Test-Time Adaptation via Self-Reinforced Optimal Transport for Zero-Shot OOD Detection with Vision–Language Models

Mingcai Chen[1,2,3] **Heng-yang Lu**[4] **Yuntao Du**[5] **Baoming Zhang**[3] **Hao Zhou**[1,2,6*]

[1]Nanjing University of Posts and Telecommunications
[2]The State Key Laboratory of Tibetan Intelligence
[3]State Key Laboratory for Novel Software Technology, Nanjing University
[4]School of Artificial Intelligence and Computer Science, Jiangnan University
[5]Joint SDU–NTU Centre for Artificial Intelligence Research (C-FAIR)
& School of Software, Shandong University
[6]Tongda College of Nanjing University of Posts and Telecommunications

Abstract

Pre-trained vision-language models (VLMs) enable zero-shot classification by matching images with text prompts, yet their raw confidence scores are unreliable for out-of-distribution (OOD) detection. We identify a confidence bottleneck: zero-shot VLMs produce posteriors whose maximum confidence yields much weaker ID/OOD separation than supervised classifiers trained on the target label space. To theoretically explain this gap, we relate VLM confidence to a well-separated supervised reference through a Markov-kernel transformation and analyze single-threshold detection using the Youden/KS index. Our analysis shows why raw VLM confidence is unlikely to recover the clean separation achieved by supervised classifiers. To address, motivated by the observation that class-wise posterior distributions contain useful structure beyond maximum confidence, we propose Self-Reinforced Optimal Transport (SROT), a label-free test-time framework that reshapes VLM posteriors with optimal transport and adaptively trains a lightweight OOD detector from rectified pseudo-labels. The detector further feeds open-set evidence back into posterior alignment, forming a self-reinforcing loop between distribution-level rectification and feature-level OOD learning. Extensive experiments on standard zero-shot OOD detection benchmarks show that SROT achieves state-of-the-art performance, with ablations validating posterior alignment, adaptive detector training, and self-reinforcement.

1 Introduction

Pre-trained vision-language models (VLMs), such as CLIP [Radford et al., 2021, Cherti et al., 2023], have become a standard interface for zero-shot visual recognition. By comparing image embeddings with text prompt embeddings, a VLM can classify images from task-specific label names without labeled examples or task-specific training. This flexibility makes VLMs attractive for open-world recognition, where a deployed model must not only recognize in-distribution (ID) categories but also reject out-of-distribution (OOD) inputs. Such zero-shot OOD detection is crucial in risk-sensitive applications, including healthcare, autonomous driving, and large-scale visual monitoring [Miyai et al., 2025a, Rasheed et al., 2022, Muhammad et al., 2020].

*Correspondence to: Hao Zhou <haozhou@njupt.edu.cn>.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

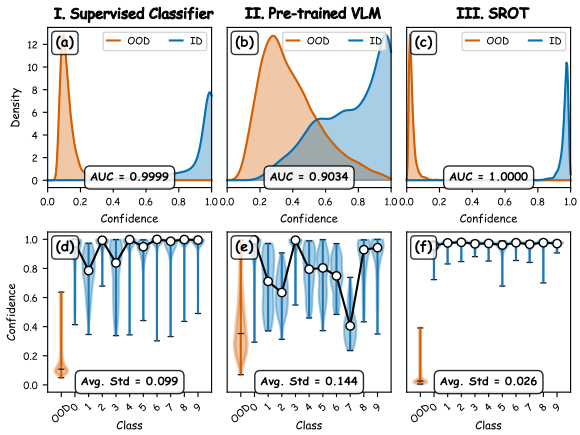


Figure 1: Motivation of SROT. We compare I. a supervised classifier, II. a zero-shot VLM, and III. our SROT on a mixture of Oxford-IIIT Pet [Parkhi et al., 2012] as ID and iNaturalist [Horn et al., 2018] as OOD. Top row (a–c): normalized confidence (e.g., MSP) distributions and AUC. Bottom row (d–f): per-class violin plots of confidence with mean per-class standard deviation (random ten classes and OOD).

Most existing OOD detection methods for supervised classifiers rely on a simple principle: compute a scalar confidence score and reject samples whose confidence falls below a threshold. Maximum softmax probability (MSP) [Hendrycks and Gimpel, 2017, Liang et al., 2018] is a representative example. For supervised classifiers trained on the target label space, this strategy is often effective because ID samples tend to receive high confidence while OOD samples tend to receive low confidence. However, it is unclear whether the same confidence-thresholding principle should transfer to pre-trained VLMs under the zero-shot setting.

We answer this question from both theoretical and empirical perspectives. The supervised classifier provides a natural reference point: when trained on the target label space, its confidence is directly optimized to separate ID classes and typically yields a clear confidence gap between ID and OOD samples. Sec. 3.2 relates the raw confidence of a zero-shot VLM to the confidence of a well-separated supervised reference through a Markov kernel [Williamson and Cranko, 2024]. This kernel captures the distortion: samples that are confident under the supervised reference may remain likely to be confident under the pre-trained VLM, but the mapping can blur the ID/OOD confidence gap. Using the Youden/KS index, which measures the best ID/OOD separation achievable by a single-threshold confidence rule, we show that such a transformed confidence signal cannot achieve better threshold separability than the supervised reference. Our empirical observation agrees with this analysis. As shown in the first row of Fig. 1, a supervised classifier yields nearly separable confidence distributions for ID and OOD samples, whereas zero-shot CLIP produces substantial overlap between them. This gap indicates that the failure of zero-shot VLM OOD detection is not merely caused by an improperly selected threshold, rather, raw confidence forms a bottleneck for separating ID and OOD samples.

To cope with this limitation, we look beyond the confidence score and examine the class-wise confidence distribution. The second row of Fig. 1 shows that the supervised classifier maintains consistently high confidence across ID classes while keeping OOD confidence clearly separated. In contrast, the zero-shot VLM exhibits less stable class-wise confidence behavior: some classes receive high confidence, whereas others are assigned weaker or more dispersed confidence. This observation suggests that useful OOD evidence is not only contained in each sample's confidence, but also in the posterior allocation over classes. Therefore, instead of directly thresholding raw confidence, we aim to utilize the class-wise posterior distribution to produce a more reliable OOD detection signal.

This observation motivates **SROT**, a **S**elf-**R**einforced **O**ptimal **T**ransport framework for zero-shot OOD detection. At test time, although labels are unavailable, an unlabeled batch naturally forms a posterior matrix over samples and prompt classes, revealing how the pre-trained VLM allocates probability mass across the candidate classes. SROT uses optimal transport to align this posterior toward a class marginal while preserving image-text semantic preference through the transport cost, thereby converting the raw prompt-normalized posterior into a more reliable OOD detection signal. The rectified posterior is then used to generate adaptive pseudo-labels for training a lightweight detector on frozen VLM image features. Finally, the detector feeds open-set evidence back into the posterior alignment process, forming a self-reinforcing loop between distribution-level rectification and feature-level OOD learning. This design exploits the class-wise posterior structure observed in Fig. 1 while avoiding reliance on raw confidence alone. As shown in Fig. 1(c,f), SROT improves ID/OOD separation and yields more stable class-wise confidence behavior.

We evaluate SROT on a broad collection of zero-shot OOD detection benchmarks, including CUB-200-2011 [Wah et al., 2011], Food-101 [Bossard et al., 2014], Oxford-IIIT Pet [Parkhi et al., 2012], Stanford Cars [Krause et al., 2013], and ImageNet variants [Deng et al., 2009, Wang et al., 2019, Hendrycks et al., 2021b, Recht et al., 2019, Hendrycks et al., 2021a] as ID datasets, together with iNaturalist [Horn et al., 2018], LSUN [Yu et al., 2015], Textures [Cimpoi et al., 2014], and Places [Zhou et al., 2018] as OOD datasets. Across these settings, SROT achieves competitive OOD detection performance while requiring only lightweight test-time adaptation on frozen VLM features.

Our contributions are summarized as follows:

- We identify a confidence bottleneck in zero-shot VLM OOD detection. Through a Youden/KS analysis that relates VLM confidence to a well-separated supervised reference via a Markov kernel, we explain why raw VLM confidence is likely to yield weaker ID/OOD separation than supervised-model confidence.
- We propose SROT, a label-free test-time framework that reshapes VLM posteriors with optimal transport, generates adaptive pseudo-labels from the rectified posterior, and trains a lightweight OOD detector with self-reinforced feedback.
- We conduct extensive experiments on standard zero-shot OOD detection benchmarks, where SROT achieves state-of-the-art performance. Ablation studies verify the effectiveness of posterior alignment, adaptive detector training, and self-reinforcement schema.

2 Related Work

2.1 Test-Time Adaptation

Test-time adaptation (TTA) allows deployed models to update on unlabeled test streams to mitigate distribution shift. Early approaches adapt models via self-supervised auxiliary losses [Sun et al., 2020] or by updating normalization layers via entropy minimization [Wang et al., 2021]. To address continual streams, methods incorporate anti-forgetting mechanisms [Wang et al., 2022] or perform selective, sample-efficient updates to avoid collapse [Niu et al., 2022]. Recent work further explores backpropagation-free classifier adjustments [Iwasawa and Matsuo, 2021] and stability-focused objectives [Niu et al., 2023]. While these methods primarily optimize for accuracy under distribution shift, considerations of reliability, such as calibrated uncertainty and robust confidence on unseen inputs, remain underexplored. A recent exception incorporates explicit calibration into the TTA objective [Tan et al., 2025], yet leveraging TTA primarily for OOD robustness, rather than pure accuracy, is still an open area.

2.2 Zero-Shot OOD Detection

Zero-shot OOD detection leverages pre-trained models without task-specific fine-tuning. Early post-hoc approaches, designed for supervised classifiers, derive OOD scores from model confidence [Hendrycks and Gimpel, 2017, Liang et al., 2018], energy functions [Liu et al., 2020], or feature-space distances [Lee et al., 2018]. With the advent of VLMs, zero-shot OOD detection is performed by thresholding image–text similarities to a set of ID class prompts [Radford et al., 2021]. Subsequent methods enhance this baseline by generating negative or pseudo-OOD prompts. For instance, Esmaeilpour et al. [2022] train a text generator to produce candidate unknown class names, while Wang et al. [2023] learn a negation prompt to represent the "not-ID" concept. Building on this, recent work explicitly employs negative prompts or labels to improve separability [Jiang et al., 2024, Nie et al., 2024, Ansari et al., 2025]. Alternative strategies incorporate Bayesian scoring [Fu et al., 2025], maximum concept matching [Miyai et al., 2025b], or adapt to long-tailed distributions [Miao et al., 2024]. Liu et al. [2025] also employs OT for zero-shot OOD detection. Unlike this approach, which uses OT primarily as a static scoring function, we introduce an OT-based adaptation process that establishes a self-reinforcing loop with a learnable detector, specifically designed to reshape the posterior allocation that makes raw VLM confidence poorly separable for OOD detection.

2.3 TTA for Zero-Shot OOD Detection

Recent efforts combine TTA with OOD detection, enabling models to adapt decision boundaries on the fly. Methods include online pseudo-ID/OOD mining with anti-forgetting mechanisms [Yang

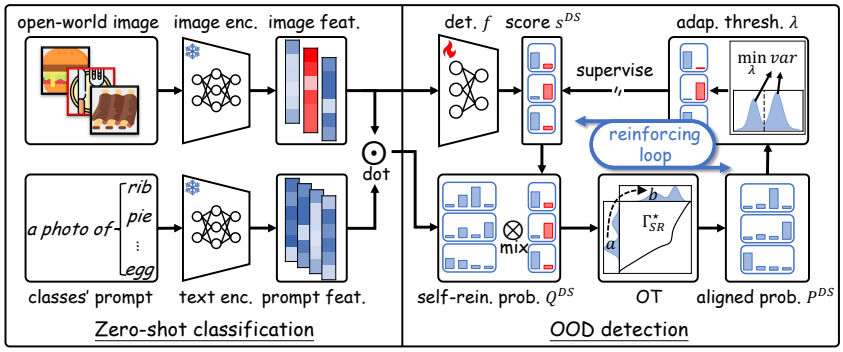


Figure 2: Overview of the proposed pipeline SROT. The method first conforms CLIP posteriors using entropic optimal transport (OT), then derives pseudo-labels from a batch of test samples to train a lightweight detector. The trained detector's OOD probabilities are subsequently used to steer the OT process, reinforcing separation between ID and OOD samples.

et al., 2023], model-free TTA that updates decision criteria from test streams [Zhang et al., 2023], and dynamic dictionaries that accumulate OOD features without fine-tuning [Yang et al., 2025]. In the zero-shot VLM setting, Cao et al. [2025] formalize zero-shot TTA and propose AdaND, which freezes the classifier and trains a detector on pseudo-labels while injecting Gaussian noise. A core challenge in unsupervised adaptation is reliance on the sub-optimal pre-trained model's posterior. Our approach directly addresses this by coupling OT-based rectification with detector training in a self-reinforcing loop, explicitly reducing the gap between the VLM and the downstream task for improved OOD separability.

3 Methodology

3.1 Problem Setup

Let $\mathcal{X}$ be the underlying image space and $\mathcal{Y} = \{1, \dots, K\}$ the set of ID classes defined by text prompts. We assume a VLM image encoder $\mathcal{E} : \mathcal{X} \to \mathbb{R}^d$ and normalized prompt embeddings $\{u_k\}_{k=1}^K$ with $\|\mathcal{E}(x)\| = \|u_k\| = 1$. At test time, unlabeled mini-batches $\mathcal{B}_t = \{x_b\}_{b=1}^B$ are sampled from a mixture of ID and OOD distribution. The model outputs a prediction $\widehat{y}(x) \in \{-1\} \cup \mathcal{Y}$, where -1 denotes OOD. With temperature $\tau > 0$, the CLIP zero-shot posterior is

$$P_{\text{clip}}(x)_k \;=\; \frac{\exp\left(\langle \mathcal{E}(x), u_k \rangle / \tau\right)}{\sum_{j=1}^K \exp\left(\langle \mathcal{E}(x), u_j \rangle / \tau\right)}. \tag{1}$$

To assess both ID recognition and OOD rejection, we report three complementary metrics following Cao et al. [2025]: Acc_S, Acc_N, and their harmonic mean Acc_H. Let

$$\text{Acc}_\text{S} = \frac{1}{|\mathcal{D}_\text{id}|} \sum_{(x_i, y_i) \in \mathcal{D}_\text{id}} \mathbb{1}\{\widehat{y}_i = y_i\}, \;\; \text{Acc}_\text{N} = \frac{1}{|\mathcal{D}_\text{ood}|} \sum_{(x_i, y_i) \in \mathcal{D}_\text{ood}} \mathbb{1}\{\widehat{y}_i = -1\}, \;\; \text{Acc}_\text{H} = 2 \cdot \frac{\text{Acc}_\text{S}\text{Acc}_\text{N}}{\text{Acc}_\text{S} + \text{Acc}_\text{N}},$$

where $\mathcal{D}$ is the evaluation set and define $\mathcal{D}_\text{id} = \{(x_i, y_i) \in \mathcal{D} : y_i \in \mathcal{Y}\}$ and $\mathcal{D}_\text{ood} = \{(x_i, y_i) \in \mathcal{D} : y_i = -1\}$, where $\mathbb{1}$ is the indicator function. For Acc_S, flagging ID samples as OOD counts as an error. Notably, while our primary focus is OOD detection, we report Acc_S because it captures a practical aspect of OOD detection: Flagging a correctly classified ID sample versus a misclassified ID sample as OOD has distinct real-world impacts. Therefore, Acc_S provides complementary insight into how well the model balances ID classification and OOD rejection.

3.2 Theoretical analysis

We study single-threshold acceptance on model's confidence when moving from a source pre-training environment s to a target environment t while keeping the model fixed. With P as a model's posterior,

for $e \in \{s,t\}$, define $C_e(x) := \max_{k\in\{1,\dots,K\}} P(x)_k \in [0,1]$. Let $Z \in \{\mathrm{ID},\mathrm{OOD}\}$ indicate membership. Accepting x as ID iff $C_e(x) \geq \theta$ (for $\theta \in (0,1)$) yields

$$\mathrm{TPR}_e(\theta) := \mathbb{P}(C_e \geq \theta \mid Z = \mathrm{ID}), \;\; \mathrm{FPR}_e(\theta) := \mathbb{P}(C_e \geq \theta \mid Z = \mathrm{OOD}), \tag{2}$$

and the single-threshold Youden/KS index

$$J_e^\star := \sup_{\theta\in(0,1)} \big[\mathrm{TPR}_e(\theta) - \mathrm{FPR}_e(\theta)\big]. \tag{3}$$

All probabilities/expectations are over the joint law of (C_s, C_t, Z) induced by the data stream.

Theorem 1 (No-improvement under confidence shift)**.** *Under the label-independence and stochastic monotonicity assumptions detailed in Appx. 1-2,*

$$J_t^\star \;\leq\; J_s^\star. \tag{4}$$

See Sec. A in the Appx. for the detailed proof.

Remarks. The theorem formalizes the confidence bottleneck of zero-shot VLMs. We view the raw VLM confidence on the target data as a Markov-kernel transformation of a better-separated reference confidence, such as the confidence produced by a supervised classifier trained on the target label space. Consequently, the best threshold over raw VLM confidence is upper-bounded by the separability of the reference confidence, and in practice can be worse when the transformation blurs the ID/OOD confidence gap. This explains the empirical degradation of zero-shot CLIP in Fig. 1: the failure is not merely due to an improperly chosen threshold, but to the raw confidence itself being a sub-optimal detection signal. Therefore, improving OOD detection requires reshaping the VLM posterior before thresholding. This motivates the OT-based rectification below, which exploits the class-wise posterior allocation rather than relying only on the maximum confidence score.

3.3 Our Method: SROT

We propose a label-free framework for zero-shot OOD detection in which OT (Optimal transport finds the minimal total cost to redistribute mass from one probability distribution to another while preserving the total amount of mass) serves as an initial posterior-rectification mechanism rather than the final detector. SROT first uses entropic OT to obtain a cleaner OOD signal from the VLM posterior, then converts this signal into adaptive pseudo-labels for training a lightweight detector on frozen image features. The detector extends the decision beyond class posterior probabilities and, in turn, provides open-set evidence to refine the OT cost.

3.3.1 Optimal-Transport Rectification

For each incoming batch we use entropic OT as a "probability redistributor" to minimally reassign CLIP posteriors toward a target class prior. The resulting OT-rectified rows serve as recalibrated confidences. This enforces an expected class marginal in a cost-aware way consistent with CLIP's similarities, improving ID/OOD separability.

With a batch $\mathcal{B}_t$ of B samples, let $Q \in [0,1]^{B\times K}$ be their CLIP posteriors, where $Q_{b,:} = P_{\text{clip}}(x_b)^\top$ is the softmax distribution over K classes for x_b. Define the cost matrix $M = -\log Q$ so that high CLIP confidence induces low transport cost. The OT plan $\Gamma \in \mathbb{R}_+^{B\times K}$ redistributes probability mass from samples (rows) to classes (columns) while matching prescribed marginals: Set sample marginal $a = \frac{1}{B}\mathbf{1}_B$ to weight each sample equally and set class marginal $b = \frac{1}{K}\mathbf{1}_K$ to impose a class-balanced prior that counteracts pre-training biases. Note that the uniform class marginal does not assume that every test batch is perfectly class-balanced. Rather, it provides a neutral batch-level constraint in the absence of label information, preventing the rectified posterior from becoming overly concentrated on a small subset of classes. When task priors are known, b can be replaced by that prior, e.g., using the estimated class frequencies in a dataset, allowing flexible transport plan. Then we solve the entropic OT problem

$$\Gamma^\star = \operatorname*{argmin}_{\Gamma\in\mathcal{U}(a,b)} \langle \Gamma, M\rangle + \varepsilon \mathrm{KL}\left(\Gamma \middle| ab^\top\right), \tag{5}$$

where $\mathcal{U}(a,b) = \Gamma \in \mathbb{R}_+^{B\times K} : \Gamma\mathbf{1}_K = a,\ \Gamma^\top\mathbf{1}_B = b$, $\varepsilon > 0$ controls entropy regularization, and KL is the Kullback–Leibler divergence. The cost term preserves CLIP's semantic structure and the

entropy term ensures a smooth, numerically stable solution. We compute $\Gamma^\star$ efficiently via Sinkhorn iterations [Cuturi, 2013], which alternates row/column scalings, adding minimal overhead.

From $\Gamma^\star$, the rectified per-sample distribution is obtained by row normalization:

$$\widetilde{P}_{b,:} = \Gamma^\star_{b,:}/a_b \in \Delta^{K-1}. \tag{6}$$

For example, if $K{=}3$ (cat, dog, car) and CLIP overpredicts "cat", excess mass accumulates on cat. With $b = (1/3, 1/3, 1/3)$, OT redistributes that excess toward dog/car where supported by the similarities, preventing cat from dominating the marginal. Genuine ID samples retain or sharpen their peaks while OOD samples remain diffuse. Thus, OT conforms CLIP's outputs with a balanced prior.

Given the rectified distribution $\widetilde{P}$, we score OOD as

$$s^{\mathrm{OT}}(x_b) = 1 - \max_{k\in[K]} \widetilde{P}_{b,k}, \tag{7}$$

which increases the confidence gap between ID (sharper maxima) and OOD (diffuse) samples and provides a robust basis for downstream detector training.

3.3.2 Training Detector with Pseudo-Labels

OT rectification improves the posterior-level OOD signal, but it remains limited to class probabilities induced by frozen VLM. To go beyond this posterior-only view, we train a lightweight detector on frozen image features using pseudo-labels derived from the OT-rectified scores. This allows SROT to capture feature-level ID/OOD regularities while remaining fully label-free.

Let $\mathcal{E}(x)$ be the image embedding and let detector $f : \mathbb{R}^d \to \mathbb{R}^2$ (a linear layer) output logits. Define

$$\big(p_{\mathrm{ood}}(x),\, p_{\mathrm{id}}(x)\big) = \mathrm{softmax}\big(f(\mathcal{E}(x))\big), \tag{8}$$

where $p_{\mathrm{ood}}(x)$ and $p_{\mathrm{id}}(x)$ represent the probabilities of x being an OOD and ID sample, respectively.

The pseudo-labels employed to train this detector are generated by thresholding the OT-rectified OOD scores, as defined in Eq. (7). To ensure stable threshold selection, we maintain a sliding window $\mathcal{W}_t$ of OOD scores from recent samples, and adaptively determine the optimal threshold λ for partitioning at each time step t.

$$\mathcal{I}_t(\lambda) = \{x \in \mathcal{W}_t :\ s_t(x) \le \lambda\},\ \ \mathcal{O}_t(\lambda) = \{x \in \mathcal{W}_t :\ s_t(x) > \lambda\}. \tag{9}$$

where $s_t(\cdot)$ is the current OOD score (initially from Eq. (7)). We select λ_t by minimizing the sum of within-set score variances (akin to Otsu's criterion):

$$\min_{\lambda}\ \frac{1}{|\mathcal{I}_t(\lambda)|}\sum_{x\in\mathcal{I}_t(\lambda)}\big[s_t(x) - \bar{s}_{\mathcal{I}_t(\lambda)}\big]^2 + \frac{1}{|\mathcal{O}_t(\lambda)|}\sum_{x\in\mathcal{O}_t(\lambda)}\big[s_t(x) - \bar{s}_{\mathcal{O}_t(\lambda)}\big]^2, \tag{10}$$

where $\bar{s}_{\mathcal{S}} = \frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}} s_t(x)$. Notably, we only need to record a single scalar (the OOD score) per sample, introducing negligible computational overhead.

With threshold λ, assign binary pseudo-labels (1 denotes OOD and 0 denotes ID) on the current mini-batch $\mathcal{B}_t$ as

$$\tilde{y}(x_b) = \mathbf{1}\big[s_t(x_b) > \lambda_t\big] \in \{0, 1\}. \tag{11}$$

Finally, train f by cross-entropy (CE) on $\mathcal{B}_t$ while keeping the VLM frozen:

$$\mathcal{L}_{\text{det}} = \frac{1}{B}\sum_{x_b\in\mathcal{B}_t} \mathrm{CE}\Big(\mathrm{softmax}(f(\mathcal{E}(x_b))),\, \tilde{y}(x_b)\Big). \tag{12}$$

3.3.3 Reinforcing OT with Open-Set Prior and Detector's OOD Scores

The initial OT rectification is agnostic to feature-level OOD regularities. After the detector stabilizes on confident ID cases, we steer OT using its OOD beliefs and a simple open-set prior. Revise the cost matrix constructed in Sec. 3.3.1 to be a convex combination of the CLIP posterior and an open-set prior $\boldsymbol{\pi} \in \Delta^{K-1}$:

$$Q^{\mathrm{DS}}_{b,k} = \big(1 - p_{\mathrm{ood}}(x_b)\big)\, P_{\mathrm{clip}}(x_b)_k + p_{\mathrm{ood}}(x_b)\, \pi_k, \text{ where } M^{\mathrm{DS}} = -\log Q^{\mathrm{DS}}. \tag{13}$$

Table 1: Zero-shot TTA for OOD detection results for CUB-200-2011, STANFORD-CARS, Food-101, and Oxford-IIIT Pet as the ID datasets. The **bold** indicates the best performance on each dataset.

ID	Method	iNaturalist			LSUN			Texture			Places			Avg.		
		$\mathrm{Acc_S}$	$\mathrm{Acc_N}$	$\mathrm{Acc_H}$	$\mathrm{Acc_S}$	$\mathrm{Acc_N}$	$\mathrm{Acc_H}$	$\mathrm{Acc_S}$	$\mathrm{Acc_N}$	$\mathrm{Acc_H}$	$\mathrm{Acc_S}$	$\mathrm{Acc_N}$	$\mathrm{Acc_H}$	$\mathrm{Acc_S}$	$\mathrm{Acc_N}$	$\mathrm{Acc_H}$
CUB-200-2011	ZS-CLIP	38.13	88.06	53.22	38.10	87.86	53.15	37.56	79.11	50.94	38.00	87.81	53.04	37.95	85.71	52.59
	Tent	37.02	46.95	41.40	38.61	55.55	45.56	34.98	41.77	38.07	40.41	74.83	52.48	37.75	54.78	44.38
	SoTTA	41.67	84.37	55.79	42.08	86.83	56.69	41.44	77.58	54.02	42.04	86.52	56.59	41.81	83.82	55.77
	TPT	37.41	89.57	52.78	37.49	89.67	52.87	36.88	81.67	50.81	37.44	89.45	52.79	37.30	87.59	52.31
	AdaND	52.34	96.40	67.84	52.41	93.91	67.27	51.82	81.24	63.28	51.82	91.51	66.17	52.10	90.77	66.14
	SROT	**54.36**	**98.63**	**70.09**	**54.38**	**98.31**	**70.03**	**54.59**	**99.16**	**70.41**	**54.43**	**96.17**	**69.52**	**54.44**	**98.07**	**70.01**
STANFORD-CARS	ZS-CLIP	50.18	96.62	66.05	53.48	98.81	69.40	53.59	99.05	69.55	53.36	98.05	69.11	52.65	98.13	68.53
	Tent	44.12	52.33	47.88	54.27	94.51	68.95	54.60	97.37	69.97	54.33	96.65	69.56	51.83	85.22	64.09
	SoTTA	51.51	92.84	66.26	54.81	97.57	70.19	55.06	98.50	70.64	54.75	96.96	69.98	54.03	96.47	69.27
	TPT	49.24	96.97	65.31	52.40	98.83	68.49	52.75	99.27	68.89	52.42	98.39	68.40	51.70	98.36	67.77
	AdaND	62.80	99.79	77.09	62.73	99.82	77.04	62.91	99.75	77.16	62.76	99.29	76.91	62.80	99.66	77.05
	SROT	**63.84**	**99.93**	**77.91**	**63.82**	**99.91**	**77.89**	**63.82**	**99.95**	**77.90**	**63.82**	**99.49**	**77.76**	**63.82**	**99.82**	**77.87**
Food-101	ZS-CLIP	80.60	94.76	87.11	80.75	96.08	87.75	80.51	93.12	86.36	80.62	94.62	87.06	80.62	94.65	87.07
	Tent	75.83	25.09	37.70	82.86	85.10	83.97	82.54	87.03	84.73	82.26	80.13	81.18	80.87	69.34	71.90
	SoTTA	81.84	84.09	82.95	82.49	93.34	87.58	82.05	90.10	85.89	82.44	91.62	86.79	82.20	89.79	85.80
	TPT	79.70	94.93	86.65	79.92	96.19	87.30	79.70	93.86	86.20	79.76	95.14	86.77	79.77	95.03	86.73
	AdaND	**86.50**	99.87	**92.71**	**86.40**	99.64	**92.55**	**86.44**	96.51	91.20	**86.42**	99.40	92.46	**86.44**	98.85	92.23
	SROT	86.40	**99.92**	92.67	86.38	**99.66**	**92.55**	86.33	**97.71**	**91.67**	86.38	**99.48**	**92.47**	86.37	**99.19**	**92.34**
Oxford-IIIT Pet	ZS-CLIP	78.58	88.30	83.16	79.75	87.30	83.35	80.20	91.16	85.33	79.59	84.17	81.82	79.53	87.73	83.41
	Tent	80.07	78.09	79.07	81.19	68.30	74.19	81.48	74.72	77.95	80.64	62.51	70.43	80.84	70.91	75.41
	SoTTA	80.07	83.54	81.77	81.78	83.83	82.79	82.09	87.52	84.72	81.49	81.25	81.37	81.36	84.03	82.66
	TPT	77.56	89.71	83.19	78.87	89.82	83.99	79.17	92.26	85.22	78.62	87.32	82.74	78.56	89.78	83.78
	AdaND	85.81	98.78	91.84	85.82	98.19	91.59	85.86	98.68	91.82	85.88	96.58	90.92	85.84	98.06	91.54
	SROT	**86.31**	**98.87**	**92.16**	**86.26**	**98.57**	**92.01**	**86.31**	**98.98**	**92.21**	**86.31**	**97.54**	**91.58**	**86.30**	**98.49**	**91.99**

If classwise OOD propensities $\boldsymbol{\rho} \in \mathbb{R}^K$ are available, set $\pi_k = \rho_k / \sum_{j=1}^{K} \rho_j$. By default, we use the non-informative prior $\pi_k = \frac{1}{K}$. Then replace M in Eq. (5) with M^{DS}:

$$\Gamma^{\star}_{\mathrm{DS}} = \underset{\Gamma \in \mathcal{U}(a,b)}{\operatorname{argmin}} \langle \Gamma, M^{\mathrm{DS}} \rangle + \varepsilon \, \mathrm{KL}\big(\Gamma \,\|\, ab^{\top}\big), \tag{14}$$

This obtains $\widetilde{P}^{\mathrm{DS}}$, get the new OOD scores $s^{\mathrm{DS}}(x_b) = 1 - \max_{k \in [K]} \widetilde{P}^{\mathrm{DS}}_{b,k}$, and refresh λ_t and pseudo-labels on the next window in every epoch. This closes a feedback loop: OT balances the VLM posteriors to supply trustworthy guidance for detector training, while the detector, which learns the underlying feature-OOD relationship, introduces adaptive open-set feature knowledge to refine the OT cost matrix, forming a mutually beneficial iterative optimization loop.

3.3.4 Final Detection and Prediction

At inference, we declare a sample OOD if $p_{\mathrm{ood}}(x) > \lambda_{\mathrm{det}}$ with p_{ood} and λ_{det} computed using Eq. (8) and Eq. (10), otherwise we make ID classification with the frozen CLIP head: $\arg\max_k P_{\mathrm{clip}}(x)_k$.

4 Experiments

4.1 Experimental Settings

Datasets We evaluate SROT in the TTA setting for zero-shot OOD detection. Following prior work, we treat each of the following datasets in turn as ID: CUB-200-2011 [Wah et al., 2011], Stanford Cars [Krause et al., 2013], Food-101 [Bossard et al., 2014], Oxford-IIIT Pet [Parkhi et al., 2012], ImageNet [Deng et al., 2009], ImageNet-Sketch (ImageNet-K) [Wang et al., 2019], ImageNet-A [Hendrycks et al., 2021b], ImageNet-V2 [Recht et al., 2019], and ImageNet-R [Hendrycks et al., 2021a]. For OOD, we use four standard sources: iNaturalist [Horn et al., 2018], LSUN [Yu et al., 2015], Textures [Cimpoi et al., 2014], and Places [Zhou et al., 2018]. Our primary metrics are $\mathrm{Acc_S}$, $\mathrm{Acc_N}$, and $\mathrm{Acc_H}$ as introduced in Sec. 3. These jointly expose any closed-set trade-offs induced by adaptation. Unless otherwise noted, all methods share identical preprocessing, batching, and evaluation code for comparability.

Backbone and hyperparameters We adopt CLIP equipped with a ViT-B/16 image encoder and a Transformer-based text encoder [Radford et al., 2021]. Both encoders remain frozen. Adaptation is restricted to a lightweight linear head for OOD scoring, optimized online with Adam (learning rate 5×10^{-4}, weight decay 0). We stabilize updates by injecting small stochastic perturbations at a

Table 2: Zero-shot TTA for OOD detection results for ImageNet and its variants as the ID datasets. The **bold** indicates the best performance on each dataset.

ID	Method	iNaturalist			LSUN			Texture			Places			Avg.		
		$\mathrm{Acc_S}$	$\mathrm{Acc_N}$	$\mathrm{Acc_H}$	$\mathrm{Acc_S}$	$\mathrm{Acc_N}$	$\mathrm{Acc_H}$	$\mathrm{Acc_S}$	$\mathrm{Acc_N}$	$\mathrm{Acc_H}$	$\mathrm{Acc_S}$	$\mathrm{Acc_N}$	$\mathrm{Acc_H}$	$\mathrm{Acc_S}$	$\mathrm{Acc_N}$	$\mathrm{Acc_H}$
ImageNet	ZS-CLIP	54.01	86.53	66.51	53.43	83.96	65.30	52.71	78.52	63.08	53.35	80.50	64.17	53.38	82.38	64.77
	Tent	48.56	35.74	41.18	55.44	75.54	63.95	54.94	70.93	61.92	55.76	73.98	63.59	53.67	64.05	57.66
	SoTTA	53.15	62.68	57.52	53.16	68.76	59.96	53.64	68.05	59.99	53.60	69.16	60.39	53.39	67.16	59.47
	TPT	52.58	88.93	66.09	51.91	86.09	64.77	51.11	80.01	62.38	51.80	82.89	63.76	51.85	84.48	64.25
	AdaND	63.26	96.87	76.54	61.34	89.44	72.77	62.45	83.54	71.47	61.92	84.82	71.58	62.24	88.67	73.09
	SROT	**66.02**	**99.10**	**79.25**	**64.69**	**95.66**	**77.18**	**65.74**	**86.24**	**74.61**	**64.74**	**85.07**	**73.53**	**65.30**	**91.52**	**76.14**
ImageNet-K	ZS-CLIP	34.17	83.46	48.49	33.46	81.20	47.39	32.61	75.57	45.56	33.40	77.10	46.61	33.41	79.33	47.01
	Tent	30.46	26.86	28.55	36.57	71.82	48.46	36.37	66.63	47.06	36.87	70.32	48.38	35.07	58.91	43.11
	SoTTA	36.18	61.70	45.61	36.28	67.19	47.12	35.91	65.31	46.34	36.57	67.09	47.34	36.23	65.32	46.60
	TPT	32.16	86.52	46.89	31.55	83.86	45.85	30.74	77.39	44.00	31.56	80.05	45.27	31.50	81.95	45.50
	AdaND	40.97	93.54	56.98	40.25	85.06	54.64	38.31	74.43	50.58	39.60	79.57	52.88	39.78	83.15	53.77
	SROT	**45.91**	**99.80**	**62.89**	**45.67**	**99.03**	**62.51**	**45.10**	**89.53**	**59.98**	**45.39**	**89.76**	**60.29**	**45.52**	**94.53**	**61.42**
ImageNet-A	ZS-CLIP	34.73	80.69	48.56	34.20	78.83	47.70	33.97	76.60	47.07	33.96	75.11	46.77	34.22	77.81	47.53
	Tent	34.99	77.19	48.15	34.83	77.05	47.97	34.36	75.19	47.17	34.60	73.83	47.12	34.70	75.81	47.60
	SoTTA	36.85	76.83	49.81	36.47	77.08	49.51	35.60	75.37	48.36	36.07	73.87	48.47	36.25	75.79	49.04
	TPT	34.12	81.17	48.04	33.20	80.23	46.97	33.12	79.92	46.83	33.05	77.00	46.25	33.37	79.58	47.02
	AdaND	43.59	91.19	58.98	41.96	80.93	55.27	45.04	79.97	57.62	42.85	72.13	53.76	43.36	81.06	56.41
	SROT	**43.73**	**97.40**	**60.36**	**42.13**	**90.96**	**57.59**	**45.72**	**87.96**	**60.17**	**43.47**	**81.71**	**56.75**	**43.76**	**89.51**	**58.72**
ImageNet-V2	ZS-CLIP	48.01	85.72	61.55	47.37	83.23	60.38	46.81	77.54	58.38	47.39	79.41	59.36	47.39	81.47	59.92
	Tent	47.94	76.98	59.08	48.28	80.50	60.36	47.56	74.47	58.05	48.34	77.37	59.50	48.03	77.33	59.25
	SoTTA	48.24	78.59	59.78	47.80	78.67	59.47	47.27	74.82	57.94	48.26	76.05	59.05	47.89	77.03	59.06
	TPT	46.63	88.37	61.05	46.12	85.58	59.94	45.21	79.14	57.55	46.02	81.95	58.94	46.00	83.76	59.37
	AdaND	56.32	97.06	71.28	54.78	86.64	67.12	57.28	80.61	66.97	55.81	79.24	65.49	56.05	85.89	67.72
	SROT	**58.10**	**97.68**	**72.86**	**57.22**	**89.93**	**69.94**	**59.10**	**84.10**	**69.42**	**58.34**	**76.18**	**66.08**	**58.19**	**86.97**	**69.58**
ImageNet-R	ZS-CLIP	61.99	94.39	74.83	61.82	88.95	72.94	60.91	77.05	68.04	61.68	84.86	71.44	61.60	86.31	71.81
	Tent	65.22	91.45	76.14	65.06	85.61	73.93	63.33	69.99	66.49	64.93	82.38	72.62	64.64	82.36	72.30
	SoTTA	66.78	86.98	75.55	66.71	83.99	74.36	65.92	72.69	69.14	66.60	80.53	72.91	66.50	81.05	72.99
	TPT	60.95	94.80	74.20	60.85	89.98	72.60	59.98	77.79	67.73	60.67	85.79	71.08	60.61	87.09	71.40
	AdaND	72.21	99.59	83.72	71.02	95.94	81.62	70.44	81.43	75.54	70.85	92.14	80.10	71.13	92.28	80.25
	SROT	**73.18**	**99.77**	**84.43**	**72.15**	**98.84**	**83.41**	**71.88**	**94.70**	**81.73**	**72.07**	**94.54**	**81.79**	**72.32**	**96.96**	**82.84**

fixed cadence following Cao et al. [2025]. This design yields little computational overhead and stable behavior under shift. See the Appx. E for details.

4.1.1 Compared Methods

We conduct a comprehensive comparison between our SROT and a suite of representative approaches: ZS-CLIP (zero-shot inference without adaptation) [Radford et al., 2021], TENT (entropy-minimization TTA) [Wang et al., 2021], SoTTA (self-training TTA for zero-shot VLMs) [Gong et al., 2023], TPT (test-time prompt tuning) [Shu et al., 2022], and AdaND (adaptive noise detection with online thresholding) [Cao et al., 2025]. For comparisons on classic OOD detection metrics, we compare with MaxLogit [Hendrycks et al., 2019], Energy [Liu et al., 2020], MCM [Ming et al., 2022], CLIPN [Wang et al., 2023], LAPT [Zhang et al., 2024], AdaNeg [Zhang and Zhang, 2024], and NegLabel [Jiang et al., 2024]. Hyperparameters for baselines follow official settings or faithful re-implementations.

4.2 Main Results

Results for fine-grained IDs (CUB-200-2011, Stanford Cars, Food-101, Oxford-IIIT Pet) and ImageNet-family IDs are summarized in Tables 1 and 2, respectively. Across all nine ID datasets, our method improves $\mathrm{Acc_H}$ over the strongest prior TTA baseline (AdaND) while simultaneously increasing both $\mathrm{Acc_S}$ and $\mathrm{Acc_N}$. Averaged over all IDs, we obtain a $+2.52$ absolute gain in $\mathrm{Acc_H}$, accompanied by $+1.81$ in $\mathrm{Acc_S}$ and $+4.07$ in $\mathrm{Acc_N}$. Entropy-minimization TTA tends to exhibit brittleness when confronted with severe distribution shift, while prompt-only parameter updates frequently lead to insufficient correction of model calibration. By associating the rectification strength with the p_{ood}, SROT regularizes pseudo-labels and counters overconfident OOD assignments, consistently improving $\mathrm{Acc_N}$ without degrading $\mathrm{Acc_S}$. Overall, the evidence supports self-reinforced OT for OOD detection as a principled and efficient route: It maintains zero-shot generalization and delivers sizable gains where distribution gap is substantive.

The one compressed margin appears on Food-101, where our average $\mathrm{Acc_H}$ is within 0.1–0.4 of AdaND. We attribute this small margin to saturation, as Food-101 is comparatively easy for CLIP-style zero-shot recognition (highest $\mathrm{Acc_S}$), leaving limited headroom.

Table 3: Zero-shot OOD detection results for ImageNet as the ID dataset. The **bold** indicates the best performance on each dataset.

Method	iNaturalist		LSUN		Texture		Places		Avg.	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Max-Logit	89.31	61.66	87.43	64.39	71.68	86.61	85.95	63.67	83.59	69.08
Energy	85.09	81.08	84.24	79.02	65.56	93.65	83.38	75.08	79.57	82.21
MCM	94.61	30.91	92.57	37.59	86.11	57.77	89.77	44.69	90.77	42.74
CLIPN	95.27	23.94	93.93	26.17	90.93	40.83	92.28	33.45	93.10	31.10
NegLabel	99.49	1.91	95.49	20.53	90.22	43.56	91.64	35.59	94.21	25.40
LAPT	99.63	1.16	96.01	19.12	92.01	33.01	91.06	40.32	94.68	23.40
AdaNeg	99.71	**0.59**	97.44	9.50	94.55	34.34	94.93	31.27	96.66	18.92
AdaND	98.91	4.19	95.86	17.08	93.01	21.76	94.55	20.95	95.58	16.00
SROT	**99.86**	**0.59**	**99.02**	**3.72**	**97.97**	**9.49**	**96.75**	**12.60**	**98.40**	**6.60**

To further validate SROT's robustness, we analyze classic OOD detection metrics, AUROC and FPR95, in Table 3. SROT establishes clear state-of-the-art performance across all OOD datasets, demonstrating a consistent pattern of superior discriminative power and false positive control. It outperforms all baselines by meaningful margins in both metrics, with the most striking improvement seen in FPR95: SROT drastically reduces false positives compared to even the strongest prior baselines, a critical advantage for safety-sensitive applications where erroneous OOD acceptance carries risks. Notably, SROT maintains this lead across diverse OOD types from fine-grained natural samples (iNaturalist) to texture and scene datasets, proving its generalizability beyond specific data distributions. Unlike baselines that often trade off AUROC for FPR95 or vice versa, SROT achieves both higher overall detection accuracy and stricter false positive control, a reflection of its self-reinforced OT rectification effectively sharpening ID/OOD separability without compromising calibration. These results confirm that SROT's core mechanism delivers not just balanced performance but also robust, reliable OOD scoring that scales across varied out-of-distribution scenarios.

4.3 Ablation Study

4.3.1 Components

The ablation results in Table 4 demonstrate that both OT rectification and the self-reinforced cost are essential and mutually complementary for achieving robust TTA on CUB-200-2011. Removing OT leads to the largest performance drop, confirming its role as the primary source of noise robustness: reintroducing OT (row "W/o self-reinforced OT") increases $\mathrm{Acc_H}$ by +2.75 (from 66.37 to 69.12). Adding the self-reinforced cost on top of OT yields an additional gain of +0.95 (69.12 $\rightarrow$ 70.07), indicating that coupling the OT cost with the detector's OOD estimates further improves calibration without compromising clean accuracy. Overall, these results demonstrate that OT and the self-reinforcement mechanism jointly contribute to model robustness.

Table 4: Ablation studies for using CUB-200-2011 as the ID dataset. Results are averaged over four OOD datasets.

Method	$\mathrm{Acc_S}$	$\mathrm{Acc_N}$	$\mathrm{Acc_H}$
W/o OT	52.23	91.16	66.37
W/o self-reinforced OT	53.65	97.15	69.12
SROT	54.51	98.07	70.07

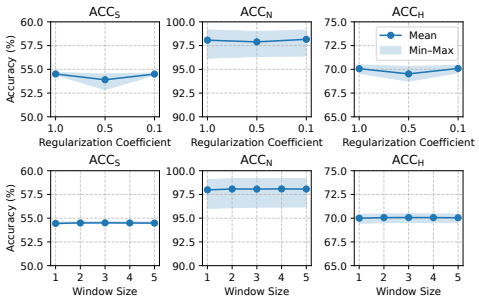


Figure 3: Sensitivity to OT regularization and window size on CUB-200-2011 with four OOD datasets.

4.3.2 Hyperparameters

SROT involves minimal hyperparameters and shows low sensitivity to its key hyperparameters, as seen in Fig. 3. Varying the window size from 1 to 5 results in a narrow $\mathrm{Acc_H}$ range of 69.46–70.07, with a size of 3 yielding peak performance and improved adaptation stability. We find that a window size greater than 1 stabilizes adaptation, so we retain this schema. The OT regularization coefficient exhibits a similarly confined $\mathrm{Acc_H}$ range of 69.52–70.09.

5 Conclusion

This work investigates the challenge of out-of-distribution detection in zero-shot vision–language models. We demonstrate both theoretically and empirically that conventional OOD detection methods, primarily designed for supervised models, fail to generalize effectively to pre-trained VLMs under zero-shot settings. To address, we introduce SROT, a novel TTA framework designed to recalibrate VLM posteriors and amplify ID/OOD separability in a label-free manner. Our approach establishes a mutually reinforcing loop between optimal transport-based posterior rectification and a lightweight detector, enabling iterative improvement without updating the backbone. Through extensive experiments on standard benchmarks, SROT consistently achieves state-of-the-art performance, improving both detection accuracy and false positive control across diverse OOD scenarios. Beyond detection metrics, this work underscores the importance of moving beyond accuracy-centric evaluation of pre-trained models toward a more rigorous assessment of their robustness in downstream tasks. By explicitly addressing the confidence miscalibration inherent in web-trained VLMs, SROT provides a pathway toward more reliable and trustworthy deployment in open-world applications.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62502233, 62602293, and 62602397; the Basic Research Program of Jiangsu Province under Grant No. BK20250650 and BK20250683; the Jiangsu Province Youth Science and Technology Talent Support Project under Grant No. JSTJ-2025-641; the Fundamental Research (Natural Science) Project of Jiangsu Higher Education Institutions under Grant Nos. 25KJB520033 and 25KJB520040; the Natural Science Research Startup Foundation for Recruited Talents of Nanjing University of Posts and Telecommunications under Grant No. NY224061; the Natural Science Foundation of Nanjing University of Posts and Telecommunications under Grant Nos. NY224058 and NY225124; the Shandong Provincial Natural Science Foundation under Grant Nos. ZR2025QC1570 and ZR2026LLX017; and the Open Project of the State Key Laboratory for Novel Software Technology, Nanjing University, under Grant No. KFKT2025B17 and KFKT2025B68.

Competing interests. The authors declare no competing interests.

References

Amirhossein Ansari, Ke Wang, and Pulei Xiong. Negrefine: Refining negative label-based zero-shot ood detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 573–582, 2025.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, volume 8694 of *Lecture Notes in Computer Science*, pages 446–461. Springer, 2014. doi: 10.1007/978-3-319-10599-4_29.

Chentao Cao, Zhun Zhong, Zhanke Zhou, Tongliang Liu, Yang Liu, Kun Zhang, and Bo Han. Noisy test-time adaptation in vision-language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 3606–3613. IEEE Computer Society, 2014. doi: 10.1109/CVPR.2014.461.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2292–2300, 2013.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model CLIP. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 6568–6576. AAAI Press, 2022. doi: 10.1609/AAAI.V36I6.20610.

Hao Fu, Naman Patel, Prashanth Krishnamurthy, and Farshad Khorrami. Clipscope: Enhancing zero-shot ood detection with bayesian scoring. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5346–5355. IEEE, 2025.

Taesik Gong, Yewon Kim, Taeckyung Lee, Sorn Chottananurak, and Sung-Ju Lee. Sotta: Robust test-time adaptation on noisy data streams. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019.

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 *2021*

IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pages 8320–8329. IEEE, 2021a. doi: 10.1109/ICCV48922.2021.00823.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 15262–15271. Computer Vision Foundation / IEEE, 2021b. doi: 10.1109/CVPR46437.2021.01501.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist species classification and detection dataset. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8769–8778. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00914.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems*, 2021.

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

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pages 554–561. IEEE Computer Society, 2013. doi: 10.1109/ICCVW.2013.77.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 2018.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2020.

Yu Liu, Hao Tang, Haiqi Zhang, Jing Qin, and Zechao Li. Ot-detector: delving into optimal transport for zero-shot out-of-distribution detection. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25, 2025. ISBN 978-1-956792-06-5. doi: 10.24963/ijcai.2025/184.

Wenjun Miao, Guansong Pang, Jin Zheng, and Xiao Bai. Long-tailed out-of-distribution detection via normalized outlier distribution adaptation. *Advances in Neural Information Processing Systems*, 37:132106–132132, 2024.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems*, 2022.

Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Yueqian Lin, Qing Yu, Go Irie, Shafiq Joty, Yixuan Li, Hai Helen Li, Ziwei Liu, Toshihiko Yamasaki, and Kiyoharu Aizawa. Generalized out-of-distribution detection and beyond in vision language model era: A survey. *Trans. Mach. Learn. Res.*, 2025, 2025a.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Gl-mcm: Global and local maximum concept matching for zero-shot out-of-distribution detection. *International Journal of Computer Vision*, 133(6):3586–3596, 2025b.

Khan Muhammad, Amin Ullah, Jaime Lloret, Javier Del Ser, and Victor Hugo C De Albuquerque. Deep learning for safe autonomous driving: Current challenges and future directions. *IEEE Transactions on Intelligent Transportation Systems*, 22(7):4316–4336, 2020.

Jun Nie, Yonggang Zhang, Zhen Fang, Tongliang Liu, Bo Han, and Xinmei Tian. Out-of-distribution detection with negative prompts. In *The twelfth international conference on learning representations*, 2024.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning*, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 3498–3505. IEEE Computer Society, 2012. doi: 10.1109/CVPR.2012.6248092.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Khansa Rasheed, Adnan Qayyum, Mohamed Ghaly, Ala I. Al-Fuqaha, Adeel Razi, and Junaid Qadir. Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Comput. Biol. Medicine*, 149:106043, 2022. doi: 10.1016/J.COMPBIOMED.2022.106043.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR, 2019.

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.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9229–9248. PMLR, 2020.

Mingkui Tan, Guohao Chen, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Peilin Zhao, and Shuaicheng Niu. Uncertainty-calibrated test-time model adaptation without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(8):6274–6289, 2025.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.

Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. CLIPN for zero-shot OOD detection: Teaching CLIP to say no. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 1802–1812. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00173.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual Test-Time Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7191–7201, 2022.

Robert C. Williamson and Zac Cranko. Information processing equalities and the information-risk bridge. *J. Mach. Learn. Res.*, 25:103:1–103:53, 2024.

Puning Yang, Jian Liang, Jie Cao, and Ran He. AUTO: Adaptive outlier optimization for online test-time OOD detection. *arXiv preprint arXiv:2303.12267*, 2023.

Yifeng Yang, Lin Zhu, Zewen Sun, Hengyu Liu, Qinying Gu, and Nanyang Ye. OODD: test-time out-of-distribution detection with dynamic dictionary. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 30630–30639. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02852.

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

Yabin Zhang and Lei Zhang. Adaneg: Adaptive negative proxy guided OOD detection with vision-language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL `http://papers.nips.cc/paper_files/paper/2024/hash/4462db5eee6823b2abad0d1f955e187a-Abstract-Conference.html`.

Yabin Zhang, Wenjie Zhu, Chenhang He, and Lei Zhang. LAPT: label-driven automated prompt tuning for OOD detection with vision-language models. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXII*, volume 15130 of *Lecture Notes in Computer Science*, pages 271–288. Springer, 2024. doi: 10.1007/978-3-031-73220-1_16. URL `https://doi.org/10.1007/978-3-031-73220-1_16`.

Yifan Zhang, Xue Wang, Tian Zhou, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Model-free test-time adaptation for out-of-distribution detection. *arXiv preprint arXiv:2311.16420*, 2023.

Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6): 1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009.

A Theoretical Analysis

Let $C_s, C_t \in [0,1]$ denote the model's *confidence* in the source (s) and target (t) environments, respectively. Let $Z \in \{\mathrm{ID}, \mathrm{OOD}\}$ indicate in-/out-of-distribution membership (For simplicity, we refer to it directly as a label in the appendix). All probabilities/expectations are over the joint law of (C_s, C_t, Z).

For $\theta \in (0,1)$ define source/target true- and false-positive rates:

$$\mathrm{TPR}_e(\theta) := \mathbb{P}(C_e \geq \theta \mid Z = \mathrm{ID}), \;\; \mathrm{FPR}_e(\theta) := \mathbb{P}(C_e \geq \theta \mid Z = \mathrm{OOD}), \; e \in \{s,t\}. \tag{15}$$

The single-threshold Youden's J index is

$$J_e^\star \;:=\; \sup_{\theta \in (0,1)} \Big[\mathrm{TPR}_e(\theta) - \mathrm{FPR}_e(\theta)\Big]. \tag{16}$$

Assumption 1 (Label-independence given source confidence)**.**

$$C_t \;\perp\!\!\!\perp\; Z \,\big|\, C_s.$$

Equivalently, for every Borel $B \subseteq [0,1]$ there exists a (Markov) kernel

$$R(u, B) := \mathbb{P}(C_t \in B \mid C_s = u)$$

that does not depend on Z.

The label-independence assumption implies that the target environment's confidence C_t depends on the in-distribution/out-of-distribution label Z only through the source confidence C_s. This implies that the shift alters the marginal distribution of confidences but preserves the conditional calibration of the target relative to the source.

Assumption 2 (Stochastic monotonicity)**.** *For each $\theta \in (0,1)$, the* conditional survival function

$$h_\theta(u) := \mathbb{P}(C_t \geq \theta \mid C_s = u) = R(u, [\theta, 1])$$

is nondecreasing in $u \in [0,1]$.

The stochastic monotonicity assumption requires that the probability of the target confidence C_t exceeding a threshold θ increases (or remains constant) as the source confidence C_s increases. This assumption holds when the source and target environments are positively correlated. Specifically, samples with high source confidence retain high confidence in the target environment after adaptation or shifting.

It is worth noting that these assumptions do not imply that raw VLM confidence is well calibrated or sufficiently separable for OOD detection. The label-independence assumption only states that, once the reference confidence C_s is fixed, the target confidence C_t is generated through the same transition mechanism for ID and OOD samples. The stochastic monotonicity assumption further requires that higher reference confidence tends to produce higher target confidence in expectation. Together, these assumptions preserve a weak population-level ordering between confidence signals, but they do not prevent substantial overlap between ID and OOD confidence distributions. Indeed, because the Markov kernel $R(u, B)$ maps confidence values probabilistically rather than deterministically, it naturally allows sample-level fluctuations, distributional spreading, and confidence-gap blurring. Thus, the assumptions are compatible with the empirical behavior of zero-shot VLMs: raw confidence may remain positively related to a better-separated reference signal while still being too distorted to support strong single-threshold OOD detection.

These assumptions may not hold under all distribution shifts. In particular, label-independence may be violated when ID and OOD samples with the same source confidence exhibit different target confidence distributions. Similarly, stochastic monotonicity may fail when the shift disrupts the ordering of confidence scores, such that higher source confidence does not consistently correspond to stochastically higher target confidence. Therefore, the theoretical results should be interpreted as conditional guarantees under an idealized, label-independent, and order-preserving confidence transfer mechanism, rather than as universal guarantees across arbitrary distribution shifts.

Lemma A.1 (Variational KS characterization)**.** *Let μ, ν be probability measures on $[0,1]$ and let*

$$\mathcal{M} \;:=\; \{\, f : [0,1] \to [0,1] \text{ measurable, nondecreasing} \,\}.$$

Define

$$A \;:=\; \sup_{\theta\in[0,1]} \big[\mu([\theta,1]) - \nu([\theta,1])\big], \qquad B \;:=\; \sup_{f\in\mathcal{M}} \big(\mathbb{E}_\mu f - \mathbb{E}_\nu f\big).$$

Then $A = B$. Moreover, the supremum over $\mathcal{M}$ is achieved by threshold indicators $f_\theta(u) = \mathbf{1}\{u \geq \theta\}$ whenever $\theta \mapsto \mu([\theta,1]) - \nu([\theta,1])$ attains its supremum; in general there exists a sequence of thresholds approaching the supremum.

Proof. Step 1 (Subset inequality $A \leq B$). For each θ, the indicator $f_\theta(u) = \mathbf{1}\{u \geq \theta\}$ lies in $\mathcal{M}$. Hence

$$B \;=\; \sup_{f\in\mathcal{M}} (\mathbb{E}_\mu f - \mathbb{E}_\nu f) \;\geq\; \sup_{\theta\in[0,1]} (\mathbb{E}_\mu f_\theta - \mathbb{E}_\nu f_\theta) \;=\; \sup_{\theta\in[0,1]} \big[\mu([\theta,1]) - \nu([\theta,1])\big] \;=\; A.$$

Step 2 (Layer-cake inequality $B \leq A$). Fix $f \in \mathcal{M}$. Define its Lebesgue–Stieltjes measure α by $\alpha([a,b]) := f(b) - f(a)$ for $0 \leq a < b \leq 1$. Since $0 \leq f \leq 1$ and f is nondecreasing, α is a finite positive measure with total mass $\alpha([0,1]) = f(1) - f(0) \leq 1$. The layer-cake (Choquet) representation gives, for all $u \in [0,1]$,

$$f(u) \;=\; f(0) + \int_{[0,1]} \mathbf{1}\{u \geq \theta\}\, d\alpha(\theta).$$

Therefore

$$\begin{aligned}
\mathbb{E}_\mu f - \mathbb{E}_\nu f &= \int_{[0,1]} \big(\mu([\theta,1]) - \nu([\theta,1])\big)\, d\alpha(\theta) \\
&\leq \Big(\sup_{\theta\in[0,1]} \big[\mu([\theta,1]) - \nu([\theta,1])\big]\Big)\, \alpha([0,1]) \\
&\leq A.
\end{aligned}$$

Taking the supremum over $f \in \mathcal{M}$ yields $B \leq A$.

Combining Steps 1 and 2 gives $A = B$. If $\theta^\star \in [0,1]$ maximizes $\theta \mapsto \mu([\theta,1]) - \nu([\theta,1])$, then $f_{\theta^\star}$ attains B. If the supremum is not attained (e.g., due to a jump), choose θ_n with $\mu([\theta_n,1]) - \nu([\theta_n,1]) \uparrow A$; then f_{θ_n} attains B arbitrarily closely. □

Theorem 1 (No-improvement under confidence shift)**.** *Under Assumptions 1–2,*

$$J_t^\star \;\leq\; J_s^\star.$$

Proof. Fix $\theta \in (0,1)$. By the law of total probability and Assumption 1,

$$\mathrm{TPR}_t(\theta) = \mathbb{E}\big[\mathbb{P}(C_t \geq \theta \mid C_s) \mid Z = \mathrm{ID}\big] = \mathbb{E}\big[h_\theta(C_s) \mid Z = \mathrm{ID}\big],$$

and similarly $\mathrm{FPR}_t(\theta) = \mathbb{E}[h_\theta(C_s) \mid Z = \mathrm{OOD}]$. Hence

$$\mathrm{TPR}_t(\theta) - \mathrm{FPR}_t(\theta) = \mathbb{E}[h_\theta(C_s) \mid \mathrm{ID}] - \mathbb{E}[h_\theta(C_s) \mid \mathrm{OOD}]. \tag{17}$$

By Assumption 2, $h_\theta \in \mathcal{M}$ for each θ, so Lemma A.1 gives

$$\mathrm{TPR}_t(\theta) - \mathrm{FPR}_t(\theta) \leq \sup_{f\in\mathcal{M}} \big(\mathbb{E}[f(C_s) \mid \mathrm{ID}] - \mathbb{E}[f(C_s) \mid \mathrm{OOD}]\big) = J_s^\star.$$

Taking $\sup_\theta$ of the left-hand side (definition of $J_t^\star$) yields $J_t^\star \leq J_s^\star$. □

Example of a shift satisfying the assumptions. We give a simple score-level construction that satisfies Assumptions 1 and 2. The example is formulated at the confidence-score level, rather than as a general image-space transformation, because the assumptions concern the Markov kernel

$$R(u, B) = \mathbb{P}(C_t \in B \mid C_s = u). \tag{18}$$

Assume $C_s \in (0,1)$ almost surely. Let ε be a real-valued random variable independent of (C_s, Z), and let $a > 0$ and $b \in \mathbb{R}$. Define the target-environment confidence score by

$$C_t = \sigma(a\ \mathrm{logit}(C_s) + b + \varepsilon)\,, \tag{19}$$

where

$$\sigma(r) = \frac{1}{1+\exp(-r)}, \qquad \mathrm{logit}(u) = \log\frac{u}{1-u}. \tag{20}$$

This construction represents a target shift that rescales, biases, and randomly perturbs the source confidence on the logit scale. For example, choosing $0 < a < 1$ compresses confidence differences, while choosing $b < 0$ lowers target confidence globally.

We verify the two assumptions. First, conditional on $C_s = u$, the distribution of C_t is determined only by u and by the noise distribution of ε. Since ε is independent of Z, the conditional law

$$\mathbb{P}(C_t \in B \mid C_s = u, Z) \tag{21}$$

does not depend on Z. Hence $C_t \perp Z \mid C_s$, satisfying Assumption 1.

Second, for any threshold $\theta \in (0,1)$ and any $u \in (0,1)$,

$$\begin{aligned} h_\theta(u) &:= \mathbb{P}(C_t \geq \theta \mid C_s = u) \\ &= \mathbb{P}(\sigma(a\ \mathrm{logit}(u) + b + \varepsilon) \geq \theta) \\ &= \mathbb{P}(\varepsilon \geq \mathrm{logit}(\theta) - a\ \mathrm{logit}(u) - b) \\ &= 1 - F_\varepsilon(\mathrm{logit}(\theta) - a\ \mathrm{logit}(u) - b)\,, \end{aligned} \tag{22}$$

where F_ε is the cumulative distribution function of ε. Since $a > 0$ and $\mathrm{logit}(u)$ is strictly increasing on $(0,1)$, the argument of F_ε is nonincreasing in u. Therefore $h_\theta(u)$ is nondecreasing in u, satisfying Assumption 2.

Thus, this example captures a confidence-level target degradation: high source confidence remains stochastically associated with high target confidence, while the target confidence may be compressed, shifted, and corrupted by independent noise. Under this construction, the assumptions of Theorem 1 hold.

B Algorithmic Description of SROT

Algorithm 1 summarizes the online inference and adaptation procedure of SROT. We use $\mathrm{Split}(\mathcal{S})$ to denote the adaptive variance-minimizing threshold in Eq. (10).

Algorithm 1: Self-Reinforced Optimal Transport (SROT)

Require: Frozen image encoder $\mathcal{E}$; text embeddings $\{u_k\}_{k=1}^K$; unlabeled test batches $\{\mathcal{B}_t\}_{t=1}^T$; detector f_θ; temperature τ; OT regularization ε; warm-up length T_0; window size L; learning rate η.

Ensure: Open-set predictions $\widehat{y}(x) \in \{-1, 1, \ldots, K\}$.

Initialize detector parameters θ

Initialize score windows $\mathcal{W}^{\mathrm{OT}} \leftarrow \emptyset$ and $\mathcal{W}^{\mathrm{det}} \leftarrow \emptyset$

Set class prior $b \leftarrow \frac{1}{K}\mathbf{1}_K$ and open-set prior $\pi \leftarrow \frac{1}{K}\mathbf{1}_K$

for $t = 1, \ldots, T$ **do**

  Receive an unlabeled batch $\mathcal{B}_t = \{x_i\}_{i=1}^B$

  Extract normalized image features: $z_i \leftarrow \mathcal{E}(x_i)$

  Compute the frozen VLM posterior:

$$Q_{ik} \leftarrow \frac{\exp(\langle z_i, u_k\rangle/\tau)}{\sum_{j=1}^K \exp(\langle z_i, u_j\rangle/\tau)}$$

  Obtain detector scores:

$$r_i \leftarrow \big[\mathrm{softmax}\big(f_\theta(z_i)\big)\big]_{\mathrm{ood}}$$

  if $t \leq T_0$ **then**

    $\overline{Q} \leftarrow Q$ ▷ Initial OT rectification

  else

    $\overline{Q}_{ik} \leftarrow (1 - r_i)Q_{ik} + r_i\pi_k$ ▷ Detector-guided rectification

  end if

  Set $a \leftarrow \frac{1}{B}\mathbf{1}_B$ and $M \leftarrow -\log(\overline{Q} + \delta)$

Solve entropic OT using Sinkhorn iterations:

$$\Gamma^\star \leftarrow \arg\min_{\Gamma \in \mathcal{U}(a,b)} \langle \Gamma, M \rangle + \varepsilon \mathrm{KL}\big(\Gamma \,\big\|\, ab^\top\big)$$

Normalize the transport plan: $\widetilde{P}_{i,:} \leftarrow \Gamma^\star_{i,:}/a_i$

Compute OT-based OOD scores: $s_i^{\mathrm{OT}} \leftarrow 1 - \max_k \widetilde{P}_{ik}$

Update the recent-score window:

$$\mathcal{W}^{\mathrm{OT}} \leftarrow \mathrm{KeepLast}_L\left(\mathcal{W}^{\mathrm{OT}} \cup \{s_i^{\mathrm{OT}}\}_{i=1}^B\right)$$

Select $\lambda_t^{\mathrm{OT}} \leftarrow \mathrm{Split}(\mathcal{W}^{\mathrm{OT}})$

Generate pseudo-labels:

$$\widetilde{y}_i \leftarrow \mathbb{1}\left[s_i^{\mathrm{OT}} > \lambda_t^{\mathrm{OT}}\right]$$

if $t \le T_0$ **then**

  $o_i \leftarrow \widetilde{y}_i$ ▷ Use OT during warm-up

else

  Update the detector-score window:

$$\mathcal{W}^{\mathrm{det}} \leftarrow \mathrm{KeepLast}_L\left(\mathcal{W}^{\mathrm{det}} \cup \{r_i\}_{i=1}^B\right)$$

  Select $\lambda_t^{\mathrm{det}} \leftarrow \mathrm{Split}(\mathcal{W}^{\mathrm{det}})$

  $o_i \leftarrow \mathbb{1}[r_i > \lambda_t^{\mathrm{det}}]$

end if

for $i = 1, \ldots, B$ **do**

  if $o_i = 1$ **then**

    $\widehat{y}(x_i) \leftarrow -1$ ▷ OOD prediction

  else

    $\widehat{y}(x_i) \leftarrow \arg\max_k Q_{ik}$ ▷ ID prediction

  end if

end for

Update only the detector:

$$\theta \leftarrow \theta - \eta \nabla_\theta \frac{1}{B} \sum_{i=1}^{B} \mathrm{CE}\left(\mathrm{softmax}(f_\theta(z_i)), \widetilde{y}_i\right)$$

end for

return $\{\widehat{y}(x)\}$

C Stream-based Setting

Our primary experiments follow the batch-based test-time adaptation setting, where a model can adapt using mini-batches of unlabeled test data, consistent with prior work such as AdaND [Cao et al., 2025]. However, in real-world deployment, data often arrives sequentially as a stream, requiring immediate prediction for each sample and periodic updates. To validate our method's applicability in such streaming scenarios, we conduct an additional experiment where samples arrive one at a time. In this setup, we make an OOD prediction for each sample immediately upon arrival using the current detector state. The detector is updated only after accumulating a full batch of samples. This mimics a realistic streaming environment where latency constraints prevent batch-level processing for every individual sample.

Table 5 and Table 7 show the complete results under stream-based setting, our method has small variance. Several experimental configurations yielding identical accuracy values. This phenomenon is primarily attributed to the stability of the model's discrete classification decisions during the adaptation process. Although the model parameters are updated incrementally as it processes the data stream (e.g., within the 0–127 sample window), the resulting shifts in the underlying class probability distributions are often insufficient to cross the decision boundaries required to alter the Top-1 label assignments. Consequently, the sample-wise predictions remain largely unchanged between the initial zero-shot model and its adapted counterparts. This indicates that while test-time adaptation may refine the model's prediction confidence or internal feature representations, it does not necessarily trigger widespread classification "flips", leading to the observed numerical equivalence in overall accuracy across both settings.

Table 5: Comparisons between batch-based and stream-based detection. Zero-shot TTA for OOD detection results for CUB-200-2011, STANFORD-CARS, Food-101, and Oxford-IIIT Pet as the ID datasets.

ID	Setting	iNaturalist			LSUN			Texture			Places			Avg		
		Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H
CUB-200-2011	Batch-data	54.43	98.65	70.15	54.45	98.30	70.08	54.66	99.16	70.47	54.50	96.16	69.57	54.51	98.07	70.07
	Stream-data	54.36	98.63	70.09	54.38	98.31	70.03	54.59	99.16	70.41	54.43	96.17	69.52	54.44	98.07	70.01
STANFORD-CARS	Batch-data	63.83	99.93	77.90	63.81	99.91	77.88	63.81	99.95	77.89	63.81	99.49	77.75	63.81	99.82	77.86
	Stream-data	63.84	99.93	77.91	63.82	99.91	77.89	63.82	99.95	77.90	63.82	99.49	77.76	63.82	99.82	77.87
Food-101	Batch-data	86.38	99.92	92.66	86.36	99.66	92.53	86.31	97.73	91.67	86.36	99.48	92.46	86.35	99.20	92.33
	Stream-data	86.40	99.92	92.67	86.38	99.66	92.55	86.33	97.71	91.67	86.38	99.48	92.47	86.37	99.19	92.34
Oxford-IIIT Pet	Batch-data	86.27	98.87	92.14	86.26	98.53	91.99	86.28	98.98	92.19	86.28	97.55	91.57	86.27	98.48	91.97
	Stream-data	86.31	98.87	92.16	86.26	98.57	92.01	86.31	98.98	92.21	86.31	97.54	91.58	86.30	98.49	91.99

Table 6: Robustness to homogeneous initial test-stream prefixes on ImageNet-1K versus iNaturalist. The baseline uses the standard randomly mixed stream. All values are percentages.

Initial prefix	Acc_S ↑	Acc_N ↑	Acc_H ↑	AUROC↑	FPR95↓
Baseline	66.02	99.10	79.25	99.86	0.59
OOD-only prefix					
10 batches	65.62	99.01	78.93	99.81	0.76
20 batches	65.73	98.46	78.83	99.79	0.78
50 batches	65.74	97.23	78.44	99.79	0.69
ID-only prefix					
10 batches	65.17	98.81	78.54	99.78	1.15
20 batches	64.46	98.75	78.00	99.75	1.38
50 batches	62.33	98.73	76.42	99.72	1.42

D Robustness to Homogeneous Test Streams

We further evaluate SROT under challenging test-stream compositions in which the first 10, 20, or 50 consecutive batches contain exclusively OOD or ID samples. After this homogeneous prefix, the stream returns to the standard randomly mixed setting. We conduct the experiments on ImageNet-1K as ID and iNaturalist as OOD, keeping all other settings unchanged.

As shown in Table 6, SROT remains robust under both types of homogeneous prefixes: AUROC remains at least 99.72%, while FPR95 remains below 1.5%. OOD-only prefixes reduce Acc_H by only 0.32–0.81 points relative to the baseline. Their main effect is a gradual reduction in Acc_N as the prefix becomes longer, whereas Acc_S remains within 0.40 points of the baseline.

ID-only prefixes have a larger effect, reducing Acc_H by 0.71–2.83 points. In particular, the 50-batch prefix reduces Acc_S by 3.69 points, while Acc_N remains close to the baseline. This asymmetry is consistent with the adaptive threshold initially partitioning a score distribution containing only one population, which may introduce biased pseudo-labels before both ID and OOD samples become available. Nevertheless, the overall degradation remains limited after the stream returns to the mixed setting, establishing the precise recovery dynamics would require time-resolved evaluation. These results demonstrate substantial robustness to temporary homogeneous periods, while also identifying long ID-only prefixes as a challenging case and a direction for improving early-stage adaptation.

E Experimental Details

E.1 Design Choices and Hyperparameters

Table 8 summarizes the main hyperparameters used in our test-time adaptation method. We perform test-time adaptation with a phase transition: We switch to OOD detector after 10 epochs of adaptation. This transition is controlled by the warm-up step parameter. It ensures sufficient model adaptation before enabling OOD detection. For OT computation, the window size is set to 3 batches, meaning we compute OT over the concatenated outputs of the current batch and the previous 2 batches. The entropic regularization parameter (ϵ) is set to 1 for most experiments, with an adjustment to 0.1 for ImageNet variants to better align with their dataset characteristics. For the noise injection please refer to AdaND Cao et al. [2025].

Additional implementation details:

Table 7: Comparisons between batch-based and stream-based detection. Zero-shot TTA for OOD detection results for ImageNet and its variants as the ID datasets.

ID	Setting	iNaturalist			LSUN			Texture			Places			Avg		
		Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H
ImageNet	Batch-data	66.02	99.10	79.25	64.69	95.66	77.18	65.74	86.24	74.61	64.74	85.07	73.53	65.30	91.52	76.14
	Stream-data	66.02	99.10	79.25	64.69	95.66	77.18	65.74	86.24	74.61	64.74	85.07	73.53	65.30	91.52	76.14
ImageNet-K	Batch-data	45.91	99.80	62.89	45.67	99.03	62.51	45.10	89.53	59.98	45.39	89.76	60.29	45.52	94.53	61.42
	Stream-data	45.91	99.80	62.89	45.67	99.03	62.51	45.10	89.53	59.98	45.39	89.76	60.29	45.52	94.53	61.42
ImageNet-A	Batch-data	43.67	97.44	60.31	42.25	91.07	57.72	45.69	88.11	60.18	43.47	81.65	56.73	43.77	89.57	58.73
	Stream-data	43.73	97.40	60.36	42.13	90.96	57.59	45.72	87.96	60.17	43.47	81.71	56.75	43.76	89.51	58.72
ImageNet-V2	Batch-data	58.10	97.68	72.86	57.22	89.93	69.94	59.10	84.10	69.42	58.34	76.18	66.08	58.19	86.97	69.58
	Stream-data	58.10	97.68	72.86	57.22	89.93	69.94	59.10	84.10	69.42	58.34	76.18	66.08	58.19	86.97	69.58
ImageNet-R	Batch-data	73.15	99.78	84.41	72.15	98.84	83.41	71.91	94.71	81.75	72.07	94.50	81.77	72.32	96.96	82.83
	Stream-data	73.18	99.77	84.43	72.15	98.84	83.41	71.88	94.70	81.73	72.07	94.54	81.79	72.32	96.96	82.84

Table 8: Main hyperparameters used in our experiments.

Category	**Hyperparameter**	**Value**
Test-Time Adaptation	Batch size	128
	Noise injection type	Gaussian
	Gaussian noise rate	0.125
	Warm-up step (phase transition)	10
	Learning rate	0.0005
	Weight decay	0
	Optimizer	Adam
	Adam β_1	0.9
Optimal Transport	OT window size	3 batches
	Entropic regularization (ϵ)	1 or 0.1

- The OT window size of 3 batches ensures we capture recent distributional characteristics while maintaining computational efficiency.
- For Sinkhorn's algorithm, we set convergence parameters to maximum iterations = 1000 and stopping threshold = 10^{-9}.
- The threshold for OOD detection is determined adaptively by minimizing the variance of outlier scores over a grid search in $[0, 1)$ with step size 0.01.

E.2 Environment

The experiments were conducted with Python 3.9.23 and PyTorch 2.7.1+cu128 (CUDA available; CUDA version 12.8) on Ubuntu 25.04, using NVIDIA GeForce RTX 5090 GPUs (driver 580.95.05) and an AMD Ryzen 9 9950X3D CPU.

E.3 Extra Time Cost of SROT

The computational overhead introduced by SROT is practical and acceptable for real-world deployment, particularly given our TTA framework is already designed to be low-budget. As shown in Table 9, SROT incurs 34.51 seconds of overhead on CUB-200-2011 over iNaturalist, leading to a total runtime of 142.18 seconds per run. For ImageNet over iNaturalist, SROT adds 56.51 seconds of overhead, resulting in a total runtime of 278.31 seconds per run. This incremental cost is dominated by lightweight entropic optimal transport computations (amortized across batches) and is fully offset by substantial improvements in zero-shot OOD detection performance.

Table 9: Running Time Comparison (CUB-200-2011/ImageNet over iNaturalist)

	CUB-200-2011 $\rightarrow$ iNaturalist	ImageNet $\rightarrow$ iNaturalist
SROT Component	34.518s	56.512s
Full Pipeline	142.182s	4m38.314s (278.314s)

F Limitations

Although SROT improves zero-shot OOD detection by reshaping VLM posteriors and learning feature-level OOD evidence, it has several limitations. First, the OT rectification relies on a prescribed class marginal, which is set to a uniform prior by default. While this choice is label-free and generally effective, it may be sub-optimal when the true ID class distribution is highly imbalanced or when a test batch contains only a small subset of classes. Second, SROT uses pseudo-labels derived from rectified OOD scores to train the detector. Although the adaptive threshold reduces sensitivity to fixed hyperparameters, noisy pseudo-labels may still affect detector training when ID and OOD samples are extremely entangled. Third, our method assumes access to unlabeled test batches or streams, since both OT alignment and adaptive threshold estimation benefit from batch-level statistics. Performance may degrade for very small batches or purely single-sample inference. Finally, SROT keeps the VLM frozen and does not address failures caused by severely mismatched prompts or missing semantic coverage in the candidate label set. Future work could explore more flexible prior estimation, robust pseudo-label selection, and prompt-aware posterior rectification for broader open-world deployment.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: [N/A]

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: App. F

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: Assumptions 1–2 and the complete proof are provided in Appendix A.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: App. E

   Guidelines:

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

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

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

   Answer: [Yes]

Justification: The code will be upload to Github.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Sec. E

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: We follow the common practice of previous works and do not report error bars or statistical significance.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Sec. E.2

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: [N/A]

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [No]

    Justification: We focus purely on algorithmic research and do not analyze specific societal positive or negative impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: This paper involves no high-risk data or models that require specific responsible release safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: [N/A]

    Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: See our code

    Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.