	Open-IRT	<b>99.31</b>	<b>3.73</b>	<b>58.69</b>	<b>94.05</b>	<b>27.23</b>	<b>52.23</b>	<b>93.21</b>	<b>22.32</b>	<b>56.85</b>	68.01	86.12	44.28

Table 5: Experiments with backbones ViT-B32 in CLIP model while the backbone ViT-B16 experiments are in the main text.

### C.5 Analysis different OOD ratios.

Table 6 presents additional experiments on different OOD ratios as detailed in the experiments section. We utilize the CIFAR-10C and CIFAR-100C datasets as ID datasets, while employing MNIST, SVHN, Tiny-ImageNet, and CIFAR-100C/10C as OOD datasets. The variable *RATIO* denotes that the number of OOD samples is  $10,000 \times RATIO$ . The results shown in Table 6 indicate that Open-IRT consistently outperforms other methods in a range of OOD ratios. To further assess robustness, we varied the OOD ratio between 0 and 1 with a step size of 0.05. As shown in Fig. 5, Open-IRT consistently maintained stable results, with fluctuations below 8.39% in AUROC and 22.60% in FPR95. These findings empirically indicate that Open-IRT is highly robust across a wide range of OOD sample ratios. The consistent results suggest that Open-IRT is robust to varying levels of OOD contamination, highlighting its effectiveness in maintaining accuracy even when faced with increasing proportions of OOD data. This is crucial for real-world applications with the unpredictable data distribution.

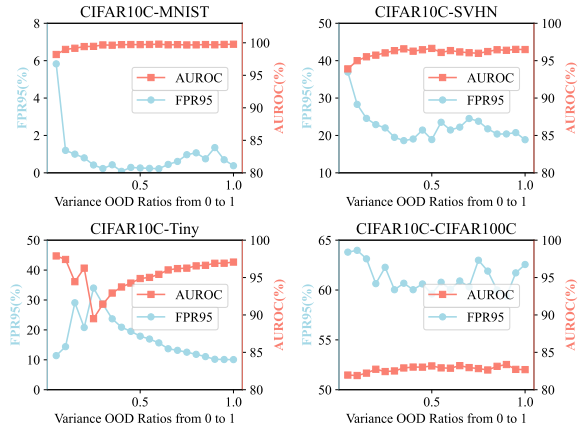


Figure 5: Analyze the FPR95 and AUROC with different OOD ratios of experiments on CLIP with CIFAR-10C as ID dataset.

Method			MNIST			SVHN			Tiny-ImageNet			CIFAR-100C/10-C		
			AUC $\uparrow$	FPR $\downarrow$	HM $\uparrow$	AUC $\uparrow$	FPR $\downarrow$	HM $\uparrow$	AUC $\uparrow$	FPR $\downarrow$	HM $\uparrow$	AUC $\uparrow$	FPR $\downarrow$	HM $\uparrow$
CIFAR-10C/CLIP	RATIO = 0.4	ZS-Eval [15]	92.02	84.36	75.60	90.22	63.12	74.19	81.27	54.11	67.17	82.66	67.60	69.09
		TPT [14]	91.99	84.78	75.77	90.20	63.53	74.42	81.24	54.25	67.43	82.66	67.80	69.32
		TPT-C [14]	83.30	69.45	74.29	65.49	74.30	53.66	70.57	75.93	64.18	75.05	75.63	68.03
		ROSITA [15]	99.34	2.53	84.68	95.36	27.50	78.02	93.17	23.05	76.48	<b>83.18</b>	60.93	69.27
		Open-IRT	<b>99.71</b>	<b>0.08</b>	<b>85.66</b>	<b>96.27</b>	<b>19.05</b>	<b>80.12</b>	<b>93.74</b>	<b>20.90</b>	<b>77.66</b>	83.03	<b>60.03</b>	<b>69.14</b>
	RATIO = 0.6	ZS-Eval [15]	91.95	84.92	75.59	90.07	63.96	74.12	86.52	41.05	71.14	82.51	67.86	68.97
		TPT [14]	91.93	85.33	75.81	90.06	64.31	74.37	86.50	41.20	71.61	82.51	68.02	69.21
		TPT-C [14]	85.62	65.95	74.75	61.01	81.58	51.21	70.17	66.98	53.49	70.18	85.88	65.33
		ROSITA [15]	99.37	3.02	83.90	95.36	28.20	79.46	95.09	17.05	78.37	82.79	64.63	<b>69.65</b>
		Open-IRT	<b>99.79</b>	<b>0.22</b>	<b>85.00</b>	<b>96.35</b>	<b>21.46</b>	<b>80.76</b>	<b>95.43</b>	<b>15.66</b>	<b>79.59</b>	<b>82.89</b>	<b>60.07</b>	69.12
	RATIO = 0.8	ZS-Eval [15]	91.94	84.99	75.59	89.96	64.26	74.15	89.52	32.39	73.24	82.57	67.96	68.91
		TPT [14]	91.92	85.47	75.76	89.94	64.56	74.38	89.50	32.53	73.68	82.57	68.09	69.13
		TPT-C [14]	83.21	67.03	75.05	60.84	69.47	50.63	73.00	61.66	49.33	63.76	93.05	51.99
		ROSITA [15]	99.44	2.91	83.89	95.46	28.10	79.84	95.77	14.38	79.05	<b>82.94</b>	64.59	<b>69.67</b>
		Open-IRT	<b>99.68</b>	<b>1.07</b>	<b>84.68</b>	<b>96.22</b>	<b>21.77</b>	<b>80.60</b>	<b>96.56</b>	<b>11.83</b>	<b>80.55</b>	82.65	<b>61.90</b>	69.50
CIFAR-100C/CLIP	RATIO = 0.4	ZS-Eval [15]	77.81	99.92	48.36	64.85	98.60	45.71	53.05	87.28	39.45	63.26	93.28	44.01
		TPT [14]	77.81	99.93	48.18	64.86	98.58	45.78	53.01	87.30	39.43	53.02	87.30	39.43
		TPT-C [14]	51.32	100.0	26.09	13.56	100.0	10.64	52.15	82.83	16.22	55.17	88.00	19.29
		ROSITA [15]	77.84	90.15	47.95	78.92	78.13	47.02	<b>71.95</b>	<b>79.48</b>	<b>49.17</b>	68.37	<b>89.30</b>	46.81
		Open-IRT	<b>97.91</b>	<b>15.69</b>	<b>61.22</b>	<b>91.23</b>	<b>48.20</b>	<b>57.06</b>	70.13	81.10	48.50	<b>97.44</b>	90.25	<b>47.80</b>
	RATIO = 0.6	ZS-Eval [15]	77.89	99.93	48.36	64.68	98.58	45.48	62.39	79.23	42.81	63.19	93.40	44.03
		TPT [14]	77.89	99.93	48.30	89.17	64.84	57.85	62.36	79.20	42.74	66.63	91.23	48.24
		TPT-C [14]	50.36	100.0	25.55	11.40	99.98	8.85	59.60	79.46	19.68	55.72	<b>86.35</b>	18.32
		ROSITA [15]	94.28	25.92	57.73	82.24	68.27	48.43	79.49	63.87	52.86	<b>68.53</b>	88.81	47.87
		Open-IRT	<b>98.31</b>	<b>11.60</b>	<b>61.13</b>	<b>91.64</b>	<b>45.93</b>	<b>58.17</b>	<b>82.15</b>	<b>60.18</b>	<b>54.77</b>	67.48	90.64	<b>48.02</b>
	RATIO = 0.8	ZS-Eval [15]	77.88	99.93	48.31	64.74	98.68	45.57	68.28	72.98	46.49	63.20	93.41	44.06
		TPT [14]	77.88	99.93	48.39	64.75	98.66	45.66	68.26	72.94	46.57	<b>68.25</b>	<b>72.94</b>	46.57
		TPT-C [14]	48.84	100.0	28.96	9.73	100.0	6.96	60.85	82.54	17.66	60.85	82.54	17.67
		ROSITA [15]	95.69	23.31	58.30	78.54	72.56	44.80	85.42	50.75	56.18	67.89	90.83	<b>47.69</b>
		Open-IRT	<b>89.63</b>	<b>9.17</b>	<b>60.72</b>	<b>93.00</b>	<b>42.08</b>	<b>58.15</b>	<b>86.29</b>	<b>48.15</b>	<b>56.33</b>	66.95	91.35	47.55

Table 6: open-set Single-Image Test Time Adaptation results with different ratios. We leverage the CIFAR-10C and CIFAR-100C as ID datasets, MNIST, SVHN, Tiny-ImageNet and CIFAR-100C/CIFAR-10C as OOD datasets, while AUC, FPR and HM mean AUROC, FPR95 and  $Acc_{HM}$ , respectively. Different from AUROC and  $Acc_{HM}$ , the FPR95 metric is the lower the better. The results indicated in bold represent the best results.

## D Limitation

While Open-IRT has shown strong results on ID and OOD detection tasks, the current study focuses primarily on classification tasks, so extending Open-IRT to other tasks such as object detection or semantic segmentation may require additional modifications to accommodate different spatial or contextual dependencies. Although the method is effective in handling typical OOD detection tasks, its performance on more complex or highly noisy data may still require further research.

## E Social Impact

Open-IRT has the potential to make significant contributions in areas where real-time adaptation to changing environments is crucial. By improving the robustness and adaptability of AI systems, particularly in the detection of out-of-distribution (OOD) samples, Open-IRT can be leveraged in fields such as healthcare, autonomous driving, and robotics. For instance, in healthcare, Open-IRT could help diagnostic models maintain high accuracy even when faced with new medical data from different populations or imaging devices. In autonomous systems, the ability to quickly adapt to new objects or environmental changes can enhance safety and reliability. Moreover, the flexibility of Open-IRT, especially in handling fluctuating OOD ratios, makes it well-suited for deployment in dynamic, real-world settings, where the data distribution may vary over time. This adaptability can lead to more trustworthy AI systems, promoting widespread adoption in critical applications where safety and reliability are paramount.

## References

- [1] Abdul Samadh , J., Gani , M. H., Hussein , N., Khattak , M. U., Naseer , M. M., Shahbaz Khan , F., & Khan , S. H. (2024) Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. *NeurIPS* **36**.
- [2] Fisher , R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of eugenics* **7**(2): 179–188.
- [3] Ganin , Y., Ustinova , E., Ajakan , H., Germain , P., Larochelle , H., Laviolette , F., March , M., & Lempitsky , V. (2016) Domain-adversarial training of neural networks. *JMLR* **17**(59):1–35.
- [4] Gao , Z., Zhang , X.-Y., & Liu , C.-L. (2024) Unified entropy optimization for open-set test-time adaptation. In *CVPR* pages 23975–23984.
- [5] Hendrycks , D. & Dietterich , T. (2019) Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*
- [6] Karmanov , A., Guan , D., Lu , S., El Saddik , A., & Xing , E. (2024) Efficient test-time adaptation of vision-language models. In *CVPR* pages 14162–14171.
- [7] Khattak , M. U., Rasheed , H., Maaz , M., Khan , S., & Khan , F. S. (2023) Maple: Multi-modal prompt learning. In *CVPR* pages 19113–19122.
- [8] Le , Y. & Yang , X. (2015) Tiny imagenet visual recognition challenge. *CS 231N* **7**(7):3.
- [9] LeCun , Y., Bottou , L., Bengio , Y., & Haffner , P. (1998) Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**(11):2278–2324.
- [10] Li , Y., Xu , X., Su , Y., & Jia , K. (2023) On the robustness of open-world test-time training: Self-training with dynamic prototype expansion. In *ICCV* pages 11836–11846.
- [11] Netzer , Y., Wang , T., Coates , A., Bissacco , A., Wu , B., Ng , A. Y., & others (2011) Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop 2011*, pp. 4.
- [12] Peng , X., Usman , B., Kaushik , N., Hoffman , J., Wang , D., & Saenko , K. (2017) Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*
- [13] Radford , A., Kim , J. W., Hallacy , C., Ramesh , A., Goh , G., Agarwal , S., Sastry , G., Askell , A., Mishkin , P., Clark , J., & others (2021) Learning transferable visual models from natural language supervision. In *ICML* pages 8748–8763.
- [14] Shu , M., Nie , W., Huang , D.-A., Yu , Z., Goldstein , T., Anandkumar , A., & Xiao , C. (2022) Test-time prompt tuning for zero-shot generalization in vision-language models. *NeurIPS* **35**:14274–14289.
- [15] Sreenivas , M. & Biswas , S. (2024) Effectiveness of vision language models for open-world single image test time adaptation. *arXiv preprint arXiv:2406.00481*
- [16] Wang , H., Li , Y., Yao , H., & Li , X. (2023) Clipn for zero-shot ood detection: Teaching clip to say no. In *ICCV* pages 1802–1812.
- [17] Zhang , C., Stepputtis , S., Sycara , K., & Xie , Y. (2025) Dual prototype evolving for test-time generalization of vision-language models. *NeurIPS* **37**:32111–32136.