EXPLORING A PRINCIPLED FRAMEWORK FOR MULTI-LABEL OUT-OF-DISTRIBUTION DETECTION WITH PRE-TRAINED VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-distribution (OOD) detection enhances the reliability of machine learning models by identifying unexpected inputs from unknown classes. Recent progresses in large-scale vision–language pre-training have enabled zero-shot OOD detection using only in-distribution (ID) class names. Nevertheless, previous studies predominantly focus on detecting OOD inputs in the multi-class classification task, leaving the more realistic *multi-label* setting greatly under-explored. To mitigate this research gap, this paper investigates zero-shot multi-label OOD detection with pre-trained vision-language models. We first theoretically demonstrate that existing methods are fundamentally mismatched with the multi-label setting, as their scoring mechanisms fail to effectively aggregate evidence from multiple co-occurring ID labels. Motivated by this, we propose a probabilistic framework that formulates multi-label OOD detection through pairwise image-label association modeling, directly yielding a principled OOD scoring function that estimates the probability of an input being associated with at least one ID label in a *training-free* manner. Comprehensive evaluations on standard multi-label benchmarks, including MS-COCO and PASCAL VOC, empirically manifest that our method achieves state-of-the-art performance across a variety of OOD detection setups.

1 INTRODUCTION

Despite significant advances in machine learning that have enabled a broad range of real-world applications (Zhao et al., 2019; Caruana & Niculescu-Mizil, 2006; Masana et al., 2022), most existing models are developed under a *closed-world* assumption, whereby training and test samples are drawn from the same distribution. In practical deployments, however, models often operate in an *open-world* environment and may encounter samples from classes unseen during training, commonly referred to as out-of-distribution (OOD) data. Such inputs can undermine model reliability and lead to severe prediction errors. A trustworthy discriminative model should therefore not only correctly classify known in-distribution (ID) samples but also identify OOD inputs as unknown, motivating the study of OOD detection (Lang et al., 2023; Salehi et al., 2021; Yang et al., 2021).

Recent progress in large-scale language–image pre-training, exemplified by CLIP (Radford et al., 2021), has extended OOD detection from conventional single-modal models to vision-language models (VLMs). An early approach (Esmaeilpour et al., 2022) trains a captioner to generate candidate OOD labels and matches these labels against input images. However, on large-scale datasets such as ImageNet-1k, whose ID label space contains many classes, the captioner may fail to generate sufficiently informative OOD candidates, resulting in limited detection performance. MCM (Ming et al., 2022) instead represents each ID class using a textual prototype and measures OOD uncertainty according to the scaled similarity between an image feature and its closest ID prototype. This work establishes a foundation for zero-shot OOD detection with pre-trained VLMs, but it uses only textual information from the predefined ID label space and thus does not fully exploit the semantic knowledge encoded by the language model. NegLabel (Jiang et al., 2024) addresses this limitation by selecting negative labels from external lexical resources, such as WordNet (Miller, 1995), according to their semantic relationships with the ID classes. These negative concepts provide additional evidence for distinguishing ID and OOD inputs.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

However, these methods are designed primarily for the conventional multi-class setting, where each image is assumed to belong to exactly one ID class. The more realistic *multi-label* setting, where an ID image may contain multiple objects and therefore be associated with several valid ID labels, remains largely under-explored. This gap sharply contrasts with the substantial progress made in leveraging pre-trained VLMs to zero-shot multi-label classification (Xie et al., 2026; Guo et al., 2023; Feng et al., 2025; Miller et al., 2025; Liu et al., 2024; Wang et al., 2025). To address it, we are motivated to explore a new zero-shot OOD detection paradigm that extends beyond mutually exclusive multi-class recognition to multi-label data.

Despite this compelling motivation, a central challenge remains: how can we design an OOD scoring function that coherently aggregates evidence from multiple co-occurring ID labels? In the multi-class setting, existing methods (Ming et al., 2022; Jiang et al., 2024) typically leverage contrastive scoring mechanisms where candidate labels compete through a shared normalization term. We formally show that such mechanisms lead to *fundamental limitations*: they either inherently penalize multi-object scenes through probability competition or exhibit severe saturation effects that functionally ignore secondary objects. In other words, the presence of additional recognizable ID objects contributes only marginally to the measured ID-ness of an input or even artificially reduces it.

To resolve this dilemma, we formulate multi-label OOD detection as probabilistic inference over pairwise image–label associations, where we introduce a binary latent variable to indicate whether an image is associated with a particular label. Under this formulation, an input is considered OOD if and only if it is associated with none of the known ID labels. This perspective allows to aggregate evidence from multiple co-occurring ID objects without imposing mutual exclusivity or competition among valid labels, thereby naturally accommodating the non-exclusive structure of multi-label data.

To deal with the computational intractability of inference under the unknown posterior distribution, we cast our method within the framework of noise-contrastive estimation (Gutmann & Hyvärinen, 2012). This yields a principled OOD scoring function that estimates the probability that an input is associated with at least one ID label in a *training-free* manner, for which we further establish a formal consistency guarantee. Extensive experiments demonstrate the empirical effectiveness of our method. For example, our method achieves 17.46% FPR95 and 17.51% FPR95 on MS-COCO and PASCAL VOC, respectively, outperforming RenyiNeg (Peng et al., 2026a) by 6.82% and 9.03%.

2 PRELIMINARY

Notation. Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input space and the categorical label space[1], respectively. Let $\mathbb{P}_Y$ denote the marginal distribution over $\mathcal{Y}$, and use $y \sim \mathbb{P}_Y$ to indicate a drawn sample. Let $\mathbb{P}_{X_\text{I}}$ and $\mathcal{Y}_\text{I} \triangleq \{y_1, \ldots, y_K\} \subset \mathcal{Y}$ represent the ID data distribution and *known* ID label space, respectively. In the *multi-label* setting, each ID data can be assigned with *multiple* labels in $\mathcal{Y}_\text{I}$ simultaneously. At test time, the environment can incur some unknown OOD distributions $\mathbb{P}_{X_\text{O}}$, where each OOD data is associated with *none* of the labels in $\mathcal{Y}_\text{I}$ (Wang et al., 2021; Zolfi et al., 2024).

Post-hoc Detection Strategy. Concurrently, OOD detection follows a training-free scoring mechanism, *i.e.,* given a pre-trained ID classification model parameterized by $\boldsymbol{\theta}$ and a scoring function S, $\mathbf{x}$ is detected as ID data if and only if $S(\mathbf{x}; \boldsymbol{\theta}) \geq \beta$, for some given threshold β:

$$g(\mathbf{x}) = \text{ID}, \text{ if } S(\mathbf{x}; \boldsymbol{\theta}) \geq \beta; \text{ otherwise, } g(\mathbf{x}) = \text{OOD}, \tag{1}$$

where β is chosen to correctly classify a high fraction of ID data (e.g., 95%).

CLIP-based Models adopt a dual-stream architecture (Radford et al., 2021) with one text encoder $f_\mathcal{T}$ and one image encoder $f_\mathcal{X}$ to map inputs of two modalities into a hyper-spherical space $\mathbb{S}^{d-1} \triangleq \left\{\mathbf{z} \in \mathbb{R}^d | \, \|\mathbf{z}\|_2 = 1\right\}$. Given any image $\mathbf{x} \in \mathcal{X}$ and any label $y \in \mathcal{Y}$, the corresponding representation can be extracted as follows:

$$\mathbf{z} = \mathbf{f}(\mathbf{x}; \boldsymbol{\theta}_\text{img}) \in \mathbb{S}^{d-1}, \quad \mathbf{r}_y = \mathbf{f}(\Lambda(y); \boldsymbol{\theta}_\text{text}) \in \mathbb{S}^{d-1},$$

where $\Lambda(\cdot)$ is a user-defined prompt template, $\mathbb{S}^{d-1} \triangleq \left\{\mathbf{z} \in \mathbb{R}^d | \, \|\mathbf{z}\|_2 = 1\right\}$, and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_\text{img}, \boldsymbol{\theta}_\text{text}\}$.

[1] In this paper, the categorical label space is defined as a set of all possible class names, diverging from the use of binary decision vectors in conventional multi-label setups.

Training-free Zero-shot OOD Detectors Studied. The pioneering work, MCM (Ming et al., 2022), treats the prompt of ID labels as concept prototypes and measures the ID-ness of the input image by comparing the similarity between the input image and the concept prototypes in the feature space learned by pre-trained VLMs, i.e.,

$$S_{\text{MCM}}(\mathbf{x};\boldsymbol{\theta}) = \max_{y\in\mathcal{Y}_{\text{I}}} \frac{\exp(\mathbf{z}^\top \mathbf{r}_y/\tau)}{\sum_{\hat{y}\in\mathcal{Y}_{\text{I}}}\exp(\mathbf{z}^\top \mathbf{r}_{\hat{y}}/\tau)}, \tag{2}$$

where $\tau > 0$ is a temperature hyper-parameter. Built upon this, NegLabel (Jiang et al., 2024) introduces a L-sized set of negative labels[2] $\mathcal{Y}_{\text{neg}} \triangleq \{y_{K+1}, \ldots, y_{K+L}\}$ to formulate the OOD scoring function of $\mathbf{x}$ as the model's prediction confidence that $\mathbf{x}$ belongs to $\mathcal{Y}_{\text{I}}$, i.e.,

$$S_{\text{NegLabel}}(\mathbf{x};\boldsymbol{\theta}) = \sum_{y\in\mathcal{Y}_{\text{I}}} \frac{\exp(\mathbf{z}^\top \mathbf{r}_y/\tau)}{\sum_{\hat{y}\in\mathcal{Y}_{\text{I}}\cup\mathcal{Y}_{\text{neg}}}\exp(\mathbf{z}^\top \mathbf{r}_{\hat{y}}/\tau)}. \tag{3}$$

Due to space constraints, the introduction of related work and the proofs of all theorems are deferred to Appendix A and Appendix B, respectively.

3 Motivation

Despite remarkable progress in this field, there remains a critical research gap in developing zero-shot OOD detectors based on pre-trained VLMs for multi-label classification. One might expect that existing multi-class OOD detection methods, particularly MCM and NegLabel, could easily transfer to the multi-label domain. We formally show that such a naive adaptation is fundamentally flawed.

Theorem 1. *Suppose an ID input $\mathbf{x}$ is associated with a set of $T \geq 2$ ID labels, denoted by $\ell(\mathbf{x}) \subseteq \mathcal{Y}_{\text{I}}$. If $\max_{y\in\mathcal{Y}_{\text{I}}} \mathbf{z}^\top \mathbf{r}_y - \min_{\hat{y}\in\ell(\mathbf{x})} \mathbf{z}^\top \mathbf{r}_{\hat{y}} \leq \epsilon$, then, for any temperature $\tau > 0$, we have*

$$S_{MCM}(\mathbf{x};\boldsymbol{\theta}) \leq \frac{\exp(\epsilon/\tau)}{T}. \tag{4}$$

Remark 1. Theorem 1 shows that MCM is structurally mismatched with the multi-label setting. When an ID input $\mathbf{x}$ contains $T \geq 2$ ID labels, $S_{\text{MCM}}(\mathbf{x};\boldsymbol{\theta})$ is strictly upper-bounded by a factor inversely proportional to T. This introduces a structural paradox where the presence of multiple ID objects artificially suppresses the measurable ID-ness of $\mathbf{x}$, which inevitably causes complex, multi-label ID images to be falsely flagged as OOD inputs.

Theorem 2. *Suppose an ID input $\mathbf{x}$ is associated with a set of $T \geq 2$ ID labels, denoted by $\ell(\mathbf{x}) \subseteq \mathcal{Y}_{\text{I}}$. Let $\hat{y} \in \ell(\mathbf{x})$ be a secondary ID label of $\mathbf{x}$[3] and $\eta_y(\mathbf{x}) = \mathbf{z}^\top \mathbf{r}_y$ for any $y \in \mathcal{Y}$, then we have*

$$\frac{\partial S_{NegLabel}(\mathbf{x};\boldsymbol{\theta})}{\partial \eta_{\hat{y}}(\mathbf{x})} > 0, \tag{5}$$

Furthermore, as the function $U(\mathbf{x},\hat{y}) = \sum_{y\in\mathcal{Y}_{\text{I}}\setminus\{\hat{y}\}} \exp\left[\eta_y(\mathbf{x})/\tau\right]$ approaches infinity, the partial derivative in Eq. (5) vanishes, i.e.,

$$\lim_{U(\mathbf{x},\hat{y})\to\infty} \frac{\partial S_{NegLabel}(\mathbf{x};\boldsymbol{\theta})}{\partial \eta_{\hat{y}}(\mathbf{x})} = 0. \tag{6}$$

Remark 2. Theorem 2 formally characterizes the *saturation effect* that bottlenecks NegLabel under the multi-label setting. While NegLabel theoretically benefits from the presence of a secondary ID label $\hat{y}$ because of a strictly positive partial derivative in Eq. (5), Eq. (6) reveals that this benefit can be severely marginalized in practice. In particular, when a primary ID object is highly prominent in the input $\mathbf{x}$, its exponential similarity score dominates $U(\mathbf{x},\hat{y})$, driving the partial gradient w.r.t. the secondary label $\hat{y}$ to vanish. Therefore, NegLabel fails to effectively aggregate concurrent evidence from multiple valid ID labels, functionally collapsing into a single-label scoring mechanism.

[2] By definition, *negative* labels are those semantically *irrelevant/dissimilar* to *all* ID labels.

[3] A secondary ID label of $\mathbf{x}$ refers to a ground-truth ID class of $\mathbf{x}$, but acts as an auxiliary element rather than the main focal point of $\mathbf{x}$.

4 METHODOLOGY

The theoretical observations delineated in Section 3 establish two fundamental desiderata for an effective multi-label OOD scoring function. First, the scoring mechanism must eschew competition among valid ID labels, since multiple ID concepts may simultaneously appear in the same image. Second, it should aggregate label-wise evidence in a non-saturating manner so that secondary and even less salient ID objects can still contribute meaningfully even when a dominant ID object is present. This directly motivates our methodological design.

At the core of our method, we introduce a binary latent variable $C \in \{0, 1\}$ to model the association between an image $\mathbf{x}$ and a label y, governed by the conditional probability mass function $q_{C|XY}$[4]. An image-label pair $(\mathbf{x}, y)$ is considered as positive ($C = 1$) *if and only if* $\mathbf{x}$ is associated with the label y, and negative ($C = 0$) otherwise. Building upon standard multi-label OOD detection paradigms (Wang et al., 2021; Zolfi et al., 2024) where the label set of OOD data has no intersection with ID labels $\mathcal{Y}_{\text{I}}$, it is natural to formulate the OOD scoring function for the multi-label setting as

$$S(\mathbf{x}) = \Pr\left(g(\mathbf{x}) = \text{ID}\right) = 1 - \Pr\left(g(\mathbf{x}) = \text{OOD}\right) = 1 - \prod_{y \in \mathcal{Y}_{\text{I}}} q\left(C = 0 \mid X = \mathbf{x}, Y = y\right). \quad (7)$$

Remark 3. Let the label set $\ell(\mathbf{x}) \subseteq \mathcal{Y}$ be the ground-truth multi-label annotation of an ID image $\mathbf{x}$. Each annotated label $y \in \ell(\mathbf{x})$ can induce a positive image-label pair $(\mathbf{x}, y)$, whereas each unannotated label $\hat{y} \notin \ell(\mathbf{x})$ can induce a negative image-label pair $(\mathbf{x}, \hat{y})$. In other words, a single ID image may contribute to multiple positive and negative pairs, meaning that the formulation of $S(\mathbf{x})$ in Eq. (7) does not impose mutual exclusivity among ID labels.

To set up an appropriate posterior $q_{C|XY}$ for computing $S(\mathbf{x})$ in Eq. (7), we first model how image-label pairs are generated under the two states of the latent variable C. When $C = 1$, the image $\mathbf{x}$ and the label y are associated, and thus the positive pair $(\mathbf{x}, y)$ can be drawn from the true joint distribution $\mathbb{P}_{XY}$[5]. In contrast, when $C = 0$, we, motivated by noise contrastive estimation (Gutmann & Hyvärinen, 2012), model the negative pair $(\mathbf{x}, y)$ as being independently sampled from their respective marginal distributions. Formally, the overall generation process is defined as

$$q\left(X = \mathbf{x}, Y = y \mid C = 1\right) \triangleq p_{XY}\left(\mathbf{x}, y\right), \quad (8)$$

$$q\left(X = \mathbf{x}, Y = y \mid C = 0\right) \triangleq p_X\left(\mathbf{x}\right) p_Y\left(y\right), \quad (9)$$

where p_Y denote the *mass* functions of the marginal distributions $\mathbb{P}_Y$, p_{XY} and p_X as the probability *density* functions of the joint distribution $\mathbb{P}_{XY}$ and the marginal distribution $\mathbb{P}_X$, respectively.

Supposing that we are given 1 positive pair (drawn from the joint data distribution $\mathbb{P}_{XY}$, i.e. the same input provided to X and Y) for every N negative pairs (drawn from the product of marginals $\mathbb{P}_X \otimes \mathbb{P}_Y$; independent randomly drawn inputs provided to X and Y), then we have the priors on the latent variable C as follows:

$$q\left(C = 1\right) = \frac{1}{N+1}, \quad q\left(C = 0\right) = \frac{N}{N+1}. \quad (10)$$

By simple manipulation and Bayes' rule, the posterior for the state $C = 0$ is given by:

$$\begin{aligned} & q\left(C = 0 \mid X = \mathbf{x}, Y = y\right) \\ = & \frac{q\left(X = \mathbf{x}, Y = y \mid C = 0\right) q\left(C = 0\right)}{q\left(X = \mathbf{x}, Y = y \mid C = 0\right) q\left(C = 0\right) + q\left(X = \mathbf{x}, Y = y \mid C = 1\right) q\left(C = 1\right)} \\ = & \frac{N \cdot p_X\left(\mathbf{x}\right) p_Y\left(y\right)}{p_{XY}\left(\mathbf{x}, y\right) + N \cdot p_X\left(\mathbf{x}\right) p_Y\left(y\right)} \\ = & \frac{N \cdot p_Y\left(y\right)}{p_{Y|X}\left(y|\mathbf{x}\right) + N \cdot p_Y\left(y\right)}. \end{aligned} \quad (11)$$

However, directly calculating $q\left(C = 0 \mid X = \mathbf{x}, Y = y\right)$ in Eq. (11), which is built upon the conditional distribution $\mathbb{P}_{Y|X}$ and the marginal distribution $\mathbb{P}_Y$, can be computationally intractable since

[4] In the rest of this paper, we write $q_{C|XY}(0|\mathbf{x}, y) = q\left(C = 0 \mid X = \mathbf{x}, Y = y\right)$.

[5] $\mathbb{P}_{XY}$ has different interpretations in multi-class and multi-label settings. See Appendix C for discussion.

the two underlying distributions are unknown in nearly all practical applications. In response to this challenge, our key idea[6] is to replace the unknown mass function $p_{Y|X}(y|\mathbf{x})$ with the estimated one $\hat{p}_{Y|X}(y|\mathbf{x};\boldsymbol{\theta})$ using the pre-trained VLM parameters $\boldsymbol{\theta}$ for a tractable estimator of the OOD scoring function $S(\mathbf{x})$ in Eq. (7), i.e.,

$$S(\mathbf{x};\boldsymbol{\theta}) = 1 - \prod_{i=1}^{K} q\left(C = 0 \mid X = \mathbf{x}, Y = y_i;\boldsymbol{\theta}\right), \tag{12}$$

where

$$q\left(C = 0 \mid X = \mathbf{x}, Y = y;\boldsymbol{\theta}\right) = \frac{N \cdot p_Y\left(y\right)}{\hat{p}_{Y|X}(y|\mathbf{x};\boldsymbol{\theta}) + N \cdot p_Y\left(y\right)}. \tag{13}$$

To tractably derive $\hat{p}_{Y|X}(y|\mathbf{x};\boldsymbol{\theta})$ in Eq. (13), we, motivated by prior studies (Zhang & Zhang, 2024; Do et al., 2021), parameterize $\hat{p}_{Y|X}(y|\mathbf{x};\boldsymbol{\theta})$ as an energy-based variational family that uses a critic $h_{\boldsymbol{\theta}}(\cdot,\cdot) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ parameterized by $\boldsymbol{\theta}$ and is scaled by the marginal mass function p_Y, i.e.,

$$\hat{p}_{Y|X}(y|\mathbf{x};\boldsymbol{\theta}) = \frac{p_Y(y)\exp h_{\boldsymbol{\theta}}(\mathbf{x},y)}{\sum_{\hat{y}\in\mathcal{Y}} p_Y(\hat{y})\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y})} = \frac{p_Y(y)\exp h_{\boldsymbol{\theta}}(\mathbf{x},y)}{\Phi_{\boldsymbol{\theta}}(\mathbf{x})}, \tag{14}$$

where the normalization constant $\Phi_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{\hat{y}\in\mathcal{Y}} p_Y(\hat{y})\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y})$ can be efficiently approximated by the classical Monte-Carlo importance sampling technique as in Peng et al. (2024). To be specific, given M i.i.d. samples $\{\hat{y}_j\}_{j=1}^{M}$ drawn from $\mathbb{P}_Y$, the law of large numbers implies

$$\Phi_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{\hat{y}\sim\mathbb{P}_Y}\left[\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y})\right] = \lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M}\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y}_i), \tag{15}$$

$$\frac{\hat{p}_{Y|X}(y|\mathbf{x};\boldsymbol{\theta})}{p_Y(y)} = \lim_{M\to\infty}\frac{M\exp h_{\boldsymbol{\theta}}(\mathbf{x},y)}{\sum_{i=1}^{M}\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y}_i)}. \tag{16}$$

Thanks to Eq. (16), we can derive a Monte-Carlo estimator for $S(\mathbf{x};\boldsymbol{\theta})$ in Eq. (12) as follows:

$$\hat{S}(\mathbf{x};\boldsymbol{\theta}) = 1 - \prod_{i=1}^{K}\frac{\lambda\cdot\sum_{j=1}^{M}\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y}_j)}{\exp h_{\boldsymbol{\theta}}(\mathbf{x},y_i) + \lambda\cdot\sum_{j=1}^{M}\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y}_j)}, \tag{17}$$

where $\lambda = N/M$ is a hyper-parameter. Following prior works (Ming et al., 2022; Jiang et al., 2024; Peng et al., 2026a) and for simplicity, we set $h_{\boldsymbol{\theta}}(\mathbf{x},y) = \mathbf{z}^\top\mathbf{r}_y/\tau$.

Theorem 3. *Let $\{\hat{y}_j\}_{j=1}^{M}$ be a set of M i.i.d. samples drawn from $\mathbb{P}_Y$, and define*[7]

$$\Delta(\mathbf{x};\boldsymbol{\theta}) = \left|S(\mathbf{x};\boldsymbol{\theta}) - \hat{S}(\mathbf{x};\boldsymbol{\theta})\right|.$$

Then, for any $\mathbf{x}\in\mathcal{X}$ and $\boldsymbol{\theta}$, the expected estimation error satisfies:

$$\mathbb{E}\left[\Delta(\mathbf{x};\boldsymbol{\theta})\right] \le K\frac{N}{(N+1)^2}\left(e^{4/\tau} - e^{2/\tau}\right)\frac{1}{\sqrt{M}},$$

where the expectation is taken over $\{\hat{y}_j\}_{j=1}^{M}\sim\mathbb{P}_Y^M$.

Remark 4. Theorem 3 provides a theoretical guarantee for our proposed tractable approximation. It demonstrates that for any input image $\mathbf{x}$ and pre-trained VLM parameters $\boldsymbol{\theta}$, the Monte-Carlo estimator $\hat{S}(\mathbf{x};\boldsymbol{\theta})$ in Eq. (17) is consistent. The expected absolute deviation from the intractable true OOD scoring function $S(\mathbf{x};\boldsymbol{\theta})$ decays at a rate of $\mathcal{O}(1/\sqrt{M})$, bounded by deterministic constants intrinsic to the VLM's embedding space and the chosen hyper-parameters.

[6] As we will demonstrate later, the mass function p_Y can cancel out so that there is no need to estimate it.

[7] Note that $\Delta(\mathbf{x};\boldsymbol{\theta})$ inherently depends on $\{\hat{y}_j\}_{j=1}^{M}$, but we omit this dependence for notational brevity.

Algorithm 1: Multi-label OOD Detection with Pre-trained Vision-Language Models

Input: Test input $\mathbf{x}$, Pre-trained VLM parameters $\boldsymbol{\theta}$, ID labels $\mathcal{Y}_{\text{I}}$, wild corpus dataset $\mathcal{D}$

Output: OOD scoring $\hat{S}_{ours}(\mathbf{x};\boldsymbol{\theta})$

1. Computing $\text{Rep}(\tilde{y})$ via Eq.(18) for each $\tilde{y} \in \mathcal{D}$

2. Reordering $\mathcal{D}$ according to the decreasing value of $\text{Rep}(\tilde{y})$ to have $\hat{\mathcal{D}} = \{\tilde{y}_{(1)}, \ldots, \tilde{y}_{(T)}\}$

3. Computing $\hat{S}_{\text{ours}}(\mathbf{x};\boldsymbol{\theta})$ via Eq. (19)

Table 1: OOD detection results on MS-COCO with ViT-B/16 CLIP as encoder. ↑ indicates larger values are better and vice versa. The best results in the last two columns are shown in bold.

Dataset	ImageNet		Textures		NUS-WIDE		Objects		Average	
Metric	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Methods requiring training (or fine-tuning)										
MSP	93.56	22.87	95.03	16.10	93.64	29.96	95.61	22.63	94.46	22.89
MaxLogit	95.51	16.12	97.00	10.67	96.35	17.09	96.96	14.69	96.44	14.16
JointEnergy	96.39	11.07	98.10	7.82	97.47	10.88	97.07	13.87	97.26	10.91
Training-free Zero-Shot Methods										
Energy	76.09	86.21	58.36	96.29	69.45	92.74	82.10	72.56	71.50	86.95
MaxLogit	82.06	75.57	69.93	91.70	77.05	85.19	86.22	61.18	78.82	78.41
MCM	90.82	43.68	90.91	38.42	92.40	39.87	91.44	42.96	91.40	41.23
GL-MCM	91.99	37.16	91.59	34.33	93.55	32.38	92.19	38.25	92.33	35.53
EOE	90.68	48.07	92.45	29.82	93.95	32.17	93.40	35.08	92.62	36.29
NegLabel	95.68	23.21	94.36	20.80	94.95	21.30	88.46	51.47	93.36	29.20
RenyiNeg	96.54	17.94	95.26	15.87	96.13	16.09	89.55	47.20	94.37	24.28
Ours	97.17	15.10	97.41	9.29	97.12	13.05	93.87	32.39	**96.39**	**17.46**

5 Practical Consideration

As shown in Eq. (17), the proposed $\hat{S}(\mathbf{x};\boldsymbol{\theta})$ requires the construction of $\{\hat{y}_i\}_{i=1}^{M}$. In view of this, a straightforward approach is to directly use the off-the-shelf wild corpus dataset $\mathcal{D} = \{\tilde{y}_1, \ldots, \tilde{y}_T\}$, e.g., WordNet (Miller, 1995). However, since wild data is available in abundance (Katz-Samuels et al., 2022), the large-scale nature of $\mathcal{D}$ make it impractical to caching the entire $\mathcal{D}$. In response, we consider selecting M the most representative samples in $\mathcal{D}$, where $M = L \ll T$ for simplicity and in alignment with prior works (Jiang et al., 2024; Peng et al., 2026a). Intuitively, representative samples are expected to reside in dense regions. Accordingly, we measure the representativeness of a sample $\tilde{y}_i \in \mathcal{D}$ by the density of its α-nearest neighbors[8], i.e.,

$$\text{Rep}(\tilde{y}_i) = -\log \sum_{\tilde{y}_j \in \mathcal{M}(\tilde{y}_i,\alpha)} \left\| \mathbf{r}_{\tilde{y}_i} - \mathbf{r}_{\tilde{y}_j} \right\|_2^2, \tag{18}$$

where $\tilde{y}_j \in \mathcal{M}(\tilde{y}_i, \alpha)$ is one of the α-nearest neighbors of $\tilde{y}_i$ in $\mathcal{D}$. Intuitively, a higher $\text{Rep}(\tilde{y}_i)$ suggests a more compact local structure, therefore indicating that the sample $\tilde{y}_i$ is more representative. From this perspective, we reorder $\mathcal{D}$ with regard to the decreasing representativeness measured by Eq. (18). Denote the reordered dataset as $\hat{\mathcal{D}} = \{\tilde{y}_{(t)}\}_{t=1}^{T}$, the representativeness-based selection criteria is then designed as $\hat{y}_i = \tilde{y}_{(i)}, \forall i = 1, \ldots, L$.

Finally, the practical formulation of our proposed OOD scoring function is given as follows:

$$\hat{S}_{\text{ours}}(\mathbf{x};\boldsymbol{\theta}) = 1 - \prod_{i=1}^{K} \frac{\lambda \cdot \sum_{j=1}^{M} \exp h_{\boldsymbol{\theta}}(\mathbf{x}, \tilde{y}_{(j)})}{\exp h_{\boldsymbol{\theta}}(\mathbf{x}, y_i) + \lambda \cdot \sum_{j=1}^{M} \exp h_{\boldsymbol{\theta}}(\mathbf{x}, \tilde{y}_{(j)})}, \tag{19}$$

For clarity, we summarize our algorithmic details in Algorithm 1.

6 Experiment

Implementation. Unless otherwise specified, we utilize the CLIP-B/16 architecture for zero-shot OOD detection. In line with previous studies (Jiang et al., 2024; Peng et al., 2026a), we formu-

[8] In practice, we use the Faiss library (Douze et al., 2024) for efficient nearest neighbor search.

Table 2: OOD detection results on PASCAL VOC with ViT-B/16 CLIP as encoder. ↑ indicates larger values are better and vice versa. The best results in the last two columns are shown in bold.

Dataset	ImageNet		Textures		NUS-WIDE		Objects		Average	
Metric	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Methods requiring training (or fine-tuning)										
MSP	90.79	46.39	97.38	12.77	96.54	16.80	97.00	14.99	95.43	22.74
MaxLogit	91.34	43.04	97.79	10.55	96.94	13.97	97.33	13.05	95.85	20.15
JointEnergy	91.29	43.11	98.25	8.10	97.34	13.12	97.65	11.51	96.13	18.96
Training-free Zero-Shot Methods										
Energy	81.94	78.71	75.85	85.34	77.86	82.77	90.74	50.54	81.60	74.34
MaxLogit	84.57	69.64	81.61	76.72	82.03	68.64	92.40	41.18	85.15	64.05
MCM	89.10	49.21	93.72	32.43	92.51	32.44	93.23	35.49	92.14	37.39
GL-MCM	90.45	43.63	94.38	30.69	93.12	30.68	93.51	34.05	92.87	34.76
EOE	88.46	61.04	95.16	22.23	93.17	32.73	94.69	29.82	92.87	36.45
NegLabel	95.80	26.08	97.53	13.65	94.89	24.60	92.00	46.25	95.03	27.64
RenyiNeg	96.10	24.66	97.35	13.48	94.73	24.66	92.36	43.36	95.14	26.54
Ours	97.14	14.86	98.70	5.60	95.94	17.04	93.63	32.54	**96.35**	**17.51**

late our text prompts as 'The nice <label>.' and construct the in-the-wild corpus dataset $\mathcal{D}$ using WordNet (Miller, 1995). During testing, all images are resized to a fixed resolution of 224 × 224. Regarding hyper-parameters, we set $\tau = 0.004$, $L = 10K$, $\alpha = 100$, and $\lambda = 1$.

Baseline. We compare our method with MSP (Hendrycks & Gimpel, 2016), ODIN (Liang et al., 2017), MaxLogit (Hendrycks et al., 2019), Energy (Liu et al., 2020), JointEnergy (Wang et al., 2021), Gradnorm (Huang et al., 2021), Vim (Wang et al., 2022), VOS (Du et al., 2022), NPOS (Tao et al., 2023), ZOC (Esmaeilpour et al., 2022), CLIPN (Wang et al., 2023), LoCoOp (Miyai et al., 2023), LSN (Nie et al., 2024), LAPT (Zhang et al., 2025), NegPrompt (Li et al., 2024), Mahalanobis (Lee et al., 2018), EOE (Cao et al., 2024), MCM (Ming et al., 2022), GL-MCM (Miyai et al., 2025), NegLabel (Jiang et al., 2024), RenyiNeg (Peng et al., 2026a).

6.1 Main Results

ID Datasets. Following Zolfi et al. (2024), we consider the following two multi-label ID datasets: MS-COCO 2017 (Lin et al., 2014) and PASCAL-VOC 2012 (Everingham et al., 2015). MS-COCO comprises 117,266 training, 4,952 validation, and 40,670 test images spanning 80 classes. PASCAL-VOC contains 5,717 training, 5,823 validation, and 10,991 test images across 20 classes. Evaluations are performed separately on each dataset, using their respective test set as ID data.

OOD Datasets. We inherit the setup from Wang et al. (2021) by considering Textures (Cimpoi et al., 2014) and a 20-category subset of ImageNet-22K (Deng et al., 2009) as OOD datasets. Since these two OOD datasets only contain single-class images and therefore present an oversimplified scenario, we follow Zolfi et al. (2024) by leveraging a 200-category subset of Objects365 (Shao et al., 2019) and a 54-category subset of NUS-WIDE (Chua et al., 2009); both of these contain images with multiple objects to better reflect the complexity of the multi-label setting.

Table 1 and Table 2 report the performance of our approach and existing competitive baselines on MS-COCO and PASCAL-VOC, respectively. From the tables, we can observe that advanced training-free zero-shot methods generally underperform learning-based methods, demonstrating the difficulty of achieving effective OOD detection without additional training or fine-tuning. In contrast, our approach substantially narrows this performance gap and consistently outperforms existing training-free baselines across the two benchmarks.

6.2 Ablation Study

Architectures. In principle, our method is generic to the choice of visual encoder. We evaluate our method with different visual encoder architectures, including ViT-B/32, ViT-L/14, ResNet-101, and report the corresponding OOD detection results in Table 3. On the one hand, the performance of OOD detection can be enhanced by more powerful visual encoders. On the other hand, our method consistently outperforms the most recent RenyiNeg regardless of the backbone architecture used, which implies the better generalization of our method over RenyiNeg.

Table 3: OOD detection results with different CLIP architectures on MS-COCO.↑ indicates larger values are better and vice versa. The best results in the last two columns are shown in bold.

Backbone	Method	ImageNet		Textures		NUS-WIDE		Objects		Average	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
ViT-B/32	NegLabel	94.48	26.27	93.54	26.58	94.40	25.58	88.30	55.66	92.68	33.52
	RenyiNeg	95.09	22.82	95.02	17.94	95.78	18.87	89.65	49.63	93.88	27.52
	Ours	96.55	16.96	96.87	11.01	95.93	19.42	92.56	39.90	**95.48**	**21.82**
ViT-L/14	NegLabel	94.87	27.34	92.80	25.41	94.53	22.91	90.36	43.89	93.14	29.89
	RenyiNeg	95.57	23.54	94.32	18.94	95.80	16.95	91.08	41.09	94.19	25.13
	Ours	96.41	17.20	97.54	9.24	96.47	17.78	94.29	30.11	**96.18**	**18.58**
ResNet-101	NegLabel	94.82	24.41	93.42	26.86	94.24	25.28	87.55	54.74	92.51	32.82
	RenyiNeg	94.57	23.83	95.35	16.70	96.09	16.68	88.27	52.45	93.57	27.41
	Ours	96.39	16.90	96.89	11.12	95.27	21.96	92.87	38.05	**95.36**	**22.01**

Table 4: OOD detection results with a larger input resolution on MS-COCO, where a ViT L/14 CLIP encoder is adopted. ↑ indicates larger values are better and vice versa. The best results in the last two columns are shown in bold.

Resolution	**Method**	**ImageNet**		**Textures**		**NUS-WIDE**		**Objects**		**Average**	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
336×336	NegLabel	94.80	27.76	93.27	23.94	94.63	22.79	88.63	51.72	92.83	31.55
	RenyiNeg	95.94	21.43	94.71	17.00	95.96	16.82	90.16	45.83	94.19	25.27
	Ours	96.50	16.50	97.63	9.10	96.79	15.09	93.99	31.77	**96.23**	**18.11**

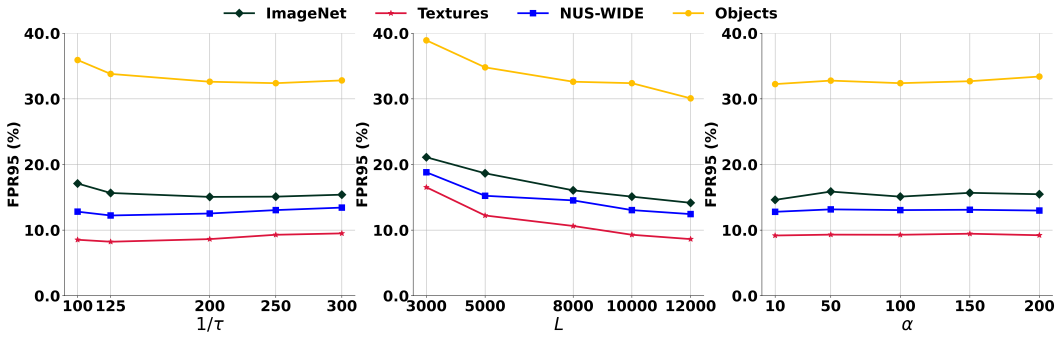


Figure 1: Hyper-parameter analysis on MS-COCO w.r.t τ (left), L (middle), and α (right).

Input Size. In principle, our method is generic to the input resolution. We evaluate our method with a larger input size, i.e., 336×336, and report the corresponding OOD detection results in Table 4. On the one hand, the performance of OOD detection can be enhanced by a larger input size. On the other hand, our method consistently outperforms the most recent NegLabel and RenyiNeg regardless of the input resolution, which implies the better generalization of ours over NegLabel and RenyiNeg.

Hyper-parameters. We evaluate the hyper-parameters most essential to our algorithmic design, including the temperature τ, L, and α. The corresponding results are plotted in Figure 1.

Due to space limitation, more ablation studies can be found in Appendix F.

6.3 Extension

Full-spectrum OOD Detection. We evaluate our approach on full-spectrum OOD detection (Yang et al., 2023), a more realistic setting that requires detecting semantic shifts while remaining robust to covariate shifts. Using PASCAL VOC as a case study, we treat a mixture of Clipart1k (Inoue et al., 2018), Watercolor2k (Inoue et al., 2018), and Comic2k (Inoue et al., 2018) as covariate-shifted ID data. Table 5 presents the experimental results on the same OOD datasets utilized in Section 6.1.

Multi-class OOD detection. Following prior work (Ming et al., 2022; Jiang et al., 2024; Peng et al., 2026a), We evaluate our method on the ImageNet-1K benchmark (Deng et al., 2009), where the validation set of ImageNet-1K is designated as the ID dataset while iNaturalist (Van Horn et al.,

Table 5: Full-spectrum OOD detection on PASCAL VOC with ViT-B/16 CLIP as encoder. ↑ indicates larger values are better and vice versa. The best results in the last two columns are in bold.

Dataset	ImageNet		Textures		NUS-WIDE		Objects		Average	
Metric	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Methods requiring training (or fine-tuning)										
MSP	81.04	81.10	92.29	46.60	90.01	57.32	91.27	55.86	88.65	60.22
MaxLogit	82.43	81.03	93.79	37.52	91.08	54.51	91.96	55.16	89.81	57.05
JointEnergy	83.79	75.66	97.87	18.63	92.56	50.10	91.75	56.63	91.49	50.51
Zero-Shot Training-free Methods										
Energy	54.33	98.24	58.94	90.64	41.15	99.12	56.50	95.44	52.73	95.86
MaxLogit	77.81	93.24	85.02	78.48	80.17	93.07	86.97	80.51	82.49	86.32
MCM	82.28	75.51	89.68	55.94	90.06	49.73	92.45	41.52	88.62	55.67
GL-MCM	89.70	49.63	89.48	46.83	89.19	56.38	88.57	55.54	89.23	52.09
EOE	81.47	80.93	92.19	39.61	87.43	69.27	93.59	38.99	88.67	57.20
NegLabel	92.41	42.33	94.92	29.88	88.25	55.34	91.04	47.96	91.65	43.88
RenyiNeg	92.22	41.71	95.23	26.93	88.28	54.86	91.24	46.44	91.74	42.49
Ours	95.62	22.49	97.83	9.98	90.68	47.97	92.35	44.26	**94.12**	**31.17**

Table 6: OOD detection results on ImageNet-1K, where a ViT B/16 CLIP encoder is adopted. ↑ indicates larger values are better and vice versa. The best results in the last two columns are in bold.

Dataset	iNaturalist		SUN		Places		Textures		Average	
Metric	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Methods requiring training (or fine-tuning)										
MSP	87.44	58.36	79.73	73.72	79.67	74.41	79.69	71.93	81.63	69.61
ODIN	94.65	30.22	87.17	54.04	85.54	55.06	87.85	51.67	88.80	47.75
Energy	95.33	26.12	92.66	35.97	91.41	39.87	86.76	57.61	91.54	39.89
GradNorm	72.56	81.50	72.86	82.00	73.70	80.41	70.26	79.36	72.35	80.82
ViM	93.16	32.19	87.19	54.01	83.75	60.67	87.18	53.94	87.82	50.20
VOS	94.62	28.99	92.57	36.88	91.23	38.39	86.33	61.02	91.19	41.32
NPOS	96.19	16.58	90.44	43.77	89.44	45.27	88.80	46.12	91.22	37.93
LSN	95.83	21.56	94.35	26.32	91.25	34.48	90.42	38.54	92.96	30.22
CLIPN	95.27	23.94	93.93	26.17	92.28	33.45	90.93	40.83	93.10	31.10
LoCoOp	96.86	16.05	95.07	23.44	91.98	32.87	90.19	42.28	93.52	28.66
LAPT	99.63	1.16	96.01	19.12	92.01	33.01	91.06	40.32	94.68	23.40
NegPrompt	98.73	6.32	95.55	22.89	93.34	27.60	91.60	35.21	94.81	23.01
Training-free Zero-Shot Methods										
Mahalanobis	55.89	99.33	59.94	99.41	65.96	98.54	64.23	98.46	61.50	98.94
Energy	85.09	81.08	84.24	79.02	83.38	75.08	65.56	93.65	79.57	82.21
ZOC	86.09	87.30	81.20	81.51	83.39	73.06	76.46	98.90	81.79	85.19
MCM	94.59	32.20	92.25	38.80	90.31	46.20	86.12	58.50	90.82	43.93
GL-MCM	96.71	15.16	93.41	29.16	90.37	37.07	83.11	55.85	90.90	35.06
EOE	97.52	12.29	95.73	20.04	92.95	30.16	85.64	57.53	92.96	30.09
NegLabel	99.49	1.91	95.49	20.53	91.64	35.59	90.22	43.56	94.21	25.40
RenyiNeg	99.64	1.29	95.71	17.94	91.90	33.98	90.79	39.45	94.51	23.17
Ours	99.48	1.74	96.08	16.77	92.03	32.76	90.97	38.99	**94.64**	**22.57**

2018), SUN (Xiao et al., 2010), Places365 (Zhou et al., 2017), and Textures (Cimpoi et al., 2014) are considered as OOD datasets. Table 6 shows that the empirical advantages of our method over the state-of-the-art still hold under the multi-class setup, which highlights its superiority and versatility.

7 Conclusion

This paper investigates training-free zero-shot OOD detection in the practically important yet underexplored multi-label setting. We formally prove that existing multi-class scoring mechanisms are fundamentally mismatched with multi-label data. To get out of this dilemma, we formulate multi-label OOD detection as probabilistic inference over pairwise image–label associations. This formulation yields a principled OOD score that measures the probability that an input is associated with at least one known ID label. We further derive a tractable training-free estimator based on noise-contrastive estimation and establish its consistency under Monte-Carlo approximation. Extensive experiments empirically manifest that our method consistently outperforms existing training-free zero-shot detectors across both single-object and multi-object benchmarks. We carefully discuss the limitations of our work and directions for future research in Appendix G.

ETHICS STATEMENT

Our study relies solely on publicly available datasets and models. No private or personally identifiable information was used. The work aims to advance the scientific understanding of OOD detection while upholding principles of transparency, fairness, and responsible research.

REPRODUCIBILITY STATEMENT

All the pre-trained CLIP-based models used in this paper are publicly accessible. We provide detailed proofs in the appendix. We believe that the implementation details provides in the main content are sufficient for reproduction. The code of this paper will be released upon acceptance.

REFERENCES

Chentao Cao, Zhun Zhong, Zhanke Zhou, Yang Liu, Tongliang Liu, and Bo Han. Envisioning outlier exposure by large language models for out-of-distribution detection. *arXiv preprint arXiv:2406.00806*, 2024.

Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pp. 161–168, 2006.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pp. 1–9, 2009.

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*, pp. 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*, pp. 248–255. Ieee, 2009.

Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31, 2018.

Kien Do, Truyen Tran, and Svetha Venkatesh. Clustering by maximizing mutual information across views. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9928–9938, 2021.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Sharon Li. Towards unknown-aware learning with virtual outlier synthesis. In *International Conference on Learning Representations*, volume 1, pp. 5, 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, pp. 6568–6576, 2022.

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

Chun-Mei Feng, Kai Yu, Xinxing Xu, Salman Khan, Rick Siow Mong Goh, Wangmeng Zuo, and Yong Liu. Text to image for multi-label image recognition with joint prompt-adapter learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Zixian Guo, Bowen Dong, Zhilong Ji, Jinfeng Bai, Yiwen Guo, and Wangmeng Zuo. Texts as images in prompt tuning for multi-label image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2808–2817, 2023.

Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of machine learning research*, 13 (2), 2012.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

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

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.

Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5001–5009, 2018.

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.

Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*, pp. 10848–10865. PMLR, 2022.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning*, pp. 5436–5446. PMLR, 2020.

Hao Lang, Yinhe Zheng, Yixuan Li, Jian Sun, Fei Huang, and Yongbin Li. A survey on out-of-distribution detection in nlp. *arXiv preprint arXiv:2305.03236*, 2023.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. 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.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Yicheng Liu, Jie Wen, Chengliang Liu, Xiaozhao Fang, Zuoyong Li, Yong Xu, and Zheng Zhang. Language-driven cross-modal classifier for zero-shot multi-label image recognition. In *Forty-first International Conference on Machine Learning*, 2024.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

Andrey Malinin and Mark Gales. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019.

Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2022.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Kevin Miller, Aditya Gangrade, Samarth Mishra, Kate Saenko, and Venkatesh Saligrama. Sparc: Score prompting and adaptive fusion for zero-shot multi-label recognition in vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4313–4321, 2025.

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. *Advances in Neural Information Processing Systems*, 36:76298–76310, 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. *International Journal of Computer Vision*, 133(6):3586–3596, 2025.

Peyman Morteza and Yixuan Li. Provable guarantees for understanding out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7831–7840, 2022.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.

Jun Nie, Yonggang Zhang, Zhen Fang, Tongliang Liu, Bo Han, and Xinmei Tian. Out-of-distribution detection with negative prompts. In *The Twelfth International Conference on Learning Representations*, 2024.

Bo Peng, Yadan Luo, Yonggang Zhang, Yixuan Li, and Zhen Fang. Conjnorm: Tractable density estimation for out-of-distribution detection. In *The Twelfth International Conference on Learning Representations*, 2024.

Bo Peng, Jie Lu, Yonggang Zhang, Guangquan Zhang, and Zhen Fang. Distributional prototype learning for out-of-distribution detection. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 1104–1114, 2025.

Bo Peng, Jie Lu, Zhen Fang, and Guangquan Zhang. A close look at negative label guided out-of-distribution detection in pre-trained vision-language models. In *Forty-third International Conference on Machine Learning*, 2026a.

Bo Peng, Jie Lu, Guangquan Zhang, and Zhen Fang. An information-theoretical framework for understanding out-of-distribution detection with pretrained vision-language models. *Advances in Neural Information Processing Systems*, 38:1930–1957, 2026b.

Bo Peng, Jie Lu, Guangquan Zhang, and Zhen Fang. Debiased negative mining improves out-of-distribution detection with pre-trained vision-language models. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 1171–1181, 2026c.

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

Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *arXiv preprint arXiv:2110.14051*, 2021.

Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning*, pp. 8491–8501. PMLR, 2020.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8430–8439, 2019.

Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *Advances in Neural Information Processing Systems*, 35:30569–30582, 2022a.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pp. 20827–20840. PMLR, 2022b.

Leitian Tao, Xuefeng Du, Xiaojin Zhu, and Yixuan Li. Non-parametric outlier synthesis. *arXiv preprint arXiv:2303.02966*, 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*, pp. 8769–8778, 2018.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4921–4930, 2022.

Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don't know? *Advances in Neural Information Processing Systems*, 34:29074–29087, 2021.

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*, pp. 1802–1812, 2023.

Shouwen Wang, Qian Wan, Junbin Gao, and Zhigang Zeng. Language-driven multi-label zero-shot learning with semantic granularity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1968–1978, 2025.

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*, pp. 3485–3492. IEEE, 2010.

Ming-Kun Xie, Zhiqiang Kou, Zhongnian Li, Gang Niu, and Masashi Sugiyama. Unlocking the power of co-occurrence in clip: A dualprompt-driven method for training-free zero-shot multi-label classification. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *International Journal of Computer Vision*, 131(10):2607–2622, 2023.

Yabin Zhang and Lei Zhang. Adaneg: Adaptive negative proxy guided ood detection with vision-language models. *Advances in Neural Information Processing Systems*, 37:38744–38768, 2024.

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*, pp. 271–288. Springer, 2025.

Yonggang Zhang, Jie Lu, Bo Peng, Zhen Fang, and Yiu-ming Cheung. Learning to shape in-distribution feature space for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 37:49384–49402, 2024.

Zihan Zhang and Xiang Xiang. Decoupling maxlogit for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3388–3397, 2023.

Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

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.

Alon Zolfi, Guy Amit, Amit Baras, Satoru Koda, Ikuya Morikawa, Yuval Elovici, and Asaf Shabtai. Yolood: Utilizing object detection concepts for multi-label out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5788–5797, 2024.

A Related Work

A.1 Multi-label Classification

Multi-Label classification aims to predict a set of labels for an input. Current approaches for adapting VLMs to multi-label classification generally fall into two distinct paradigms: fine-tuning and external knowledge-enhanced zero-shot adaptation. The first category, despite freezing the VLM backbone, necessitates labeled data to optimize auxiliary learnable parameters. DualCoOp (Sun et al., 2022a) leverages multi-label annotations to learn class-specific positive and negative prompt contexts. Given that images may be costly to collect while texts are cheap to generate, TaI-DPT (Guo et al., 2023) proposes texts as images in prompt tuning to adapt CLIP to multi-label image classification based on only the textual modality. Differently, T2I-PAL (Feng et al., 2025) bridges the modality gap in TaI-DPT using text-generated images alongside integrated prompting and adapter mechanisms. Despite their effectiveness, this reliance on offline supervised training limits their flexibility in data-scarce scenarios. In contrast, the second category boosts zero-shot performance by incorporating rich semantic priors from external knowledge sources, primarily Large Language Models (LLMs). Methods like CoMC (Liu et al., 2024), RCNn (Wang et al., 2025), and DualPrompt (Xie et al., 2026) employ LLMs to generate image captions, class descriptions, and co-occurrence relationships, respectively, which then serve to guide the frozen VLM.

A.2 Traditional Multi-class Out-of-distribution Detection

The popularity of OOD detection is motivated by the empirical observation (Nguyen et al., 2015) that neural networks tend to be over-confident in OOD data. One line of work performs OOD detection by devising post-hoc scoring functions, including confidence-based methods (Hendrycks et al., 2019; Ming et al., 2022; Zhang & Xiang, 2023), energy-based methods (Liu et al., 2020), distance-based approaches (Peng et al., 2025; Lee et al., 2018; Sun et al., 2022b; Zhang et al., 2024; Sastry & Oore, 2020; Morteza & Li, 2022; Peng et al., 2024), gradient-based approach (Huang et al., 2021), and Bayesian approaches (Kristiadi et al., 2020; Malinin & Gales, 2019). Another line of work addresses OOD detection by fine-tuning a pre-trained discrimination model with training-time regularizations that help the model learn ID/OOD discrepancy following the guideline of outlier exposure (Hendrycks et al., 2018). For instance, the discriminative model is regularized to produce lower confidence (Lee et al., 2017; Malinin & Gales, 2018), smaller feature magnitudes (Liu et al., 2020) or higher energy (Dhamija et al., 2018) for outlier points. More recently, some works have considered a practical scenario where the auxiliary outliers can be arbitrarily different from the real OOD data, therefore distributionally augmenting the observed OOD data. Besides, the given OOD samples tend to include unlabeled ID counterparts. Due to this, WOOD (Katz-Samuels et al., 2022) formulates learning with noisy OOD samples as a constrained optimization problem.

A.3 VLM-based Multi-class Out-of-distribution Detection

The core of VLM-based OOD detection lies in how to leverage textual supervision with pre-trained VLMs to assist OOD detection on the visual domain. On the one hand, the pioneering work, MCM (Ming et al., 2022), defines textual features as concept prototypes for each ID class and uses the scaled distance between visual features and the closest ID prototype to measure OOD uncertainty. Instead of relying on textual information from only ID label space, ZOC (Esmaeilpour et al., 2022) applies VLMs to discern OOD instances by training a captioner that generates potential OOD labels. Nevertheless, this captioner often fails to produce effective OOD labels, particularly for ID datasets containing many classes. Differently, NegLabel (Jiang et al., 2024) incorporates additional negative class names mined from available data sources as negative proxies. Considering the nonalignment between target visual OOD distribution and the generated negative textual OOD distribution, AdaNeg (Zhang & Zhang, 2024) leverages the benefits of test-time adaptation to generate adaptive proxies by exploring potential OOD images during testing. More recently, Peng et al. (2026b) understand CLIP-based post-hoc OOD detection from an information-theoretical perspective while Peng et al. (2026c) improve negative mining strategy in CLIP-based post-hoc OOD detection from a positive-unlabeled perspective. On the other hand, CLIP-based OOD detection can also be improved by prompt representation learning. In particular, LoCoOp (Miyai et al., 2023) learns ID text prompts by pushing them away from the portions of CLIP local features that have ID-irrelevant nuisances (e.g., backgrounds). CLIPN (Wang et al., 2023) and LSN (Nie et al., 2024)

design a learnable "no" prompt and a "no" text encoder to capture negation semantics within images. Differently, LAPT (Zhang et al., 2025) initializes prompts with negative labels, followed by tuning prompts with cross-modal and cross-distribution mixing.

A.4 Multi-label Out-of-distribution Detection

Despite substantial progress in OOD detection for multi-class classification, the multi-label setting remains relatively under-explored, with existing work primarily focusing on supervised models. JointEnergy (Wang et al., 2021) aggregates label-wise energy scores from a supervised multi-label classifier to capture evidence across categories, while YolOOD (Zolfi et al., 2024) exploits spatial objectness and class confidence through an adapted object-detection architecture. These studies highlight the importance of label-wise evidence aggregation and spatial modeling, but leave open how to perform multi-label OOD detection without task-specific supervised training. We address this gap with a training-free, zero-shot framework based on frozen pre-trained VLMs. Beyond this distinction in supervision, our approach derives the detection score from a probabilistic model of image–label associations, explicitly modeling the event that at least one ID label is present. This formulation grounds the aggregation of non-exclusive label evidence in the multi-label OOD criterion, without requiring labeled ID training images or bounding-box supervision.

B Proofs of Main Theorems

B.1 Proof of Theorem 1

Proof. Let $y^* = \arg\max_{y\in\mathcal{Y}_I} \mathbf{z}^\top \mathbf{r}_y$. Under the assumption $\max_{y\in\mathcal{Y}_I} \mathbf{z}^\top \mathbf{r}_y - \min_{\hat{y}\in\ell(\mathbf{x})} \mathbf{z}^\top \mathbf{r}_{\hat{y}} \le \epsilon$, we have, for every $y \in \ell(\mathbf{x})$, $\mathbf{z}^\top \mathbf{r}_{y^*} - \mathbf{z}^\top \mathbf{r}_y \le \epsilon$, and hence $\mathbf{z}^\top \mathbf{r}_y \ge \mathbf{z}^\top \mathbf{r}_{y^*} - \epsilon$.

Since $\ell(\mathbf{x}) \subseteq \mathcal{Y}_I$, $|\ell(\mathbf{x})| = T$, and $\tau > 0$, the denominator of the MCM score satisfies

$$\sum_{y\in\mathcal{Y}_I} \exp\left(\frac{\mathbf{z}^\top \mathbf{r}_y}{\tau}\right) \ge \sum_{y\in\ell(\mathbf{x})} \exp\left(\frac{\mathbf{z}^\top \mathbf{r}_y}{\tau}\right) \ge \sum_{y\in\ell(\mathbf{x})} \exp\left(\frac{\mathbf{z}^\top \mathbf{r}_{y^*} - \epsilon}{\tau}\right) = T \exp\left(\frac{\mathbf{z}^\top \mathbf{r}_{y^*} - \epsilon}{\tau}\right).$$

By the definition of y^*, substituting this lower bound into the MCM score yields

$$S_{\mathrm{MCM}}(\mathbf{x};\boldsymbol{\theta}) = \frac{\exp(\mathbf{z}^\top \mathbf{r}_{y^*}/\tau)}{\sum_{\hat{y}\in\mathcal{Y}_I} \exp(\mathbf{z}^\top \mathbf{r}_{\hat{y}}/\tau)} \le \frac{\exp(\mathbf{z}^\top \mathbf{r}_{y^*}/\tau)}{T\exp((\mathbf{z}^\top \mathbf{r}_{y^*} - \epsilon)/\tau)} = \frac{\exp(\epsilon/\tau)}{T}.$$

□

B.2 Proof of Theorem 2

Proof. Recall that the NegLabel score is defined as

$$S_{\mathrm{NegLabel}}(\mathbf{x};\boldsymbol{\theta}) = \sum_{y\in\mathcal{Y}_I} \frac{\exp(\eta_y(\mathbf{x})/\tau)}{\sum_{\hat{y}\in\mathcal{Y}_I\cup\mathcal{Y}_{\mathrm{neg}}} \exp(\eta_{\hat{y}}(\mathbf{x})/\tau)},$$

where $\eta_y(\mathbf{x}) = \mathbf{z}^\top \mathbf{r}_y$ and $\tau > 0$. Fix a secondary ID label $\hat{y} \in \ell(\mathbf{x})$, and define

$$a = \exp\left(\frac{\eta_{\hat{y}}(\mathbf{x})}{\tau}\right),$$

$$U(\mathbf{x},\hat{y}) = \sum_{y\in\mathcal{Y}_I\setminus\{\hat{y}\}} \exp\left(\frac{\eta_y(\mathbf{x})}{\tau}\right),$$

$$V(\mathbf{x}) = \sum_{y\in\mathcal{Y}_{\mathrm{neg}}} \exp\left(\frac{\eta_y(\mathbf{x})}{\tau}\right).$$

For notational simplicity, write $U = U(\mathbf{x},\hat{y})$ and $V = V(\mathbf{x})$. Then

$$S_{\mathrm{NegLabel}}(\mathbf{x};\boldsymbol{\theta}) = \frac{U + a}{U + a + V}.$$

When differentiating with respect to $\eta_{\hat{y}}(\mathbf{x})$, both U and V are constant, while

$$\frac{\partial a}{\partial \eta_{\hat{y}}(\mathbf{x})} = \frac{a}{\tau}.$$

Applying the quotient rule yields

$$\frac{\partial S_{\text{NegLabel}}(\mathbf{x};\boldsymbol{\theta})}{\partial \eta_{\hat{y}}(\mathbf{x})} = \frac{a}{\tau}\frac{(U+a+V)-(U+a)}{(U+a+V)^2} = \frac{aV}{\tau(U+a+V)^2}.$$

Since $a>0$, $\tau>0$, and $V>0$ whenever the negative-label set is nonempty, it follows that

$$\frac{\partial S_{\text{NegLabel}}(\mathbf{x};\boldsymbol{\theta})}{\partial \eta_{\hat{y}}(\mathbf{x})} > 0.$$

This proves the first claim.

For the second claim, observe that

$$\frac{\partial S_{\text{NegLabel}}(\mathbf{x};\boldsymbol{\theta})}{\partial \eta_{\hat{y}}(\mathbf{x})} = \frac{aV}{\tau(U+a+V)^2}.$$

Holding a, V, and τ fixed, we have

$$0 \le \frac{aV}{\tau(U+a+V)^2} \le \frac{aV}{\tau U^2}.$$

Because

$$\lim_{U\to\infty} \frac{aV}{\tau U^2} = 0,$$

the squeeze theorem gives

$$\lim_{U(\mathbf{x},\hat{y})\to\infty} \frac{\partial S_{\text{NegLabel}}(\mathbf{x};\boldsymbol{\theta})}{\partial \eta_{\hat{y}}(\mathbf{x})} = 0.$$

Therefore, although increasing the similarity associated with the secondary ID label always increases the NegLabel score, its marginal contribution vanishes at the rate $\mathcal{O}\big(U(\mathbf{x},\hat{y})^{-2}\big)$ when the remaining ID-label evidence becomes dominant. □

B.3 Proof of Theorem 3

Proof. For fixed $\mathbf{x}\in\mathcal{X}$ and $\boldsymbol{\theta}$, let $\Phi_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{\hat{y}\sim P_Y}\left[\exp\big(h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y})\big)\right]$, and its Monte Carlo estimator

$$\widehat{\Phi}_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{M}\sum_{j=1}^{M}\exp\big(h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y}_j)\big), \qquad \hat{y}_1,\dots,\hat{y}_M \overset{\text{i.i.d.}}{\sim} P_Y.$$

For each label $y_i\in\mathcal{Y}$, let $a_i = \exp\big(h_{\boldsymbol{\theta}}(\mathbf{x},y_i)\big)$. Substituting the energy-based parameterization

$$\widehat{p}_{Y|X}(y_i \mid \mathbf{x};\boldsymbol{\theta}) = \frac{p_Y(y_i)\exp\big(h_{\boldsymbol{\theta}}(\mathbf{x},y_i)\big)}{\Phi_{\boldsymbol{\theta}}(\mathbf{x})}$$

into the posterior in Eq. (11), the marginal mass $p_Y(y_i)$ cancels, yielding

$$q_i = q(C=0 \mid X=\mathbf{x}, Y=y_i;\boldsymbol{\theta}) = \frac{N\Phi_{\boldsymbol{\theta}}(\mathbf{x})}{a_i + N\Phi_{\boldsymbol{\theta}}(\mathbf{x})}.$$

Similarly, because $\lambda = N/M$, the Monte Carlo posterior appearing in Eq. (13) is

$$\widehat{q}_i = \frac{N\widehat{\Phi}_{\boldsymbol{\theta}}(\mathbf{x})}{a_i + N\widehat{\Phi}_{\boldsymbol{\theta}}(\mathbf{x})}.$$

Consequently,

$$S(\mathbf{x};\boldsymbol{\theta}) = 1 - \prod_{i=1}^{K} q_i, \quad \widehat{S}(\mathbf{x};\boldsymbol{\theta}) = 1 - \prod_{i=1}^{K}\widehat{q}_i.$$

Since the image and text representations are unit-normalized, i.e., $\|\mathbf{z}\|_2 = \|\mathbf{r}_y\|_2 = 1$. Thus,

$$-1 \le \mathbf{z}^\top \mathbf{r}_y \le 1.$$

Under the choice $h_{\boldsymbol{\theta}}(\mathbf{x}, y) = \mathbf{z}^\top \mathbf{r}_y / \tau$, we therefore have

$$e^{-1/\tau} \le \exp\big(h_{\boldsymbol{\theta}}(\mathbf{x}, y)\big) \le e^{1/\tau}.$$

Let $b_- = e^{-1/\tau}$, $b_+ = e^{1/\tau}$. It follows that

$$a_i,\ \Phi_{\boldsymbol{\theta}}(\mathbf{x}),\ \widehat{\Phi}_{\boldsymbol{\theta}}(\mathbf{x}) \in [b_-, b_+].$$

For each i, define

$$g_i(t) = \frac{Nt}{a_i + Nt}.$$

Its derivative is

$$g_i'(t) = \frac{N a_i}{(a_i + Nt)^2}.$$

For every $t \in [b_-, b_+]$, we have $a_i + Nt \ge b_- + N b_- = (N+1) b_-$ and

$$|g_i'(t)| \le \frac{N b_+}{(N+1)^2 b_-^2} = \frac{N}{(N+1)^2} e^{3/\tau}.$$

By the mean value theorem, then we have

$$|q_i - \widehat{q}_i| \le \frac{N}{(N+1)^2} e^{3/\tau} \left| \Phi_{\boldsymbol{\theta}}(\mathbf{x}) - \widehat{\Phi}_{\boldsymbol{\theta}}(\mathbf{x}) \right|.$$

Because $q_i, \widehat{q}_i \in [0, 1]$, the telescoping-product inequality gives

$$\begin{aligned} \Delta(\mathbf{x}; \boldsymbol{\theta}) &= \left| S(\mathbf{x}; \boldsymbol{\theta}) - \widehat{S}(\mathbf{x}; \boldsymbol{\theta}) \right| \\ &= \left| \prod_{i=1}^{K} q_i - \prod_{i=1}^{K} \widehat{q}_i \right| \\ &\le \sum_{i=1}^{K} |q_i - \widehat{q}_i| \\ &\le K \frac{N}{(N+1)^2} e^{3/\tau} \left| \Phi_{\boldsymbol{\theta}}(\mathbf{x}) - \widehat{\Phi}_{\boldsymbol{\theta}}(\mathbf{x}) \right|. \end{aligned} \tag{20}$$

To obtain the high-probability bound, let $W_j = \exp\big(h_{\boldsymbol{\theta}}(\mathbf{x}, \hat{y}_j)\big)$. Then $W_1, \dots, W_M$ are i.i.d.,

$$W_j \in [b_-, b_+], \qquad \Phi_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{\hat{y}_j \sim P_Y}[W_j], \qquad \widehat{\Phi}_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{M} \sum_{j=1}^{M} W_j.$$

By Hoeffding's inequality, with probability at least $1 - \delta$,

$$\left| \widehat{\Phi}_{\boldsymbol{\theta}}(\mathbf{x}) - \Phi_{\boldsymbol{\theta}}(\mathbf{x}) \right| \le (b_+ - b_-) \sqrt{\frac{\ln(2/\delta)}{2M}} = \left(e^{1/\tau} - e^{-1/\tau} \right) \sqrt{\frac{\ln(2/\delta)}{2M}}.$$

Substituting this inequality into Eq. (20) yields

$$\Delta(\mathbf{x}; \boldsymbol{\theta}) = K \frac{N}{(N+1)^2} \left(e^{4/\tau} - e^{2/\tau} \right) \sqrt{\frac{\ln(2/\delta)}{2M}}.$$

This proves the first claim.

For the expected error, taking expectations over $\{\hat{y}_j\}_{j=1}^{M} \sim \mathbb{P}_Y^M$ in Eq. (20) gives

$$\mathbb{E}[\Delta(x; \boldsymbol{\theta})] \le K \frac{N}{(N+1)^2} e^{3/\tau} \mathbb{E}\left[\left| \widehat{\Phi}_{\boldsymbol{\theta}}(\mathbf{x}) - \Phi_{\boldsymbol{\theta}}(\mathbf{x}) \right| \right].$$

By the Cauchy–Schwarz inequality,

$$\mathbb{E}\left[\left|\widehat{\Phi}_{\boldsymbol{\theta}}(\mathbf{x}) - \Phi_{\boldsymbol{\theta}}(\mathbf{x})\right|\right] \leq \sqrt{\mathbb{E}\left[\left(\widehat{\Phi}_{\boldsymbol{\theta}}(\mathbf{x}) - \Phi_{\boldsymbol{\theta}}(\mathbf{x})\right)^2\right]} = \sqrt{\mathrm{Var}\left(\widehat{\Phi}_{\boldsymbol{\theta}}(\mathbf{x})\right)} = \sqrt{\frac{\mathrm{Var}(W_1)}{M}}.$$

Since $W_1 \in [b_-, b_+]$, Popoviciu's inequality gives

$$\mathrm{Var}(W_1) \leq \frac{(b_+ - b_-)^2}{4}.$$

Hence,

$$\mathbb{E}\left[\left|\widehat{\Phi}_{\boldsymbol{\theta}}(\mathbf{x}) - \Phi_{\boldsymbol{\theta}}(\mathbf{x})\right|\right] \leq \frac{b_+ - b_-}{2\sqrt{M}} \leq \frac{b_+ - b_-}{\sqrt{M}} = \frac{e^{1/\tau} - e^{-1/\tau}}{\sqrt{M}}$$

such that

$$\mathbb{E}[\Delta(\mathbf{x}; \boldsymbol{\theta})] \leq K \frac{N}{(N+1)^2} e^{3/\tau} \left(e^{1/\tau} - e^{-1/\tau}\right) \frac{1}{\sqrt{M}} = K \frac{N}{(N+1)^2} \left(e^{4/\tau} - e^{2/\tau}\right) \frac{1}{\sqrt{M}}.$$

Thus the expected estimation error converges to zero at the rate $\mathcal{O}(M^{-1/2})$, completing the proof. □

C Interpretation of $\mathbb{P}_{XY}$ in multi-class and multi-label settings.

$\mathbb{P}_{XY}$ in the multi-class setting In the multi-class setting, each ID image $\mathbf{x}$ is associated with exactly one ID label. To be specific, let $\ell(\mathbf{x}) \in \mathcal{Y}_{\mathrm{I}} \subset \mathcal{Y}$ be the ground-truth label of an input $\mathbf{x}$, then the joint data distribution $\mathbb{P}_{XY}$ is unambiguous, i.e.,

$$p_{Y|X}(y \mid x) = \mathbb{1}\{y = \ell(x)\}, \quad p_{XY}(x, y) = p_X(x)\mathbb{1}\{y = \ell(x)\}.$$

Thus, $(X, Y) \sim \mathbb{P}_{XY}$ is an ordinary labeled pair consisting of an image and its unique class label.

$\mathbb{P}_{XY}$ in the multi-label setting In the multi-label setting, the complete annotation of an image is a set $\ell(x) \subseteq \mathcal{Y}$, or, more generally, by a random set variable $\mathcal{L}$. Since $\mathcal{Y}$ remains a single categorical variable, $\mathbb{P}_{XY}$ cannot directly represent the complete set-valued annotation. Instead, it should be interpreted as a distribution over positive image–label pairs such that a precise construction is

$$\mathbf{x} \sim \mathbb{P}_X, \quad \ell \sim \mathbb{P}_{\mathcal{L}|X}, \quad y \sim \mathbb{Q}_{Y|X\mathcal{L}},$$

where the probability mass function $q_{Y|X\mathcal{L}}$ should satisfy

$$q_{Y|X\mathcal{L}}(y \mid \mathbf{x}, \ell) = 0 \quad \text{for every } y \notin \ell, \qquad \sum_{y \in \ell} q_{Y|X\mathcal{L}}(y \mid \mathbf{x}, \ell) = 1.$$

The induced joint distribution is

$$p_{XY}(\mathbf{x}, y) = p_X(\mathbf{x}) \sum_{\ell : y \in \ell} p_{\mathcal{L}|X}(\ell \mid \mathbf{x}) p_{Y|X\mathcal{L}}(y \mid \mathbf{x}, \ell) = p_X(\mathbf{x}) q_{Y|X\mathcal{L}}(y \mid \mathbf{x}, \ell(\mathbf{x})),$$

where the last equation holds because each input $\mathbf{x}$ has a deterministic ground-truth label set $\ell(\mathbf{x})$.

A naive choice is to sample uniformly among the labels associated with $\mathbf{x}$, namely

$$q_{Y|X\mathcal{L}}(y \mid x, \ell(\mathbf{x})) = \frac{\mathbb{1}\{y \in \ell(\mathbf{x})\}}{|\ell(\mathbf{x})|},$$

and hence

$$p_{Y|X}(y \mid x) = \frac{\mathbb{1}\{y \in \ell(\mathbf{x})\}}{|\ell(\mathbf{x})|},$$

Under this multi-label interpretation, drawing $(X, Y) \sim \mathbb{P}_{XY}$ means first drawing an image, and subsequently drawing one of its valid labels. It does not imply that Y contains the complete multi-label annotation.

Table 7: OOD detection results with different prompt templates on MS-COCO, where a ViT B/16 CLIP encoder is adopted. ↑ indicates larger values are better and vice versa. The best results in the last two columns are shown in bold.

Prompt	Method	ImageNet		Textures		NUS-WIDE		Objects		Average	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
<label>	NegLabel	93.14	32.68	86.08	62.55	86.12	57.88	89.96	49.64	88.83	50.69
	RenyiNeg	92.36	35.71	88.24	55.27	86.35	57.12	91.11	44.88	89.51	48.25
	Ours	94.90	25.09	89.76	44.18	88.90	46.45	90.77	46.24	**91.08**	**40.49**
a <label> in a video game	NegLabel	91.46	42.20	85.06	73.07	86.44	63.39	89.58	51.76	88.13	57.60
	RenyiNeg	90.52	45.18	87.38	60.32	87.61	56.35	90.57	47.28	89.02	52.28
	Ours	93.94	28.51	88.92	50.09	89.76	45.07	91.01	44.35	**90.91**	**42.00**
itap of a <label>	NegLabel	90.43	43.29	81.21	76.31	87.50	49.37	91.43	40.93	87.64	52.48
	RenyiNeg	89.71	45.19	84.60	65.04	86.81	52.10	92.13	37.20	88.31	49.88
	Ours	93.77	28.48	87.69	52.93	90.37	36.66	92.28	37.13	**91.03**	**38.80**
a origami <label>	NegLabel	92.73	30.97	89.30	47.84	89.90	41.84	86.85	55.57	89.69	44.05
	RenyiNeg	91.95	33.86	90.09	43.95	90.19	40.50	87.71	51.68	89.98	42.50
	Ours	93.83	27.32	89.65	42.52	91.06	35.80	87.97	52.60	**90.63**	**39.56**
art of the <label>	NegLabel	94.55	25.19	86.52	55.83	85.05	57.39	86.48	57.28	88.15	48.92
	RenyiNeg	94.08	26.59	88.21	47.77	85.14	55.80	87.34	52.19	88.69	45.59
	Ours	95.91	19.23	89.22	44.04	87.78	46.58	87.12	55.65	**90.01**	**41.38**

D Evaluation Metrics

The performance of OOD detection is evaluated via two widely used metrics: 1) the false positive rate of OOD data is measured when the true positive rate of ID data reaches 95% (FPR95); 2) the area under the receiver operating characteristic curve (AUROC) is computed to quantify the probability of the ID case receiving a higher score than the OOD case.

E More Implementation Details

For the training-based baseline in Table 1 and Table 2, we train a linear multi-label classifier for each ID dataset. Each classifier is randomly initialized and trained on its corresponding multi-label dataset using a sigmoid activation function. We use the SGD optimizer with a batch size of 2048, a base learning rate of 0.1, momentum of 0.9, and no weight decay. The learning rate follows a cosine decay schedule with a 10-epoch warmup period. For each ID dataset, we select the checkpoint that achieves the highest mAP on the validation split.

F Ablation Study on Prompt Template

We evaluate our method with different prompt template on MS-COCO, as shown in Table 7.

G Limitations and Outlook

Despite its theoretical motivation and empirical effectiveness, our framework has several limitations that suggest directions for future research.

First, our probabilistic formulation adopts a product-based aggregation of pairwise image–label associations without explicitly modeling dependencies among labels. Since objects often exhibit strong co-occurrence patterns, incorporating such dependencies could improve the estimation of image-level ID confidence. Future work could explore dependency-aware aggregation while preserving the training-free nature of our framework.

Second, our method inherits the representational limitations of pre-trained VLMs and depends on the coverage of the external text corpus. Global image representations may provide insufficient evidence for small, occluded, or less salient objects, while a fixed lexical resource may inadequately represent specialized concepts. Incorporating region-level visual evidence and exploring more diverse or adaptive corpus construction are promising directions for improving robustness.