Intra-Image Mining and Symmetric Maximum Concept Matching for Few Shot Out-of-Distribution Detection

Kaixiang Chen[1,2], **Pengfei Fang**[1,2*], **Hui Xue**[1,2*]

[1]School of Computer Science and Engineering, Southeast University

[2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China

{kxchen, fangpengfei, hxue}@seu.edu.cn

Abstract

Recent vision-language model (VLM)-based methods have achieved promising results in zero-shot out-of-distribution (OOD) detection by effectively leveraging the local patch features. However, the zero-shot nature inherently comes with two limitations: *1) imperfect local feature prototypes; 2) lack of OOD prototypes.* In this paper, we propose **I**ntra-**I**mage **M**ining (IIM), a lightweight framework designed to overcome these limitations in a few-shot manner. IIM is motivated by the fact that local patches within an image often exhibit diverse semantics, with some patches deviating from the main class concept. Therefore, for each image, we first select the top-k class prototype-related patches as positive samples and leverage them to refine and optimize the local feature prototype. Then, the next top-k among the remaining patches are selected as negatives—serving as OOD signals to construct OOD prototypes. This process yields coherent local positives and challenging negatives, effectively enhancing the model's local feature discrimination. Besides, we propose a novel inference strategy named **S**ymmetric **M**aximum **C**oncept **M**atching (S-MCM). While existing approaches typically adopt an image-to-text scheme—comparing the image features to textual class prototypes—S-MCM further incorporate a text-to-image perspective, leading to more reliable OOD detection. We also propose two benchmarks to analyze the impact of semantic diversity within ID dataset. Built on a frozen VLM, IIM, in conjunction with S-MCM, achieves consistent gains in OOD detection on ImageNet-1k and other benchmarks, outperforming prior methods in FPR95 and AUROC across various few-shot settings.

Code — https://github.com/pSGAme/IIM-SMCM

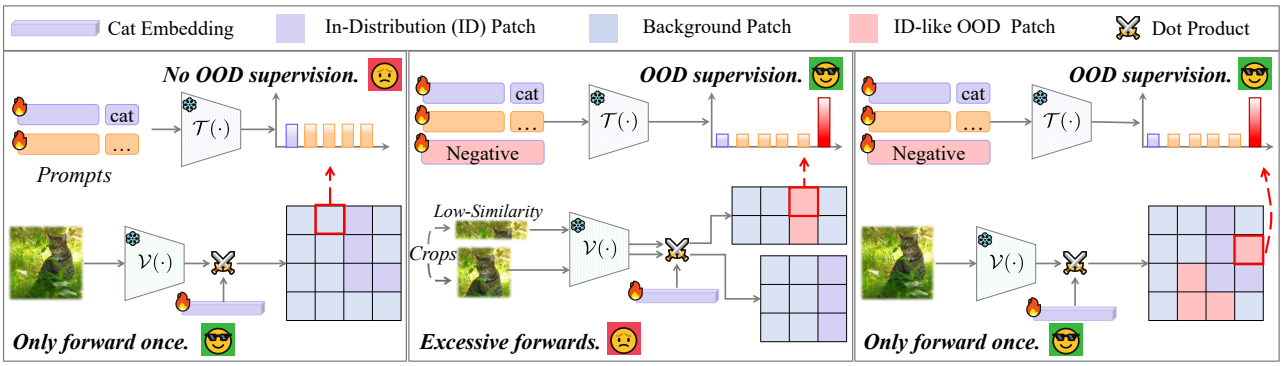
Introduction

Out-of-distribution (OOD) detection aims to identify inputs that fall outside the set of known classes encountered during training—a capability that is critical for the safe and reliable deployment of machine learning systems in real-world applications such as autonomous driving (Zhou et al. 2025a; Ni et al. 2025; Zhang et al. 2025; Zeng et al. 2025c,b), medical diagnosis (Su et al. 2025; Wu et al. 2024), and target recognition (Zhou et al. 2025b; Chen, Fang, and Xue 2025; Li et al. 2025b,a; Zhao 2024). While single-modal supervised learning methods have demonstrated strong performance in OOD detection (Sun et al. 2022; Wei et al. 2022), their reliance on extensive computational resources and labeled data severely limits their scalability. Recent advances in vision-language models (VLMs), such as CLIP (Radford et al. 2021), have exhibited impressive generalization by aligning images and texts in a shared semantic space, making them particularly attractive for the cost-effective **zero-shot** OOD detection. MCM (Ming et al. 2022) is one of the first to leverage CLIP for zero-shot OOD detection by defining the OOD score as the maximum similarity to any in-distribution (ID) class. Building on this, GL-MCM (Miyai et al. 2025) further incorporating local features into the matching process, achieving competitive results.

However, it is important to recognize that zero-shot methods, by their very nature, come with inherent limitations, which manifest in two main aspects: ***1) Imperfect local feature prototype in the textual branch*** — the textual class prototype is obtained by encoding a fixed textual prompt (e.g., "`a photo of a <category>`"), resulting in a global class representation, which is inherently suboptimal to localized image patches, as it fails to capture fine-grained local semantics. Consequently, the full potential of methods like GL-MCM cannot be fully realized. ***2) Lack of OOD Prototypes*** — although ID prototypes are available during evaluation in zero-shot settings, the inherent absence of OOD prototypes restricts the model's capacity to effectively differentiate between ID and OOD concepts.

To address the aforementioned issues, recent research has shifted its focus toward few-shot OOD detection. To mitigate the **first issue**, based on prompt tuning, LoCoOp (Miyai et al. 2023) introduces to identify class-irrelevant regions and pushing them away from all class prototypes (Fig.1-a), thereby enhancing the prototypes' ability to capture fine-grained local semantics. However, due to the lack of OOD prototypes, this is achieved by maximizing the entropy between these regions and the prototypes, which deviates from the intended goal of explicit semantic repulsion. To tackle the **second issue**, ID-Like (Bai et al. 2024) proposes mining ID-like outliers from multiple cropped views of ID images, selecting the most class-dissimilar crop as OOD signals, and introducing negative prompts into the text encoder to build OOD prototypes. While effective, this approach in-

*Co-corresponding authors



(a) LoCoOp (NeurIPS'23) (b) Local-Prompt (ICLR'25) (c) IIM (Ours)

Figure 1: Illustration of (a) LoCoOp, (b) Local-Prompt, and (c) our proposed IIM. The cat embedding is obtained by encoding learnable prompts. IIM incorporates an intra-image mining mechanism to extract both positive and challenging negative patches from a single image. This strategy effectively introduces supervision from OOD prototypes and enhances the local feature representations in the textual branch, all without incurring additional forward passes.

curs substantial computational overhead, as each crop requires a separate forward pass. The recent state-of-the-art method Local-Prompt (Zeng et al. 2025a) (Fig.1-b) draws inspiration from ID-Like, but goes a step further: it extracts the **most class-aligned patches** from the **most dissimilar crop** as OOD signals. By this way, Local-Prompt elegantly addresses both of the aforementioned zero-shot issues and resolves the mismatch in LoCoOp between objective and implementation. However, the computational inefficiency caused by the excessive forward passes still exists.

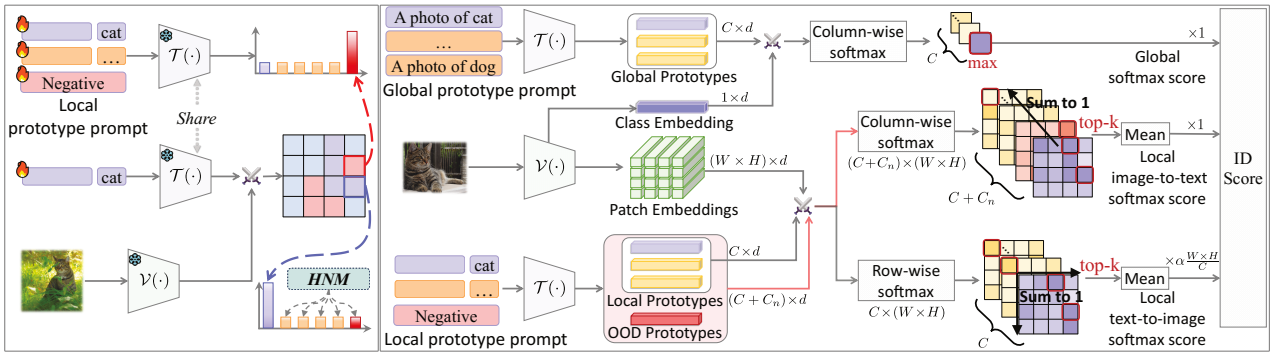
In this paper, we propose **I**nra-**I**mage **M**ining (IIM), a *simple yet effective* framework that simultaneously refines local feature prototypes and constructs OOD prototypes efficiently. IIM builds on the intuition that local patches within an image may exhibit diverse semantics—some of which may deviate from the main class concept. Based on this insight, for each image, we identify the top-k patches most aligned with the class prototype as positives and leverage them to optimize the local feature prototype. These are then excluded, and the next top-k among the remaining patches are selected as negatives—serving as ID-like OOD signals to construct OOD prototypes. It is worth noting that we do not treat all remaining patches as negatives. To avoid learning trivial OOD prototypes that are far from ID distributions, we selectively choose those whose features lie within a semantically meaningful range—close enough to ID to be challenging, This aligns with reality, as truly hard-to-distinguish OOD samples often lie near the boundary of the ID distribution. This process yields semantically coherent positives and diverse, challenging negatives, effectively enhancing the model's local feature discrimination. Compared to Local-Prompt, our method is more efficient in generating OOD signals, as each image requires only a single forward pass.

Besides, we propose a novel OOD evaluation method named **S**ymmetric **M**aximum **C**oncept **M**atching (S-MCM). The *motivation* is rooted in revisiting how the ID set is defined. Conventionally, class prototypes are treated as the ID set, and OOD detection is performed by measuring the similarity between the test image and these prototypes—a setting we refer to as image-to-text. In contrast, S-MCM introduces a complementary perspective: it treats the set of patch features from the test image as the "ID set" and evaluates whether each class prototype is well aligned with this set. We refer to this as text-to-image. By unifying both directions into evaluation, S-MCM achieves significant performance gains in OOD detection.

To summarize, our main contributions are outlined as follows: *1)* We propose IIM, a simple framework that simultaneously refines local feature prototypes and constructs OOD prototypes effectively and efficiently. *2)* We propose S-MCM, a novel OOD inference strategy symmetrically treat the class prototypes and image patches as ID sets, leading to a more reliable OOD detection. *3)* We present two new benchmarks—ImageNet-500-Easy and ImageNet-500-Hard—which offer a simple yet effective means to analyze the impact of semantic diversity within ID dataset.

Related Work

Efficient Out-of-Distribution Detection with CLIP. The goal of out-of-distribution (OOD) detection is to identify inputs that deviate from the training distribution, which is crucial for ensuring the reliability of deployed models. With the success of vision-language models (VLMs) such as CLIP (Radford et al. 2021), VLM-based zero-shot OOD detection has emerged as a promising research direction. For instance, MCM (Ming et al. 2022) utilizes global visual and textual representations from CLIP and defines the OOD score as the maximum similarity to any in-distribution (ID) class. GL-MCM (Miyai et al. 2025) further incorporates the most responsive local feature on the visual feature map to enhance discrimination. R-MCM (Zeng et al. 2025a) generalizes this idea by considering the top-k most responsive local features, improving robustness to local variations. Considering the inherent limitation (*e.g.*, lack of OOD prototypes) in zero-shot paradigm, few-shot learning offers a practical compromise for VLM-based OOD detection with minimal resource cost. LoCoOp (Miyai et al. 2023) introduces few-shot prompt tuning to refine local feature proto-



(a) IIM (Training) (b) S-MCM (Evaluation)

Figure 2: Overview of our proposed IIM (left), and evaluation strategy, S-MCM (right). In training, the trainable prompts interact solely with local patches to generate reliable local prototypes. During evaluation, the local prototypes are used to calculate image-to-text and text-to-image scores, whereas the global feature prototypes are used to compute the MCM score.

types by identifying ID-irrelevant regions and pushing them away from all ID textual embeddings. ID-Like (Bai et al. 2024) proposes setting ID-like OOD outliers as the most class-dissimilar crops. Local-Prompt (Zeng et al. 2025a) advances this idea by extracting the most class-aligned patches from these crops to serve as stronger OOD signals.

Few-Shot Prompt Tuning of CLIP. As a milestone in vision-language learning, CLIP (Radford et al. 2021) has significantly enhanced various downstream tasks through few-shot prompt tuning techniques. Notable examples include CoOp (Zhou et al. 2022) and VPT (Jia et al. 2022), which appends contextual learnable embeddings to the text encoder and the image encoder, respectively. Additionally, MaPLe (Khattak et al. 2023) and MMRL (Guo and Gu 2025) build upon CoOp and VPT by introducing multi-modal prompting to better align vision and language representations, achieving great success. Our IIM builds on CoOp but addresses its limitation of relying solely on global image features while ignoring local patch information.

Preliminary

Few-Shot OOD Detection. Out-of-distribution (OOD) detection aims to determine whether a given input image $x \in \mathcal{X}$ is drawn from the in-distribution (ID) or from an out-of-distribution (OOD) source. The detector learns a scoring function $S : \mathcal{X} \to \mathbb{R}$, where a higher score $S(x)$ indicates a higher likelihood that x is ID, and an input is classified as OOD if $S(x) < \tau$ for some threshold τ:

$$D(x) = \begin{cases} \text{ID}, & S(x) \geq \tau \\ \text{OOD}, & S(x) < \tau \end{cases} \tag{1}$$

In the few-shot setting, the model has access to a limited support set $\mathcal{D}_{\text{support}} = \bigcup_{c \in \mathcal{Y}_{\text{ID}}} \{(x_i^c, y_i^c)\}_{i=1}^K$, where $\mathcal{Y}_{\text{ID}}$ denotes the set of known ID classes and K is the number of labeled samples per class. No OOD samples are available during training. The goal is to utilize this small set of labeled ID data to detect unseen OOD examples $x_{\text{ood}} \sim \mathcal{P}_{\text{OOD}}$, where $\mathcal{Y}_{\text{OOD}} \cap \mathcal{Y}_{\text{ID}} = \emptyset$.

Zero-Shot OOD Detection with CLIP. CLIP (Radford et al. 2021) is a vision-language model (VLM) that consists of an image encoder $\mathcal{V}(\cdot)$ (ViT by default) (Dosovitskiy et al. 2021) and a text encoder $\mathcal{T}(\cdot)$. Given an image x , the image encoder produces both global and local representations: $\mathcal{V}(x) = [v^g, v^l]$, where $v_x^g \in \mathbb{R}^d$ denotes the global class token embedding, with d being the hidden dimension, and $v^l \in \mathbb{R}^{(W \times H) \times d}$ represents the local patch embeddings, where W and H are the width and height of the feature map, respectively. Given a textual prompt (e.g., "`a photo of a <category>`") and its corresponding embedding t_c, where $c \in \{1, 2, ..., C\}$ denotes the c-th category, the MCM (Ming et al. 2022) score is defined as:

$$S_{\text{MCM}}(x) = \max_c \frac{s(v^g, t_c)}{\sum_{i=1}^C s(v^g, t_i)}, \tag{2}$$

where $s(v, t_c) = \exp\big(\text{sim}(v, \mathcal{T}(t_c))/T\big)$ denotes the exponential similarity between the visual feature v and the class prototype $\mathcal{T}(t_c) \in \mathbb{R}^d$, sim denotes the cosine similarity, T is a temperature factor. GL-MCM (Miyai et al. 2025) further incorporates local patch embeddings into consideration:

$$S_{\text{GL-MCM}}(x) = S_{\text{MCM}}(x) + \max_{c,h} \frac{s(v_h^l, t_c)}{\sum_{i=1}^C s(v_h^l, t_i)}, \tag{3}$$

where $h \in \{1, 2, ..., W \times H\}$. Local-Prompt (Zeng et al. 2025a) proposes a general form of GL-MCM, *i.e.*, R-MCM:

$$S_{\text{R-MCM}}(x) = S_{\text{MCM}}(x) + \overline{\mathbb{T}}_k^{(h,c)} \{ \frac{s(v_h^l, t_c)}{\sum_{i=1}^C s(v_h^l, t_i) + \sum_{\tilde{i}=1}^{C_n} s(v_h^l, t_{\tilde{i}})} \}, \tag{4}$$

where $\overline{\mathbb{T}}_k^{(h,c)}$ is the mean of top-k elements across c and h, $t_{\tilde{i}}$ represents the negative prompts for constructing OOD prototypes, which need to be optimized during training, and C_n is the number of OOD prototypes. However, in zero-shot R-MCM, negative prompts are inherently not optimized and should not be used. To clearly distinguish this variant, we denote the version without negative prompts as R-MCM*.

Method

In this section, we will first provide a detailed explanation of proposed **I**nra-**I**mage **M**ining (IIM), as shown in Fig. 2-a. Then we will describe our proposed **S**ymmetric **M**aximum **C**oncept **M**atching (S-MCM), as shown in Fig. 2-b.

Training: Intra-Image Mining

In GL-MCM and R-MCM, $\mathcal{T}(t_c)$ interacts with both v^g and v^l. However, since v^l captures only partial image content, aligning it with the global class prototype $\mathcal{T}(t_c)$ is suboptimal. Moreover, as Eq. (4) shows, introducing OOD prototypes provides clearer semantic contrast, as OOD samples tend to align more closely with these prototypes. Next, we detail how our IIM constructs the local and OOD prototypes.

Positive Patches. Given an image x with label y, we select the top-k patches that are most aligned with the class prototype as positives. We define the sum of the similarity scores between these patches and the class prototype of y as

$$S_k^y(x) = \widehat{\mathbb{T}}_k^{(h)}\{s(v_h^l, t_y^l)\}. \tag{5}$$

Here, $\widehat{\mathbb{T}}_k^{(h)}$ denotes the sum of the top-k elements across h. $s(\cdot)$ is defined in Eq. (2). $t_y^l = [z_1, \ldots, z_L; z_y]$ serves as the local prototype prompt that is optimized during training, where z_i ($i \in 1, \ldots, L$) are the learnable embeddings of the context words, L denotes the number of context tokens. z_y is the embedding corresponding to the y-th class name.

Intuitively, we want $S_k^y(x)$ to be large, ensuring that the selected top-k patches are strongly aligned with class y while remaining distant from other classes. Inspired by hard negative mining (Chen, Gong, and Zhang 2024), we go a step further: we additionally identify the top-k patches with highest similarity to each $c \neq y$, and the similarities, denoted as $S_k^c(x)$. We then explicitly push these patches away from the non-y class prototypes

ID-like OOD Patches. Constructing meaningful OOD prototypes relies on access to OOD signals during training. Local-Prompt (Zeng et al. 2025a) obtains these signals by extracting the most class-aligned patches from the most class-dissimilar crop. However, this approach incurs substantial computational overhead, as each crop requires its own forward pass. Moreover, its effectiveness depends heavily on randomness: if the "most class-dissimilar" crop still resembles the ground-truth class, its most class-aligned patches may not serve as meaningful OOD signals.

To address these limitations, we propose a simple yet effective alternative. Rather than generating a separate most class-dissimilar crop ($\mathcal{P}_{dis}$), we first remove the top-k most class-aligned patches from the original image. The remaining region, inherently containing less class-relevant information, and thus can be set to $\mathcal{P}_{dis}$. We then extract the top-k most class-aligned patches from $\mathcal{P}_{dis}$ as OOD signals. This strategy ensures the OOD signals come from less relevant regions, offering robust and informative OOD signals without additional forward passes and computational overhead.

As a result, we use the the top-$2k{\sim}k$ patches that are most aligned with the class prototype as ID-like OOD signals:

$$I_{2k\sim k}^y(x) = \{q \mid s(v_q^l, t_y^l) \in \mathbb{T}_{2k\sim k}^{(h)}\{s(v_h^l, t_y^l)\}\}, \tag{6}$$

where $\mathbb{T}_{2k\sim k}^{(h)}$ is the set of top-$2k{\sim}k$ elements, and $I_{2k\sim k}^y(x)$ denotes the index set of these ID-like OOD patches. We define the sum of the similarity scores between these patches and the class prototype of c as:

$$\widetilde{S}_{2k\sim k}^c(x) = \sum_{p \in I_{2k\sim k}^y(x)} s(v_p^l, \mathcal{T}(t_c^l)). \tag{7}$$

Negative Prompts. We establish a set of negative prompts to generate OOD prototypes, where the number of negative prompts is denoted by C_n, and the c-th negative prompt is denoted as $t_{\widetilde{c}}^l$. As we mentioned before, rather than treating all remaining patches as OOD patches, we selectively choose those whose features fall within a semantically meaningful range — sufficiently close to the ID distribution to serve as challenging OOD patches. The purpose of this design is to avoid trivial OOD prototypes that lie far from the ID distribution, ensuring meaningful OOD supervision.

Training Losses

Positive Term. The loss for the positive patches is:

$$\mathcal{L}_{pos}(x, y) = -\log \frac{S_k^y(x)}{\sum_{c=1}^{C} S_k^c(x) + \sum_{\widetilde{c}=1}^{C_n} S_k^{\widetilde{c}}(x)}. \tag{8}$$

Negative Term. The loss for the ID-Like OOD patches is:

$$\mathcal{L}_{neg}(x, y) = -\log \frac{\sum_{\widetilde{c}=1}^{C_n} \widetilde{S}_{2k\sim k}^{\widetilde{c}}(x)}{\sum_{c=1}^{C} \widetilde{S}_{2k\sim k}^{c}(x) + \sum_{\widetilde{c}=1}^{C_n} \widetilde{S}_{2k\sim k}^{\widetilde{c}}(x)}. \tag{9}$$

Diversity Regularization. We not only encourage greater diversity in negative prompts but also extend this principle to enhance the diversity of class prompts:

$$\begin{aligned}\mathcal{L}_{div} = &\frac{1}{C_n(C_n-1)/2} \sum_{\widetilde{i}=1}^{C_n} \sum_{\widetilde{j}=\widetilde{i}+1}^{C_n} \text{sim}(\mathcal{T}(t_{\widetilde{i}}^l), \mathcal{T}(t_{\widetilde{j}}^l)) \\ &+ \frac{1}{C(C-1)/2} \sum_{i=1}^{C} \sum_{j=i+1}^{C} \text{sim}(\mathcal{T}(t_i^l), \mathcal{T}(t_j^l))\end{aligned} \tag{10}$$

Overall Loss. The overall loss is formulated as:

$$\mathcal{L} = \mathcal{L}_{pos} + \lambda_{neg} \cdot \mathcal{L}_{neg} + \lambda_{div} \cdot \mathcal{L}_{div}, \tag{11}$$

where λ_{neg} and λ_{div} are two hyper-parameters control the weights of the two losses.

Evaluate: Symmetric Maximum Concept Matching

When revisiting GL-MCM and R-MCM, we can observe that both methods approach OOD detection from the perspective of determining whether a test image belongs to the ID set — that is, whether it aligns with the distribution of ID class prototypes (*image-to-text*). In contrast, one can also view the problem from the opposite direction: treating the patches of a test image as an "ID set" and assessing whether these class prototypes belong to the distribution defined by these patches, i.e., *text-to-image*. Based on such an intuition, we propose the S-MCM score $S_{\text{S-MCM}}$, defined as:

$$\begin{aligned}S_{i2t} &= \overline{\mathbb{T}}_k^{(h,c)}\{\frac{s(v_h^l, t_c^l)}{\sum_{c=1}^{C} s(v_h^l, t_c^l) + (\sum_{\widetilde{c}=1}^{C_n} s(v_h^l, t_{\widetilde{c}}^l))}\}, \\ S_{t2i} &= \frac{WH}{C} \cdot \overline{\mathbb{T}}_k^{(h,c)}\{\frac{s(v_h^l, t_c^l)}{\sum_{h=1}^{W\cdot H} s(v_h^l, t_c^l)}\}, \\ S_{\text{S-MCM}}(x) &= S_{\text{MCM}}(x) + S_{i2t} + \alpha \cdot S_{t2i}.\end{aligned} \tag{12}$$

S_{i2t} also introduces an similarity measure with negative prompts in its denominator, which is highlighted using parentheses. In S_{i2t}, the multiplication by $\frac{WH}{C}$ ensures that S_{t2i} and S_{i2t} have comparable average magnitudes. The parameter α controls the weight, allowing users to balance the relative importance of S_{t2i} and S_{i2t}.

ID Dataset	Method	OOD Dataset								Average	
		iNaturalist		SUN		Places		Texture			
		FPR95↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑
ImageNet -1k	0-*shot*										
	MCM	30.91	94.61	37.59	92.57	44.69	89.77	57.77	86.11	42.74	90.77
	GL-MCM	15.18	96.71	30.42	93.09	38.85	89.90	57.93	83.63	35.47	90.83
	R-MCM*	12.59	97.23	27.43	93.86	35.85	90.76	58.40	83.56	33.57	91.35
	4-*shot*										
	LoCoOp	21.67	95.69	22.98	95.07	31.41	92.10	49.79	87.85	31.46	92.68
	Local-Prompt	**9.65**	**97.87**	20.40	95.57	29.39	92.67	51.20	88.00	27.66	93.53
	IIM	9.98	97.47	**16.73**	**96.37**	**27.53**	**92.97**	**38.44**	**92.14**	**23.17**	**94.74**
	16-*shot*										
	LoCoOp	16.05	96.86	23.44	95.07	32.87	91.98	42.28	90.19	28.66	93.52
	Local-Prompt	**8.71**	**98.10**	23.97	94.85	32.50	92.32	47.93	89.04	28.27	93.58
	IIM	9.24	97.94	**15.94**	**96.56**	**26.47**	**93.45**	**33.82**	**93.11**	**21.37**	**95.27**
ImageNet -500-Easy (1∼ 500)	0-*shot*										
	MCM	14.85	97.23	8.08	98.29	15.41	96.16	20.46	96.18	14.70	96.79
	GL-MCM	4.30	98.83	7.82	98.17	15.32	95.87	27.91	93.26	13.84	96.53
	R-MCM*	3.12	98.99	6.45	98.50	13.56	96.38	25.26	94.05	12.10	96.98
	4-*shot*										
	LoCoOp	4.18	98.88	3.28	99.15	10.90	97.04	15.45	96.42	8.45	97.87
	Local-Prompt	2.97	98.88	4.10	98.95	11.23	97.01	14.76	96.68	8.26	97.88
	IIM	**2.38**	**98.96**	**2.62**	**99.34**	**9.48**	**97.29**	**9.87**	**98.16**	**6.09**	**98.44**
	16-*shot*										
	LoCoOp	3.80	99.06	3.09	99.23	9.77	97.33	12.82	96.94	7.37	98.14
	Local-Prompt	1.87	99.12	3.39	99.11	10.17	97.26	12.53	97.02	6.99	98.13
	IIM	**1.86**	**99.27**	**2.53**	**99.56**	**9.05**	**97.51**	**8.30**	**98.26**	**5.44**	**98.65**

Table 1: Comparison with state-of-the-art (SOTA) methods on ImageNet-1k and ImageNet-500-Easy. Here, * denotes that negative prompts are *excluded* during evaluation in S-MCM or R-MCM. The best results are shown in bold.

Experiments

Experimental Setup

Datasets. Following existing works (Zeng et al. 2025a; Miyai et al. 2023), we use ***ImageNet-1K*** (Deng et al. 2009) and ***ImageNet-100*** as the ID datasets. Based on ImageNet-1K, we observe that the first 500 classes, which mainly contain natural categories, are relatively easier for OOD separation, whereas the last 500 classes, primarily composed of man-made objects, form a denser and more ambiguous feature space. Although this class-wise split is relatively coarse, it provides a simple yet effective way to analyze the impact of semantic diversity within the ID data. We therefore evaluate the two halves (1–500 vs. 501–1000) independently as ID datasets under the same OOD setting, and refer to them as ***ImageNet-500-Easy*** and ***ImageNet-500-Hard***, respectively. The corresponding OOD datasets are ***iNaturalist*** (Van Horn et al. 2018), ***SUN*** (Xiao et al. 2010), ***Texture*** (Cimpoi et al. 2014) and ***Place365*** (Zhou et al. 2017).

Implementation Details. We perform prompt tuning on a pre-trained CLIP (ViT-B/16) with model dimension $d = 512$. Following Local-Prompt (Zeng et al. 2025a), the prompt length L is 16 and the temperature T is fixed at 1. For top-k selection, we set $k = 50$ for both training and testing, and for OOD prototypes, the number $C_n = 300$. Hyperparameters are set as $\lambda_{\text{neg}} = 5$, $\lambda_{\text{div}} = 0.5$, and $\alpha = 1.0$ in S-MCM. We train for 50 epochs with SGD, learning rate 2.5×10^{-4}, and batch size 32. Experiments run on a single NVIDIA 4090 GPU, except ImageNet-1k on a single H100. Results are averaged over three runs with seeds 1, 2, and 3.

Comparison Methods. To evaluate the effectiveness of IIM and S-MCM, we compare them with two categories of methods: zero-shot and local-patch-based few-shot OOD detection methods. For zero-shot methods, we include MCM (Ming et al. 2022), GL-MCM (Miyai et al. 2025), and R-MCM (Zeng et al. 2025a). For few-shot baselines, we consider LoCoOp (Miyai et al. 2023) and Local-Prompt (Zeng et al. 2025a). The superscript * indicates the *exclusion* of negative prompts (e.g., R-MCM*, IIM*).

Evaluation Metrics. We use the following metrics: (1) the false positive rate of OOD images when the true positive rate of in-distribution images is at 95% (FPR95), (2) the area under the receiver operating characteristic curve (AUROC).

Main Results

The OOD detection performances are summarized in Tab. 1 and Tab. 2, yielding several key insights: ***1) R-MCM* achieves the best performance among zero-shot performance.*** In fact, MCM and GL-MCM can be regarded as simplified variants of R-MCM*, which makes its superior performance reasonable and expected. ***2) Our method consistently outperforms existing SOTA methods.*** In particular, under the 16-shot setting with ImageNet-1k as the ID dataset, IIM achieves an average FPR95 that is 6.90% better than that of Local-Prompt. Furthermore, under the 4-shot

ID Dataset	Method	OOD Dataset								Average	
		iNaturalist		SUN		Places		Texture			
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
ImageNet -500-Hard (501∼ 1k)	0-*shot*										
	MCM	30.48	93.86	50.92	89.45	51.25	88.61	64.06	82.51	49.18	88.61
	GL-MCM	20.49	95.75	38.56	91.28	43.40	89.54	60.95	82.05	40.85	89.66
	R-MCM*	17.41	96.29	37.43	91.51	41.85	89.97	59.71	82.39	39.10	90.04
	4-*shot*										
	LoCoOp	25.67	94.44	38.34	90.99	42.86	89.42	48.25	87.66	38.78	90.63
	Local-Prompt	17.21	96.51	37.11	92.18	41.20	90.79	55.88	85.46	37.85	91.24
	IIM	**12.51**	**97.26**	**26.14**	**94.16**	**32.87**	**92.25**	**39.70**	**90.51**	**27.81**	**93.55**
	16-*shot*										
	LoCoOp	28.49	94.11	38.03	92.84	41.82	89.97	45.33	88.26	38.42	91.30
	Local-Prompt	14.70	96.91	36.23	92.46	40.32	91.00	53.52	86.10	36.19	91.62
	IIM	**10.35**	**97.62**	**24.41**	**94.34**	**31.74**	**92.50**	**38.71**	**90.66**	**26.30**	**93.78**
ImageNet -100	0-*shot*										
	MCM	18.13	96.77	36.45	94.54	34.52	94.36	41.22	92.25	32.58	94.48
	GL-MCM	3.45	98.36	16.75	96.91	19.72	96.26	36.90	92.26	19.21	95.95
	R-MCM*	2.87	98.93	17.59	96.93	18.93	96.50	37.32	92.52	19.18	96.22
	4-*shot*										
	LoCoOp	8.02	98.21	13.59	97.31	20.20	95.73	25.73	94.87	16.88	96.53
	Local-Prompt	6.99	98.05	24.34	96.03	26.04	95.49	26.32	93.67	23.42	95.81
	IIM*	**2.72**	**99.18**	**10.10**	**98.05**	**15.27**	**97.05**	**19.26**	**96.45**	**11.84**	**97.68**
	16-*shot*										
	LoCoOp	5.04	98.49	10.47	97.58	16.55	96.29	17.44	96.20	12.38	97.14
	Local-Prompt	4.94	98.57	20.06	96.73	22.29	96.19	31.09	95.54	19.60	96.76
	IIM*	**1.21**	**99.52**	**7.34**	**98.43**	**11.76**	**97.38**	**13.42**	**97.44**	**8.43**	**98.21**

Table 2: Comparison with state-of-the-art (SOTA) methods on ImageNet-500-Hard and ImageNet-100.

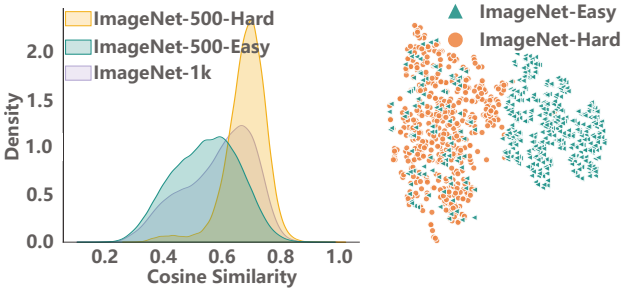


Figure 3: Pairwise similarity distribution (left) and t-SNE visualization (right) of ImageNet-500-Easy/Hard.

HNM	NP	$\mathcal{L}_{neg}$	$\mathcal{L}_{div}$	FPR95↓	AUROC↑
✓	✓			30.02	92.42
✓				30.17	92.14
				31.37	91.11
✓	✓	✓		28.33	93.27
✓	✓		✓	29.55	92.72
✓	✓	✓	✓	**27.81**	**93.55**

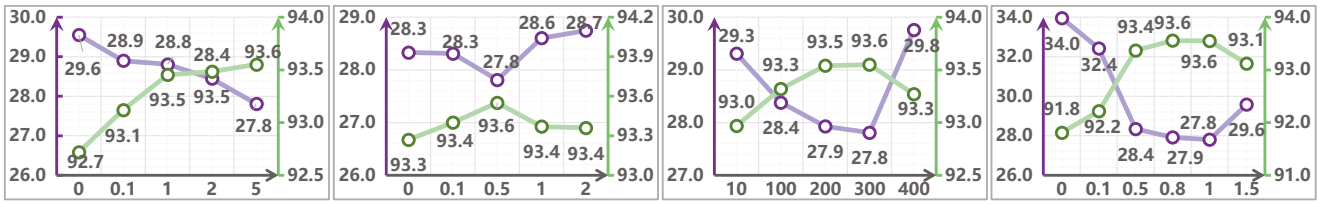
Table 3: Ablation study on each component. The first row means using $\mathcal{L}_{pos}$. NP: negative prompts; HNM: hard negative mining

setting with ImageNet-500-Hard as the ID dataset, IIM reduces the average FPR95 by 10.04%, representing a substantial improvement. These results clearly demonstrate the remarkable effectiveness of our IIM and S-MCM. ***3) The two 500-class subsets of ImageNet exhibit significantly different levels of OOD difficulty.*** To explain this, we visualize the text class prototypes (encoded by CLIP) for two subsets in Fig. 3. As shown in Fig. 3 (left), ImageNet-500-Hard prototypes exhibit much higher similarity, leading to a denser distribution of class centers in the feature space, as shown in Fig. 3 (right). Such increased density reduces the model's ability to distinguish OOD samples, as overly similar class prototypes can blur decision boundaries and hinder effective OOD separation. Therefore, we recommend ImageNet-500-Hard as a challenging benchmark for evaluating OOD detection performance in the future.

Ablation Studies

We perform ablation studies on a 4-shot model trained with ImageNet-500-Hard as the ID dataset to evaluate the effectiveness of each component, unless otherwise noted.

Effectiveness of Each Component. The ablation results are shown in Tab. 3. Key observations include: ***1) Hard Negative Mining (HNM) is essential.*** It yields a modest FPR95 reduction of 1.20%. ***2) Negative prompts alone offer limited gains.*** Using only negative prompts reduces FPR95 by just 0.15%. However, when combined with $\mathcal{L}_{neg}$, the improvement becomes significant (1.84%), underscoring the value of



(a) λ_{neg} (b) λ_{div} (c) C_n (d) α

Figure 4: Sensitivity analysis of λ_{neg}, λ_{div}, C_n, and α, where the purple curve represents FPR95(↓) and the green curve represents AUROC(↑).

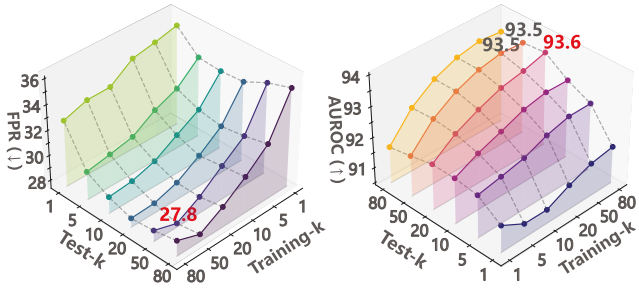


Figure 5: Ablation study of training k and test k.

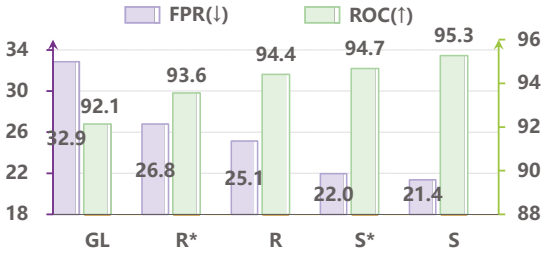


Figure 6: Ablation study of OOD evaluators using ImageNet-1K as the in-distribution dataset.

OOD signals. ***3) $\mathcal{L}_{div}$ provides additional benefit.*** Its inclusion reduces FPR95 by around 0.5%. We also report the performance as λ_{neg} and λ_{div} vary, as shown in Fig. 4 (a-b), the best performance is achieved when $\lambda_{neg}=5$ and $\lambda_{div}=0.5$.

Number of Negative Prototypes and α in S-MCM. The number of negative prompts C_n stands for the OOD prototypes learns from the OOD signals, while α denote the importance of the text-to-image score. Results in Fig. 4 (c) and (d) indicate that the performance is achieved optimal when $C_n=300$ and $\alpha=1.0$.

Training and Testing k. The regional number k determines how selectively the model attends to localized regions within the input. **During training**, the top-k regions are used to learn ID-related local prototype prompts, while the top-$2k \sim k$ regions help learn negative prompts from OOD-like areas. **During testing**, the top-k ID-related regions are used to compute the OOD score. As shown in Fig. 5, both excessively large and small values of k degrade performance—larger k introduces noise from irrelevant regions, while smaller k misses important cues. The optimal k for both training and testing is 50, aligning with our analysis.

Method	Epoch	Batch Size	Time	Batch Time
Local-Prompt*	30	256	58 min	17115 ms
Local-Prompt	50	32	275 min	5331 ms
LoCoOp*	50	32	**19 min**	**350 ms**
IIM	50	32	26 min	496 ms

Table 4: Comparison of training efficiency. * indicates the original training configuration.

Different Evaluation Methods. We compare our proposed S-MCM with existing evaluators (MCM, GL-MCM, and R-MCM) under the 16-shot setting based on the IIM framework. To assess the effect of negative prompts, we also evaluate S-MCM* and R-MCM*, where such prompts are removed. As shown in Fig. 6, S-MCM outperforms existing strategies, with further gains from incorporating negative prompts. An exception occurs on ImageNet-100, where testing with negative prompts slightly degrades performance. We attribute this to the relatively small number of classes, which may lead to less discriminative OOD prototypes.

Training Efficiency. In addition to performance, we also assess training efficiency. For both baselines, we adopt their official configurations and further reimplement Local-Prompt under our training setup for fairness. Despite using more epochs and a smaller batch size, IIM shortens the training time by 32 minutes compared to Local-Prompt. When trained under identical settings, Local-Prompt takes 275 minutes to converge, due to its multiple forward passes.

Conclusion

In this paper, we address two key issues in zero-shot OOD detection: imperfect local prototypes and the lack of OOD prototypes. We propose IIM, which optimizes local prototype prompts and OOD prototype prompts by learning from the top-k and top-$2k \sim k$ class-related patches, respectively. In addition, we introduce an evaluation metric, S-MCM, that considers both image-to-text and text-to-image similarities. Extensive experiments demonstrate the effectiveness of our method. Finally, we present two new benchmarks—ImageNet-500-Easy and ImageNet-500-Hard—which offer a simple yet effective means to analyze the impact of semantic diversity within ID dataset.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62476056 and 62306070) and the Social Development Science and Technology Project of Jiangsu Province (No. BE2022811). This work was also supported in part by the Southeast University Start-Up Grant for New Faculty under Grant 4009002309. Furthermore, the work was supported by the Big Data Computing Center of Southeast University and the SEU Innovation Capability Enhancement Plan for Doctoral Student (CXJH_SEU 25133). This work was also supported by "the Fundamental Research Funds for the Central Universities (2242025K30024)".

References

Bai, Y.; Han, Z.; Cao, B.; Jiang, X.; Hu, Q.; and Zhang, C. 2024. Id-like prompt learning for few-shot out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17480–17489.

Chen, K.; Fang, P.; and Xue, H. 2025. DePro: Domain Ensemble using Decoupled Prompts for Universal Cross-Domain Retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, 958–967.

Chen, K.; Gong, T.; and Zhang, L. 2024. Camera-Aware Recurrent Learning and Earth Mover's Test-Time Adaption for Generalizable Person Re-Identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1): 357–370.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Guo, Y.; and Gu, X. 2025. MMRL: Multi-Modal Representation Learning for Vision-Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, 25015–25025.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, 709–727. Springer.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19113–19122.

Li, Z.; Chen, Z.; Wen, H.; Fu, Z.; Hu, Y.; and Guan, W. 2025a. Encoder: Entity mining and modification relation binding for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5101–5109.

Li, Z.; Fu, Z.; Hu, Y.; Chen, Z.; Wen, H.; and Nie, L. 2025b. FineCIR: Explicit Parsing of Fine-Grained Modification Semantics for Composed Image Retrieval. *https://arxiv.org/abs/2503.21309*.

Ming, Y.; Cai, Z.; Gu, J.; Sun, Y.; Li, W.; and Li, Y. 2022. Delving into Out-of-Distribution Detection with Vision-Language Representations. In *Advances in Neural Information Processing Systems*.

Miyai, A.; Yu, Q.; Irie, G.; and Aizawa, K. 2023. LoCoOp: Few-Shot Out-of-Distribution Detection via Prompt Learning. In *Thirty-Seventh Conference on Neural Information Processing Systems*.

Miyai, A.; Yu, Q.; Irie, G.; and Aizawa, K. 2025. GL-MCM: Global and Local Maximum Concept Matching for Zero-Shot Out-of-Distribution Detection. *International Journal of Computer Vision*, 1–11.

Ni, C.; Zhao, G.; Wang, X.; Zhu, Z.; Qin, W.; Huang, G.; Liu, C.; Chen, Y.; Wang, Y.; Zhang, X.; et al. 2025. Recondreamer: Crafting world models for driving scene reconstruction via online restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1559–1569.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.

Su, X.; Mao, Q.; Wu, Z.; Lin, X.; You, S.; Liao, Y.; and Xu, C. 2025. Large language models driven neural architecture search for universal and lightweight disease diagnosis on histopathology slide images. *npj Digital Medicine*, 8(1): 682.

Sun, Y.; Ming, Y.; Zhu, X.; and Li, Y. 2022. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, 20827–20840. PMLR.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.

Wei, H.; Xie, R.; Cheng, H.; Feng, L.; An, B.; and Li, Y. 2022. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*, 23631–23644. PMLR.

Wu, Z.; Xu, H.; Long, Y.; You, S.; Su, X.; Long, J.; Luo, Y.; and Xu, C. 2024. Detecting any instruction-to-answer interaction relationship: Universal instruction-to-answer navigator for med-vqa. In *Forty-first International Conference on Machine Learning*.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from

abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 3485–3492. IEEE.

Zeng, F.; Cheng, Z.; Zhu, F.; Wei, H.; and Zhang, X.-Y. 2025a. Local-prompt: Extensible local prompts for few-shot out-of-distribution detection. In *The Thirteenth International Conference on Learning Representations*.

Zeng, S.; Chang, X.; Xie, M.; Liu, X.; Bai, Y.; Pan, Z.; Xu, M.; and Wei, X. 2025b. FutureSightDrive: Thinking Visually with Spatio-Temporal CoT for Autonomous Driving. *arXiv preprint arXiv:2505.17685*.

Zeng, S.; Qi, D.; Chang, X.; Xiong, F.; Xie, S.; Wu, X.; Liang, S.; Xu, M.; and Wei, X. 2025c. JanusVLN: Decoupling Semantics and Spatiality with Dual Implicit Memory for Vision-Language Navigation. *arXiv preprint arXiv:2509.22548*.

Zhang, H.; Zhang, X.; Liu, Y.; Gao, S.; and Ma, D. 2025. Event-Triggered Adaptive Tracking Control for USV Based on Enhanced Optimized Backstepping Technique. *ISA Transactions*.

Zhao, Z. 2024. Balf: Simple and efficient blur aware local feature detector. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3362–3372.

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

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, S.; Nie, J.; Zhao, Z.; Cao, Y.; and Lu, X. 2025a. FocusTrack: One-Stage Focus-and-Suppress Framework for 3D Point Cloud Object Tracking. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 7366–7375.

Zhou, S.; Zhang, X.; Chu, X.; Zhang, B.; Zhao, Z.; and Lu, X. 2025b. FastPillars: A Deployment-friendly Pillar-based 3D Detector. *IEEE Transactions on Circuits and Systems for Video Technology*.