TriPrompt: Progressive Local Prompting for Few-shot Out-of-Distribution Detection

Ziyou Xiang[1] **Chaowei Fang**[2,3] **Zhihong Wu**[1] **Jiliang Li**[4] **Di Xu**[5] **De Cheng**[6]
[1]Xidian University, Xi'an, China
[2]School of Software Engineering, Xi'an Jiaotong University, China
[3]State Key Lab. of Novel Software Technology, Nanjing University, China
[4]Xi'an Jiaotong University, China
[5]Huawei Technologies Ltd.
[6]School of Telecommunication and Engineering, Xidian University, China

Abstract

Few-shot out-of-distribution (OOD) detection is essential for deploying recognition models in open-world environments, where a reliable model should correctly recognize in-distribution (ID) samples while rejecting samples from unseen categories. Recent CLIP-based methods have shown promising few-shot OOD detection ability by exploiting transferable vision–language representations. However, many methods rely mainly on global image features or emphasize only the most class-discriminative local cues, making them less effective for OOD detection where ID and OOD samples share visually similar local patterns. Although local prompting introduces patch-level reasoning, existing methods often treat local evidence as a binary distinction between ID-related and non-ID cues, leaving the ambiguous transition region between reliable ID evidence and clear non-ID evidence insufficiently modeled. To address this issue, we propose TriPrompt, a progressive local prompting framework for few-shot OOD detection. TriPrompt decomposes patch-level evidence into three complementary components within a unified local feature space: Core-ID cues that provide reliable class-defining evidence, Far-ID cues that capture ambiguous transition-region evidence, and OOD cues learned from weak non-ID proxies mined directly from patch features. This patch-native design enables suppression-aware local reasoning without requiring additional OOD data or repeated multi-crop feature extraction. At inference time, TriPrompt combines global CLIP confidence with Core-ID, Far-ID, and OOD local responses to reduce overconfident ID assignment on hard OOD samples. Experiments on ImageNet-1K OOD benchmarks show that TriPrompt consistently improves few-shot OOD detection performance and remains compatible with mainstream CLIP-based scoring methods.

1 Introduction

Out-of-distribution (OOD) detection aims to determine whether a test sample belongs to one of the known in-distribution (ID) categories or should be rejected as unknown [Hendrycks and Gimpel, 2017]. This capability is essential for deploying machine learning systems in open-world environments, where unseen categories, abnormal inputs, and distribution shifts inevitably appear at test time. A reliable OOD detector should therefore satisfy two requirements simultaneously: it should

Corresponding authors are: De Cheng (dcheng@xidian.edu.cn) and Chaowei Fang (chaoweifang@outlook.com).

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

preserve strong ID recognition performance on known categories, and it should avoid overconfident predictions on samples outside the training label space.

Recent vision–language models, especially CLIP [Radford et al., 2021], provide a strong foundation for OOD detection due to their transferable visual representations and open-vocabulary text alignment. A growing number of CLIP-based zero-shot and few-shot OOD detection methods have been developed by designing better text prompts, calibrating ID/OOD scores, or exploiting the semantic structure of the visual–language embedding space [Ming et al., 2022, Jiang et al., 2024, Chen et al., 2024, Miyai et al., 2023, Bai et al., 2024]. However, many existing methods mainly rely on global image representations. Although global features are effective for capturing image-level semantics, they may overlook subtle local discrepancies that are crucial for OOD detection [Xiao et al., 2025, Cheng et al., 2026, Asokan et al., 2025], especially when ID and OOD categories are visually similar or share common object parts.

This limitation has motivated recent studies to incorporate local prompting or patch-level modeling into CLIP-based few-shot OOD detection [Miyai et al., 2023, 2025, Zeng et al., 2025, Lafon et al., 2024]. By exploiting local patch features, these methods improve the detector's sensitivity to regional evidence and provide a more detailed view of ID/OOD differences. Nevertheless, current local modeling remains incomplete. Most existing approaches implicitly treat local evidence as a relatively simple distinction between class-related ID cues and non-ID or background-like cues. Such a binary view is insufficient for hard OOD detection, because local patches often do not fall cleanly into either side. In particular, some local regions are not reliable class-defining ID evidence, but they are also not clear non-ID evidence. They may still share strong similarity with ID categories while being compatible with OOD patterns, making them difficult to interpret reliably.

This observation suggests an important gap in existing local CLIP-based OOD detection: the transition region between reliable ID evidence and clear non-ID evidence is rarely modeled explicitly. For ID samples, a detector should rely more on discriminative local regions that truly support known-class recognition. For OOD samples, however, visually similar local regions may activate ID prompts and lead to overconfident false acceptance. If these ambiguous local cues are absorbed into the ID side, the ID decision region may be over-expanded; if they are simply treated as OOD evidence, useful but less discriminative ID information may be suppressed. Therefore, effective local OOD reasoning requires a more fine-grained decomposition of patch-level evidence beyond the conventional ID/OOD opposition.

Motivated by this analysis, we propose **TriPrompt**, a progressive local prompting framework for CLIP-based few-shot OOD detection. The key idea of TriPrompt is to model patch-level evidence along the ID-to-OOD transition within the same local feature space used during inference. Instead of treating local responses as a binary opposition between ID and OOD, TriPrompt decomposes them into three complementary components: *Core-ID*, *Far-ID*, and *OOD*. The Core-ID branch captures reliable class-defining local evidence for ID recognition. The OOD branch learns class-shared non-ID prompts from weak patch-native proxies explored from regions with low Core-ID responses, since real OOD patches are unavailable during training. The Far-ID branch focuses on ambiguous boundary cues near the ID–OOD transition, which are neither stable Core-ID evidence nor clear non-ID evidence. By explicitly modeling these hard-to-disambiguate cues, TriPrompt can suppress overconfident ID assignment on hard OOD samples while preserving useful ID recognition ability.

TriPrompt has two notable characteristics. First, it provides a more faithful local evidence decomposition for OOD detection by jointly modeling reliable ID cues, non-ID cues, and the critical transition-region cues between them. This makes the local branch more suitable for OOD detection than methods that rely only on strong ID evidence or clear negative evidence. Second, TriPrompt is patch-native: all three branches are constructed from the patch tokens produced by a single CLIP forward pass. Thus, OOD-related local supervision can be explored directly from the inference feature space, avoiding repeated random cropping and repeated feature extraction used in some prior local methods. At inference time, TriPrompt combines global CLIP confidence with local suppression-aware reasoning, where OOD prompts and Far-ID prompts jointly reduce overconfidence on visually similar OOD samples.

Extensive experiments on ImageNet-1K with multiple OOD benchmarks demonstrate the effectiveness of TriPrompt. The proposed framework consistently improves OOD detection performance and remains compatible with mainstream CLIP-based scoring methods. These results indicate that

explicitly modeling the ID–OOD transition region is a promising direction for strengthening local OOD reasoning in CLIP-based detectors.

We summarize our main contributions as follows:

- We propose TriPrompt, a progressive local prompting framework for CLIP-based OOD detection. TriPrompt decomposes patch-level evidence into Core-ID, Far-ID, and OOD components within a unified local feature space, explicitly modeling ambiguous transition-region cues rather than treating local evidence as a strict ID/OOD binary opposition.
- We develop a patch-native OOD prompt learning strategy that explores non-ID proxies directly from local patch features. This design provides efficient OOD-related supervision without requiring additional OOD data or repeated multi-crop feature extraction.
- Extensive experiments on ImageNet-1K OOD benchmarks demonstrate that TriPrompt achieves strong OOD detection performance and remains compatible with mainstream CLIP-based scoring methods.

2 Related Work

2.1 OOD Detection

OOD detection is essential for reliable deployment in open-world visual systems, where unknown classes and abnormal inputs inevitably appear at test time [Hendrycks and Gimpel, 2017, Liang et al., 2018, Liu et al., 2020, Xu et al., 2025, Ling et al., 2025]. In realistic scenarios, however, real OOD samples are often unavailable during training. This has motivated work on negative semantic modeling and OOD-oriented prompt learning, which aims to anticipate unknown information that may appear at inference time. In the zero-shot setting, some methods construct unknown or negated semantics in the text space to improve detection [Esmaeilpour et al., 2022, Wang et al., 2023, Jiang et al., 2024]. In supervised settings, other methods explicitly synthesize abnormal samples or abnormal features to facilitate boundary learning, for example by generating virtual OOD features in the ID embedding space [Du et al., 2022, Tao et al., 2023]. More recent approaches exploit background information in ID samples to model negative semantics, such as LoCoOp and ID-like [Miyai et al., 2023, Bai et al., 2024]. Although these methods improve ID–OOD separation, most existing approaches still operate mainly on global representations and have limited ability to model fine-grained regional anomalies or local semantic shifts.

2.2 Prompt Tuning for Vision-Language Models

Prompt tuning for vision–language models has developed rapidly in recent years and has become a widely used parameter-efficient adaptation strategy [Radford et al., 2021, Zhou et al., 2022b,a, Koleilat et al., 2025, Li et al., 2025, Guo et al., 2026]. In OOD detection, prompt-based methods further adapt this paradigm to improve semantic separation between ID and OOD samples. In addition to global prompt tuning, recent studies incorporate local visual information into VLM-based recognition. In Vision Transformers [Dosovitskiy et al., 2020], local tokens encode region-level features and are widely used in dense prediction tasks such as segmentation and object detection [Cheng et al., 2022, Strudel et al., 2021, Chen et al., 2023]. In OOD detection, GL-MCM Miyai et al. [2025] extends MCM with local feature scoring, while GalLop [Lafon et al., 2024] and Local-Prompt [Zeng et al., 2025] further introduce local prompts and patch-level alignment to improve sensitivity to fine-grained visual patterns. Nevertheless, current local prompt designs still focus mainly on highly ID-related core features and low-ID-related background features, leaving transition-region features partially similar to both ID and OOD underexplored.

3 Method

3.1 Problem Definition

We study downstream prompt learning for CLIP in the setting of few-shot out-of-distribution (OOD) detection. Let the downstream training set be $\mathbb{D} = \{(\mathbf{x}, y)\}$, where $\mathbf{x}$ denotes an input image and

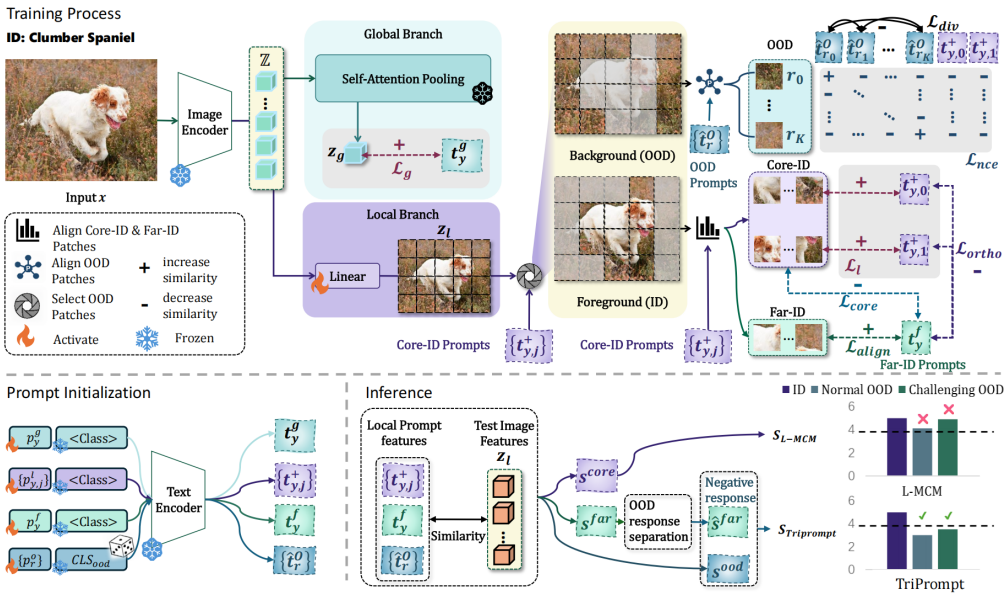

Figure 1: Overview of TriPrompt. TriPrompt is built upon a frozen CLIP backbone and integrates a global prompt branch with Core-ID, OOD, and Far-ID local prompts to model reliable ID evidence, weak non-ID proxies, and ambiguous transition-region cues.

$y \in \{1, \ldots, C\}$ denotes the corresponding label from the set of C known in-distribution (ID) classes. Only ID samples are available during training, and no real OOD samples are observed.

During inference, the model receives a mixture of ID samples from known classes and OOD samples from unseen semantic categories. The objective is twofold: (i) ID samples should be correctly classified into one of the C known classes; (ii) OOD samples should be rejected rather than incorrectly assigned to any ID class. The main challenge is that prompt tuning must improve ID recognition while preserving the ability to detect unseen OOD samples under limited ID-only supervision.

3.2 Method Overview

We propose TriPrompt, a patch-native prompting framework for CLIP-based few-shot OOD detection. TriPrompt keeps the CLIP backbone frozen and extends the model only at the prompt and objective levels. Given an input image, TriPrompt uses local CLIP patch features to model three types of local evidence: reliable ID evidence, weak non-ID evidence, and ambiguous transition-region evidence. These evidence types are modeled by the Core-ID, OOD, and Far-ID prompts, respectively.

CLIP consists of a visual encoder and a text encoder, and prompt learning adapts the text encoder with a small set of learnable tokens while keeping the backbone frozen. Starting from CLIP patch features, TriPrompt decomposes local responses into different reliability regimes for OOD-aware prompt learning.

First, TriPrompt learns Core-ID prompts to capture reliable local ID evidence. For each class, a set of positive local prompts is optimized with few-shot ID samples to model discriminative object regions.

Second, TriPrompt introduces class-shared OOD prompts to model weak non-ID evidence. Motivated by SSOD [Pei et al., 2023], TriPrompt mines local ID-image regions that consistently receive low Core-ID responses. These regions are treated as weak non-ID proxies that provide conservative patch-level supervision for local non-ID suppression, enabling TriPrompt to learn OOD-sensitive cues from the same patch-level feature space used during inference.

Third, TriPrompt introduces Far-ID prompts to model ambiguous local evidence. A binary separation between Core-ID and OOD evidence is often insufficient because many patches are neither reliable class-defining regions nor clear weak non-ID regions. Due to this ambiguity, the model often struggles to assign such patches to either ID or OOD categories, making them challenging in practice. Therefore,

TriPrompt models these regions with learnable Far-ID prompts. Far-ID serves as an ambiguity-aware suppression component that helps correct misjudgments on ambiguous local features.

During inference, TriPrompt computes local responses from the Core-ID, OOD, and Far-ID prompts using the same CLIP patch features. The final OOD score combines the global prompt score with the three local responses. In this way, TriPrompt complements global CLIP-based OOD scoring with patch-level evidence decomposition without modifying the CLIP backbone.

3.3 Prompted CLIP Representation and Core-ID Prompt Learning

Prompted CLIP representation. Formally, let $\mathcal{T}$ and $\mathcal{V}$ denote the CLIP text encoder and image encoder, respectively. A learnable prompt is written as $\mathbf{p} = [p^1, \ldots, p^V]$, where V denotes the number of context tokens. All class names are embedded into the same fixed template, namely "A photo of a", and the learnable prompt is inserted before the class token. Let $\mathbf{a}_c$ denote the embedding of class name c. The prompted text feature of class c is defined as $\mathbf{t}_c = \mathcal{T}([\mathbf{p}, \mathbf{a}_c])$. Given an input image $\mathbf{x}$, the image encoder produces a global image feature together with a set of patch features $\mathbb{Z} = \{\mathbf{z}_i\}_{i=1}^{L}$, where L denotes the number of patches.

The prompted model uses one learnable global prompt, N learnable Core-ID local prompts $\{\mathbf{p}_j^+\}_{j=1}^{N}$ for each class, N_o class-shared OOD prompts $\{\hat{\mathbf{p}}_r\}_{r=1}^{N_o}$, and one Far-ID prompt $\mathbf{p}^f$ for each class.

Global and Core-ID local prompt learning. The global branch and the sparse Core-ID local branch fully follow Gallop Lafon et al. [2024], so we omit their details here and defer the complete formulation to the appendix. The text features of the Core-ID prompts, which are reused in the following subsections, are written as $\mathbf{t}_{c,j}^+ = \mathcal{T}([\mathbf{p}_j^+, \mathbf{a}_c])$ for the j-th prompt of class c.

3.4 OOD Prompt Learning from Patch-Native Weak Non-ID Proxies

The OOD branch uses background patches with the lowest ID responses as proxies for OOD information and models them using class-shared OOD prompts. Let the class-shared OOD prompts be $\hat{\mathcal{P}} = (\hat{\mathbf{p}}_1, \ldots, \hat{\mathbf{p}}_{N_o})$, and let the corresponding text features be $\hat{\mathbf{t}}_1^o, \ldots, \hat{\mathbf{t}}_{N_o}^o$, where each $\hat{\mathbf{t}}_r^o$ is produced by the CLIP text encoder from the r-th OOD prompt. For the subsequent OOD, Far-ID, and inference modules, a patch-wise class-level Core-ID score and a patch-wise OOD score are additionally defined as

$$s_{c,i}^{core} = \frac{1}{N}\sum_{j=1}^{N}\langle \mathbf{z}_i, \mathbf{t}_{c,j}^+ \rangle, \qquad s_{r,i}^{ood} = \langle \mathbf{z}_i, \hat{\mathbf{t}}_r^o \rangle, \tag{1}$$

OOD alignment loss. For each weak non-ID proxy, the strongest responding OOD prompt is treated as the target explanation, while the positive Core-ID prompts of the ground-truth class together with all OOD prompts form the normalization pool:

$$\mathcal{L}_{nce} = -\frac{1}{K_t}\sum_{i\in\Omega^{bg}} \log \frac{\max_r \exp(s_{r,i}^{ood}/\tau)}{\exp(s_{y,i}^{core}/\tau) \; + \; \sum_{r=1}^{N_o}\exp(s_{r,i}^{ood}/\tau)}. \tag{2}$$

Here, Ω^{bg} contains the K_t patches with the lowest $s_{y,i}^{core}$ responses, where y denotes the ground-truth class and i denotes the patch index. This objective encourages low-Core-ID regions to align with the class-shared OOD prompts. Each patch is aligned only with the highest-response OOD prompt to reduce overlap among different OOD prompts.

OOD diversity regularization. If all OOD prompts collapse into a single direction, the learned OOD space becomes excessively limited. Therefore, a diversity regularization term is introduced:

$$\mathcal{L}_{div} = \frac{1}{N_o(N_o-1)}\sum_{r\neq r'}\langle \hat{\mathbf{t}}_r^o, \hat{\mathbf{t}}_{r'}^o \rangle^2. \tag{3}$$

The final OOD objective combines proxy alignment and prompt diversity:

$$\mathcal{L}_{ood} = \mathcal{L}_{nce} + \lambda_{div}\mathcal{L}_{div}. \tag{4}$$

These two terms encourage discriminative and diverse OOD prompts.

3.5 Far-ID Prompt Learning

Far-ID identifies ambiguous features that exhibit both ID-related and OOD-related characteristics and aligns such features with Far-ID prompts. One Far-ID prompt is introduced for each class with text feature $\mathbf{t}_c^f$, where $\mathbf{t}_c^f$ is produced by the CLIP text encoder from the Far-ID prompt of class c. For subsequent reuse, the patch-wise Far-ID response of class c on patch i is defined as

$$s_{c,i}^{far} = \langle \mathbf{z}_i, \mathbf{t}_c^f \rangle. \tag{5}$$

To prevent the Far-ID loss from affecting the alignment between the OOD prompts and the visual features, we detach the local visual features and the OOD-branch logits involved in this loss.

Far-ID region mining. To avoid overlap with the weak non-ID proxies already used by the OOD branch, Far-ID regions are mined only from patches outside Ω^{bg}. Within the remaining regions, the negative absolute OOD-ID gap

$$g_i = -\left| \max_r s_{r,i}^{ood} - s_{y,i}^{core} \right| \tag{6}$$

is used to identify local regions located within the ambiguity-prone regime between the main Core-ID support and the learned non-ID directions. The top-K_f non-background patches with the largest g_i values are then selected as Far-ID regions and denoted by Ω^{far}, while the top-K_f patches with the highest Core-ID responses are selected as reference Core-ID regions and denoted by Ω^{core}.

Far-ID alignment loss. On the mined Far-ID regions, a prompt-level contrastive loss is optimized in which the class-specific Far-ID prompt competes with the positive Core-ID prompts of the ground-truth class:

$$\mathcal{L}_{align} = -\frac{1}{K_f} \sum_{i \in \Omega^{far}} \log \frac{\exp(s_{y,i}^{far}/\tau)}{\sum_{j=1}^{N} \exp(\langle \mathbf{z}_i, \mathbf{t}_{y,j}^+ \rangle/\tau) + \exp(s_{y,i}^{far}/\tau)}. \tag{7}$$

This term aligns ambiguity-prone regions with the Far-ID prompt while separating them from the Core-ID prompts.

Core-ID suppression. To prevent Far-ID from collapsing onto highly discriminative ID regions, the Far-ID response on Ω^{core} is additionally suppressed:

$$\mathcal{L}_{core} = \frac{1}{K_f} \sum_{i \in \Omega^{core}} \log \left(1 + \exp \left(\frac{s_{y,i}^{far}}{\tau} \right) \right). \tag{8}$$

Orthogonality regularization. The Far-ID prompt is further regularized to remain separated from the positive local prompts of the same class:

$$\mathcal{L}_{ortho} = \frac{1}{N} \sum_{j=1}^{N} \langle \mathbf{t}_y^f, \mathbf{t}_{y,j}^+ \rangle^2. \tag{9}$$

The final Far-ID objective combines Far-ID alignment, Core-ID suppression, and prompt-level separation:

$$\mathcal{L}_{far} = \mathcal{L}_{align} + \mathcal{L}_{core} + \lambda_{ortho} \mathcal{L}_{ortho}. \tag{10}$$

The final training objective is the sum of the global, Core-ID, OOD, and Far-ID losses.

3.6 Inference

OOD detection. Based on Regional-MCM Zeng et al. [2025], local TriPrompt introduces two mechanisms to enhance local OOD detection: Far-ID OOD separation and top-K_{neg} selection over negative responses. Before incorporating Far-ID responses into inference, Far-ID OOD separation is first applied. For each patch i and class c, the strongest learned OOD response on that patch is compared against the corresponding Core-ID response, and the resulting probability is used to correct the Far-ID score:

$$\tilde{s}_{c,i}^{far} = \frac{\max_r \exp(s_{r,i}^{ood}/\tau)}{\max_r \exp(s_{r,i}^{ood}/\tau) + \exp(s_{c,i}^{core}/\tau)} \cdot \exp(s_{c,i}^{far}/\tau). \tag{11}$$

Table 1: Results when ImageNet-1k is used as ID data. We compare full-data and few-shot tuning methods. † represents the result by our re-implementation. For the few-shot (4-shot/16-shot) setting, **Bold** and underline denote the best and second-best results in each column, respectively. The same convention is used in all subsequent tables, where best and second-best are determined within each shot setting.

Method	iNaturalist		SUN		Places		Texture		**Average**	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Full data										
MSP [Hendrycks and Gimpel, 2017]	54.05	87.43	73.37	78.03	72.98	78.03	68.85	79.06	67.31	80.64
NPOS [Tao et al., 2023]	16.58	96.19	43.77	90.44	45.27	89.44	46.12	88.80	37.93	91.22
LAPT [Zhang et al., 2024]	**1.16**	**99.63**	**19.12**	**96.01**	**33.01**	**92.01**	**40.32**	**91.06**	**23.40**	**94.68**
Few-shot					4-*shot*					
CoOp [Zhou et al., 2022b]	18.95	95.52	29.58	92.90	38.72	89.64	48.03	85.87	33.82	90.98
LoCoOp [Miyai et al., 2023]	21.67	95.69	22.98	95.07	31.41	92.10	49.79	87.85	31.46	92.68
ID-like [Bai et al., 2024]	8.98	97.35	42.03	91.64	44.00	90.57	**25.27**	**94.32**	30.07	93.68
Local-Prompt [Zeng et al., 2025]	9.65	97.87	20.40	95.57	29.39	92.67	51.20	88.00	27.66	93.53
TriPrompt(Ours)	**6.81**	**98.52**	**16.62**	**96.27**	**25.94**	**92.91**	37.38	90.26	**21.69**	**94.49**
					16-*shot*					
CoOp [Zhou et al., 2022b]	14.60	96.62	28.48	92.65	36.49	89.98	43.13	88.03	30.67	91.82
LoCoOp [Miyai et al., 2023]	16.05	96.86	23.44	95.07	32.87	91.98	42.28	90.19	28.66	93.52
GalLop [Lafon et al., 2024]	13.70	97.10	24.90	94.00	32.50	91.30	38.40	90.40	27.30	93.20
FAGL [Lu et al., 2025]	14.49	96.48	27.65	93.46	31.09	92.44	**29.50**	**92.93**	25.68	93.82
SCT [Yu et al., 2024]	13.94	95.86	20.55	95.33	29.86	92.24	41.51	89.06	26.47	93.37
Local-Prompt [Zeng et al., 2025]	8.71	98.10	23.97	94.85	32.50	92.32	47.93	89.04	28.27	93.58
APT [Miao et al., 2025]	9.70	97.79	20.12	95.52	28.54	**92.84**	45.78	88.43	26.03	93.64
TriPrompt(Ours)	**8.00**	**98.22**	**16.02**	**96.02**	**26.02**	92.80	31.58	91.30	**20.41**	**94.59**

Table 2: Compatibility evaluation of TriPrompt with global zero-shot OOD detectors on ImageNet-1k.

Method	iNaturalist		SUN		Places		Texture		**Average**	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MCM [Ming et al., 2022]	32.20	94.59	38.80	92.25	46.20	90.31	58.50	86.12	43.93	90.82
MCM + TriPrompt$_{local}$	6.81	98.43	15.20	96.19	25.41	92.88	34.61	90.67	20.51	94.54
NegLabel [Jiang et al., 2024]	1.91	99.49	20.53	95.49	35.59	91.64	43.56	90.22	25.40	94.21
NegLabel + TriPrompt$_{local}$	1.51	99.58	13.26	96.79	26.82	93.12	36.26	91.10	19.46	95.15
CSP [Chen et al., 2024]	1.54	99.60	13.66	96.66	29.32	92.90	25.52	**93.86**	17.51	95.76
CSP + TriPrompt$_{local}$	**1.13**	**99.62**	**10.48**	**97.11**	**22.97**	**93.43**	**21.40**	93.31	**14.00**	**95.87**

We denote the corrected Far-ID responses together with the OOD responses as negative responses. For each patch, only the top-K_{neg} negative responses with the highest values are retained as suppression terms, which avoids excessive interference from irrelevant categories. The local TriPrompt score is defined as

$$S_{\mathrm{TriPrompt}}(\mathbf{x}) = \mathcal{M}_{K_{inf}} \left\{ \frac{\max_c \exp(s^{core}_{c,i}/\tau)}{\sum_{c'} \exp(s^{core}_{c',i}/\tau) + \sum_{c'_{neg} \in \mathrm{TopK}_{c_{neg}}(s^{neg}_{c_{neg},i}; K_{neg})} s^{neg}_{c'_{neg},i}} \right\}_{i=1}^{L}. \quad (12)$$

Here, s^{neg} denotes the negative responses obtained by concatenating the corrected Far-ID responses and the OOD responses, and c_{neg} denotes the total number of negative-response categories, i.e., the sum of all class-wise Far-ID prompts and all OOD prompts. The final detector combines the local TriPrompt score with the global MCM Ming et al. [2022] score as $S_{\mathrm{MCM}}(\mathbf{x}) + \alpha S_{\mathrm{TriPrompt}}(\mathbf{x})$, where α controls the contribution of the local term.

ID classification. For ID classification, TriPrompt follows Gallop and reuses only the global and Core-ID branches under the same classification setting.

4 Experiments

4.1 Experimental Setup

Datasets. Following MOS [Huang and Li, 2021] and MCM [Ming et al., 2022], we use ImageNet-1K [Deng et al., 2009] as the ID dataset and evaluate OOD detection on iNaturalist [Van Horn et al., 2018], SUN [Xiao et al., 2010], Places [Zhou et al., 2017], and Textures [Cimpoi et al., 2014], with OOD categories selected to be disjoint from ImageNet-1K.

Table 4: Comparison results on near-OOD detection tasks using ImageNet-10 and ImageNet-20 as mutually exclusive ID/OOD sets. We report 16-shot results for few-shot methods.

Method	ID / OOD	ImageNet-10 / ImageNet-20		ImageNet-20 / ImageNet-10		**Average**	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MCM [Ming et al., 2022]		5.00	98.71	12.51	97.70	8.75	98.21
GL-MCM [Miyai et al., 2025]		10.10	98.04	9.00	98.62	9.55	98.33
LoCoOp [Miyai et al., 2023]		11.20	97.49	12.00	97.79	11.60	97.64
Local-Prompt [Zeng et al., 2025]		**3.90**	**99.06**	6.20	98.84	5.05	98.95
TriPrompt(Ours)		**3.90**	99.04	**2.80**	**99.26**	**3.35**	**99.15**

4.2 Comparison with State-of-the-Art Methods

OOD Detection Performance. Tables 1, 2, and 4 show that TriPrompt performs strongly under both standard OOD and near-OOD evaluation protocols.

ID Accuracy. Table 3 reports 16-shot ID classification accuracy on ImageNet-1K. Overall, TriPrompt maintains competitive ID accuracy while focusing on improving OOD detection.

Table 3: ID classification accuracy under the 16-shot setting on ImageNet1k.

Method	**ID ACC ↑**
LoCoOp [Miyai et al., 2023]	71.7
GalLop [Lafon et al., 2024]	**75.1**
Local-Prompt [Zeng et al., 2025]	74.2
TriPrompt (Ours)	74.3

4.3 Ablation Studies

Module Effectiveness. For a cleaner attribution, all OOD ablations below use the local score alone, without the global term $S_{\mathrm{MCM}}(\mathbf{x})$ that the final detector adds. Table 5 shows that both the OOD and Far-ID modules are effective, and combining them gives the strongest overall OOD performance. This confirms that explicit non-ID suppression and transition-region modeling provide complementary benefits.

Table 5: Module effectiveness on ImageNet-1K ID with four OOD datasets and their average.

Shot	Setting	iNaturalist		SUN		Places		Textures		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
4-shot	Core-ID	9.05	97.90	20.66	95.02	31.30	91.46	52.50	84.00	28.38	92.09
4-shot	Core-ID + OOD	7.02	98.31	18.88	95.60	**30.09**	91.63	46.40	86.75	25.60	93.07
4-shot	Core-ID + Far-ID	6.61	**98.40**	**18.79**	95.69	30.50	**91.64**	45.04	**87.82**	25.23	**93.39**
4-shot	Core-ID + Far-ID + OOD	**6.60**	**98.40**	18.97	**95.72**	30.37	91.61	**44.81**	87.69	**25.19**	93.35
16-shot	Core-ID	7.30	**98.34**	20.01	94.83	28.09	91.88	41.60	87.19	24.25	93.06
16-shot	Core-ID + OOD	**7.10**	98.30	17.75	95.48	**27.55**	**92.00**	37.25	89.06	22.41	93.71
16-shot	Core-ID + Far-ID	8.66	98.02	18.30	95.49	29.12	91.63	**34.42**	**90.29**	22.62	**93.86**
16-shot	Core-ID + Far-ID + OOD	7.91	98.17	**17.61**	**95.57**	28.36	91.80	35.50	89.85	**22.34**	93.85

Far-ID Loss Effectiveness. Table 6 shows that the full Far-ID objective is the most effective setting overall. This verifies that alignment, Core-ID suppression, and prompt-level separation are complementary in learning useful transition-region prompts.

Table 6: Effectiveness of different loss combinations in the Far-ID module on ImageNet-1K ID with four OOD datasets and their average. For a direct comparison of the Far-ID loss terms, only the Far-ID prompts are used as negative-prompt candidates at inference (the OOD responses are removed from s^{neg}).

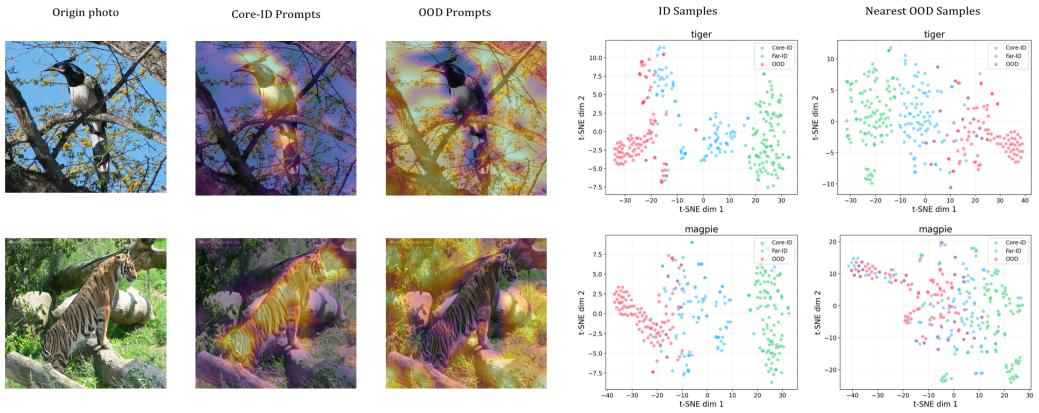

Figure 2: Visualization results of TriPrompt. The heatmap and t-SNE results show that Core-ID and OOD prompts capture foreground/background evidence, while Far-ID captures transition-region features between ID and OOD.

Shot	Setting	iNaturalist		SUN		Places		Textures		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
4-shot	$\mathcal{L}_{align}$ only	8.52	97.91	**18.50**	**95.78**	**29.32**	**91.80**	48.95	86.12	26.32	92.90
4-shot	$\mathcal{L}_{align} + \mathcal{L}_{core}$	8.62	97.94	18.57	95.71	29.65	91.77	48.28	86.34	26.28	92.94
4-shot	$\mathcal{L}_{align} + \mathcal{L}_{ortho}$	**7.26**	98.19	18.56	95.66	31.33	91.32	45.43	**87.94**	25.64	93.28
4-shot	Full objective	7.34	**98.23**	18.82	95.70	30.84	91.53	**45.12**	87.84	**25.53**	**93.32**
16-shot	$\mathcal{L}_{align}$ only	8.03	98.10	19.72	95.24	30.33	91.30	36.03	90.12	23.53	93.69
16-shot	$\mathcal{L}_{align} + \mathcal{L}_{core}$	**7.37**	**98.30**	19.76	95.22	29.84	91.46	35.46	89.67	23.11	93.67
16-shot	$\mathcal{L}_{align} + \mathcal{L}_{ortho}$	10.73	97.51	18.97	95.46	30.58	91.28	34.70	**90.69**	23.75	93.73
16-shot	Full objective	8.57	97.98	**18.16**	**95.48**	**29.30**	**91.58**	**34.13**	90.46	**22.54**	**93.87**

Effectiveness of TriPrompt Inference Design. Table 7 demonstrates the effectiveness of the two components in the TriPrompt inference design. This confirms the effectiveness of the proposed TriPrompt inference strategy.

Table 7: Effectiveness of the TriPrompt inference design.

Shot	Setting	iNaturalist		SUN		Places		Textures		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
4-shot	w/o Far-ID OOD separation	6.62	98.39	20.96	95.23	33.59	91.03	45.50	**87.75**	26.67	93.10
4-shot	w/o top-K_{neg} selection	8.05	98.12	19.57	95.34	30.78	**91.61**	49.61	85.31	27.00	92.60
4-shot	Full TriPrompt	**6.60**	**98.40**	**18.97**	**95.72**	**30.37**	**91.61**	**44.81**	87.69	**25.19**	**93.35**
16-shot	w/o Far-ID OOD separation	9.43	97.90	19.59	95.26	31.05	91.09	**34.49**	90.48	23.64	93.68
16-shot	w/o top-K_{neg} selection	10.29	97.72	19.33	95.42	32.02	90.94	34.91	**90.99**	24.14	93.77
16-shot	Full TriPrompt	**7.91**	**98.17**	**17.61**	**95.57**	**28.36**	**91.80**	35.50	89.85	**22.34**	**93.85**

4.4 Visualization

Prompt Response Heatmaps. We first visualize the GT-class Core-ID heatmaps and the heatmaps of the strongest OOD prompt on ImageNet-1K evaluation samples that are never exposed during training. As shown in Fig. 2, Core-ID and OOD prompts consistently focus on foreground and background regions, respectively, showing that the two branches generalize to unseen test images and capture complementary local evidence.

t-SNE Visualization of Three Feature Types. We further visualize Core-ID, OOD, and Far-ID patch features after OOD separation with t-SNE on both ID test samples and the 50 nearest OOD samples for the current class. Fig. 2 shows that Far-ID indeed captures transition-region features between ID and OOD, and on nearest OOD samples it further attends to non-Core-ID cues that are difficult for the OOD prompts alone to capture.

5 Conclusion

We presented TriPrompt, a CLIP-based few-shot OOD detection framework that progressively models Core-ID, OOD, and Far-ID cues in a shared patch-level feature space. By explicitly modeling transition-region features between ID and OOD, TriPrompt improves local OOD reasoning while remaining compatible with existing global CLIP-based scoring methods. Extensive experiments and visualizations show that TriPrompt improves fine-grained OOD detection and effectively captures complementary local evidence beyond standard ID-focused prompt learning. We hope this progressive decomposition of local evidence can inspire future OOD methods to move beyond binary ID–OOD separation and develop more nuanced uncertainty modeling for open-world recognition.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Key R&D Program of China under Grant No. 2023YFA1008600, in part by the National Natural Science Foundation of China under Grants 62576262, in part by the Natural Science Foundation Basic Research Program of Shaanxi Province under grant 2026JC-YXQN-202, in part by the Fundamental Research Funds for the Central Universities under Grant QTZX25083. This work was also supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62376206. The authors declare no competing interests.

References

Mothilal Asokan, Kebin Wu, and Fatima Albreiki. FineLIP: Extending CLIP's reach via fine-grained alignment with longer text inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yichen Bai, Zongbo Han, Bing Cao, Xiaoheng Jiang, Qinghua Hu, and Changqing Zhang. Id-like prompt learning for few-shot out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17480–17489, 2024.

Mengyuan Chen, Junyu Gao, and Changsheng Xu. Conjugated semantic pool improves ood detection with pre-trained vision-language models. *arXiv preprint arXiv:2410.08611*, 2024.

Qiang Chen, Xiaokang Chen, Jian Wang, Shan Zhang, Kun Yao, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. Group detr: Fast detr training with group-wise one-to-many assignment. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6610–6619, 2023. doi: 10.1109/ICCV51070.2023.00610.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1280–1289, 2022. doi: 10.1109/CVPR52688.2022.00135.

De Cheng, Yubo Li, Chaowei Fang, Shizhou Zhang, Nannan Wang, and Xinbo Gao. Isolating interference factors for robust cloth-changing person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6):6089–6106, 2026. doi: 10.1109/TPAMI.2026.3655110.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. VOS: learning what you don't know by virtual outlier synthesis. In *International Conference on Learning Representations*, 2022.

Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 6568–6576, 2022.

Yufei Guo, Yinan Wu, Haoran Duan, Guiguang Ding, and Jungong Han. PhaseLoRA: Control-regime-conditioned low-rank adaptation for continuous-action vision-language-action policies. *arXiv preprint arXiv:2608.15285*, 2026.

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.

Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 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.

Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Biomedcoop: Learning to prompt for biomedical vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14766–14776, June 2025.

Marc Lafon, Elias Ramzi, Clément Rambour, Nicolas Audebert, and Nicolas Thome. Gallop: Learning global and local prompts for vision-language models. In *European Conference on Computer Vision*, pages 264–282. Springer, 2024.

Haoyang Li, Liang Wang, Chao Wang, Jing Jiang, Yan Peng, and Guodong Long. Dpc: Dual-prompt collaboration for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25623–25632, June 2025.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Zhiwei Ling, Yachen Chang, Hailiang Zhao, Xinkui Zhao, Kingsum Chow, and Shuiguang Deng. Cadref: Robust out-of-distribution detection via class-aware decoupled relative feature leveraging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4968–4977, June 2025.

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

Xinhua Lu, Runhe Lai, Yanqi Wu, Kanghao Chen, Wei-Shi Zheng, and Ruixuan Wang. Fa: Forced prompt learning of vision-language models for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1152–1161, October 2025.

Wenjun Miao, Guansong Pang, Zihan Wang, Jin Zheng, and Xiao Bai. Auxiliary prompt tuning of vision-language models for few-shot out-of-distribution detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–10, 2025. doi: 10.1109/ICCV51701.2025.00454.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35:35087–35102, 2022.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. In *Advances in Neural Information Processing Systems*, 2023.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Gl-mcm: Global and local maximum concept matching for zero-shot out-of-distribution detection. *IJCV*, 2025.

Sen Pei, Jiaxi Sun, Pengda Qin, Qi Chen, Xing Wu, and Xuna Wang. Image background serves as good proxy for out-of-distribution data. In *International Conference on Learning Representations*, 2023. URL https://api.semanticscholar.org/CorpusID:259316200.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Robin Strudel, Ricardo Garcia Pinel, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7242–7252, 2021. URL https://api.semanticscholar.org/CorpusID:234470051.

Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *International Conference on Learning Representations*, 2023.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812, 2023.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. FLAIR: VLM with fine-grained language-informed image representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Zhuo Xu, Xiang Xiang, and Yifan Liang. Overcoming shortcut problem in vlm for robust out-of-distribution detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15402–15412, 2025. doi: 10.1109/CVPR52734.2025.01435.

Geng Yu, Jianing Zhu, Jiangchao Yao, and Bo Han. Self-calibrated tuning of vision-language models for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 37: 56322–56348, 2024.

Fanhu Zeng, Zhen Cheng, Fei Zhu, Hongxin Wei, and Xu-Yao Zhang. Local-prompt: Extensible local prompts for few-shot out-of-distribution detection. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ew3VifXaxZ.

Yabin Zhang, Wenjie Zhu, Chenhang He, and Lei Zhang. Lapt: Label-driven automated prompt tuning for ood detection with vision-language models. In *European conference on computer vision*, pages 271–288. Springer, 2024.

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

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

Table 8: Hyper-parameters used in TriPrompt.

Hyper-parameter	**Description**	**Value**
Backbone	Pretrained CLIP backbone	ViT-B/16
Optimizer	Prompt and projection optimization	SGD
Batch size	Training batch size	128
Learning rate	SGD learning rate	0.002
N	Core-ID prompts per class	2
N_o	Class-shared OOD prompts	300
K_t	Weak non-ID proxy patches	50
K_f	Far-ID mining patches	30
τ	Temperature	1.0
λ_{div}	OOD diversity weight	30.0
λ_{ortho}	Far-ID orthogonality weight	0.5
α	Local score weight at inference	3.0
K_{inf}	Top local scores for aggregation	20
K_{neg}	Retained negative scores at inference	$C/2$

A Implementation Details

We use CLIP ViT-B/16 as the pretrained backbone and keep both the image encoder and the text encoder frozen throughout training. Unless otherwise stated, we follow the Local-Prompt training protocol and optimize the learnable prompts together with the local linear projection using SGD without warm-up.

Training is conducted in three stages. We first optimize the global prompt and the Core-ID local prompts. We then freeze the backbone, the learned ID prompts, and the local projection, and train the class-shared OOD prompts. Finally, we freeze the backbone, the local projection, the Core-ID prompts, and the OOD prompts, and optimize only the Far-ID prompts. We keep K_t and N_o the same as in Local-Prompt and do not further tune them. When TriPrompt is combined with zero-shot methods such as NegLabel that change the distribution of global scores and use an inference temperature different from the local branch, we re-adjust α for compatibility. Table 8 summarizes the hyper-parameters used in the main experiments.

B Global and Core-ID Prompt Learning

For completeness, we briefly summarize the inherited global and Core-ID branches that are used as the base of TriPrompt and fully follow Gallop. Given an input image $\mathbf{x}$, the CLIP image encoder produces a global feature $\mathbf{z}^g$ and a set of local patch features $\mathbb{Z} = \{\mathbf{z}_i\}_{i=1}^{L}$. The learnable global prompt produces the class text feature $\mathbf{t}_c^g$ for each class c, and the global branch is optimized by

$$\mathcal{L}_g = -\log \frac{\exp(\langle \mathbf{z}^g, \mathbf{t}_y^g \rangle / \tau)}{\sum_{c=1}^{C} \exp(\langle \mathbf{z}^g, \mathbf{t}_c^g \rangle / \tau)}. \tag{13}$$

For the sparse Core-ID local branch, let $\{\mathbf{p}_j^+\}_{j=1}^{N}$ denote the N learnable Core-ID prompts, and let $\mathbf{t}_{c,j}^+$ be the corresponding text feature for class c. For each class-prompt pair (c, j), we keep only the top-K patch responses:

$$s_{c,j}^+ = \mathcal{M}_K \left(\{ \langle \mathbf{z}_i, \mathbf{t}_{c,j}^+ \rangle \}_{i=1}^{L} \right), \tag{14}$$

where $\mathcal{M}_K(\cdot)$ denotes the average of the largest K values. Each Core-ID prompt is then optimized with

$$\mathcal{L}_l = -\frac{1}{N} \sum_{j=1}^{N} \log \frac{\exp(s_{y,j}^+ / \tau)}{\sum_{c=1}^{C} \exp(s_{c,j}^+ / \tau)}. \tag{15}$$

These two branches provide the inherited global confidence and reliable local ID evidence on top of which TriPrompt builds the OOD and Far-ID modules.

C Limitation and Future Work

TriPrompt still relies on heuristic patch mining rules to construct weak non-ID proxies and ambiguity-prone Far-ID regions, and its hyper-parameters are currently selected for the ImageNet-1K setting. Future work may study more adaptive region mining strategies, stronger calibration across different global OOD scoring methods, and broader validation on other backbones and open-world recognition scenarios.

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: The abstract and introduction state the main contribution as a progressive local prompting framework for few-shot OOD detection and claim improved OOD performance with compatibility to mainstream CLIP-based scoring methods; these claims are matched by the method, experiments, and conclusion.

   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: The appendix includes a dedicated "Limitation and Future Work" section discussing heuristic patch mining, sensitivity to the ImageNet-1K setting, and future directions for broader validation and stronger calibration.

   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: [N/A]

Justification: The paper does not present formal theoretical theorems or proofs; it is an empirical method paper.

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: The paper specifies the backbone, datasets, evaluation settings, training stages, major hyper-parameters, and inference design in the main text and appendix, which together provide the information needed to reproduce the reported experiments.

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 is provided in the supplementary material together with the information needed to reproduce the main experiments, while the data are based on public benchmarks cited in the paper.

   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: The experimental section and appendix describe the datasets, few-shot settings, optimizer, backbone, prompt configuration, training stages, and key hyper-parameters used in both training and inference.

   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: The paper reports point estimates for OOD detection and ID classification metrics, but it does not provide error bars, confidence intervals, or statistical significance tests.

   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: [No]

   Justification: The paper currently does not report hardware type, memory, runtime, or total compute cost.

   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: The work uses standard public benchmarks and pretrained models for method development and evaluation, and we are not aware of any aspect of the work that conflicts with the NeurIPS Code of Ethics.

   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: The current paper does not include a dedicated discussion of broader societal impacts, although the method may be relevant to reliability and safety in open-world recognition systems.

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: The paper does not release high-risk generative models, scraped datasets, or similarly sensitive assets that would require special 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: [No]

Justification: The paper cites the public datasets and pretrained models it builds on, but it does not explicitly state the corresponding licenses or terms of use in the manuscript.

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: [N/A]

Justification: The paper does not introduce a new dataset, benchmark, or released model asset as a standalone contribution.

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: The work does not involve crowdsourcing or research with human subjects.

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: The paper does not involve human subjects research, so IRB approval is not applicable.

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: LLMs are not an important, original, or non-standard component of the proposed method or experimental pipeline.

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.