NegAS: Negative Label Guided Attention and Scoring for Out-of-Distribution Object Detection with Vision-Language Models

Yingjie Zhang[1], Shuai Li[1] *, and Peng Wang[1]

Northwestern Polytechnical University, Xi'An, China
zyjzyj@mail.nwpu.edu.cn, {cssli, peng.wang}@nwpu.edu.cn

Abstract. Out-of-Distribution (OOD) detection is essential for ensuring the robustness and reliability of object detection systems deployed in safety-critical applications. While prior research has mainly focused on uni-modal detectors or vision-language model (VLM) based classifiers, the potential of VLM-based object detectors in OOD scenarios remains underexplored. In this work, we take the first step toward building OOD object detection methods upon VLMs. We identify two key challenges: (i) learning discriminative features to separate in-distribution (ID) from OOD instances, and (ii) designing scoring functions consistent with VLM probabilistic outputs. Hence, we introduce Negative Label Guided Attention and Scoring (NegAS). To address (i), we propose a negative label guided attention module (NegA), where LLM-generated, visually-similar but semantically-different negative labels are used to guide attention toward potential OOD background regions. To address (ii), we introduce a novel sigmoid-based OOD scoring function (NegS) that leverages both ID and negative labels, producing strong responses for ID instances and suppressed responses for OOD ones. Extensive experiments demonstrate that our approach improves OOD detection performance by a large margin while maintaining competitive ID accuracy, *e.g.*, reducing the FPR95 by 11.4% on the COCO dataset and 25.5% on the OpenImages dataset compared to the baseline model. While initially designed for dense VLM detectors like YOLO-World, we successfully adapt NegAS to Grounding DINO, a query-based VLM transformer, achieving significant improvements, demonstrating the generalizability of our framework. The code is available at `https://github.com/yjzzyj/NegAS`.

Keywords: Out-of-Distribution Object Detection · Vision-Language Model · Negative Label Guided Attention and Scoring

1 Introduction

Deep learning based object detection methods have achieved remarkable progresses in recent years. However, their accuracy and reliability often degrade in real-world scenarios. A critical issue arises when these models encounter objects

* Corresponding Author

of classes that were not present during training, known as Out-of-Distribution (OOD) data. In such cases, models tend to misclassify OOD objects as in-distribution (ID) ones, which can lead to severe consequences in safety-critical domains such as autonomous driving and medical image analysis. This makes OOD object detection (simply referred to as OOD detection) an essential research problem for enhancing the overall trustworthiness and reliability of deep learning detection methods.

OOD classification has been extensively studied in traditional uni-modal classification models, though such approaches generally do not incorporate textual information. Recently, vision-language models (VLMs), such as CLIP [26], which align textual and visual spaces, have demonstrated strong generalizability and are gradually becoming the mainstream for classification tasks. Consequently, OOD classification methods built upon CLIP have attracted significant attention. Existing research in this direction mainly explores how to exploit the alignment between text and vision to improve OOD classification performance, and can broadly be grouped into two categories. The first category, such as LoCoOp [23], NegPrompt [17], LSN [25], IDPrompt [1], GalLoP [15], focuses on designing negative prompts to enhance feature discriminability, to separate ID and OOD samples more effectively in the feature space. The second, such as MCM [22], GL-MCM [24], NegLabel [12], EOE [2], emphasizes developing improved OOD scoring functions that enable more accurate and reliable distinction between ID and OOD instances.

Motivated by the advances in OOD classification, the field of OOD detection has recently gained growing interest. Representative approaches—VOS [6], STUD [5], SyncOOD [19], WFS [35]—largely follow the Faster R-CNN [27] framework and utilize visual features only. More recently, vision language object detection models have demonstrated superior generalization compared to traditional uni-modal detectors. For example, YOLO-World [3] aligns local image regions with text queries and achieves strong performance in downstream detection tasks. Nevertheless, its ability to address OOD detection remains largely unexplored. In this paper, we bridge this gap by investigating OOD object detection methods specially built upon VLMs with two mainly-focused challenges.

The first challenge is to better separate ID and OOD samples in the feature space. YOLO-World introduces a text-guided attention layer that uses ID labels to emphasize foreground features and align vision–language representations. However, it treats all background regions uniformly. We argue that background regions vary in importance. We observe that certain background areas may contain potential OOD objects. Focusing more on these regions can yield a clearer decision boundary between ID and OOD instances. To this end, we propose a Negative label guided Attention (NegA) module that uses curated negative labels to reweight background features, guiding the network toward characteristics that contradict ID labels or resemble OOD labels. Besides, we design an LLM-based algorithm to generate high-quality negative labels that possess sufficient semantic differences from the ID labels, ensuring they provide meaningful and diverse guidance information.

The second challenge lies in designing an appropriate OOD scoring function. Most OOD methods are built on Faster R-CNN [27] trained with softmax-based losses, yielding softmax-based scores. In contrast, VLM detectors are trained with sigmoid-based losses, making a direct transfer suboptimal. We address this with a Negative label guided Scoring (NegS) function tailored to VLM-based OOD object detection. Aligned with the probabilistic formulation of VLMs, it explicitly models the relative separation between ID and OOD samples, reducing overconfident predictions on unseen data.

We implement NegAS on two representative VLM detectors with different architectures—a dense detector YOLO-World and a query-based detector Grounding DINO—and conduct experiments on both. Results on these two baselines consistently validate the effectiveness of NegAS for OOD detection.

Our main contributions are as follows:

- To the best of our knowledge, this is the first work to explore the potential of vision-language model (VLM) detectors for OOD object detection.
- We propose NegAS, which consists of a Negative label guided Attention (NegA) module and Scoring (NegS) function for VLM-based OOD object detection to enhance the separability between ID and OOD samples. Furthermore, we open-source our complete prompt templates and the comprehensive sets of curated negative labels to ensure full reproducibility.
- We implement NegAS on YOLO World, a dense VLM detector, and Grounding DINO, a query-based VLM detection transformer. Extensive experiments demonstrate the effectiveness of NegAS. This bridges the architectural gap between dense and query-based VLM detectors, demonstrating the generalizability of NegAS.

2 Related Works

2.1 Pretrained vision-language models

Thanks to the Transformer architecture [31], pretrained vision-language models (VLMs) such as CLIP [26] have profoundly influenced vision tasks. Trained with contrastive learning, CLIP exhibits strong cross-modal alignment and generalization, becoming the de facto foundation for many vision-language models. Its robust representation ability has enabled a wide range of vision-language applications and motivated extensive research on adapting VLMs to downstream tasks.

2.2 VLM object detectors

Built upon pretrained VLMs, recent advances have enabled object detectors to move beyond closed-set taxonomies. ViLD [9] and GLIP [16] pioneered knowledge distillation and grounding-based approaches for open-vocabulary detection. Grounding DINO [20] further integrates grounded pretraining into

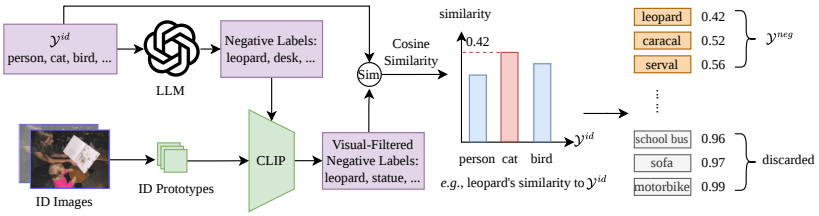

Fig. 1: Negative Label Mining. Given ID category texts, we prompt an LLM to generate candidate negative labels. Each candidate is compared to every ID label by computing their cosine similarity in the text embedding space. The candidates are sorted by their maximum similarity to the ID label set, and the K labels with the lowest maximum similarity are selected to form $\mathcal{Y}^{\text{neg}}$. Labels with high similarity (*e.g.*, "school bus", "sofa") are discarded to avoid semantic overlap with ID categories.

a transformer-based detector with language-guided query selection and cross-modality decoding, achieving state-of-the-art zero-shot performance. YOLO-World [3] extends the real-time YOLO framework with reparameterizable vision-language fusion and a prompt-then-detect paradigm, enabling practical open-vocabulary detection in real-world scenarios. These models highlight a trend toward tighter vision-language integration and stronger open-set generalization.

2.3 Out-of-distribution object detection

Early approaches attempt to synthesize OOD samples, either by perturbing in-distribution (ID) data or generating images with generative models. VOS [6] perturbs ID distributions to create pseudo-OOD samples, training models with binary objectives to distinguish ID and OOD. SyncOOD [19] feeds specially designed prompts into large language models (LLMs) to generate OOD labels, which are then used to edit images within bounding boxes via generative models. WFS [35] introduces the Open-Domain Unknown Object Detection (ODU-OD) task, targeting accurate detection of unknown objects across multiple unseen domains with diverse image styles. SIREN [4] explicitly shapes feature spaces by enforcing compactness for ID categories and separability for OOD regions. These efforts highlight the challenge of extending detectors beyond the semantic space of their training data.

2.4 Prompt learning

Prompt learning, initially introduced in NLP, has recently been adapted to vision tasks as a flexible means of aligning VLMs with downstream objectives. CoOp [38] adds learnable text prompts in CLIP, significantly improving its performance on image classification. CoCoOp [37] extends this idea with conditional prompts that generalize better to unseen classes. MaPLe [13] introduces vision prompts to refine prompt tuning strategies to enable better vision-language alignment.

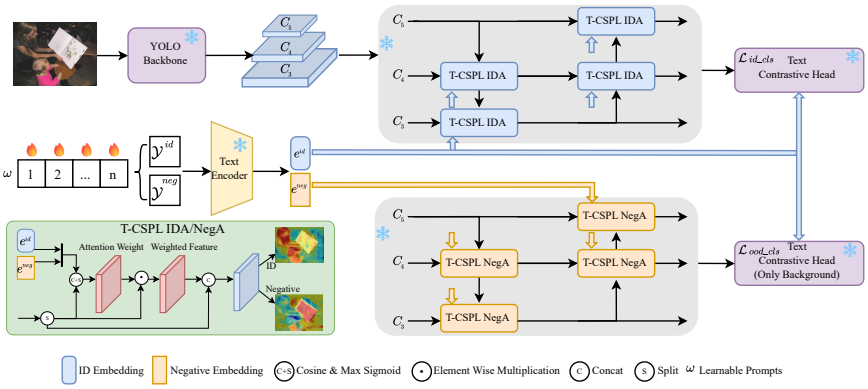

Fig. 2: The framework of our method. Based on the original YOLO-World, we add an additional NegA-based OOD branch. NegA leverages the interaction between the text embeddings of LLM-curated negative labels and visual features to enhance potential OOD background regions. The enhanced features are then fed into the detection head to compute the classification loss for the background regions only. Through backpropagation, the learnable prompts in the text encoder are updated, helping the model learn a better decision boundary. During inference, the NegA attention branch is discarded (no additional inference cost), while the negative text embeddings are retained for computing the NegS OOD score.

In object detection, PromptDet [8] introduces task-specific prompts for open-vocabulary detection, showing the effectiveness of prompt learning in bridging the gap between vision and language. These advances make prompt learning a key technique for adapting large VLMs to diverse tasks.

3 Methodology

We consider the closed-set fine-tuning setting for OOD detection: given a predefined set of ID categories, a dense VLM detector (*e.g.*, YOLO-World) is fine-tuned on ID training data, and at test time must distinguish ID objects from OOD objects unseen during training.

As shown in Figure 2, our method is built upon YOLO-World [3], a widely used VLM-based dense object detector. In this section, we first introduce how negative labels are generated in Sec. 3.1. Next, we explain how these negative labels guide the model to focus on important background regions in Sec. 3.2 and describe the training loss functions in Sec. 3.3. Finally, we present the NegS score computation based on negative labels in Sec. 3.4.

3.1 Negative label selection

NegLabel [12] proposes to select negative labels from a large lexical database. However, we find that many words selected in this way are meaningless. To

improve the quality of negative labels, we resort to large language models (LLMs) for negative label generation. As illustrated in Figure 1, given a collection of in-distribution (ID) category texts $\mathcal{Y}_N^{id}$ with N entries, designed prompts are provided to the LLM (*e.g.*, GPT-4o [11]) to produce a novel set of negative category texts, denoted as $\mathcal{C}_M^{neg}$ with M entries. These negative texts describe objects that share certain visual characteristics with ID categories while differing semantically. Formally,

$$\mathcal{C}_M^{neg} = \mathcal{LLM}(\mathcal{Y}_N^{id}; Prompts), \tag{1}$$

where $\mathcal{LLM}$ denotes the large language model. The used $Prompts$ and selection procedure are described below.

Prompt design. We provide the LLM with the following instructions.

System Prompts: *"You are an AI visual assistant. You should not make up information or provide false details. The user will provide you with some In-Distribution (ID) class names in text, and you should respond with the corresponding OOD class names..."*.

User Prompts: *"The OOD class names should be of following features: 1. The OOD class names should be diverse and commonly seen in the real world... Now please list corresponding negative label names for ID names: [person, bird, ...]"*.

The complete prompt templates for each dataset and the full list of generated negative labels are provided in the supplementary materials.

Negative label mining. Given the ID label set $\mathcal{Y}_N^{id} = \{y_j^{id}\}_{j=1}^N$, we prompt an LLM to generate a candidate pool $\mathcal{C} = \{c_i\}_{i=1}^{|\mathcal{C}|}$. Let $\mathcal{T}(\cdot) \in \mathbb{R}^C$ denote the text-encoder embedding and $\text{sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$ be cosine similarity.

Visual similarity filtering. To ensure visual similarity to ID objects, we first build visual prototypes for each ID class from the ID dataset (*e.g.*, by averaging CLIP [26] image-encoder features of instances in that class). Then, leveraging CLIP, we compute the image–text similarity between each candidate label c_i and the set of ID prototypes. We retain the M candidates with a similarity over δ (default 0.5), forming a visually similar candidate set $\mathcal{C}_M^{neg} \subseteq \mathcal{C}$.

Semantic dissimilarity ranking. For each candidate $c_i \in \mathcal{C}_M^{neg}$, we measure its semantic closeness to the ID label set by the maximum text-text similarity:

$$s_i \;=\; \max_{1 \le j \le N} \text{sim}\big(\mathcal{T}(c_i),\, \mathcal{T}(y_j^{id})\big). \tag{2}$$

We define a dissimilarity score as $d_i \;=\; 1 - s_i$, where dissimilarity score $d_i = 0$ indicates that c_i is identical (or extremely close) to some ID label in the text-embedding space. Finally, inspired NegLabel [12], we select the K candidates with the largest dissimilarity score d_i (equivalently, the smallest similarity s_i) to form the negative label set $\mathcal{Y}_K^{neg}$, and discard candidates with large s_i to reduce semantic overlap with ID labels. This procedure yields negative labels that are visually similar while semantically distinct from the ID label set.

3.2 Negative label guided attention

Review of Text-Guided CSPLayer (YOLO-World). YOLO-World introduces Max-Sigmoid Attention (MSA) into CSPLayer to inject linguistic information into multi-scale visual features. Let $T_N = \{t_j\}_{j=1}^{N} \in \mathbb{R}^{N\times C} = \mathcal{T}(\mathcal{Y}_N)$ be the text embeddings and $I_l \in \mathbb{R}^{C\times H_l\times W_l}$ be the visual feature at FPN level $l \in \{3,4,5\}$. For each spatial location (m,n), denote $I_l^{(m,n)} \in \mathbb{R}^C$ as its feature vector. The Max-Sigmoid Attention $\hat{I}_l = MSA(I_l, T_N)$ is defined as:

$$A_l^{(m,n)} = \sigma\left(\max_{1\le j\le N} sim(t_j, I_l^{(m,n)})\right), \quad A_l \in \mathbb{R}^{H_l\times W_l}, \quad t_j \in \mathbb{R}^C, \tag{3}$$

$$\hat{I}_l = I_l \odot A_l, \tag{4}$$

where $A_l^{(m,n)}$ is a scalar attention weight and $\odot$ denotes element-wise multiplication with broadcasting. Thus, regions closely associated with the texts are enhanced, while irrelevant regions are suppressed.

YOLO-World with NegA (ours). Our goal is to leverage negative labels to *amplify potential OOD evidence in background regions* while preserving the original ID detection capability of YOLO-World, thereby learning a more precise decision boundary between ID and OOD samples. We impose two design principles: (i) keep the YOLO-World architecture and parameters frozen to retain its generalization ability; (ii) apply negative-label guidance *only* on background regions to avoid conflicting with ID semantics.

Dual-branch neck with shared parameters. As shown in Figure 2, we build two parallel neck branches with identical architecture and shared parameters. The ID branch follows the original YOLO-World and produces ID-enhanced features, while the OOD branch is conditioned on negative text embeddings and produces OOD-enhanced features:

$$\hat{I}_l^{id} = MSA(I_l, T_N^{id}), \qquad \hat{I}_l^{ood} = MSA(I_l, T_K^{neg}), \tag{5}$$

where $T_N^{id} = \mathcal{T}(\mathcal{Y}_N^{id}) \in \mathbb{R}^{N\times C}$ and $T_K^{neg} = \mathcal{T}(\mathcal{Y}_K^{neg}) \in \mathbb{R}^{K\times C}$ are embeddings for ID and negative label sets, respectively. Both $\hat{I}_l^{id}$ and $\hat{I}_l^{ood}$ are forwarded to the same detection head for classification, but the OOD supervision is restricted to background regions (details below). At each forward pass, we perturb negative embeddings with Gaussian noise (default $\mu = 0$, $\sigma = 0.05$). This perturbation acts as a semantic-coverage regularizer for the finite set of generated negative labels by locally expanding their embedding neighborhoods.

Background mask construction. We construct a binary background mask $M_l \in \{0,1\}^{H_l\times W_l}$ from ground-truth (GT) bounding boxes during training. For each FPN level $l \in \{3,4,5\}$ with stride $\{8,16,32\}$, we project GT boxes onto the feature-map grid by dividing box coordinates by the stride. For each grid location (m,n), we set $M_l^{(m,n)} = 0$ if (m,n) falls *inside* any projected GT box (foreground), and $M_l^{(m,n)} = 1$ otherwise (background). This mask depends only on GT annotations. The construction is intentionally conservative: any cell

inside a projected box is treated as foreground, preventing negative-label guidance from penalizing ID regions near object boundaries.

Training vs. inference. The OOD branch serves as *training-time background augmentation*: it modulates attention using negative labels and provides additional supervision to shape a clearer ID–OOD boundary. During training, $\hat{I}_l^{ood}$ is used to compute background-only OOD loss, while $\hat{I}_l^{id}$ is used for standard detection. **At inference, we discard the NegA (OOD) branch for feature extraction** and keep only the ID branch to produce detections. **Negative text embeddings are still retained** and are used only for computing the test-time OOD score (NegS in Section 3.4) together with ID-branch features, introducing no additional inference overhead in the neck. During inference, we compute per-box classification scores $s^{id} \in [0,1]^N$ and $\tilde{s}^{neg} \in [0,1]^K$ using ID-branch features and (T_N^{id}, T_K^{neg}), then obtain the OOD score by Equation (10). The inference pipeline is provided in supplementary materials.

3.3 Training objectives

Both $\hat{I}_l^{id}$ and $\hat{I}_l^{ood}$ are forwarded through the same detection head to predict scores over the N ID classes $\mathcal{Y}_N^{id}$. Let the set of predictions from the two branches be $\mathcal{P}^{id} = \{P_i^{id}\}_{i\in\Omega}$ and $\mathcal{P}^{ood} = \{P_i^{ood}\}_{i\in\Omega}$, where Ω indexes all predictions across all FPN levels $l \in \{3,4,5\}$. For index $i \in \Omega$, we define $P_i^{id} = (bbox_i^{id}, s_i^{id})$ and $P_i^{ood} = (bbox_i^{ood}, s_i^{ood})$, where $bbox_i \in \mathbb{R}^4$ denotes the bounding box coordinates and $s_i \in [0,1]^N$ denotes sigmoid probabilities over the N ID classes.

Both branches share the same classifier; the difference lies in how their features are produced and how their losses are applied.

Losses. Let $target_i \in [0,1]^N$ be the soft target probability for the i^{th} prediction, generated by the TaskAlignedAssigner adopted in YOLO-World. For the ID branch, we compute the standard Binary Cross Entropy (BCE) classification loss between the soft targets and the predicted sigmoid probabilities:

$$\mathcal{L}_{id\_cls} = \sum_{i\in\Omega} \mathrm{BCE}(s_i^{id}, target_i), \tag{6}$$

For the OOD branch, we apply the classification loss *only* on background regions using the background mask constructed in Section 3.2. Since each prediction $P_i^{ood} \in \mathcal{P}^{ood}$ is generated from a specific scale level $l(i)$ and spatial grid location $(m(i), n(i))$, its corresponding binary mask value $m_i = M_{l(i)}^{(m(i),n(i))} \in \{0,1\}$. The OOD classification loss is formulated as:

$$\mathcal{L}_{ood\_cls} = \sum_{i\in\Omega} m_i \,\mathrm{BCE}(s_i^{ood}, target_i). \tag{7}$$

This formulation masks out foreground predictions ($m_i = 0$), ensuring that only background predictions ($m_i = 1$) contribute to the OOD loss.

Total objective. The overall training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{det} + \alpha\,\mathcal{L}_{ood\_cls}, \qquad \mathcal{L}_{det} = \mathcal{L}_{id\_cls} + \mathcal{L}_{box}, \tag{8}$$

where α is a balancing coefficient (set to 3 by default), and $\mathcal{L}_{box}$ represents the standard bounding box regression loss computed exclusively on the ID branch.

3.4 Test-time OOD detection

Negative label guided OOD scoring (NegS). Existing OOD detectors often employ energy-based scores defined on softmax logits [6, 21]. For example, the energy used in VOS [6] is defined as:

$$E(P^{id}) = -\log \sum_{j=1}^{N} e^{z[j]}, \quad S(P^{id}) = \frac{e^{-E(P^{id})}}{1 + e^{-E(P^{id})}}, \tag{9}$$

where $z[j]$ is the predicted logits for ID class j prior to the softmax activation; $S(P^{id})$ is used as the final OOD score.

However, traditional energy scores are theoretically derived from the softmax partition function (*i.e.*, the denominator $\sum e^{z[j]}$), which assumes mutually exclusive classes. In contrast, open-vocabulary VLM detectors typically employ independent *sigmoid* activations for multi-label classification. Consequently, applying softmax-based energy formulations to sigmoid logits is not directly compatible. More importantly, conventional energy scores do not exploit the negative-label semantics that we explicitly inject during training.

We therefore propose NegS, a sigmoid-based OOD scoring function assisted by negative labels. Given ID and negative label sets $\mathcal{Y}_N^{id}$ and $\mathcal{Y}_K^{neg}$, we compute their respective text embeddings $T_N^{id} = \mathcal{T}(\mathcal{Y}_N^{id})$ and $T_K^{neg} = \mathcal{T}(\mathcal{Y}_K^{neg})$. For each predicted box $P^{id} = (bbox^{id}, s^{id}) \in \mathcal{P}^{id}$ produced by the *ID branch*, we additionally compute negative-label sigmoid scores $\tilde{s}^{neg} \in [0,1]^K$ using the same ID-branch visual features and T_K^{neg}. NegS is defined as:

$$NegS(P^{id}) = \max_{1 \le j \le N} s^{id}[j] \; - \; \max_{1 \le k \le K} \tilde{s}^{neg}[k]. \tag{10}$$

A prediction strongly aligned with ID semantics yields a high first term, while a prediction aligned with negative semantics yields a high second term. Thus, NegS effectively separates ID objects from OOD or background regions by explicitly modeling their relative separation.

4 Experiments

4.1 Implementation details.

Datasets. Following the experimental setting in [6], we employ PASCAL VOC [7] and BDD100K [36] as the In-Distribution (ID) datasets, containing 20 and 10 categories, respectively. For the Out-of-Distribution (OOD) evaluation, we adopt MS-COCO [18] and OpenImages [14], with overlapping categories removed to ensure no intersection with the ID datasets.

Table 1: Comparison with baseline models on OOD detection performance on different in-distribution datasets. Best results are in **bold**. "YOLO-World" and "GroundingDINO" means the original YOLO-World and GroundingDINO models without modification. "YOLO-World$_{CoOp}$" denotes incorporating CoOp learnable prompts into YOLO-World and training only prompts on ID labels. "NegAS" denotes our proposed method. ↑ (↓) denotes performance improvements compared with the baseline. "YW" means YOLO-World and "GD" means Grounding DINO.

		Method	OOD Dataset				mAP (ID)↑
			MS-COCO		OpenImages		
			FPR95↓	AUROC↑	FPR95↓	AUROC↑	
VOC	Traditional Detector	Energy [21]	56.9	83.7	58.7	83.0	-
		Gram Matrices [28]	62.8	79.9	67.4	48.7	60.6
		Generalized ODIN [10]	59.6	83.1	70.3	79.2	-
		CSI [29]	59.9	81.8	57.4	83.0	59.5
		Siren-vMF [4]	75.5	76.1	78.4	71.1	60.8
		Siren-KKN [4]	64.8	78.2	66.0	74.9	60.8
		SR-VAE [34]	42.2	90.3	46.3	87.9	-
		DFFF [33]	41.3	90.8	44.5	88.7	-
		SyncOOD [19]	36.4	86.5	13.3	**95.4**	-
		WFS [35]	37.5	85.6	**11.3**	94.9	-
		VOS [6]	47.5	88.7	51.3	85.2	60.3
		SAFE [32]	47.4	80.3	20.1	92.3	60.8
	VLM	YW(baseline) [3]	21.3	93.4	48.8	81.4	69.0
		YW$_{CoOp}$(baseline)	13.9	95.7	47.0	83.3	72.3
		YW + NegAS	**9.9**$_{\downarrow 11.4}$	**95.9**$_{\uparrow 1.5}$	23.3$_{\downarrow 25.5}$	92.6$_{\uparrow 11.2}$	**72.5**
		GD(baseline) [20]	11.7	96.8	33.4	91.8	63.8
		GD+NegAS	**7.2**$_{\downarrow 4.5}$	**97.6**$_{\uparrow 0.8}$	24.3$_{\downarrow 9.1}$	92.1$_{\uparrow 0.3}$	**66.1**
BDD	Traditional Detector	Energy [21]	60.1	77.5	55.0	80.0	-
		Gram Matrices [28]	60.9	74.9	77.6	59.4	30.3
		Generalized ODIN [10]	57.3	85.2	50.2	87.2	-
		CSI [29]	47.1	84.1	37.1	88.0	29.9
		Siren-vMF	67.5	80.1	66.3	79.8	31.3
		Siren-KKN	54.0	86.6	47.3	89.0	31.3
		SR-VAE [34]	32.2	90.7	21.8	93.6	-
		DFFF [33]	30.7	90.7	22.7	92.5	-
		SyncOOD [19]	22.7	95.4	13.0	96.3	-
		WFS [35]	21.8	93.4	7.8	96.9	-
		VOS [6]	44.3	86.9	35.5	88.5	31.1
		SAFE [32]	32.6	89.0	16.0	94.6	**31.2**
	VLM	YW(baseline) [3]	40.4	83.9	8.2	97.0	22.5
		YW$_{CoOp}$(baseline)	47.0	82.0	15.1	92.6	26.1
		YW + NegAS	**20.9**$_{\downarrow 19.5}$	**92.0**$_{\uparrow 9.0}$	**3.6**$_{\downarrow 5.4}$	**98.7**$_{\uparrow 1.7}$	29.0
		GD(baseline) [20]	21.2	94.4	21.6	94.2	23.6
		GD+NegAS	**17.8**$_{\downarrow 3.4}$	**95.8**$_{\uparrow 1.4}$	**11.5**$_{\downarrow 10.1}$	**95.2**$_{\uparrow 1.0}$	25.2

Experimental settings. Our method is based on YOLO World [3], a recent open-vocabulary detection framework. For fair comparison, all models are trained under the same conditions unless otherwise specified.

We evaluate the following configurations:

- YOLO World: The official YOLO World model without any additional training or modification.
- YOLO World$_{CoOp}$: YOLO World equipped with learnable text prompts following CoOp [38], trained on two ID datasets using ID labels while keeping all other parameters frozen.
- NegAS (ours): The proposed Negative Label Guided Attention and Scoring method, trained on ID datasets using both ID and curated negative labels.

During training, with all other parameters except the prompts frozen, all models are trained using the AdamW optimizer with a learning rate of 1e-2 for 30 epochs with batch size 160 on 4 NVIDIA V100 GPUs. The image resolution is set to 640×640. For YOLO World$_{CoOp}$ and NegAS, we use prompt initialization "*a photo of*", a total of three learnable prompt tokens, each with dimension 512. We select 20 negative labels from over 200 LLM-generated candidates for VOC, and 20 from over 300 candidates for BDD.

Evaluation metrics. Following standard OOD detection practice established by prior works such as VOS [6], we report three metrics under the same evaluation pipeline: FPR95 (↓): False Positive Rate at 95% True Positive Rate; AUROC (↑): Area Under the Receiver Operating Characteristic curve; and mAP (ID) (↑): Mean Average Precision on in-distribution data.

4.2 Comparison with baselines and existing methods

We first compare NegAS with two YOLO-World based baselines: the original model without fine-tuning and the CoOp-augmented variant trained with learnable prompts. As shown in Table 1, NegAS consistently improves OOD detection metrics on VOC and BDD datasets while maintaining competitive in-distribution accuracy. On the PASCAL-VOC in-distribution dataset, NegAS achieves an FPR95 reduction from 48.8% to 23.3% when the OOD dataset is OpenImages, corresponding to an improvement of nearly 25%, and raises the AUROC from 81.4% to 92.6%. On OOD dataset MS-COCO, NegAS yields the lowest FPR95 (9.9%) and the highest AUROC (95.9%), outperforming both YOLO-World and its CoOp-enhanced variant. Similarly, on the BDD dataset, NegAS also achieves the best results.

4.3 Generalization to other VLM detectors

To demonstrate the architectural generalizability of NegAS, we extend our framework to Grounding DINO, a query-based VLM detection transformer. Specifically, we introduce an auxiliary OOD branch parallel to the original architecture, as illustrated in supplementary materials.

During training, the model is optimized using AdamW with a learning rate of 1e-4 for 20 epochs. Consistent with our YOLO-World implementation, learnable prompts are initialized as "a photo of", and Gaussian noise ($\sigma = 0.05$) is injected

Table 2: Ablation study on the effectiveness of the proposed NegA and NegS components. All variants are trained with learnable prompts. ✓ and ✗ denote the presence and absence of each component, respectively.

ID $\mathcal{D}$	NegA	NegS	OOD: MS-COCO / OpenImages	
			FPR95↓	AUROC↑
VOC	✗	✗	13.9 / 47.0	95.7 / 83.3
	✗	✓	11.4 / 33.3	96.2 / 88.8
	✓	✗	11.8 / 41.8	**96.4** / 84.4
	✓	✓	**9.9** / **23.3**	95.9 / **92.6**
BDD	✗	✗	47.0 / 15.1	82.0 / 92.6
	✗	✓	43.6 / 12.2	82.9 / 94.5
	✓	✗	25.3 / 7.2	91.1 / 97.3
	✓	✓	**20.9** / **3.6**	**92.0** / **98.7**

into the text embeddings. During inference, the auxiliary OOD branch is discarded to maintain efficiency. The negative labels bypass feature fusion entirely and are used only with ID-label enhanced queries to compute NegS.

Unlike dense detectors, query-based models like Grounding DINO lack explicit absolute spatial information, making background mask construction complex. The detailed background mask construction is provided in the supplementary materials. As shown in Table 1, NegAS significantly improves the OOD detection performance of Grounding DINO, proving that our negative-label-guided framework effectively generalizes well beyond dense VLM detectors.

4.4 Ablation study

Effectiveness of NegA and NegS. To better understand the contribution of each component, we conduct ablation experiments summarized in Table 2. All variants are trained with learnable prompts for a consistent comparison. Applying NegS consistently improves both AUROC and FPR95, confirming that incorporating negative semantics into the scoring function leads to more discriminative confidence estimation. The NegA module further enhances OOD separation, as it leverages negative label guidance to excavate latent OOD information. With both components combined, NegAS surpasses YOLO-World with learnable prompts by a large margin, validating the complementary effect of NegA and NegS in enhancing OOD detection robustness.

Comparison with softmax-based scoring functions. We compare NegS with direct adaptations of softmax-based MCM and NegLabel-style scoring on YOLO-World proposals. As shown in Table 3, NegS consistently improves FPR95 and AUROC over these scoring functions under the same detector setting, and the full NegAS further reduces FPR95. This indicates that directly transferring softmax-based OOD scoring is suboptimal for VLM detectors, which are based on sigmoid-style classification scores.

Selection of negative labels. In Table 5, we ablate the generation strategy and the number of negative labels. "object" means that the negative labels

Table 3: Comparison with direct adaptations of MCM and NegLabel-style scoring on YOLO-World proposals. Results are reported on VOC as ID and MS-COCO/OpenImages as OOD datasets.

Method	FPR95↓	AUROC↑
YW + MCM	32.3 / 56.3	86.4 / 75.7
YW + NegLabel	18.2 / 45.8	91.3 / 82.1
YW + NegS	13.5 / 38.8	94.7 / 83.6
YW_{CoOp} + MCM	19.8 / 52.7	87.8 / 76.3
YW_{CoOp} + NegLabel	13.3 / 42.7	94.5 / 87.4
YW_{CoOp} + NegS	11.4 / 33.3	96.2 / 88.8
YW + NegAS	**9.9** / **23.3**	95.9 / **92.6**

Table 4: Ablation on Gaussian noise standard deviation applied to negative text embeddings, OOD loss coefficient α, and background mask. Results on VOC (ID) with MS-COCO / OpenImages (OOD).

Std	FPR95↓	AUROC↑
0.1	15.7 / 41.9	94.4 / 87.0
0.01	11.9 / 36.0	95.3 / 86.9
0.05	**9.9** / **23.3**	**95.9** / **92.6**
0.0	9.9 / 33.4	96.7 / 89.5

(a) Noise standard deviation.

α	FPR95↓	AUROC↑
2	11.3	95.1
3	**9.9**	**95.9**
4	12.2	95.0

(b) OOD loss coefficient α.

Mask	FPR95↓	AUROC↑
✓	**9.9** / **23.3**	**95.9** / **92.6**
✗	10.9 / 32.2	95.4 / 88.8

(c) Background mask.

contain only a single "object" category. "LLM" means that we prompt a large language model to generate potential negative labels given the ID labels. "Corpus" means that we select negative labels from a large-scale corpus database [30], following the same procedure as in [12]. The results show that using a single "object" already outperforms the baseline. Labels selected from the corpus further improve the performance, but are still inferior to those chosen via the LLM.

Influence of negative label embedding noise and OOD loss factor. We investigate the influence of Gaussian noise on negative label embeddings and the OOD loss factor in Tables 4a and 4b. The benefit of the perturbation depends on how well the selected negative labels cover the target OOD categories. For VOC-COCO, the selected negative labels are closer to the COCO OOD categories, so the coverage is relatively sufficient and the perturbation brings limited gains. For VOC-OpenImages, the selected negative labels are farther from the OpenImages OOD categories, so the perturbation helps expand the local negative-label neighborhood. We observe that the OOD loss is very small. Considering that an appropriate amount of OOD correction is beneficial—insufficient correction fails to take effect, while excessive correction may disturb the original feature distribution—we investigate the influence of the OOD loss coefficient α. Empirically, when $\alpha = 3$, the method achieves the best performance.

Background mask. As shown in Table 4c, calculating OOD loss only on background regions outperforms calculating the overall image, confirming that foreground features modulated by ID and negative labels can conflict.

Table 5: Ablation study on the type and the number of the negative labels. The ID dataset is VOC.

Negative Label	Number	OOD Dataset			
		MS-COCO		OpenImages	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑
"object"	1	13.9	96.0	45.3	82.9
LLM [11]	1	12.9	96.2	42.3	85.1
	5	12.4	95.5	38.7	86.0
	10	13.6	84.9	40.6	85.2
	20	**9.9**	**95.9**	**23.3**	**92.6**
	30	15.5	93.8	38.2	87.4
Corpus [30]	1	12.7	96.1	45.5	83.0
	5	12.2	96.3	44.1	83.7
	10	13.7	95.8	45.0	84.2
	20	**11.7**	**96.6**	**42.4**	**85.0**
	30	12.0	95.8	40.8	84.6

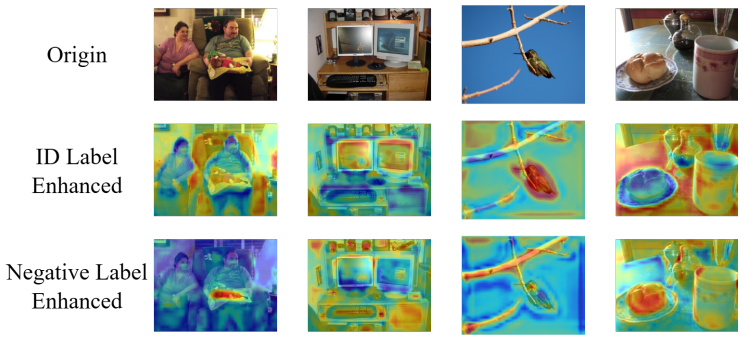

Fig. 3: Feature heatmap comparison between YOLO-World and NegAS. The feature maps are extracted from the neck layer.

4.5 Qualitative results

We visualized feature heatmaps in Figure 3 to illustrate the attention shift introduced by NegA. The ID label enhanced feature maps show high responses on ID classes (chair, TV, bird, dining table), while negative label guided attentions focus on potential OOD regions (doll, desk, branch, bread). This demonstrates that NegAS guides the model toward latent OOD information, helping to better distinguish ID and OOD samples. Red regions indicate higher attention weights. The ID label enhanced features (middle row) highlight ID object regions (*e.g.*, chair, TV, bird, dining table), while the negative label guided features (bottom row) focus on potential OOD regions (*e.g.*, doll, desk, branch, bread), demonstrating NegA's ability to attend to OOD-relevant background areas.

5 Conclusion

In this paper, we explored the OOD problem for dense vision-language detection models. First, we proposed Negative Label Guided Attention (NegA), which uses negative labels generated by an LLM to guide the model to focus more on important OOD background regions. This enhances feature diversity and learns a better decision boundary through prompt learning. Second, we introduced Negative Label Guided OOD Scoring (NegS) function during the testing phase. This function is specifically designed for sigmoid-based VLMs and enables better differentiation between ID and OOD samples. Finally, we generalized the proposed NegAS to a query-based VLM detector to investigate the generalizability. Extensive experiments on multiple benchmarks demonstrated the effectiveness and robustness of NegAS and its generalizability to query-based VLM detectors.

Acknowledgements

This work was partly supported by the Fundamental Research Funds for the Central Universities, China (G2026KY05068).

References

1. Bai, Y., Han, Z., Cao, B., Jiang, X., Hu, Q., Zhang, C.: Id-like prompt learning for few-shot out-of-distribution detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17480–17489 (2024)
2. Cao, C., Zhong, Z., Zhou, Z., Liu, Y., Liu, T., Han, B.: Envisioning outlier exposure by large language models for out-of-distribution detection. arXiv preprint arXiv:2406.00806 (2024)
3. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: Yolo-world: Real-time open-vocabulary object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16901–16911 (2024)
4. Du, X., Gozum, G., Ming, Y., Li, Y.: Siren: Shaping representations for detecting out-of-distribution objects. Advances in neural information processing systems **35**, 20434–20449 (2022)
5. Du, X., Wang, X., Gozum, G., Li, Y.: Unknown-aware object detection: Learning what you don't know from videos in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13678–13688 (2022)
6. Du, X., Wang, Z., Cai, M., Li, Y.: Vos: Learning what you don't know by virtual outlier synthesis. arXiv preprint arXiv:2202.01197 (2022)
7. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
8. Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., Xie, W., Ma, L.: Promptdet: Towards open-vocabulary detection using uncurated images. In: European conference on computer vision. pp. 701–717. Springer (2022)
9. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021)
10. Hsu, Y.C., Shen, Y., Jin, H., Kira, Z.: Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10951–10960 (2020)
11. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
12. Jiang, X., Liu, F., Fang, Z., Chen, H., Liu, T., Zheng, F., Han, B.: Negative label guided ood detection with pretrained vision-language models. arXiv preprint arXiv:2403.20078 (2024)
13. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19113–19122 (2023)
14. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International journal of computer vision **128**(7), 1956–1981 (2020)
15. Lafon, M., Ramzi, E., Rambour, C., Audebert, N., Thome, N.: Gallop: Learning global and local prompts for vision-language models. In: European Conference on Computer Vision. pp. 264–282. Springer (2024)
16. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10965–10975 (2022)

17. Li, T., Pang, G., Bai, X., Miao, W., Zheng, J.: Learning transferable negative prompts for out-of-distribution detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17584–17594 (2024)
18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
19. Liu, J., Wen, X., Zhao, S., Chen, Y., Qi, X.: Can ood object detectors learn from foundation models? In: European Conference on Computer Vision. pp. 213–231. Springer (2024)
20. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
21. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. Advances in neural information processing systems **33**, 21464–21475 (2020)
22. Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., Li, Y.: Delving into out-of-distribution detection with vision-language representations. Advances in neural information processing systems **35**, 35087–35102 (2022)
23. Miyai, A., Yu, Q., Irie, G., Aizawa, K.: Locoop: Few-shot out-of-distribution detection via prompt learning. Advances in Neural Information Processing Systems **36**, 76298–76310 (2023)
24. Miyai, A., Yu, Q., Irie, G., Aizawa, K.: Zero-shot in-distribution detection in multi-object settings using vision-language foundation models. arXiv preprint arXiv:2304.04521 (2023)
25. Nie, J., Zhang, Y., Fang, Z., Liu, T., Han, B., Tian, X.: Out-of-distribution detection with negative prompts. In: The twelfth international conference on learning representations (2024)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
27. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence **39**(6), 1137–1149 (2016)
28. Sastry, C.S., Oore, S.: Detecting out-of-distribution examples with gram matrices. In: International conference on machine learning. pp. 8491–8501. PMLR (2020)
29. Tack, J., Mo, S., Jeong, J., Shin, J.: Csi: Novelty detection via contrastive learning on distributionally shifted instances. Advances in neural information processing systems **33**, 11839–11852 (2020)
30. University, P.: About wordnet. `https://wordnet.princeton.edu/` (2010), accessed: 2025-11-09
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
32. Wilson, S., Fischer, T., Dayoub, F., Miller, D., Sünderhauf, N.: Safe: Sensitivity-aware features for out-of-distribution object detection. In: Proceedings of the ieee/cvf international conference on computer vision. pp. 23565–23576 (2023)
33. Wu, A., Chen, D., Deng, C.: Deep feature deblurring diffusion for detecting out-of-distribution objects. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13381–13391 (2023)

34. Wu, A., Deng, C.: Discriminating known from unknown objects via structure-enhanced recurrent variational autoencoder. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23956–23965 (2023)
35. Wu, A., Deng, C.: Percept, memory, and imagine: World feature simulating for open-domain unknown object detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4682–4691 (2025)
36. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2636–2645 (2020)
37. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022)
38. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)