Learning Well-Structured Logits: Leveraging Vision–Language Complementarity for Open-World Test-Time Adaptation

Jia-Qi Lin[1,2], **Yinghua Yao**[1,2], **Chang-Dong Wang**[3,4] and **Yuangang Pan**[1,2*]
[1] Center for Frontier AI Research, Agency for Science, Technology and Research, Singapore
[2]Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore
[3]School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, China
[4]Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou, China
linj1@a-star.edu.sg, eva.yh.yao@gmail.com, changdongwang@hotmail.com,
pan_yuangang@a-star.edu.sg

Abstract

Open-world test-time adaptation (OWTTA) is increasingly studied for its ability to adapt models at inference time in the presence of both domain discrepancy and semantic variance. Existing methods typically rely on either discriminative models or vision-language models (VLMs) alone, leaving their complementarity in OWTTA underexplored. In this paper, we propose a general framework that explicitly leverages the complementary strengths of discriminative models and VLMs to enable robust open-world test-time adaptation. We first define *well-structured logits* for OWTTA and provide a theoretical analysis that reveals the complementary properties of logits from discriminative models and VLMs. Building on a unified formulation, we then decompose OWTTA into two coupled modules: confidence-calibrated filtering (CCF), which provides an estimate of in-distribution membership, and semantic complementarity adaptation (SCA), which gives the refined predictions through complementary logit fusion. Extensive experiments across multiple benchmarks empirically confirm the logit complementarity between discriminative models and VLMs, and show that our method effectively leverages it to consistently improve performance.

1 Introduction

Deep learning has witnessed rapid advances and demonstrated strong performance in diverse applications. These achievements are typically realized under the assumption that the training and test data follow the same i.i.d. distribution [Gong *et al.*, 2022]. In open-world scenarios, however, this assumption is often violated as *domain discrepancy* [Qu *et al.*, 2024] and *semantic variance* [Zhao and Lee, 2024] frequently coexist, leading to degraded performance and unreliable predictions in practice. For instance, in medical image analysis, domain discrepancy may arise from differences in imaging equipment, while semantic variance can occur when the testing data include novel or unseen disease categories. This dual challenge implies that models should not only adapt to domain shifts among in-distribution (ID) samples but also mitigate the negative impact introduced by out-of-distribution (OOD) samples. To address this challenge, open-world test-time adaption (OWTTA) [Li *et al.*, 2023] has been proposed as a promising paradigm, enabling models to adapt to the target domain without requiring label data or source statistics.

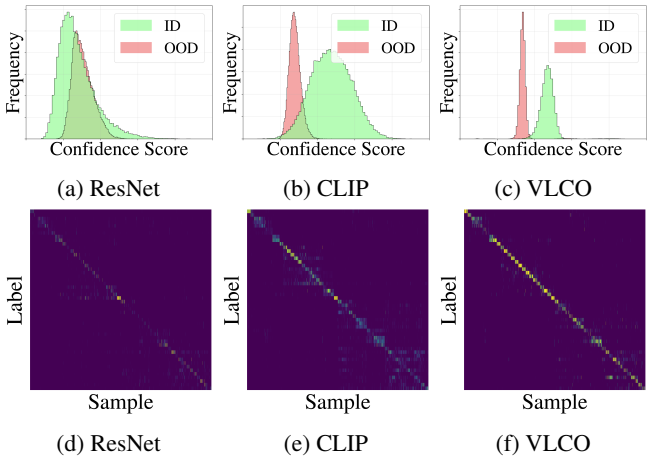


Figure 1: OWTTA visualization on ImageNet-C & MNIST. Top: confidence scores; bottom: post-softmax logits.

Conventional OWTTA methods [Gao *et al.*, 2024] typically rely on discriminative models, which can learn task-specific representations but remain constrained by their limited adaptability. Recently, vision-language models (VLMs) [Radford *et al.*, 2021], pre-trained on large-scale and diverse data, have been widely recognized for their zero-shot ability. By jointly modeling vision and language, VLMs learn powerful representations that capture rich semantics and establish effective alignment between visual inputs and textual class names. This property inherently mitigates the impact of domain discrepancy and makes VLMs a strong backbone for TTA [Shu *et al.*, 2022; Karmanov *et al.*, 2024], providing richer semantic cues that enhance generalization across domains. Despite these advantages, the role of VLMs in open-world TTA re-

*Corresponding author

mains largely underexplored. Existing methods typically rely on either a discriminative model or a VLM alone, without explicitly characterizing or effectively exploiting their implicit logit complementarity.

In this paper, we investigate how to uncover and exploit the implicit complementarity between discriminative models and vision-language models in OWTTA. We begin from a theoretical perspective by introducing the notion of *well-structured logits*, characterized by *semantic-shift separability*, *inter-class discriminativity* and *intra-class consistency*. This analysis reveals a complementary regime in which discriminative models and VLMs exhibit distinct yet compatible logit characteristics under open-world scenario, and it further motivates propositions showing that a principled logit mixing can yield well-structured logits. Guided by this insight, we propose vision-language complementarity for open-world test-time adaptation (VLCO). Specifically, we introduce a general framework for OWTTA, which contains coupled modules: confidence-calibrated filtering (CCF) and semantic complementarity adaptation (SCA). CCF provides a more reliable estimate of in-distribution membership for each sample by calibrating and fusing confidence signals from a discriminative model and a vision–language model, while SCA provides more accurate label prediction for reliable samples by integrating discriminative and vision–language cues through complementary logit fusion. As a result, VLCO operationalizes the identified complementarity and learns well-structured logits for robust open-world adaptation. The contributions of this paper are summarized as follows:

- We introduce the notion of *well-structured logits* for OWTTA by defining *semantic-shift separability*, *inter-class discriminativity* and *intra-class consistency*, and theoretically show that mixing complementary logits yields well-structured predictions.
- We propose a general framework for OWTTA that operationalizes vision–language complementarity by tightly coupling a discriminative model and a VLM through two modules, confidence-calibrated filtering (CCF) and semantic complementarity adaptation (SCA).
- Extensive experiments across multiple benchmarks empirically confirm the existence of complementary logit behaviors between discriminative models and VLMs, and demonstrate that our method yields well-structured logits and consistently improves open-world test-time adaptation performance.

2 Related Work

In this section, we review representative prior work relevant to our problem setting. To clarify distinctions among closely related adaptation paradigms, Table 1 provides an illustrative comparison of common settings.

Setting	Domain Discrepancy	Source Free	Semantic Shift	One-time Access
DA	✓	×	×	×
OSDA	✓	×	✓	×
SFDA	✓	✓	×	×
OS-SFDA	✓	✓	✓	×
TTA	✓	✓	×	✓
OWTTA	✓	✓	✓	✓

Table 1: Comparison of adaptation settings, including domain adaptation (DA), source-free domain adaptation (SFDA), open-set domain adaptation (OSDA), open-set source-free domain adaptation (OS-SFDA), test-time adaptation (TTA), and open-world test-time adaptation (OWTTA).

Test-Time Adaptation. Test-time adaptation (TTA) [Liang *et al.*, 2025; Su *et al.*, 2024a; Su *et al.*, 2022] adapts a source-trained model to target domains using only unlabeled test data. Representative approaches include entropy minimization [Liang *et al.*, 2020; Bar *et al.*, 2024; Zhang *et al.*, 2025; Wu *et al.*, 2025], distribution alignment [Wang *et al.*, 2024b; Zhang *et al.*, 2024; Su *et al.*, 2024b], and continual adaptation [Wang *et al.*, 2022; Niu *et al.*, 2022; Han *et al.*, 2025]. Most TTA methods operate under closed-set assumptions, requiring identical label spaces between source and target domains.

Open-Set Domain Adaptation. Open-set domain adaptation (OSDA) [Busto and Gall, 2017; Pham *et al.*, 2025; Choe *et al.*, 2024; Wan *et al.*, 2024] relaxes the closed-set assumption by allowing novel categories absent from the source and typically assumes access to the full target batch. Open-set source-free domain adaptation (OS-SFDA) [Yu *et al.*, 2025; Liu *et al.*, 2025; Wan *et al.*, 2024] further removes source data access, relying only on the source model and unlabeled target data. However, both settings usually assume offline optimization over the target domain, making them unsuitable for single-pass open-world test-time adaptation.

Open-World Test-Time Adaptation. Open-world test-time adaptation (OWTTA) tackles domain shift and semantic variation under single-pass test-time constraints. Unlike OOD detection [Yang *et al.*, 2024], OWTTA requires both unknown-sample separation and reliable in-domain classification. Existing methods usually combine OOD filtering with adaptation via distribution alignment [Li *et al.*, 2023] or entropy-based criteria [Gao *et al.*, 2024; Gong *et al.*, 2023]. These methods largely depend on task-specific discriminative representations, which limits robustness and generalization under open-world conditions.

Pre-trained Vision-Language Models. Large-scale vision-laguage models (VLMs), such as CLIP [Radford *et al.*, 2021], ALIGN [Jia *et al.*, 2021], and GroupViT [Xu *et al.*, 2022], are trained on massive image–text pairs via contrastive learning [Chen *et al.*, 2020] and exhibit strong generalization across diverse downstream tasks. Recent studies have explored leveraging VLMs for test-time adaptation [Osowiechi *et al.*, 2024; Wang *et al.*, 2024a; Phan *et al.*, 2024], indicating that VLMs constitute a promising semantic foundation, while leaving open how to systematically exploit their semantic representations for robust adaptation.

3 Leveraging Logit Complementarity for Open-World Test-Time Adaptation

OWTTA aims to improve prediction accuracy under distribution shifts without access to ground-truth labels, while handling unknown or noisy samples. In this setting, naive adaptation strategies that rely on instance-level predictions are highly susceptible to error accumulation and drift. Effective OWTTA therefore requires reliable and stable supervisory signals that do not depend on true labels.

We argue that the structure of model logits provides a useful lens for understanding and guiding the design of such supervision. Since test-time adaptation operates directly on model outputs, the geometric properties of logits critically influence whether adaptation remains stable under open-world distribution shifts. In particular, we characterize desirable logit structures through the following properties.

Definition 1 (Well-structured logits)**.** Let $\ell_i \in \mathbb{R}^K$ denote the pre-softmax logits vector produced by a model for input x_i. We say the logits are *well-structured* if they exhibit:

1. **Semantic-shift separability**: logits under semantic shifts (OOD) exhibit distinguishable structures from those of the original semantics (ID);
2. **Inter-class discriminativity**: the top-ranked logit is well separated from the remaining logits, providing clear separation between candidate classes;
3. **Intra-class consistency**: logits corresponding to samples from the same semantic class exhibit low variation, forming compact class-wise structures in the logit space.

Semantic-shift separability enables the model to suppress unreliable updates induced by semantically different inputs. Inter-class discriminativity mitigates misclassification of noisy samples, while intra-class consistency stabilizes adaptation by reducing the impact of unreliable instance-level predictions. Together, these properties characterize a well-structured logit space that supports reliable supervision in OWTTA.

Definition 2 (Semantic-shift separability)**.** Let s denote a score derived from logits. Semantic-shift separability refers to the ability of s to distinguish in-distribution inputs from out-of-distribution ones, where stronger separability implies that they can be more reliably partitioned based on s.

Definition 3 (Inter-class discriminativity)**.** Let $\ell_i^{(1)} \geq \ell_i^{(2)} \geq \cdots \geq \ell_i^{(K)}$ denote the sorted logits. The inter-class discriminativity of dataset $\mathcal{D}$ can be measured by the average margin

$$\Phi := \frac{1}{N}\sum_{i=1}^{N} \phi_i,$$

where $\phi_i = \ell_i^{(1)} - \ell_i^{(2)}$ denotes the instance-wise margin and larger Φ indicates a clearer decision margin.

Definition 4 (Intra-class consistency)**.** For each class k, let $\mathcal{D}_k = \{x_i | y_i = k\}$ denote the set of samples in class k. The intra-class consistency of $\mathcal{D}$ can be defined as

$$\Psi := \frac{1}{K}\sum_{k} \Psi_k, \quad \Psi_k := \frac{1}{|\mathcal{D}_k|}\sum_{x_i \in \mathcal{D}_k} \left\|\ell_i - \zeta_k\right\|_2^2,$$

where Ψ_k denotes the class-wise consistency and $\zeta_k = \frac{1}{|\mathcal{D}_k|}\sum_{x_i \in \mathcal{D}_k} \ell_i$ is the class-wise mean logit. A smaller Ψ indicates a more compact and class-consistent logit structure.

Let $s^{\mathbb{D}}$ and $s^{\mathbb{F}}$ denote the confidence score produced by two models, where $\mathbb{D}$ indexes a discriminative model and $\mathbb{F}$ indexes a vision-laguage model. Proposition 1 shows that a linear fusion of the confidence scores yields better semantic-shift separability, thereby reducing the influence of out-of-distribution samples during adaptation.

Proposition 1 (Optimal Fusion Improves Normalized Separation)**.** *Let $\beta \in \{0, 1\}$ indicate whether a sample is ID ($\beta = 1$) or OOD ($\beta = 0$). Assume the confidence scores $s^{\mathbb{D}}$ and $s^{\mathbb{F}}$ are conditionally independent given β, and each follows a two-component Gaussian mixture distribution with shared variances, namely,*

$$s^*|(\beta = \upsilon) \sim \mathcal{N}\big(\mu_{s^*,\upsilon},\, \sigma_{s^*}^2\big), \quad * \in \{\mathbb{D}, \mathbb{F}\}, \upsilon \in \{0, 1\},$$

where $\sigma_{s^*}^2$ is identical for both $\beta = 0$ and $\beta = 1$.

Let $\Delta_{s^*} = \mu_{s^*,1} - \mu_{s^*,0} > 0\ \forall * \in \{\mathbb{D}, \mathbb{F}\}$ and define the fused score $s^{\mathbb{M}} := \frac{s^{\mathbb{D}} + \alpha s^{\mathbb{F}}}{1+\alpha}, \forall \alpha > 0$. Then the fused score can achieve a strictly larger normalized separation, i.e.,

$$\frac{\mathbb{E}[s^{\mathbb{M}}|\beta = 1] - \mathbb{E}[s^{\mathbb{M}}|\beta = 0]}{\sqrt{\mathrm{Var}(s^{\mathbb{M}}|\beta)}} > \frac{\mathbb{E}[s^*|\beta = 1] - \mathbb{E}[s^*|\beta = 0]}{\sqrt{\mathrm{Var}(s^*|\beta)}},$$

$\forall * \in \{\mathbb{D}, \mathbb{F}\}$ *given* $\alpha = \Delta_{s^{\mathbb{F}}}\sigma_{s^{\mathbb{D}}}^2/(\Delta_{s^{\mathbb{D}}}\sigma_{s^{\mathbb{F}}}^2)$.

Let $\ell_i^{\mathbb{D}}, \ell_i^{\mathbb{F}} \in \mathbb{R}^K$ denote the pre-softmax logits for input x_i. By examining the logit structures of different model families (See Table 6 & Table 7), we observe a clear complementarity: the discriminative model exhibits high inter-class discriminativity but low intra-class consistency, whereas CLIP logits are less discriminative but substantially more consistent. Motivated by this complementarity, we combine logits from two models to derive reliable pseudo labels from better-structured fused logits.

We define their linear combination as $\ell_i^{\mathbb{M}} = \lambda\ell_i^{\mathbb{D}} + (1-\lambda)\ell_i^{\mathbb{F}}$ $\forall \lambda \in (0, 1)$. Next, we establish that the mixed logits $\ell_i^{\mathbb{M}}$ preserves both a guaranteed decision margin (Proposition 2) and bounded intra-class variation (Proposition 3), thereby enabling stable test-time adaptation and improved accuracy.

Proposition 2 (Lower-Bounded Decision Margin under Logit Mixing)**.** *Let $\ell_i^{\mathbb{D}}, \ell_i^{\mathbb{F}}, \ell_i^{\mathbb{M}} \in \mathbb{R}^K$ denote the pre-softmax logits vectors for input x_i, as defined above. Let $\ell_i^{(1)} \geq \ell_i^{(2)} \geq \cdots \geq \ell_i^{(K)}$ denote the sorted logits. Assume there exist positive constants $\varphi^{\mathbb{D}}, \varphi^{\mathbb{F}} > 0$ such that, $\forall x_i$ and $\forall k \neq 1$, $\ell_i^{\mathbb{D},(1)} - \ell_i^{\mathbb{D},(k)} \geq \varphi^{\mathbb{D}}$ for the discriminative model, and $|\ell_i^{\mathbb{F},(j)} - \ell_i^{\mathbb{F},(k)}| \leq \varphi^{\mathbb{F}}$ $\forall j, k$ for the vision-language model. Letting $\varphi^{\mathbb{M}} = \frac{1}{N}\sum_{i=1}^{N}(\ell_i^{\mathbb{M},(1)} - \ell_i^{\mathbb{M},(2)})$, we have $\varphi^{\mathbb{M}} > \varphi^{\mathbb{F}}$ when $\varphi^{\mathbb{D}} > \frac{(2-\lambda)}{\lambda}\varphi^{\mathbb{F}}$.*

Proposition 3 (Upper-Bounded Variation under Logit Mixing)**.** *Let $\ell_i^{\mathbb{D}}, \ell_i^{\mathbb{F}}, \ell_i^{\mathbb{M}} \in \mathbb{R}^K$ denote the pre-softmax logits vectors for input x_i, as defined above.*

Following Definition 4, the intra-class consistency of the mixed logits satisfies:

$$\Psi^{\mathbb{M}} \leq 2\lambda^2\Psi^{\mathbb{D}} + 2(1-\lambda)^2\Psi^{\mathbb{F}}.$$

Taken together, Proposition 2 and Proposition 3 show that mixing discriminative and CLIP logits preserves discriminative margins and enhances intra-class consistency. Motivated by this complementarity, we develop a framework that explicitly leverages it for OWTTA.

4 Methodology

In this section, we propose vision-language complementarity for open-world test-time adaptation (VLCO). It tackles open-world test-time adaptation by jointly leveraging a discriminative model and a vision–language model and explicitly exploiting their complementary properties.

We formulate a general framework for OWTTA by modeling the distribution $p(y|x; f^{\mathbb{D}}, f^{\mathbb{F}})$ induced by the discriminative model $f^{\mathbb{D}}$ and the VLM $f^{\mathbb{F}}$. To capture the semantic shift between samples, we introduce a binary latent variable $\beta \in \{0, 1\}$ indicating whether a sample is ID ($\beta = 1$) or OOD ($\beta = 0$). The label distribution can then be decomposed as:

$$\begin{aligned} p(y|x; f^{\mathbb{D}}, f^{\mathbb{F}}) &= \sum_{\beta\in\{0,1\}} p(y, \beta|x; f^{\mathbb{D}}, f^{\mathbb{F}}) \\ &= \sum_{\beta\in\{0,1\}} p(\beta|x; f^{\mathbb{D}}, f^{\mathbb{F}})\, p(y|x, \beta; f^{\mathbb{D}}, f^{\mathbb{F}}). \end{aligned} \tag{1}$$

The decomposition in Eq. (1) yields two coupled objectives: (i) *Confidence-Calibrated Filtering* (CCF), which estimates the ID/OOD distribution $p(\beta|x; f^{\mathbb{D}}, f^{\mathbb{F}})$, and (ii) *Semantic Complementarity Adaptation* (SCA), which refines the label distribution $p(y|x, \beta; f^{\mathbb{D}}, f^{\mathbb{F}})$ for the ID samples.

Specifically, we define $f^{\mathbb{D}} = \{f_\theta, w, b\}$, where f_θ denotes the encoder and (w, b) parameterize the linear classifier, and $f^{\mathbb{F}} = \{f_I, f_T\}$, where f_I denotes the VLM image encoder and f_T denotes the VLM text encoder. In the following, we present the detials of CCF and SCA.

4.1 Confidence-Calibrated Filtering

The CCF module estimates the reliability of each test sample by calibrating and fusing confidence signals from the discriminative model and the vision–language model.

Specifically, the confidence score of discriminative model is defined as:

$$s^{\mathbb{D}} := \max_j \frac{c_j^\top f_\theta(x)}{\|c_j\|\,\|f_\theta(x)\|},$$

where $f_\theta(x) \in \mathbb{R}^d$ is the feature embedding and $c = [c_1, \ldots, c_K]^\top \in \mathbb{R}^{K\times d}$ denotes the class prototypes, initialized by the classifier weights w.

Correspondingly, the confidence score of vision-language model is defined as:

$$s^{\mathbb{F}} := \max_j \frac{f_T(e_j)^\top f_I(x)}{\|f_T(e_j)\|\,\|f_I(x)\|},$$

where $\{e_j\}_{j=1}^K$ are the class-specific text prompts.

To model the semantic-shift separability, we assume that the confidence score s is drawn from a two-component Gaussian mixture with shared variance:

$$p(s^*) = \sum_{\upsilon\in\{0,1\}} \pi_\upsilon\, \mathcal{N}\big(s^*;\, \mu_{s^*,\upsilon},\, \sigma^2_{s^*}\big), \quad * \in \{\mathbb{D}, \mathbb{F}\},$$

where π_υ denotes the mixing coefficients. Consequently, for $\upsilon \in \{0, 1\}$, the distribution $p(\beta = \upsilon|x; f^*)$ is equivalent to $p(\beta = \upsilon|s^*)$:

$$p(\beta = \upsilon|s^*) = \frac{\pi_\upsilon \exp\big(-\frac{(s^*-\mu_{s^*,\upsilon})^2}{2\sigma^2_{s^*}}\big)}{\sum_{\upsilon'\in\{0,1\}} \pi_{\upsilon'} \exp\big(-\frac{(s^*-\mu_{s^*,\upsilon})^2}{2\sigma^2_{s^*}}\big)}. \tag{2}$$

Based on the connection between k-means and mixtures of Gaussians [Kulis and Jordan, 2012], when $\sigma^2_{s^*} \to 0$, we can have the following k-means based clustering loss:

$$\arg\min_{S^+,S^-} \frac{1}{|\mathcal{S}^+|}\sum_{i\in\mathcal{S}^+}(s_i^* - \mu_{s^*,1})^2 + \frac{1}{|\mathcal{S}^-|}\sum_{i\in\mathcal{S}^-}(s_i^* - \mu_{s^*,0})^2, \tag{3}$$

where $\mathcal{S}^+ = \{i|\beta_i = 1\}$ and $\mathcal{S}^- = \{i|\beta_i = 0\}$ denote the predicted ID and OOD clusters, respectively, and the corresponding cluster means are $\mu_{s^*,1} = \frac{1}{|\mathcal{S}^+|}\sum_{i\in\mathcal{S}^+} s_i^*$ and $\mu_{s^*,0} = \frac{1}{|\mathcal{S}^-|}\sum_{i\in\mathcal{S}^-} s_i^*$. Therefore, CCF gives a hard estimation on β via two-cluster k-means clustering on the confidence scores s^*, yielding a binary partition of samples into $\mathcal{S}^+$ and $\mathcal{S}^-$.

Confidence-Calibrated Filtering. To improve OOD filtering, we propose the confidence-calibrated filtering module $p(\beta|x; f^{\mathbb{D}}, f^{\mathbb{F}})$, which augments the discriminative model's confidence with semantic cues from vision-language model.

Since the two confidence scores, $s^{\mathbb{D}}$ and $s^{\mathbb{F}}$, are not directly comparable in scale, we transform them into calibrated evidences via a weighted average:

$$s^{\mathbb{E}} = \frac{s^{\mathbb{D}} + \alpha s^{\mathbb{F}}}{1+\alpha}.$$

According to Proposition 1, an appropriate choice of α, i.e., $\alpha = \Delta_{s^{\mathbb{F}}}\sigma^2_{s^{\mathbb{D}}}/\Delta_{s^{\mathbb{D}}}\sigma^2_{s^{\mathbb{F}}}$ yields a strictly stronger filter than either component score alone. We then substitute s in Eq. (3) with $s^{\mathbb{E}}$ for a better estimation of β. This improvement is also reflected in our visualizations (See Fig. 1), where CCF produces a cleaner separation.

4.2 Semantic Complementarity Adaptation

In this subsection, we introduce the semantic complementarity adaptation (SCA) module, which exploits complementary information from the discriminative model and the VLM to facilitate adaptation to the target domain.

For OOD samples ($\beta = 0$), we define $p(y|x, \beta = 0; f^{\mathbb{D}}, f^{\mathbb{F}}) \equiv 0$, since such samples do not correspond to any in-distribution label. Consequently, we restrict our analysis to $p(y|, \beta = 1; f^{\mathbb{D}}, f^{\mathbb{F}})$. Specifically, for the discriminative model, the likelihood for class k is given by:

$$p(y = k|x, \beta = 1; f^{\mathbb{D}}) \propto \exp\big(z_k^{\mathbb{D}}(x)\big), \tag{4}$$

where $z_k^{\mathbb{D}}(x) := w_k^\top f_\theta(x) + b_k$. For the vision-language model, its logits is defined by:

$$p(y = k|x, \beta = 1; f^{\mathbb{F}}) \propto \exp\big(z_k^{\mathbb{F}}(x)\big), \tag{5}$$

where $z_k^{\mathbb{F}}(x) := f_T(e_k)^\top f_I(x)$.

The two likelihoods encode complementary information, with $f^{\mathbb{D}}$ providing strong discriminative structure and $f^{\mathbb{F}}$ contributing semantic stability.

Semantic Complementarity Adaptation. To exploit the complementarity between the logits, we adopt an equal-weighted product-of-experts in the logit space:

$$p(y=k \mid x, \beta=1; f^{\mathbb{D}}, f^{\mathbb{F}}) = \frac{\exp(z_k^{\mathbb{M}}(x))}{\sum_{j=1}^{K} \exp(z_j^{\mathbb{M}}(x))}, \quad (6)$$

where $z_k^{\mathbb{M}}(x) := 1/2(z_k^{\mathbb{D}}(x) + z_k^{\mathbb{F}}(x))$. Eq. (6) directly recovers the mixed-logit structure analyzed in Proposition 2 and Proposition 3 with mixing coefficient $\lambda = 1/2$.

According to the empirical results in Table 7 and Table 6, we observe that $\Psi^{\mathbb{D}} > \Psi^{\mathbb{M}} > \Psi^{\mathbb{F}}$ and $\varphi^{\mathbb{D}} > \varphi^{\mathbb{M}} > \varphi^{\mathbb{F}}$. indicating that SCA is able to explore the complementary roles of discriminativity and consistency in the logits.

4.3 Optimization

Let $t \in \{1, 2, \ldots, T\}$ denote the test-time index. The overall training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{MSE}}, \quad (7)$$

where the cross-entropy term drives adaptation using pseudo-labels over $\mathcal{S}^+$:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N^+} \sum_{i \in \mathcal{S}^+} \log p\big(y = \hat{y}_i^t | x_i^t, \beta=1; f^{\mathbb{D}}, f^{\mathbb{F}}\big), \quad (8)$$

with $\hat{y}_i^t = \arg\max_j p\big(y=j|x_i^t, \beta=1; f^{\mathbb{D}}, f^{\mathbb{F}}\big)$. The second term encourages feature compactness of the discriminative model with prototypes updated in an exponential moving average (EMA) manner [Pan *et al.*, 2024]:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N^+} \sum_{i \in \mathcal{S}^+} \sum_{j=1}^{K} \xi_{ij}^t \|z_i^t - c_j^t\|_2^2, \quad (9)$$

where $\xi_{ij}^t \in \{0,1\}$ is a one-hot indicator that assigns each feature z_i^t to its nearest prototype c_j^t. The details of the proposed method are shown in Algorithm 1.

Algorithm 1 VLCO

Input: Test samples x, pre-trained discriminative model $f^{\mathbb{D}}$, vision-language model $f^{\mathbb{F}}$, test prompts e.
Initialization: Calculate text embeddings by $f_T(e)$.

```
for each test batch x^t do
    Infer ID/OOD label β using (3)
    if β = 1 then
        Update f_θ by (7), and update c by EMA
        ŷ^t = arg max_j p(y = j|x^t, β = 1; f^D, f^F)
    end if
end for
```

5 Experiments

5.1 Experiment Setup

Datasets. We conduct evaluations on blur corruption and style transfer benchmarks. For blur corruption, CIFAR10-C, CIFAR100-C, and ImageNet-C [Hendrycks and Dietterich, 2019] are used as ID datasets, while for style transfer, ImageNet-R [Hendrycks *et al.*, 2021] and VisDA-C [Peng *et al.*, 2017] are adopted as ID datasets. The OOD datasets include Gaussian Noise together with five real-world datasets that exhibit distributions different from the corresponding ID datasets, including MNIST [Deng, 2012], SVHN [Netzer *et al.*, 2011], Tiny-ImageNet [Le and Yang, 2015], CIFAR100-C, and CIFAR10-C. The open-world datasets are constructed by maintaining an equal number of ID and OOD samples.

Baselines. We introduce two-types of baselines for comparison: discriminative model-based baselines and vision-language model-based baselines. The discriminative model-based baselines adopt ResNet50 as the representative architecture, which includes: TEST [He *et al.*, 2016], TENT [Wang *et al.*, 2021], SHOT [Liang *et al.*, 2020], OSTTA [Lee *et al.*, 2023], CoTTA [Wang *et al.*, 2022], UniEnt [Gao *et al.*, 2024], OWT3 [Li *et al.*, 2023][1] The vision–language model baselines are constructed upon CLIP-ViT-B/16, including CLIP [Radford *et al.*, 2021], TDA [Karmanov *et al.*, 2024], BCA [Zhou *et al.*, 2025], C-TPT [Yoon *et al.*, 2024], WATT-S [Osowiechi *et al.*, 2024]. All baselines are tuned following their original settings for fair comparison.

Evaluation Settings. Following recent OWTTA studies [Li *et al.*, 2023], we evaluate the models using three complementary metrics: ACC_I, ACC_O, and ACC_H. Specifically, ACC_I denotes the classification accuracy on ID samples, ACC_O measures the detection accuracy on OOD samples, and ACC_H is their harmonic mean. A higher ACC_H indicates that the method performs well on both ID classification and OOD detection, rather than improving one aspect at the expense of the other.

5.2 Experimental Results and Analysis

In this section, we conduct extensive experiments on five open-world datasets to validate the effectiveness of the proposed method under the OWTTA setting. The results are summarized in Tables 2, 3, 4, and 5, with the best performance highlighted in **bold** and the second best underlined. From these tables, we have the following key observations:

(1) The proposed method consistently outperforms all the baselines, validating its effectiveness for OWTTA. In particular, on the ImageNet-C open-world dataset in Table 3, VLCO achieves the highest ACC_H, exceeding the second-best method by at least 15.33%, which demonstrates strong robustness under different open-world scenarios.

(2) VLCO mitigates unreliable outcomes in challenging OOD conditions. For example, Table 2 shows that CLIP alone can be unreliable when distinguishing Noise OOD samples on the CIFAR100-C open-world dataset. By contrast, VLCO reduces the impact of noisy and inconsistent signals, leading to more stable and reliable adaptation.

(3) VLCO effectively leverages the complementarity between discriminative models and VLMs. As shown in Table 3, discriminative-model-based methods perform relatively poorly, whereas VLCO benefits from VLM guidance and achieves a markedly higher ACC_H, suggesting that the gains stem from complementary cues rather than simple ensembling. The visualization results are reported in Fig. 1,

[1] We refer to this method as OWT3 to avoid ambiguity between the method name and the general OWTTT setting.

CIFAR10-C	Noise			MNIST			SVHN			Tiny			CIFAR100-C		
	ACC_I	ACC_O	ACC_H	ACC_I	ACC_O	ACC_H	ACC_I	ACC_O	ACC_H	ACC_I	ACC_O	ACC_H	ACC_I	ACC_O	ACC_H
TEST	69.19	99.96	81.78	63.61	91.11	74.92	63.54	90.91	74.80	61.09	77.48	68.32	61.27	76.58	68.07
TENT	42.75	61.99	50.60	63.21	72.80	67.67	70.84	82.92	76.41	69.70	76.23	72.82	69.22	74.07	71.56
SHOT	42.71	60.72	50.15	64.57	73.51	68.75	71.12	84.37	77.18	69.51	77.06	73.09	69.10	74.76	71.82
OSTTA	53.43	13.69	21.80	53.88	36.94	43.83	62.76	41.20	49.74	74.85	32.38	45.20	**76.39**	29.80	42.87
CoTTA	35.89	68.20	47.03	60.20	69.25	64.41	69.20	80.11	74.26	69.07	75.14	71.98	68.67	73.65	71.07
UniEnt	39.29	59.97	47.48	59.70	69.79	64.35	68.47	80.33	73.93	68.58	75.46	71.86	68.80	73.50	71.07
OWT3	69.41	99.96	81.93	63.89	91.10	75.11	63.72	90.97	74.94	61.05	77.56	68.32	61.37	76.47	68.09
CLIP	57.82	99.97	73.27	62.31	**99.69**	76.69	66.05	96.70	78.49	72.76	81.27	76.78	65.17	75.68	70.03
TDA	58.10	99.97	73.49	61.17	**99.69**	75.82	66.52	96.70	78.82	72.83	81.27	76.82	66.66	75.68	70.88
BCA	57.88	99.97	73.31	60.94	**99.69**	75.64	66.06	96.70	78.50	72.18	81.27	76.46	66.05	75.68	70.54
C-TPT	68.36	99.97	81.20	61.89	**99.69**	76.37	65.50	96.70	78.10	71.32	81.27	75.97	65.22	75.68	70.06
WATT-S	57.49	99.97	73.00	62.23	**99.69**	76.63	66.69	96.70	78.94	72.70	81.27	76.75	65.71	75.68	70.34
Ours	**80.93**	**100.00**	**89.46**	**84.86**	98.61	**91.22**	**81.17**	**99.16**	**89.27**	**81.83**	**84.92**	**83.35**	74.81	**81.97**	**78.23**

CIFAR100-C	Noise			MNIST			SVHN			Tiny			CIFAR10-C		
	ACC_I	ACC_O	ACC_H	ACC_I	ACC_O	ACC_H	ACC_I	ACC_O	ACC_H	ACC_I	ACC_O	ACC_H	ACC_I	ACC_O	ACC_H
TEST	25.30	**99.99**	40.38	24.76	46.81	32.39	26.34	91.65	40.92	26.20	86.68	40.24	25.79	87.86	39.88
TENT	24.45	95.75	38.95	27.50	91.49	42.29	32.05	95.17	47.95	31.33	89.64	46.43	32.08	89.93	47.29
SHOT	25.98	96.62	40.95	28.26	87.36	42.71	32.42	95.14	48.36	31.18	89.71	46.28	32.07	**90.10**	47.30
OSTTA	12.20	40.58	18.76	17.97	46.56	25.93	33.33	62.59	43.50	36.81	52.15	43.16	**42.75**	43.67	43.21
CoTTA	20.93	93.25	34.19	25.23	92.85	39.68	30.09	**95.26**	45.73	31.17	89.85	46.28	31.86	89.90	47.05
UniEnt	20.05	94.48	33.08	25.54	92.86	40.06	31.36	94.48	47.09	31.20	89.19	46.23	32.52	89.03	47.64
OWT3	25.54	**99.99**	40.69	24.51	45.51	31.86	26.48	91.70	41.09	26.34	86.68	40.40	25.93	87.82	40.04
CLIP	33.67	2.99	5.49	31.14	**99.29**	47.41	31.57	94.88	47.38	37.81	66.30	48.16	31.13	68.01	42.71
TDA	34.84	2.99	5.51	32.00	**99.29**	48.40	31.95	94.88	47.80	38.58	66.30	48.78	31.99	68.01	43.51
BCA	34.67	2.99	5.51	31.84	**99.29**	48.22	31.77	94.88	47.60	38.66	66.30	48.84	31.89	68.01	43.42
C-TPT	46.21	2.99	5.62	31.65	**99.29**	48.00	31.73	94.88	47.56	38.47	66.30	48.69	31.74	68.01	43.28
WATT-S	37.07	2.99	5.53	33.25	**99.29**	49.82	33.27	94.88	49.27	41.38	66.30	50.96	33.21	68.01	44.63
Ours	**50.87**	98.66	**67.13**	**49.01**	97.65	**65.26**	**43.91**	94.17	**59.89**	**43.81**	76.33	**55.67**	39.34	75.35	**51.69**

Table 2: Open-world test-time adaptation results on CIFAR10-C and CIFAR100-C.

ImageNet-C	Noise			MNIST			SVHN		
	ACC_I	ACC_O	ACC_H	ACC_I	ACC_O	ACC_H	ACC_I	ACC_O	ACC_H
TEST	11.79	40.85	18.30	11.33	59.57	19.04	11.62	82.04	20.36
TENT	9.31	68.49	16.39	15.26	72.67	25.22	17.28	78.79	28.34
SHOT	9.70	64.57	16.87	14.76	71.69	24.48	17.70	79.63	28.96
OSTTA	13.90	65.34	22.92	20.54	57.34	30.25	21.93	59.43	32.04
CoTTA	8.75	67.48	15.49	13.78	82.82	23.63	14.13	74.66	23.76
UniEnt	8.29	73.44	14.90	14.13	79.71	24.00	14.81	82.88	25.13
OWT3	9.76	68.03	17.07	15.41	75.29	25.58	16.45	78.67	27.21
CLIP	27.88	**100.00**	43.61	29.94	**99.96**	46.08	29.75	96.39	45.47
TDA	29.32	**100.00**	45.34	30.79	**99.96**	47.08	30.76	96.39	46.64
BCA	28.84	**100.00**	44.77	30.26	**99.96**	46.46	30.25	96.39	46.05
C-TPT	28.20	**100.00**	43.99	29.73	**99.96**	45.83	29.67	96.39	45.02
WATT-S	28.73	**100.00**	44.64	30.99	**99.96**	47.31	30.80	96.39	46.68
Ours	**45.57**	99.86	**62.58**	**46.46**	99.09	**63.26**	**45.25**	**98.49**	**62.01**

Table 3: Open-world test-time adaptation results on ImageNet-C.

ImageNet-R	Noise			MNIST			SVHN		
	ACC_I	ACC_O	ACC_H	ACC_I	ACC_O	ACC_H	ACC_I	ACC_O	ACC_H
TEST	24.71	**100.00**	39.63	23.60	99.81	38.17	22.46	98.89	36.61
TENT	16.87	93.23	28.57	19.81	84.72	32.11	20.09	87.99	32.71
SHOT	17.27	96.24	29.28	19.70	81.43	31.72	20.44	89.61	33.29
OSTTA	25.46	58.59	35.50	22.46	49.67	30.93	25.66	58.02	35.58
CoTTA	15.58	85.41	26.35	17.40	86.71	28.98	17.51	83.04	28.92
UniEnt	10.56	72.46	18.43	15.01	79.59	25.26	15.84	83.72	26.64
OWT3	16.56	92.13	28.07	19.13	83.75	31.15	19.42	86.93	31.75
CLIP	46.43	**100.00**	63.42	55.14	**99.85**	71.05	60.98	99.16	75.52
TDA	44.01	**100.00**	61.12	53.69	**99.85**	69.83	61.74	99.16	76.10
BCA	43.87	**100.00**	60.99	53.51	**99.85**	69.68	61.42	99.16	75.86
C-TPT	43.86	**100.00**	60.98	52.66	**99.85**	68.95	59.20	99.16	74.14
WATT-S	45.91	**100.00**	62.93	54.84	**99.85**	70.80	60.65	99.16	75.27
Ours	**65.62**	99.58	**79.11**	**64.03**	98.49	**77.60**	63.71	**99.52**	**77.69**

Table 4: Open-world test-time adaptation results on ImageNet-R.

VisDA-C	Noise			MNIST			SVHN		
	ACC_I	ACC_O	ACC_H	ACC_I	ACC_O	ACC_H	ACC_I	ACC_O	ACC_H
TEST	41.27	**100.00**	58.43	43.12	97.41	59.78	42.06	99.46	59.12
TENT	51.57	**100.00**	68.05	41.50	79.96	54.64	44.57	88.40	59.26
SHOT	42.39	99.94	59.53	42.69	68.08	52.48	43.45	72.69	54.33
OSTTA	39.60	54.33	45.81	46.41	29.81	36.30	44.91	45.55	45.23
CoTTA	44.01	99.56	61.04	40.71	83.91	54.82	40.89	89.86	56.20
UniEnt	54.84	95.99	69.80	41.92	80.13	55.04	42.83	88.53	57.73
OWT3	46.12	99.99	63.12	42.14	77.39	54.57	43.04	91.60	58.56
CLIP	48.51	**100.00**	65.33	59.81	**99.71**	74.77	60.08	97.35	74.30
TDA	48.37	**100.00**	65.20	60.10	**99.71**	75.00	60.09	97.35	74.31
BCA	48.26	**100.00**	65.10	60.00	**99.71**	74.92	59.98	97.35	74.23
C-TPT	48.33	**100.00**	78.90	60.16	**99.71**	75.04	60.17	97.35	74.37
WATT-S	48.60	**100.00**	65.41	59.93	**99.71**	74.86	60.06	97.35	74.29
Ours	**70.31**	99.75	**82.48**	**70.73**	99.17	**82.57**	**75.11**	**99.53**	**85.62**

Table 5: Open-world test-time adaptation results on VisDA-C.

where we show the post-softmax logits heatmaps for 50 of the 1,000 classes for clarity.

In summary, VLCO introduces a vision–language enhanced framework that systematically leverages the complementarity between CLIP and discriminative models, yielding consistently superior performance across datasets.

5.3 Discriminativity and Consistency Analysis

In this subsection, we examine the trade-off between discriminativity (Disc) and Consistency (Cons) for the proposed method. Table 6 and Table 7 report results with MNIST as the OOD dataset under blur corruption and style transfer. Existing methods typically face a clear tension in which gains in discriminativity reduce consistency under distribution shift, while stronger consistency can come at the expense of dis-

criminativity. VLCO achieves a more favorable balance by improving intra-class consistency over ResNet while retaining strong discriminativity, and enhancing discriminativity over CLIP while maintaining comparable consistency. As a result, VLCO strengthens both aspects and attains a better overall operating point across datasets, indicating that it produces more reliable and effective logits for OWTTA.

Method	CIFAR10-C			CIFAR100-C			ImageNet-C		
	Disc↑	Cons ↓	ACC_H↑	Disc↑	Cons ↓	ACC_H↑	Disc↑	Cons ↓	ACC_H↑
ResNet	**0.5190**	0.3709	74.92	**0.1382**	0.6507	32.39	**0.0201**	0.4660	19.04
CLIP	0.0339	**0.0021**	76.69	0.0057	**0.0019**	47.41	0.0018	**0.0044**	46.08
Ours	0.1375	0.0168	**91.22**	0.0275	0.0121	**65.26**	0.0043	0.0119	**63.26**

Table 6: Discriminativity and consistency analysis under blur corruptions, with MNIST as the OOD dataset.

Method	ImageNet-R			VisDA-C		
	Disc↑	Cons ↓	ACC_H↑	Disc↑	Cons ↓	ACC_H↑
ResNet	**0.0675**	0.5975	38.17	**0.2712**	0.2835	59.78
CLIP	0.0112	**0.0076**	71.05	0.0596	**0.0038**	74.77
Ours	0.0192	0.0217	**77.60**	0.0804	0.0099	**82.57**

Table 7: Discriminativity and consistency analysis under style transfer, with MNIST as the OOD dataset.

5.4 Ablation study

In this subsection, we conduct comprehensive ablation studies to assess the contributions of confidence-calibrated filtering (CCF), semantic complementarity adaptation (SCA) and the MSE loss $\mathcal{L}_{\text{MSE}}$. Specifically, we conduct experiments on the CIFAR100-C open-world dataset and evaluate performance using the ACC_H metric. The results are summarized in Table 8. From Table 8, we observe that $\mathcal{L}_{\text{MSE}}$ stabilizes adaptation by encouraging features to align with class prototypes, while CCF and SCA contribute most of the performance gains, indicating that leveraging complementary vision–language knowledge substantially improves open-world adaptation.

CCF	SCA	$\mathcal{L}_{\text{MSE}}$	Noise	MNIST	SVHN	Tiny	CIFAR10-C
×	×	×	49.07	43.60	46.41	46.16	46.39
×	×	✓	60.72	52.90	50.98	47.36	46.56
✓	×	✓	61.47	57.95	53.64	51.17	47.35
×	✓	✓	64.04	61.16	52.29	48.81	48.44
✓	✓	✓	**67.13**	**65.26**	**59.89**	**55.67**	**51.69**

Table 8: Ablation study on CIFAR100-C open-world dataset.

5.5 Running-Time Analysis

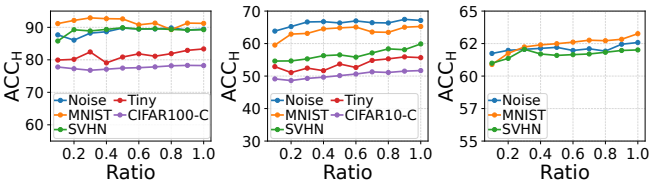
In this subsection, we compare the running time of CLIP-based methods. We use a batch size of 128 and report the average per-batch runtime in seconds on the CIFAR100-C open-world dataset. The results are summarized in Table 9. As shown in the table, the proposed method achieves a favorable trade-off between computational efficiency and adaptation performance.

CIFAR100-C	Noise	MNIST	SVHN	Tiny	CIFAR10-C
CLIP	0.1555	0.1569	0.1573	0.1556	0.1588
TDA	0.1748	0.1180	0.1195	0.1378	0.1334
BCA	0.1446	0.1256	0.1234	0.1382	0.1314
C-TPT	6.2115	6.0783	6.1201	6.3027	6.3314
WATT-S	21.1350	6.8855	7.4319	13.4887	11.0688
Ours	0.4843	0.3762	0.3663	0.4929	0.3956

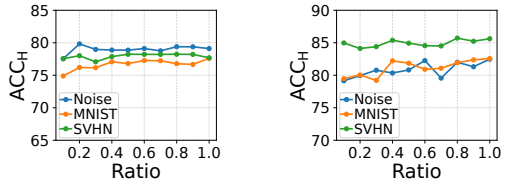
Table 9: Running-Time comparison on CIFAR100-C open-world dataset.

5.6 Analysis Under Different OOD Ratios

In this subsection, we evaluate the robustness of our method under different OOD ratios. Specifically, we vary the OOD ratio $N_{\text{OOD}}/N_{\text{ID}}$ from 0.1 to 1. The corresponding ACC_H results are reported in Fig. 2. When the OOD ratio is extremely low, we observe slight fluctuations, which is expected because ACC_O becomes sensitive in the rare-OOD regime with only a few OOD samples. Overall, the proposed method remains stable across datasets and OOD ratios, demonstrating its robustness to varying OOD proportions in open-world scenarios.



(a) CIFAR10-C (b) CIFAR100-C (c) ImageNet-C



(d) ImageNet-R (e) VisDA-C

Figure 2: Performance analysis under different OOD ratios. VLCO maintains stable performance across all open-world datasets.

5.7 Conclusion

In this paper, we investigate open-world test-time adaptation (OWTTA), where domain discrepancy and semantic variance occur simultaneously. We propose vision-language complementarity for open-world test-time adaptation (VLCO), a general framework that exploits the complementarity between discriminative models and vision–language models. VLCO consists of two components: confidence-calibrated filtering, which estimates sample reliability by calibrating and fusing confidence signals from both models, and semantic complementarity adaptation, which refines predictions on reliable samples through complementary logit fusion to enable effective and stable adaptation. Extensive experiments on diverse benchmarks further demonstrate the effectiveness and robustness of VLCO across different open-world scenarios. Further study on lightweight deployment and efficient model coordination will be explored in future work.

Acknowledgments

This work was supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-004). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

[Bar *et al.*, 2024] Yarin Bar, Shalev Shaer, and Yaniv Romano. Protected test-time adaptation via online entropy matching: A betting approach. In *NeurIPS*, 2024.

[Busto and Gall, 2017] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *ICCV*, pages 754–763. IEEE Computer Society, 2017.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, volume 119, pages 1597–1607, 2020.

[Choe *et al.*, 2024] Seun-An Choe, Ah-Hyung Shin, Keon-Hee Park, Jinwoo Choi, and Gyeong-Moon Park. Open-set domain adaptation for semantic segmentation. In *CVPR*, pages 23943–23953. IEEE, 2024.

[Deng, 2012] Li Deng. The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Process. Mag.*, 29(6):141–142, 2012.

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

[Gong *et al.*, 2022] Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. NOTE: robust continual test-time adaptation against temporal correlation. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *NeurIPS*, 2022.

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

[Han *et al.*, 2025] Jisu Han, Jaemin Na, and Wonjun Hwang. Ranked entropy minimization for continual test-time adaptation. *CoRR*, 2025.

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

[Hendrycks and Dietterich, 2019] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.

[Hendrycks *et al.*, 2021] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8320–8329, 2021.

[Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, volume 139, pages 4904–4916, 2021.

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

[Kulis and Jordan, 2012] Brian Kulis and Michael I Jordan. Revisiting k-means: new algorithms via bayesian nonparametrics. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1131–1138, 2012.

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

[Lee *et al.*, 2023] Jungsoo Lee, Debasmit Das, Jaegul Choo, and Sungha Choi. Towards open-set test-time adaptation utilizing the wisdom of crowds in entropy minimization. In *ICCV*, page 16334, 2023.

[Li *et al.*, 2023] 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 11802–11812, 2023.

[Liang *et al.*, 2020] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, volume 119, pages 6028–6039, 2020.

[Liang *et al.*, 2025] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *Int. J. Comput. Vis.*, 133(1):31–64, 2025.

[Liu *et al.*, 2025] Weiming Liu, Jun Dan, Fan Wang, Xinting Liao, Junhao Dong, Hua Yu, Shunjie Dong, and Lianyong Qi. Distinguish then exploit: Source-free open set domain adaptation via weight barcode estimation and sparse label assignment. In *CVPR*, pages 4927–4938, 2025.

[Netzer *et al.*, 2011] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

[Niu *et al.*, 2022] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *ICML*, volume 162, pages 16888–16905, 2022.

[Osowiechi *et al.*, 2024] David Osowiechi, Mehrdad Noori, Gustavo Adolfo Vargas Hakim, Moslem Yazdanpanah, Ali Bahri, Milad Cheraghalikhani, Sahar Dastani, Farzad Beizaee, Ismail Ben Ayed, and Christian Desrosiers. WATT: weight average test time adaptation of CLIP. In *NeurIPS*, 2024.

[Pan *et al.*, 2024] Yuangang Pan, Yinghua Yao, and Ivor W. Tsang. PC-X: profound clustering via slow exemplars. In *CPAL*, volume 234, pages 1–19, 2024.

[Peng *et al.*, 2017] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *CoRR*, 2017.

[Pham *et al.*, 2025] Thai-Hoang Pham, Yuanlong Wang, Changchang Yin, Xueru Zhang, and Ping Zhang. Open-set heterogeneous domain adaptation: Theoretical analysis and algorithm. In *AAAI*, pages 19895–19903, 2025.

[Phan *et al.*, 2024] Viet Hoang Phan, Tung Lam Tran, Quyen Tran, and Trung Le. Enhancing domain adaptation through prompt gradient alignment. In *NeurIPS*, 2024.

[Qu *et al.*, 2024] Sanqing Qu, Tianpei Zou, Lianghua He, Florian Röhrbein, Alois Knoll, Guang Chen, and Changjun Jiang. LEAD: learning decomposition for source-free universal domain adaptation. In *CVPR*, pages 23334–23343, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pages 8748–8763, 2021.

[Shu *et al.*, 2022] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022.

[Su *et al.*, 2022] Yongyi Su, Xun Xu, and Kui Jia. Revisiting realistic test-time training: Sequential inference and adaptation by anchored clustering. In *NeurIPS*, 2022.

[Su *et al.*, 2024a] Yongyi Su, Xun Xu, and Kui Jia. Towards real-world test-time adaptation: Tri-net self-training with balanced normalization. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *AAAI*, pages 15126–15135, 2024.

[Su *et al.*, 2024b] Yongyi Su, Xun Xu, Tianrui Li, and Kui Jia. Revisiting realistic test-time training: Sequential inference and adaptation by anchored clustering regularized self-training. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(8):5524–5540, 2024.

[Wan *et al.*, 2024] Fuli Wan, Han Zhao, Xu Yang, and Cheng Deng. Unveiling the unknown: Unleashing the power of unknown to known in open-set source-free domain adaptation. In *CVPR*, pages 24015–24024, 2024.

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

[Wang *et al.*, 2022] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *CVPR*, pages 7191–7201, 2022.

[Wang *et al.*, 2024a] Ran Wang, Hua Zuo, Zhen Fang, and Jie Lu. Towards robustness prompt tuning with fully test-time adaptation for clip's zero-shot generalization. In *ACM MM*, pages 8604–8612, 2024.

[Wang *et al.*, 2024b] Ziqiang Wang, Zhixiang Chi, Yanan Wu, Li Gu, Zhi Liu, Konstantinos N. Plataniotis, and Yang Wang. Distribution alignment for fully test-time adaptation with dynamic online data streams. In *ECCV*, volume 15082, pages 332–349, 2024.

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

[Xu *et al.*, 2022] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas M. Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, pages 18113–18123, 2022.

[Yang *et al.*, 2024] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *Int. J. Comput. Vis.*, 132(12):5635–5662, 2024.

[Yoon *et al.*, 2024] Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark A. Hasegawa-Johnson, Yingzhen Li, and Chang D. Yoo. C-TPT: calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *ICLR*, 2024.

[Yu *et al.*, 2025] Zhiqi Yu, Zhichao Liao, Jingjing Li, Zhi Chen, and Lei Zhu. Dynamic target distribution estimation for source-free open-set domain adaptation. In *AAAI*, pages 22254–22262, 2025.

[Zhang *et al.*, 2024] Zhen-Yu Zhang, Zhiyu Xie, Huaxiu Yao, and Masashi Sugiyama. Test-time adaptation in non-stationary environments via adaptive representation alignment. In *NeurIPS*, 2024.

[Zhang *et al.*, 2025] Qingyang Zhang, Yatao Bian, Xinke Kong, Peilin Zhao, and Changqing Zhang. COME: test-time adaption by conservatively minimizing entropy. In *ICLR*, 2025.

[Zhao and Lee, 2024] Na Zhao and Gim Hee Lee. Robust visual recognition with class-imbalanced open-world noisy data. In *AAAI*, pages 16989–16997, 2024.

[Zhou *et al.*, 2025] Lihua Zhou, Mao Ye, Shuaifeng Li, Nianxin Li, Xiatian Zhu, Lei Deng, Hongbin Liu, and Zhen Lei. Bayesian test-time adaptation for vision-language models. In *CVPR*, pages 29999–30009, 2025.