An Information-theoretical Framework for Understanding Out-of-distribution Detection with Pretrained Vision-Language Models

Bo Peng, Jie Lu, Guangquan Zhang, Zhen Fang*
University of Technology Sydney

Abstract

Out-of-distribution (OOD) detection, recognized for its ability to identify samples of unknown classes, provides solid advantages in ensuring the reliability of machine learning models. Among existing OOD detection methods, pre-trained vision-language models have emerged as powerful post-hoc OOD detectors by leveraging textual and visual information. Despite the empirical success, there still remains a lack of research on a formal understanding of their effectiveness. This paper bridges the gap by theoretically demonstrating that existing CLIP-based post-hoc methods effectively perform a stochastic estimation of the point-wise mutual information (PMI) between the input image and each in-distribution label. This estimation is then utilized to construct energy functions for modeling in-distribution distributions. Different from prior methods that inherently consider PMI estimation as a whole task, we, motivated by the divide-and-conquer philosophy, decompose PMI estimation into multiple easier sub-tasks by applying the chain rule of PMI, which not only reduces the estimation complexity but also provably increases the estimation upper bound to reduce the underestimation bias. Extensive evaluations across mainstream benchmarks empirically manifest that our method establishes a new state-of-the-art in a variety of OOD detection setups.

1 Introduction

Despite the significant progress in machine learning that has facilitated a broad spectrum of classification tasks [2, 69, 39], models often operate under a *closed-world* scenario, where test data stems from the same distribution as the training data. However, real-world applications often entail *open-world* scenarios in which deployed models may encounter unseen classes of samples during training, giving rise to what is known as out-of-distribution (OOD) data. These OOD instances can potentially undermine a model's stability and, in certain cases, inflict severe damage on its performance. Accordingly, a reliable discriminative model should not only correctly classify known in-distribution (ID) samples but also flag any OOD inputs as unknown. This directly motivates OOD detection [49, 63, 26] which ensures the safety of decision-critical applications [23, 71].

This paper focuses on post-hoc OOD detection, which are more practical than learning-based methods that require resource-intensive retraining. Earlier studies [18, 32, 52, 34] primarily utilized the single modality of pre-trained models, but the success of contrastive language-image pre-training (CLIP) [48] has recently shifted research toward expanding post-hoc OOD detection from single-modal to multi-modal methods. Researchers have since explored ways to better leverage multi-modal models to enhance the performance and applicability of post-hoc OOD detection. A notable method is MCM [40], which 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. This

*Correspondence to Zhen Fang (zhen.fang@uts.edu.au)

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

method has paved the way for using pre-trained vision-language models (VLMs) in post-hoc OOD detection. However, MCM relies only on textual information from the ID label space, leaving VLMs' text interpretation capabilities underutilized. To address this, NegLabel [24] introduces numerous negative labels, allowing the model to better distinguish OOD samples. A heuristic grouping strategy in NegLabel is also proposed to further enhance OOD detection performance. Despite its promising potential, it is worth noting that a formalized understanding of CLIP-based post-hoc OOD detection remains significantly lacking in the field. This prompts the research question underlying this work:

How to theoretically justify the empirical effectiveness of CLIP-based post-hoc OOD detection?

Theoretical Significance. To address this challenge, we draw inspiration from information theory and propose an information-theoretical density-based framework. In this framework, ID data is modeled as an energy-based model, where the point-wise mutual information (PMI) [4] between the input image and each ID label forms the energy functions. We argue that this analytical framework is well-suited for studying OOD detection, as OOD data, by definition, inherently diverges from ID data in terms of their underlying density distributions. Guided by this framework, we show that representative CLIP-based post-hoc OOD detection methods [40, 24] can be interpreted as stochastic Monte Carlo estimations of PMI. Furthermore, we theoretically establish the following key points: 1) introducing negative labels increases the estimation upper bound, thereby mitigating underestimation bias; and 2) the grouping strategy effectively approximates the expectation through multiple sampling, reducing estimation variance.

Algorithmic Contribution. To further facilitate PMI estimation for OOD scoring, the starting point of our method is to decompose PMI as a sum of terms by applying the chain rule on PMI. In addition to reduce the overall estimation complexity according to the *divide-and-conquer* philosophy, we prove that the decomposed PMI estimation can further increases the estimation upper bound to reduce underestimation bias without explicitly introduce a corresponding number of negative labels. Notably, NegLabel [24] has empirically found that introducing excessive negative label would degrade OOD detection performance.

2 Related work

The core of CLIP-based OOD detection lies in how to leverage texture supervision with pre-trained VLMs to assist OOD detection on the visual domain. On the one hand, the pioneering work, MCM [40], defines textual features as concept proto- types for each ID class and uses the scaled distance between visual features and the closest ID prototype to measure OOD uncertainty. Intead of relying only on textual information from the ID label space, NegLabel [24] incorporates additional negative class names mined from available data sources, such as WordNet, as negative proxies. To mitigate the nonalignment between target visual OOD distribution and the generated negative textual OOD distribution, AdaNeg [66] leverages the benefits of test-time adaptation to generate adaptive proxies by exploring potential OOD images during testing. On the other hand, CLIP-based OOD detection can also be improved by prompt representation learning. In particular, LoCoOp [41] learns ID text prompts by pushing them away from the portions of CLIP local features that have ID-irrelevant nuisances (e.g., backgrounds). CLIPN [59] and LSN [44] design a learnable "no" prompt and a "no" text encoder to capture negation semantics within images. Differently, LAPT [67] initializes prompts with negative labels [24], followed by tuning prompts with cross-modal and cross-distribution mixing. *Due to limited space, related works on traditional OOD detection are discussed in Appendix A.*

3 Preliminary

Notations. Let $\mathcal{X}$ and $\mathcal{Y}$ represent the input space and the label space, respectively. Considering two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, we represent 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. Similarly, p_Y and $p_{Y|X}$ denote the *mass* functions of marginal and conditional distributions $\mathbb{P}_Y$ and $\mathbb{P}_{Y|X}$, respectively. We write $\mathbb{P}_{X_{\text{I}}Y_{\text{I}}}$ as the joint ID distribution defined over $\mathcal{X} \times \mathcal{Y}_{\text{I}}$, where $\mathcal{Y}_{\text{I}} \triangleq \{y_1, \dots, y_K\} \subset \mathcal{Y}$ is the

space for *known* ID labels. During testing, there are some unknown OOD joint distributions $\mathbb{P}_{X_{\text{o}}Y_{\text{o}}}$ defined over $\mathcal{X} \times \mathcal{Y}_{\text{o}}$, where $\mathcal{Y}_{\text{o}} \triangleq \mathcal{Y} \setminus \mathcal{Y}_{\text{I}}$ presents the space of *unknown* OOD labels.

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, then $\mathbf{x}$ is detected as ID data if and only if $S(\mathbf{x};\boldsymbol{\theta}) \geq \lambda$, for some given threshold λ:

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

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

CLIP-based Models. Given any visual input $\mathbf{x} \in \mathcal{X}$ and any label $y \in \mathcal{Y}$, we extract features of $\mathbf{x}$ and y using an arbitrary CLIP-based model $\mathbf{f}$ that consists an image encoder $\mathbf{f}(\cdot;\boldsymbol{\theta}_{\text{img}})$ and an text encoder $\mathbf{f}(\cdot;\boldsymbol{\theta}_{\text{text}})$ as follows:

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

where $\mathcal{Q}(\cdot)$ is the text prompt template, $\mathbb{S}^{d-1} \triangleq \left\{\mathbf{r} \in \mathbb{R}^d | \, \|\mathbf{r}\|_2 = 1\right\}$, and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{\text{img}}, \boldsymbol{\theta}_{\text{text}}\}$.

CLIP-based OOD Detectors Studied. CLIP-based models, which are initially proposed for zero-shot ID classification, have recently been extended to zero-shot OOD detection where there is no need to train on ID samples. The pioneering work, MCM [40], 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 CLIP-based models, i.e.,

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

where $\tau > 0$ is a temperature hyper-parameter. Unlike MCM that only employs information from the ID label space, NegLabel [24] introduces a L-sized set of negative labels[2] $\{y_{K+1}, \ldots, y_{K+L}\}$ sourced from lexical databases, followed by randomly grouping the selected L negative labels into T *non-overlapping* subsets $\mathcal{G}_1, ..., \mathcal{G}_T$, i.e., $\mathcal{G}_i \cap \mathcal{G}_j = \emptyset, \forall i \neq j$:

$$S_{\text{NegLabel}}(\mathbf{x};\boldsymbol{\theta}) \triangleq \frac{1}{T} \sum_{t=1}^{T} \sum_{y \in \mathcal{Y}_{\text{I}}} \frac{\exp(\mathbf{z}^\top \mathbf{r}/\tau)}{\sum\limits_{y_j \in \mathcal{G}_t \cup \mathcal{Y}_{\text{I}}} \exp(\mathbf{z}^\top \mathbf{r}_j/\tau)}. \tag{3}$$

4 Theoretical Analysis

While both MCM and NegLabel have empirically emerged to be effective post-hoc OOD detectors, their inherent connections and theoretical understandings are largely lacking. To the best of our knowledge, there is limited prior work providing provable guarantees for CLIP-based post-hoc OOD detection methods from a rigorous mathematical point of view. In this section, we provide theoretical justification for CLIP-based post-hoc OOD detection from the perspective of *information-theoretical density estimation*. In particular, due to the fact that OOD data, by definition, inherently diverges from ID data by means of their data density distributions, we, following advanced density-based OOD detection methods [46, 42, 34], render the ID density function as an ideal metric for ID-OOD discrimination. Inspired by prior works [34], we consider modeling the unknown true ID density function $p_{X_{\text{I}}}$ of ID input marginal distribution $\mathbb{P}_{X_{\text{I}}}$ by resorting to the energy-based model [27, 17]:

$$\hat{p}_{X_{\text{I}}}(\mathbf{x}) = \frac{\exp\left[E(\mathbf{x})\right]}{Z} \propto \exp\left[E(\mathbf{x})\right], \quad E(\mathbf{x}) = \frac{1}{\alpha} \log \sum_{y \in \mathcal{Y}_{\text{I}}} \exp\left[\alpha \mathcal{W}(\mathbf{x}, y)\right], \tag{4}$$

where $Z = \int \exp\left[E(\mathbf{x})\right] d\mathbf{x}$ is an *input-independent* normalization function, $\alpha > 0$ is a hyper-parameter and $\mathcal{W}(\mathbf{x}, y)$ is the *point-wise mutual information* (PMI) [4] that explicitly measures the association between the input $\mathbf{x} \in \mathcal{X}$ and the label $y \in \mathcal{Y}$. As implied by the following definition of PMI, the formulation in Eq. (4) implicitly assumes that the modelled ID density $\hat{p}_{X_{\text{I}}}(\mathbf{x})$ is induced by an underlying unknown joint distribution $\mathbb{P}_{XY}$ defined over $\mathcal{X} \times \mathcal{Y}$.

[2] In accordance to Jiang et al. [24], labels $y \in \mathcal{Y}_{\text{o}}$ that have lower affinities with ID images compared to OOD images are considered as negative labels

Definition 1. The PMI between two observations $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$ is defined as follows:

$$\mathcal{W}(\mathbf{x}, y) \triangleq \log \frac{p_{XY}(\mathbf{x}, y)}{p_X(\mathbf{x}) p_Y(y)} = \log \frac{p_{Y|X}(y|\mathbf{x})}{p_Y(y)}. \tag{5}$$

Note that directly calculating $\mathcal{W}(\mathbf{x}, y)$ in Eq. (5), which is built upon the conditional distribution $\mathbb{P}_{Y|X}$ and the marginal distribution $\mathbb{P}_Y$, can be computationally intractable since the two underlying distributions are unknown in nearly all practical applications. In response to this challenge, our key idea 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 CLIP-based model parameters $\boldsymbol{\theta}$ for a tractable estimator of the modeled ID data density function $\hat{p}_{X_{\text{I}}}(\mathbf{x})$ in Eq. (4), i.e.,

$$\hat{p}_{X_{\text{I}}}(\mathbf{x}; \boldsymbol{\theta}) = \frac{\exp\left[E_{\boldsymbol{\theta}}(\mathbf{x})\right]}{Z_{\boldsymbol{\theta}}} \propto \exp\left[E_{\boldsymbol{\theta}}(\mathbf{x})\right], \quad E_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{\alpha} \log \sum_{y \in \mathcal{Y}_{\text{I}}} \exp\left[\alpha \hat{\mathcal{W}}(\mathbf{x}, y)\right], \tag{6}$$

where $Z_{\boldsymbol{\theta}} = \int \exp\left[E_{\boldsymbol{\theta}}(\mathbf{x})\right] d\mathbf{x}$ and $\hat{\mathcal{W}}(\mathbf{x}, y; \boldsymbol{\theta}) = \log \frac{\hat{p}_{Y|X}(y|\mathbf{x}; \boldsymbol{\theta})}{p_Y(y)}$ is the estimator[3] of the true PMI $\mathcal{W}(\mathbf{x}, y)$. In the following, we demonstrate that MCM and NegLabel, despite their seemingly distinct scoring functions, can be interpreted as methods for stochastically estimating PMI and therefore the energy function that effectively replicates the behavior of the ID density.

4.1 Towards Understanding MCM

To tractably derive $\hat{p}_{Y|X}(y|\mathbf{x}; \boldsymbol{\theta})$ in Eq. (6), we, motivated by prior works [47, 9], assume that $\hat{p}_{Y|X}(y|\mathbf{x}; \boldsymbol{\theta})$ belongs to a energy-based variational family that uses a critic $h_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$ and is scaled by the marginal mass function p_Y, i.e.,

$$\hat{p}_{Y|X}(y|\mathbf{x}; \boldsymbol{\theta}) = \frac{p_Y(y) \exp h_{\boldsymbol{\theta}}(\mathbf{x}, y)}{\sum_{\hat{y} \in \mathcal{Y}} p_Y(\hat{y}) \exp h_{\boldsymbol{\theta}}(\mathbf{x}, \hat{y})} = \frac{p_Y(y) \exp h_{\boldsymbol{\theta}}(\mathbf{x}, y)}{\Phi_{\boldsymbol{\theta}}(\mathbf{x})}, \tag{7}$$

where $\Phi_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{\hat{y} \sim \mathbb{P}_Y}\left[\exp h_{\boldsymbol{\theta}}(\mathbf{x}, \hat{y})\right]$ is the normalization function. Following Peng et al. [46], one can consider a Monte-Carlo method to construct a simple and analytically tractable estimator of $\Phi_{\boldsymbol{\theta}}(\mathbf{x})$ by sampling a N-sized set of *i.i.d.* samples $\hat{\mathbf{y}}_N = \{\hat{y}_1, ..., \hat{y}_N\} \sim \mathbb{P}_Y^N$, i.e.,

$$\Phi_{\boldsymbol{\theta}}(\mathbf{x}) \approx \frac{1}{N} \sum_{j=1}^{N} \exp h_{\boldsymbol{\theta}}(\mathbf{x}, \hat{y}_j), \tag{8}$$

which implies that

$$\log \frac{\hat{p}_{Y|X}(y|\mathbf{x}; \boldsymbol{\theta})}{p_Y(y)} \approx \log \frac{N \exp h_{\boldsymbol{\theta}}(\mathbf{x}, y)}{\sum_{j=1}^{N} \exp h_{\boldsymbol{\theta}}(\mathbf{x}, \hat{y}_j)}. \tag{9}$$

If we set $h_{\boldsymbol{\theta}}(\mathbf{x}, y) = \mathbf{z}^\top \mathbf{r}/\tau$ and $\hat{\mathbf{y}}_N = \mathcal{Y}_{\text{I}}$ such that $N = K$ in Eq. (9), in the extreme case where $\alpha \to +\infty$, combining Eq. (9) and Eq. (6) implies that

$$\begin{aligned}
\lim_{\alpha \to +\infty} E_{\boldsymbol{\theta}}(\mathbf{x}) &= \lim_{\alpha \to +\infty} \frac{1}{\alpha} \log \sum_{y \in \mathcal{Y}_{\text{I}}} \exp\left[\alpha \log \frac{\hat{p}_{Y|X}(y|\mathbf{x}; \boldsymbol{\theta})}{p_Y(y)}\right] \\
&= \max_{y \in \mathcal{Y}_{\text{I}}} \log \frac{\hat{p}_{Y|X}(y|\mathbf{x}; \boldsymbol{\theta})}{p_Y(y)} \\
&\approx \max_{y \in \mathcal{Y}_{\text{I}}} \log \frac{K \cdot \exp h_{\boldsymbol{\theta}}(\mathbf{x}, y)}{\sum_{j=1}^{K} \exp h_{\boldsymbol{\theta}}(\mathbf{x}, y_j)} \\
&= \log \underbrace{\max_{y \in \mathcal{Y}_{\text{I}}} \frac{\exp(\mathbf{z}^\top \mathbf{r}/\tau)}{\sum_{j=1}^{K} \exp(\mathbf{z}^\top \mathbf{r}_j/\tau)}}_{S_{\text{MCM}}(\mathbf{x}; \boldsymbol{\theta})} + \underbrace{\log K}_{\text{const}}.
\end{aligned} \tag{10}$$

Since the logarithm function is monotonically increasing, Eq. (10) implies that $S_{\text{MCM}}(\mathbf{x}; \boldsymbol{\theta})$ in Eq. (2) can be understood as a stochastic estimator of $\lim_{\alpha \to +\infty} E_{\boldsymbol{\theta}}(\mathbf{x})$ (up to a constant). Theorem 1 provides provable guarantees of how $S_{\text{MCM}}(\mathbf{x}; \boldsymbol{\theta})$ correctly recovers the true energy function $E(\mathbf{x})$ in Eq. (4) when $\alpha \to +\infty$.

[3] As we will demonstrate later, the mass function p_Y can cancel out during the calculation of $\hat{\mathcal{W}}(\mathbf{x}, y; \boldsymbol{\theta})$ so that there is no need to estimate p_Y.

Theorem 1. *Let $h_{\boldsymbol{\theta}}(\mathbf{x}, y) = \mathbf{z}^\top \mathbf{r}/\tau$ and $N = K$. If we, following prior works [47, 45], assume that $h_{\boldsymbol{\theta}}(\mathbf{x}, y) = \log \frac{p_{Y|X}(y|\mathbf{x})}{p_Y(y)} + c(\mathbf{x})$ with $c(\mathbf{x})$ as a constant term depending on $\mathbf{x}$, in the extreme case where $\alpha \to +\infty$, then we have the following[4]:*

$$\lim_{\alpha \to +\infty} E(\mathbf{x}) = \lim_{\alpha \to +\infty} E_{\boldsymbol{\theta}}(\mathbf{x}) = \lim_{N \to +\infty} \log N S_{MCM}(\mathbf{x}; \boldsymbol{\theta}). \tag{11}$$

4.2 Towards Understanding NegLabel

To study NegLabel theoretically, we make the ID data density estimation depend on multiple samples. In particular, given a set of *i.i.d.* samples $\hat{\mathbf{y}}_N = \{\hat{y}_1, ..., \hat{y}_N\} \sim \mathbb{P}_Y^N$, we can rewrite $\hat{p}_{Y|X}(y|\mathbf{x}; \boldsymbol{\theta})$ as:

$$\hat{p}_{Y|X}(y|\mathbf{x}; \boldsymbol{\theta}) = \mathbb{E}_{\hat{\mathbf{y}}_N \sim \mathbb{P}_Y^N} \left[\hat{p}_{Y|X\mathbf{Y}_N}(y|\mathbf{x}, \hat{\mathbf{y}}_N; \boldsymbol{\theta}) \right], \tag{12}$$

where $\mathbf{Y}_N = (Y_1, Y_2, \ldots, Y_N)$ is an N-dimensional random variable with each Y_i as an i.i.d. copy of the random variable Y. Motivated by Poole et al. [47], we then model the term $\hat{p}_{Y|X\mathbf{Y}_N}(y|\mathbf{x}, \hat{\mathbf{y}}_N; \boldsymbol{\theta})$ in Eq. (12) as follows[5]:

$$\hat{p}_{Y|X\mathbf{Y}_N}(y|\mathbf{x}, \hat{\mathbf{y}}_N; \boldsymbol{\theta}) = \frac{p_Y(y) \exp h_{\boldsymbol{\theta}}(\mathbf{x}, y)}{\Psi_{\boldsymbol{\theta}}(\mathbf{x}, y, \hat{\mathbf{y}}_N)/(N+1)}, \tag{13}$$

where

$$\Psi_{\boldsymbol{\theta}}(\mathbf{x}, y, \hat{\mathbf{y}}_N) = \exp h_{\boldsymbol{\theta}}(\mathbf{x}, y) + \sum_{j=1}^{N} \exp h_{\boldsymbol{\theta}}(\mathbf{x}, \hat{y}_j).$$

If we set $h_{\boldsymbol{\theta}}(\mathbf{x}, y) = \mathbf{z}^\top \mathbf{r}/\tau$, $N = K + L/T$ and $\alpha = 1$, combining Eq. (12) and Eq. (13) with Eq. (6) directly results in the following:

$$\begin{aligned} E_{\boldsymbol{\theta}}(\mathbf{x}) &= \log \sum_{y \in \mathcal{Y}_{\text{I}}} \frac{\hat{p}_{Y|X}(y|\mathbf{x}; \boldsymbol{\theta})}{p_Y(y)} \\ &= \log \sum_{y \in \mathcal{Y}_{\text{I}}} \mathbb{E}_{\hat{\mathbf{y}}_{N-1} \sim \mathbb{P}_Y^{N-1}} \left[\frac{\exp h_{\boldsymbol{\theta}}(\mathbf{x}, y)}{\Psi_{\boldsymbol{\theta}}(\mathbf{x}, y, \hat{\mathbf{y}}_{N-1})} \right] + \log N \\ &\approx \log \underbrace{\sum_{y \in \mathcal{Y}_{\text{I}}} \frac{1}{T} \sum_{t=1}^{T} \frac{\exp(\mathbf{z}^\top \mathbf{r}/\tau)}{\sum_{y_j \in \mathcal{G}_t \cup \mathcal{Y}_{\text{I}}} \exp(\mathbf{z}^\top \mathbf{r}_j/\tau)}}_{S_{\text{NegLabel}}(\mathbf{x}; \boldsymbol{\theta})} + \underbrace{\log(K + L/T)}_{\text{const}}. \end{aligned} \tag{14}$$

Since the logarithm function is monotonically increasing, Eq. (14) implies that $S_{\text{NegLabel}}(\mathbf{x}; \boldsymbol{\theta})$ can be interpreted as another Monte-Carlo estimator of $E_{\boldsymbol{\theta}}(\mathbf{x})$ (up to a constant) by sampling $\hat{\mathbf{y}}_{N-1}$ from $\mathbb{P}_Y^{N-1}$ T times with $N = K + L/T$, where, for each $y \in \mathcal{Y}_{\text{I}}$, samples from $\mathcal{G}_t \cup \mathcal{Y}_{\text{I}} \setminus \{y\}$ are instantiated as $\hat{\mathbf{y}}_{N-1}$ on the t-th round of sampling ($1 \le t \le T$)[6]. Theorem 2 provides provable guarantees of how $S_{\text{NegLabel}}(\mathbf{x}; \boldsymbol{\theta})$ correctly recover $E(\mathbf{x})$ in Eq. (4).

Theorem 2. *Let $h_{\boldsymbol{\theta}}(\mathbf{x}, y) = \mathbf{z}^\top \mathbf{r}/\tau$, $\alpha = 1$, and $N = K + L/T$. If we, following following prior works [47, 45], assume that $h_{\boldsymbol{\theta}}(\mathbf{x}, y) = \log \frac{p_{Y|X}(y|\mathbf{x})}{p_Y(y)} + c(\mathbf{x})$ with $c(\mathbf{x})$ as a constant term depending on $\mathbf{x}$, then we have the following[7]:*

$$E(\mathbf{x}) = \lim_{N \to +\infty} E_{\boldsymbol{\theta}}(\mathbf{x}) = \lim_{\substack{T \to +\infty \\ L/T \to +\infty}} \log(K + L/T) S_{NegLabel}(\mathbf{x}; \boldsymbol{\theta}). \tag{15}$$

Remark. To reveal how negative labels benefit OOD detection, let us looking to Eq.(14) where the estimated energy function $E_{\boldsymbol{\theta}}(\mathbf{x})$ is upper bounded by $\log N$ with N as the number of labels drawn from $\mathbb{P}_Y$. By introducing negative labels to take $N = K + L/T$, the estimator $E_{\boldsymbol{\theta}}(\mathbf{x})$ are allowed to capture at most $\log(K + L/T)$ nats of $E(\mathbf{x})$, which is strictly larger than $\log K$ of $S_{\text{MCM}}(\mathbf{x}; \boldsymbol{\theta})$ in Eq. (9). On the other hand, Theorem 2 states that the recovery of $E(\mathbf{x})$ requires $N = K + L/T \to +\infty$, which theoretically justified the use of negative labels in OOD scoring.

[4] We detail the derivation in Appendix B

[5] We justify this formulation in Appendix C

[6] More details can be found in Step 3 of Appendix D

[7] We detail the derivation in Appendix D.

5 Methodology

Based on the thorectical analysis in Section 4, one may conclude that MCM and NegLabel can be regarded to formulate the estimation of PMI as a whole task. Differently, inspired by the *divide-and-conquer* philosophy, we conjecture that PMI estimation could be simplified as well as improved by decoupling the task into multiple earlier sub-tasks. Central to our method, we introduce an auxiliary random variable $\tilde{X} = \mathcal{T}(X) \in \mathcal{X}$, whose realization is denoted by $\tilde{\mathbf{x}}$, to represent sub-views[8] of the random variable $X \in \mathcal{X}$ with $\mathcal{T}$ as a transformation function. To be specific, given the input $\mathbf{x}$ is an image, we can create a sub-view $\tilde{\mathbf{x}}$ by randomly either 1) occluding some of the pixels in $\mathbf{x}$ with $\mathcal{T}$ as *Cutout* [8] or 2) cropping a random portion of $\mathbf{x}$ with $\mathcal{T}$ as *Random Cropping*. In the rest of this paper, we consider the latter case as the default setting.

Theorem 3. *For any $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$, let $\tilde{\mathbf{x}} = \mathcal{T}(\mathbf{x})$ be a sub-view of the input $\mathbf{x}$, $\mathcal{W}(\mathbf{x}, y)$ can be decomposed into the following two terms[9]:*

$$\mathcal{W}(\mathbf{x}, y) = \mathcal{W}(\tilde{\mathbf{x}}, y) + \mathcal{W}(\mathbf{x}, y|\tilde{\mathbf{x}}), \tag{16}$$

where $\mathcal{W}(\mathbf{x}, y|\tilde{\mathbf{x}})$, i.e., the PMI between $\mathbf{x}$ and y conditioned on $\tilde{\mathbf{x}}$, is defined as follows:

$$\begin{aligned} \mathcal{W}(\mathbf{x}, y|\tilde{\mathbf{x}}) &\triangleq \log \frac{p_{XY|\tilde{X}}(\mathbf{x}, y|\tilde{\mathbf{x}})}{p_{X|\tilde{X}}(\mathbf{x}|\tilde{\mathbf{x}}) p_{Y|\tilde{X}}(y|\tilde{\mathbf{x}})} \\ &= \log \frac{p_{Y|X\tilde{X}}(y|\mathbf{x}, \tilde{\mathbf{x}})}{p_{Y|\tilde{X}}(y|\tilde{\mathbf{x}})}. \end{aligned} \tag{17}$$

Let $\hat{\mathcal{W}}(\tilde{\mathbf{x}}, y; \boldsymbol{\theta})$ and $\hat{\mathcal{W}}(\mathbf{x}, y|\tilde{\mathbf{x}}; \boldsymbol{\theta})$ be the estimator of $\mathcal{W}(\tilde{\mathbf{x}}, y)$ and $\mathcal{W}(\mathbf{x}, y|\tilde{\mathbf{x}})$ with the pre-trained parameters $\boldsymbol{\theta}$, respectively, Theorem 3 directly implies that we can rewrite $\hat{\mathcal{W}}(\mathbf{x}, y; \boldsymbol{\theta})$ as follows:

$$\hat{\mathcal{W}}(\mathbf{x}, y; \boldsymbol{\theta}) = \hat{\mathcal{W}}(\tilde{\mathbf{x}}, y; \boldsymbol{\theta}) + \hat{\mathcal{W}}(\mathbf{x}, y|\tilde{\mathbf{x}}; \boldsymbol{\theta}), \tag{18}$$

Parameterizing $\hat{\mathcal{W}}(\tilde{\mathbf{x}}, y; \boldsymbol{\theta})$. Let $\hat{p}_{Y|\tilde{X}}(y|\tilde{\mathbf{x}}; \boldsymbol{\theta})$ denote the estimator of $p_{Y|\tilde{X}}(y|\tilde{\mathbf{x}})$, according to Eq. (5), we can parameterize $\hat{\mathcal{W}}(\tilde{\mathbf{x}}, y; \boldsymbol{\theta})$ as

$$\hat{\mathcal{W}}(\tilde{\mathbf{x}}, y; \boldsymbol{\theta}) = \log \frac{\hat{p}_{Y|\tilde{X}}(y|\tilde{\mathbf{x}}; \boldsymbol{\theta})}{p_Y(y)}, \tag{19}$$

Given $\hat{\mathbf{y}}_{N-1} = \{\hat{y}_1, ..., \hat{y}_{N-1}\} \sim \mathbb{P}_Y^{N-1}$, following Eq. (12) and Eq. (13), $\hat{p}_{Y|\tilde{X}}(y|\tilde{\mathbf{x}}; \boldsymbol{\theta})$ takes the following form:

$$\hat{p}_{Y|\tilde{X}}(y|\tilde{\mathbf{x}}; \boldsymbol{\theta}) = \mathbb{E}_{\hat{\mathbf{y}}_{N-1} \sim \mathbb{P}_Y^{N-1}} \left[\frac{p_Y(y) \exp h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}, y)}{\Psi_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}, y, \hat{\mathbf{y}}_{N-1})/N} \right]. \tag{20}$$

Similar to Eq. (14), let $N = K + L/T$ and $h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}, y) = \tilde{\mathbf{z}}^\top \mathbf{r}/\tau$ with $\tilde{\mathbf{z}} = \mathbf{f}(\tilde{\mathbf{x}}; \boldsymbol{\theta}_{\text{img}})$, we can arrive at the following Monte-Carlo estimator of $\hat{\mathcal{W}}(\tilde{\mathbf{x}}, y; \boldsymbol{\theta})$ given by

$$\begin{aligned} \hat{\mathcal{W}}(\tilde{\mathbf{x}}, y; \boldsymbol{\theta}) &= \log \mathbb{E}_{\hat{\mathbf{y}}_{N-1} \sim \mathbb{P}_Y^{N-1}} \left[\frac{N \cdot \exp h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}, y)}{\Psi(\tilde{\mathbf{x}}, y, \hat{\mathbf{y}}_{N-1})} \right] \\ &\approx \Lambda(\tilde{\mathbf{x}}, y; \boldsymbol{\theta}) \\ &\triangleq \log \sum_{y \in \mathcal{Y}_{\text{I}}} \frac{1}{T} \sum_{t=1}^{T} \frac{\exp(\tilde{\mathbf{z}}^\top \mathbf{r}/\tau)}{\sum_{y_j \in \mathcal{G}_t \cup \mathcal{Y}_{\text{I}}} \exp(\tilde{\mathbf{z}}^\top \mathbf{r}_j/\tau)} + \log(K + L/T). \end{aligned} \tag{21}$$

Parameterizing $\hat{\mathcal{W}}(\mathbf{x}, y|\tilde{\mathbf{x}}; \boldsymbol{\theta})$. Let $\hat{p}_{Y|X\tilde{X}}(y|\mathbf{x}, \tilde{\mathbf{x}}; \boldsymbol{\theta})$ be the estimator of $p_{Y|X\tilde{X}}(y|\mathbf{x}, \tilde{\mathbf{x}})$, according to Eq. (47), we can parameterize $\hat{\mathcal{W}}(\mathbf{x}, y|\tilde{\mathbf{x}}; \boldsymbol{\theta})$ as:

$$\hat{\mathcal{W}}(\mathbf{x}, y|\tilde{\mathbf{x}}; \boldsymbol{\theta}) = \log \frac{\hat{p}_{Y|X\tilde{X}}(y|\mathbf{x}, \tilde{\mathbf{x}}; \boldsymbol{\theta})}{\hat{p}_{Y|\tilde{X}}(y|\tilde{\mathbf{x}}; \boldsymbol{\theta})}. \tag{22}$$

[8] Sub-views are those derived from the original view without introducing any external information

[9] We detail the derivation in Appendix E

Under the assumption of a similar energy-based variational family to Eq. (7), we can formulate $\hat{p}_{Y|X\tilde{X}}(y|\mathbf{x},\tilde{\mathbf{x}};\boldsymbol{\theta})$ as follows:

$$\hat{p}_{Y|X\tilde{X}}(y|\mathbf{x},\tilde{\mathbf{x}};\boldsymbol{\theta}) = \frac{\hat{p}_{Y|\tilde{X}}(y|\tilde{\mathbf{x}};\boldsymbol{\theta})\exp h_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{x}},y)}{\sum_{\hat{y}\in\mathcal{Y}}\hat{p}_{Y|\tilde{X}}(\hat{y}|\tilde{\mathbf{x}};\boldsymbol{\theta})\exp h_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{x}},\hat{y})}. \tag{23}$$

Combining Eq. (23) with Eq. (22), we have the following:

$$\begin{aligned}\hat{\mathcal{W}}(\mathbf{x},y|\tilde{\mathbf{x}};\boldsymbol{\theta}) &= \log\frac{\exp h_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{x}},y)}{\sum_{\hat{y}\in\mathcal{Y}}\hat{p}_{Y|\tilde{X}}(\hat{y}|\tilde{\mathbf{x}};\boldsymbol{\theta})\exp h_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{x}},\hat{y})}\\ &= \log\frac{\exp h_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{x}},y)}{\mathbb{E}_{\hat{y}\sim\mathbb{P}_Y}\left[\eta(\tilde{\mathbf{x}},\hat{y})\exp h_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{x}},\hat{y})\right]}\\ &\approx \log\frac{\exp h_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{x}},y)}{\sum_{j=1}^{K+L}\eta(\tilde{\mathbf{x}},\hat{y})\exp h_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{x}},y_j)} + \log(K+L),\end{aligned} \tag{24}$$

where $\eta(\tilde{\mathbf{x}},\hat{y}) = \hat{p}_{Y|\tilde{X}}(\hat{y}|\tilde{\mathbf{x}};\boldsymbol{\theta})/p_Y(\hat{y})$. We note that it is suffice to follow Eq. (6) to derive the last step of Eq. (24), where, as suggested by Theorem 1, both negative labels and ID labels are leveraged for the Monte-Carlo estimation of the expectation. Recalling that, according to Eq. (19), $\eta(\tilde{\mathbf{x}},\hat{y}) = \exp\hat{\mathcal{W}}(\tilde{\mathbf{x}},\hat{y};\boldsymbol{\theta})$, connecting Eq. (24) to Eq. (18) results in reformulating the estimated energy function $E_{\boldsymbol{\theta}}(\mathbf{x})$ in Eq. (6) with $\alpha = 1$ as follows:

$$\begin{aligned}E_{\boldsymbol{\theta}}(\mathbf{x}) &= \log\sum_{y\in\mathcal{Y}_{\text{I}}}\exp\left[\hat{\mathcal{W}}(\tilde{\mathbf{x}},y;\boldsymbol{\theta}) + \hat{\mathcal{W}}(\mathbf{x},y|\tilde{\mathbf{x}};\boldsymbol{\theta})\right]\\ &= \log\sum_{y\in\mathcal{Y}_{\text{I}}}\frac{\exp\left[\hat{\mathcal{W}}(\tilde{\mathbf{x}},y;\boldsymbol{\theta}) + h_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{x}},y)\right]}{\mathbb{E}_{\hat{y}\sim\mathbb{P}_Y}\left[\exp\left[\hat{\mathcal{W}}(\tilde{\mathbf{x}},\hat{y};\boldsymbol{\theta}) + h_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{x}},\hat{y})\right]\right]}\\ &\approx \log S_{\text{ours}}(\mathbf{x};\boldsymbol{\theta}) + \log(K+L),\end{aligned} \tag{25}$$

where, inspired by Tsai et al. [54], we define $h_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{x}},y) \triangleq \mathbf{r}^\top[\beta\tilde{\mathbf{z}} + (1-\beta)\mathbf{z}]/\kappa$ with $\beta\in(0,1)$ and $\kappa > 0$ as two hyper-parameters, and

$$S_{\text{ours}}(\mathbf{x};\boldsymbol{\theta}) \triangleq \sum_{y\in\mathcal{Y}_{\text{I}}}\frac{\exp\left[\Lambda(\tilde{\mathbf{x}},y;\boldsymbol{\theta}) + h_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{x}},y)\right]}{\sum_{j=1}^{K+L}\exp\left[\Lambda(\tilde{\mathbf{x}},y_j;\boldsymbol{\theta}) + h_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{x}},y_j)\right]}. \tag{26}$$

Similarly, we present the following theorem to reveal the provable guarantee of how $S_{\text{ours}}(\mathbf{x};\boldsymbol{\theta})$ correctly recovers the true energy function $E(\mathbf{x})$ in Eq. (4).

Theorem 4. *Let $h_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{x}},y) \triangleq \mathbf{r}^\top[\beta\tilde{\mathbf{z}}+(1-\beta)\mathbf{z}]/\kappa$, $h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}},y) = \tilde{\mathbf{z}}^\top\mathbf{r}/\tau$, $\alpha = 1$ and $N = L + K/T$. If we, following prior works [47, 36], assume that $h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}},y) = \log\frac{p_{Y|\tilde{X}}(y|\tilde{\mathbf{x}})}{p_Y(y)} + c(\tilde{\mathbf{x}})$ with $c(\tilde{\mathbf{x}})$ as a constant term depending on $\tilde{\mathbf{x}}$, and that $h_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{x}},y) = \log\frac{p_{Y|X\tilde{X}}(y|\mathbf{x},\tilde{\mathbf{x}})}{p_{Y|\tilde{X}}(y|\tilde{\mathbf{x}})} + c(\mathbf{x},\tilde{\mathbf{x}})$ with $c(\mathbf{x},\tilde{\mathbf{x}})$ as a constant term depending on $\mathbf{x}$ and $\tilde{\mathbf{x}}$, then we have the following[10]:*

$$E(\mathbf{x}) = \lim_{N\to+\infty} E_{\boldsymbol{\theta}}(\mathbf{x}) = \lim_{\substack{T\to+\infty\\ L/T\to+\infty}}\log(K+L)S_{\textit{ours}}(\mathbf{x};\boldsymbol{\theta}). \tag{27}$$

Remark. Comparing Eq. (25) with Eq. (10) and Eq. (14), one can find that $S_{\text{ours}}(\mathbf{x};\boldsymbol{\theta})$ capture at most $\log(K+L)$ nats of the true $E(\mathbf{x})$, which is strictly larger than $\log K$ in $S_{\text{MCM}}(\mathbf{x};\boldsymbol{\theta})$ and $\log(K+L/T)$ in $S_{\text{Neglabel}}(\mathbf{x};\boldsymbol{\theta})$. Although this upper bound, i.e., $\log(K+L)$, can be achieved by $S_{\text{Neglabel}}(\mathbf{x};\boldsymbol{\theta})$ in Eq. (14) by either 1) introducing TL negative labels or 2) fixing $T = 1$, we note that 1) NegLabel [24] has empirically observed the degeneration of OOD detection performance caused by excessive negative labels, and that 2) decreasing T can be in conflict with Theorem 2 where the recovery of $E(\mathbf{x})$ with $S_{\text{NegLabel}}(\mathbf{x};\boldsymbol{\theta})$ explicitly requires T to be sufficiently large.

[10] We detail the derivation in Appendix F

Table 1: OOD detection results on the ImageNet-1K dataset. ↑ indicates larger values are better and vice versa. The best results in the last two columns are shown in bold.

Method	iNaturalist		SUN		Places		Textures		Average	
	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
KNN	94.52	29.17	92.67	35.62	91.02	39.61	85.67	64.35	90.97	42.19
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	88.90	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
NegPro	98.73	6.32	95.55	22.89	93.34	27.60	91.60	35.21	94.81	23.01
HFTT	93.27	27.44	95.28	19.24	90.26	43.54	88.23	43.08	91.76	33.33
Zero-Shot Training-free Methods										
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
NegLabel	99.49	1.91	95.49	20.53	91.64	35.59	90.22	43.56	94.21	25.40
Ours (Median)	99.70	1.04	96.16	16.06	93.37	26.92	91.01	40.78	**95.07**	**21.20**
Ours (Mean)	99.64	1.04	96.32	18.45	95.81	31.15	92.15	38.79	**96.00**	**22.36**

6 Experiments

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.

Baseline Methods. We compare our method with MSP [18], ODIN [32], Energy [34], KNN [52], Gradnorm [22], Vim [57], VOS [10], NPOS [53], ZOC [13], CLIPN [59], LoCoOp [41], LSN [44], LAPT [67], NegPro [31], HFTT [30], MCM [40], NegLabel [24] and AdaNeg [66].

Implementation Details. Unless otherwise specified, we employ the CLIP ViT-B/16 model as the pre-trained VLM. We use the same NegMining algorithm as NegLabel [24] to extract top 15% dissimilar words to ID labels from WordNet as negative labels, followed by separating the negative labels into $T = 10$ groups for OOD scoring. Following NegLabel [24], we adopt the text prompt of 'The nice <label>.'. We apply the random cropping augmentation on each test-time image $\mathbf{x}$ with the scale range $(\lambda, 1.0)$ to produce the sub-view $\tilde{\mathbf{x}}$, followed by resizing $\tilde{\mathbf{x}}$ to 224×224. Regarding hyper-parameters, we set $\tau = 0.02$, $\kappa = 0.08$, $\lambda = 0.55$, $\alpha = 0.8$ and $\beta = 0.3$. All experiments are conducted with a single Tesla A100 GPU.

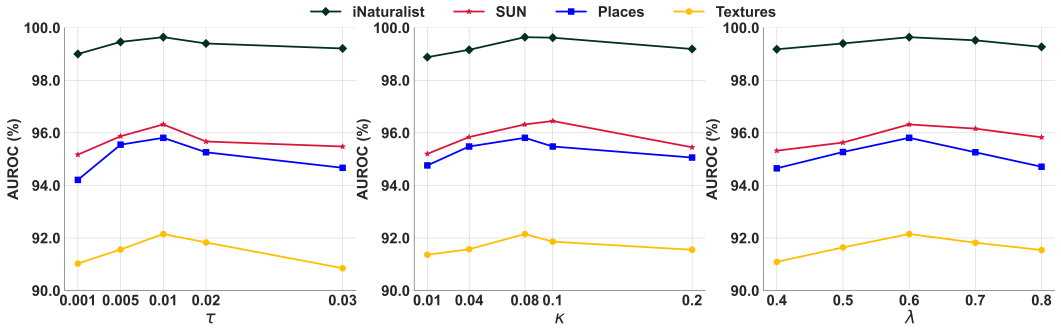

Figure 1: Ablation study on ImageNet-1K w.r.t hyper-parameters τ (left), κ (middle) and λ (right)

Table 2: Evaluation on domain-generalizable OOD detection. ↑ indicates larger values are better and vice versa. The best results in the last two columns are shown in bold per ID dataset.

ID Dataset	Method	iNaturalist		SUN		Places		Textures		Average	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
ImageNet-S	MCM	87.74	63.06	85.35	67.24	81.19	70.64	74.77	79.59	82.26	70.13
	NegLabel	99.34	2.24	94.93	22.73	90.78	38.62	89.29	46.10	93.59	27.42
	Ours	99.51	1.62	96.02	20.17	93.89	34.69	91.35	42.94	**95.52**	**25.11**
ImageNet-A	MCM	79.50	76.85	76.19	79.78	70.95	80.51	61.98	86.37	80.88	72.16
	NegLabel	98.80	4.09	89.83	44.38	82.88	60.10	80.25	64.34	87.94	43.23
	Ours	99.16	3.58	91.64	39.63	86.25	55.64	87.43	58.76	**91.23**	**39.40**

6.1 Main Results

We conduct experiments on the ImageNet dataset, demonstrating the scalability of our method. Specifically, we inherit the setup from prior work [40, 24, 66], where the ID dataset is ImageNet-1K [7] and OOD datasets include iNaturalist [55], SUN [61], Places365 [70], and Textures [5]. At test time, all images are resized to 224×224. Table 1 presents the performance of our approach and existing competitive baselines, where the proposed approach significantly outperforms existing methods. Specifically, advanced post-hoc methods generally perform better than learning-based methods especially when the SUN dataset acts as the OOD data without requiring additional training. Besides, compared with the state-of-the-art NegLabel, our method reveals 3.16% and 2.21% averaged improvement w.r.t FPR95 and AUROC on the ImageNet dataset. For advanced works, i.e., NegLabel+AdaNeg, that additionally consider visual negative proxies in OOD scoring, our improved version, i.e., Ours+AdaNeg, performs better on all four OOD datasets.

6.2 Ablation Study

We analyze the hyper-parameters most essential to our algorithmic design, including the minimum crop scale λ and two scaling temperatures τ and κ. The corresponding results are plotted in Figure 1. On the one hand, having a large or small value of the two scaling temperatures does not necessarily improve the OOD detection performance while our method consistently outperforms the state-of-the-art NegLabel when the value of τ and κ varies from 0.05 to 0.02 and from 0.04 to 0.1 respectively. On the other hand, it can be found that the aggressive cropping strategy, which corresponding to that the value λ is small, can deteriorate the OOD detection. We suspect that this is because aggressive cropping may hurt the semantics of the original image.

6.3 Extensions

Domain-generalizable OOD Detection. We consider domain generalizable OOD detection scenarios, where domain shifts occur in ID data. With ImageNet-1K as the ID data, we, following NegLabel [24], consider ImageNet-S [56] and ImageNet-A [21] as ID data receptively. The performance gain in Table 4 implies the more robustness of our method to domain shift.

Table 3: OOD detection results on the ImageNet-1K with various learned prompts, i.e., NegPro [31] and LAPT [67], respectively. Following Zhang & Zhang [66], the performance is measured by FPR95 ↓. The best results are shown in bold.

Method	iNaturalist		SUN		Places		Textures		Average	
	NegPro	LAPT	NegPro	LAPT	NegPro	LAPT	NegPro	LAPT	NegPro	LAPT
NegLabel+AdaNeg	3.87	0.58	11.35	9.98	25.45	30.47	29.79	25.25	17.62	16.32
Ours+AdaNeg	4.16	0.63	9.47	8.39	23.79	25.64	26.42	25.76	**15.96**	**15.11**

OOD Detection with Learned Prompt. While this paper, following Neglabel [24], to use a predefined prompts for ID label, we show that our method can be made stronger with the mostly recent technology of prompt learning. Empirically, we compare the results in Table 5 by using the prompts learned by either Negpro [31] or LAPT [67].

7 Conclusion

This paper presents a information-theoretic framework to characterizes and unifies the theoretical understanding of post-hoc OOD detection with pre-trained VLMs. In particular, by modeling the ID data with an energy-based model with the PMI between the input image and each ID label as energy functions, We show that representative CLIP-based post-hoc OOD detection methods implicitly work as stochastic Monte Carlo estimations of PMI for density estimation. Motivated by the *divide-and-conquer* philosophy, we decompose the original PMI into a sum of conditional and unconditional PMI terms to facilitate OOD detection, which demonstrates both theoretical and empirical superiority.

Acknowledgement

This work is supported by Australian Research Council Discovery Early Career Researcher Award (DE250100363) and Australian Laureate Fellowship (FL190100149).

References

[1] Mouïn Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. Neco: Neural collapse based out-of-distribution detection. In *The Twelfth International Conference on Learning Representations*, 2024.

[2] 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.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

[4] Kenneth Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[5] 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.

[6] Chris Cremer, Quaid Morris, and David Duvenaud. Reinterpreting importance-weighted autoencoders. *arXiv preprint arXiv:1704.02916*, 2017.

[7] 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.

[8] Terrance DeVries. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[9] 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.

[10] 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.

[11] Xuefeng Du, Zhen Fang, Ilias Diakonikolas, and Yixuan Li. How does wild data provably help ood detection? In *The Twelfth International Conference on Learning Representations*, 2024.

[12] Xuefeng Du, Yiyou Sun, and Yixuan Li. When and how does in-distribution label help out-of-distribution detection? *arXiv preprint arXiv:2405.18635*, 2024.

[13] 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.

[14] Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? In *NeurIPS*, 2022.

[15] Zhen Fang, Yixuan Li, Feng Liu, Bo Han, and Jie Lu. On the learnability of out-of-distribution detection. *Journal of Machine Learning Research*, 25, 2024.

[16] Ruiyuan Gao, Chenchen Zhao, Lanqing Hong, and Qiang Xu. Diffguard: Semantic mismatch-guided out-of-distribution detection using pre-trained diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1579–1589, 2023.

[17] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

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

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

[20] 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.

[21] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021.

[22] 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.

[23] Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37:100270, 2020.

[24] 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.

[25] 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.

[26] 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.

[27] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[28] 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.

[29] 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.

[30] Saehyung Lee, Jisoo Mok, Sangha Park, Yongho Shin, Dahuin Jung, and Sungroh Yoon. Textual training for the hassle-free removal of unwanted visual data: case studies on ood and hateful image detection. *Advances in Neural Information Processing Systems*, 37:125312–125335, 2024.

[31] 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.

[32] 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.

[33] Luping Liu, Yi Ren, Xize Cheng, and Zhou Zhao. Out-of-distribution detection with diffusion-based neighborhood.

[34] 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.

[35] Xixi Liu, Yaroslava Lochman, and Christopher Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23946–23955, 2023.

[36] Zhuang Ma and Michael Collins. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. *arXiv preprint arXiv:1809.01812*, 2018.

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

[38] 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.

[39] 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.

[40] 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.

[41] 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, 2024.

[42] 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.

[43] 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.

[44] 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.

[45] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[46] 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.

[47] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pp. 5171–5180. PMLR, 2019.

[48] 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.

[49] 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.

[50] 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.

[51] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pp. 5628–5637. PMLR, 2019.

[52] 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, 2022.

[53] Leitian Tao, Xuefeng Du, Xiaojin Zhu, and Yixuan Li. Non-parametric outlier synthesis. *arXiv preprint arXiv:2303.02966*, 2023.

[54] Tsung Wei Tsai, Chongxuan Li, and Jun Zhu. Mice: Mixture of contrastive experts for unsupervised image clustering. In *International conference on learning representations*, 2020.

[55] 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.

[56] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

[57] 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.

[58] 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.

[59] 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.

[60] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

[61] 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.

[62] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18134–18144, 2022.

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

[64] Yijun Yang, Ruiyuan Gao, and Qiang Xu. Out-of-distribution detection with semantic mismatch under masking. In *European Conference on Computer Vision*, pp. 373–390. Springer, 2022.

[65] Yeonguk Yu, Sungho Shin, Seongju Lee, Changhyun Jun, and Kyoobin Lee. Block selection method for using feature norm in out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15701–15711, 2023.

[66] Yabin Zhang and Lei Zhang. Adaneg: Adaptive negative proxy guided ood detection with vision-language models. *arXiv preprint arXiv:2410.20149*, 2024.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] David Zimmerer, Peter M Full, Fabian Isensee, Paul Jäger, Tim Adler, Jens Petersen, Gregor Köhler, Tobias Ross, Annika Reinke, Antanas Kascenas, et al. Mood 2020: A public benchmark for out-of-distribution detection and localization on medical images. *IEEE Transactions on Medical Imaging*, 41(10):2728–2738, 2022.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We clearly illustrate the focused problems and state our contribution methodologically, theoretically and empirically in the abstract and introduction.n

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitation of our method in the appendix.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: assumptions are given in the main content and see appendix for full proofs.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provide implementation details of our method in the Experiment section

   Guidelines:

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

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

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

Answer: [No]

Justification: We will release our code upon acceptance.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We provide implementation details of our method in the Experiment section.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We have provided the standard deviation of our method in the appendix.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We have included the number and type of used GPU in the Experiment section.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: I have read the ethics review guidelines before conducting research.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We have discussed the societal impact of our method in the appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: This paper does not involve the use of any pretrained language models, image generators, or scraped datasets.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: we have cited the original paper that produced the code package or dataset and the used datasets in this paper are properly licensed.

    Guidelines:

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

13. **New assets**

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

Answer: [NA]

Justification: The datasets and models used in this paper are open-source.

Guidelines:

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

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

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

Answer: [NA]

Justification: The datasets are not with crowdsourcing nor research with human subjects.

Guidelines:

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

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

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

Answer: [NA]

Justification: The datasets are not with crowdsourcing nor research with human subjects.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [No]

Justification: No LLM is used in this paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Related Works on Traditional OOD Detection

On the theoretical side, there are various attempts to explore the theoretical understanding of OOD detection. Fang et al. [14, 15] study the generalization of OOD detection by PAC learning and find a necessary condition for the learnability of OOD detection. Morteza & Li [42] provides a provable understanding of the OOD detection result by modelling the feature embedding space as a mixture of multivariate Gaussian distributions. Du et al. [12] studies the impact of ID labels on OOD detection.

On the practical side, the popularity of OOD detection is motivated by the empirical observation [43] 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 [35, 20, 68], energy-based methods [34, 58], distance-based approaches [52, 29, 50, 65, 1], gradient-based approaches [22], generative approaches [64, 16, 33], and Bayesian approaches [38]. 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 [19]. For instance, the discriminative model is regularized to produce lower confidence [28, 37] or higher energy [34] for outlier points. More recently, some works consider a more practical but challenging scenario where auxiliary outliers are contaminated with unlabelled ID counterparts. WOOD [25] formulates learning with noisy OOD data as a constrained optimization problem while SAL [11] separates candidate outliers from the unlabeled data and trains a binary classifier using the candidate outliers and the labelled ID data.

B Proof of Theorem 1

As a reminder, Theorem 1 is stated as follows:

Theorem 1. *Let $h_{\boldsymbol{\theta}}(\mathbf{x}, y) = \mathbf{z}^\top \mathbf{r}/\tau$ and $N = K$. If we, following prior works [47, 45], assume that $h_{\boldsymbol{\theta}}(\mathbf{x}, y) = \log \frac{p_{Y|X}(y|\mathbf{x})}{p_Y(y)} + c(\mathbf{x})$ with $c(\mathbf{x})$ as a constant term depending on $\mathbf{x}$, in the extreme case where $\alpha \to +\infty$, we then have the following:*

$$\lim_{\alpha\to+\infty} E(\mathbf{x}) = \lim_{\alpha\to+\infty} E_{\boldsymbol{\theta}}(\mathbf{x}) = \lim_{N\to+\infty} \log NS_{MCM}(\mathbf{x};\boldsymbol{\theta}). \tag{28}$$

Proof. **Step 1:**

$$\begin{aligned}
\lim_{\alpha\to+\infty} E_{\boldsymbol{\theta}}(\mathbf{x}) &= \lim_{\alpha\to+\infty} \frac{1}{\alpha} \log \sum_{y\in\mathcal{Y}_{\text{I}}} \exp\left[\alpha \log \frac{\hat{p}_{Y|X}(y|\mathbf{x};\boldsymbol{\theta})}{p_Y(y)}\right] \\
&= \lim_{\alpha\to+\infty} \frac{1}{\alpha} \log \sum_{y\in\mathcal{Y}_{\text{I}}} \exp\left[\alpha \log \frac{\exp h_{\boldsymbol{\theta}}(\mathbf{x}, y)}{\mathbb{E}_{\hat{y}\sim\mathbb{P}_Y}\left[\exp h_{\boldsymbol{\theta}}(\mathbf{x}, \hat{y})\right]}\right] \\
&= \lim_{\alpha\to+\infty} \frac{1}{\alpha} \log \sum_{y\in\mathcal{Y}_{\text{I}}} \exp\left[\alpha \log \frac{\frac{p_{Y|X}(y|\mathbf{x})}{p_Y(y)} \exp c(\mathbf{x})}{\mathbb{E}_{\hat{y}\sim\mathbb{P}_Y}\left[\frac{p_{Y|X}(\hat{y}|\mathbf{x})}{p_Y(\hat{y})} \exp c(\mathbf{x})\right]}\right] \\
&= \lim_{\alpha\to+\infty} \frac{1}{\alpha} \log \sum_{y\in\mathcal{Y}_{\text{I}}} \exp\left[\alpha \log \frac{p_{Y|X}(y|\mathbf{x})}{p_Y(y)}\right] = \lim_{\alpha\to+\infty} E(\mathbf{x}),
\end{aligned} \tag{29}$$

where the penultimate step of Eq. (29) is derived based on the fact that

$$\mathbb{E}_{\hat{y}\sim\mathbb{P}_Y}\left[\frac{p_{Y|X}(y|\mathbf{x})}{p_Y(\hat{y})}\right] = \sum_{\hat{y}\in\mathcal{Y}} p_Y(\hat{y}) \frac{p_{Y|X}(\hat{y}|\mathbf{x})}{p_Y(\hat{y})} = \sum_{\hat{y}\in\mathcal{Y}} p_{Y|X}(\hat{y}|\mathbf{x}) = 1. \tag{30}$$

Step 2:

Given that $h_{\boldsymbol{\theta}}(\mathbf{x}, y) = \mathbf{z}^\top \mathbf{r}_j/\tau$, and that, as implied by the law of large numbers, $\lim_{N\to+\infty} \frac{1}{N}\sum_{j=1}^{N} \exp h_{\boldsymbol{\theta}}(\mathbf{x}, y_j) = \mathbb{E}_{\hat{y}\sim\mathbb{P}_Y}\left[\exp h_{\boldsymbol{\theta}}(\mathbf{x}, \hat{y})\right]$ (for all $y_j \in \mathcal{Y}$), we have the following:

$$
\begin{aligned}
\lim_{N\to+\infty} \log N S_{\text{MCM}}(\mathbf{x};\boldsymbol{\theta}) &= \lim_{N\to+\infty} \log \max_{y\in\mathcal{Y}_{\text{I}}} \frac{N\exp(\mathbf{z}^\top\mathbf{r}/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{z}^\top\mathbf{r}_j/\tau)} \\
&= \log \max_{y\in\mathcal{Y}_{\text{I}}} \exp(\mathbf{z}^\top\mathbf{r}/\tau) - \lim_{N\to+\infty} \log \frac{1}{N}\sum_{j=1}^{N}\exp(\mathbf{z}^\top\mathbf{r}_j/\tau) \\
&= \log \max_{y\in\mathcal{Y}_{\text{I}}} \exp(\mathbf{z}^\top\mathbf{r}/\tau) - \log \lim_{N\to+\infty} \frac{1}{N}\sum_{j=1}^{N}\exp(\mathbf{z}^\top\mathbf{r}_j/\tau) \\
&= \log \max_{y\in\mathcal{Y}_{\text{I}}} \exp(\mathbf{z}^\top\mathbf{r}/\tau) - \log \mathbb{E}_{\hat{y}\sim\mathbb{P}_Y}\left[\exp(\mathbf{z}^\top\mathbf{r}/\tau)\right] \\
&= \log \max_{y\in\mathcal{Y}_{\text{I}}} \frac{\exp(\mathbf{z}^\top\mathbf{r}/\tau)}{\mathbb{E}_{\hat{y}\sim\mathbb{P}_Y}\left[\exp(\mathbf{z}^\top\mathbf{r}/\tau)\right]} \\
&= \max_{y\in\mathcal{Y}_{\text{I}}} \log \frac{\exp(\mathbf{z}^\top\mathbf{r}/\tau)}{\mathbb{E}_{\hat{y}\sim\mathbb{P}_Y}\left[\exp(\mathbf{z}^\top\mathbf{r}/\tau)\right]} \\
&= \lim_{\alpha\to+\infty} \frac{1}{\alpha} \log \sum_{y\in\mathcal{Y}_{\text{I}}} \exp\left[\alpha \log \frac{\exp(\mathbf{z}^\top\mathbf{r}/\tau)}{\mathbb{E}_{\hat{y}\sim\mathbb{P}_Y}\left[\exp(\mathbf{z}^\top\mathbf{r}/\tau)\right]}\right] \\
&= \lim_{\alpha\to+\infty} \frac{1}{\alpha} \log \sum_{y\in\mathcal{Y}_{\text{I}}} \exp\left[\alpha \log \frac{\exp h_{\boldsymbol{\theta}}(\mathbf{x}, y)}{\mathbb{E}_{\hat{y}\sim\mathbb{P}_Y}\left[\exp h_{\boldsymbol{\theta}}(\mathbf{x}, \hat{y})\right]}\right] \\
&= \lim_{\alpha\to+\infty} E_{\boldsymbol{\theta}}(\mathbf{x})
\end{aligned}
\tag{31}
$$

Step 1 and Step 2 imply this result.

□

C Justification of $\hat{p}_{Y|X\mathbf{Y}_N}(y|\mathbf{x}, \hat{\mathbf{y}}_N; \boldsymbol{\theta})$

As a reminer, $\hat{p}_{Y|X\mathbf{Y}_N}(y|\mathbf{x}, \hat{\mathbf{y}}_N; \boldsymbol{\theta})$ is given as follows:

$$
\hat{p}_{Y|X\mathbf{Y}_N}(y|\mathbf{x}, \hat{\mathbf{y}}_N; \boldsymbol{\theta}) = \frac{p_Y(y)\exp h_{\boldsymbol{\theta}}(\mathbf{x}, y)}{\exp h_{\boldsymbol{\theta}}(\mathbf{x}, y) + \sum_{j=1}^{N}\exp h_{\boldsymbol{\theta}}(\mathbf{x}, \hat{y}_j)}, \tag{32}
$$

$$
\begin{aligned}
\sum_{y\in\mathcal{Y}}\hat{p}_{Y|\tilde{X}}(y|\tilde{\mathbf{x}};\boldsymbol{\theta}) &= \sum_{y\in\mathcal{Y}}\mathbb{E}_{\hat{\mathbf{y}}_{N-1}\sim\mathbb{P}_Y^{N-1}}\left[\hat{p}_{Y|X\mathbf{Y}_{N-1}}(y|\mathbf{x},\hat{\mathbf{y}}_{N-1};\boldsymbol{\theta})\right] \\
&= \sum_{y\in\mathcal{Y}}\mathbb{E}_{\hat{\mathbf{y}}_{N-1}\sim\mathbb{P}_Y^{N-1}}\left[N\frac{p_Y(y)\exp h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}},y)}{\exp h_{\boldsymbol{\theta}}(\mathbf{x},y)+\sum_{j=1}^{N-1}\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y}_j)}\right] \\
&= \sum_{y\in\mathcal{Y}}p_Y(y)\mathbb{E}_{\hat{\mathbf{y}}_{N-1}\sim\mathbb{P}_Y^{N-1}}\left[N\frac{\exp h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}},y)}{\exp h_{\boldsymbol{\theta}}(\mathbf{x},y)+\sum_{j=1}^{N-1}\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y}_j)}\right] \\
&= \mathbb{E}_{y\sim\mathbb{P}_Y}\mathbb{E}_{\hat{\mathbf{y}}_{N-1}\sim\mathbb{P}_Y^{N-1}}\left[N\frac{\exp h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}},y)}{\exp h_{\boldsymbol{\theta}}(\mathbf{x},y)+\sum_{j=1}^{N-1}\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y}_j)}\right] \\
&= N\mathbb{E}_{\hat{\mathbf{y}}_N\sim\mathbb{P}_Y^N}\left[\frac{\exp h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}},\hat{y}_N)}{\sum_{j=1}^{N}\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y}_j)}\right] \\
&= \sum_{i=1}^{N}\mathbb{E}_{\hat{\mathbf{y}}_N\sim\mathbb{P}_Y^N}\left[\frac{\exp h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}},\hat{y}_i)}{\sum_{j=1}^{N}\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y}_j)}\right] \\
&= \mathbb{E}_{\hat{\mathbf{y}}_N\sim\mathbb{P}_Y^N}\left[\frac{\sum_{i=1}^{N}\exp h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}},\hat{y}_i)}{\sum_{j=1}^{N}\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y}_j)}\right]=1
\end{aligned}
\tag{33}
$$

Please refer to Section 5.2 in Cremer et al. [6] for more details of the derivation.

D Proof of Theorem 2

As a reminder, Theorem 2 is stated as follows:

Theorem 2. *Let $h_{\boldsymbol{\theta}}(\mathbf{x},y)=\mathbf{z}^\top\mathbf{r}/\tau$, $\alpha=1$, and $N=K+L/T$. If we, following prior works [47, 45], assume that $h_{\boldsymbol{\theta}}(\mathbf{x},y)=\log\frac{p_{Y|X}(y|\mathbf{x})}{p_Y(y)}+c(\mathbf{x})$ with $c(\mathbf{x})$ as a constant term depending on $\mathbf{x}$, we then have the following:*

$$
E(\mathbf{x})=\lim_{N\to+\infty}E_{\boldsymbol{\theta}}(\mathbf{x})=\lim_{\substack{T\to+\infty\\ L/T\to+\infty}}\log(K+L/T)S_{\textit{NegLabel}}(\mathbf{x};\boldsymbol{\theta}). \tag{34}
$$

Proof. Given that $\alpha=1$, we have

$$
E(\mathbf{x})=\log\sum_{y\in\mathcal{Y}_{\text{I}}}\exp\left[\mathcal{W}(x,y)\right]=\log\sum_{y\in\mathcal{Y}_{\text{I}}}\exp\left[\log\frac{p_{Y|X}(y|\mathbf{x})}{p_Y(y)}\right]=\log\sum_{y\in\mathcal{Y}_{\text{I}}}\frac{p_{Y|X}(y|\mathbf{x})}{p_Y(y)}. \tag{35}
$$

$$
E_{\boldsymbol{\theta}}(\mathbf{x})=\log\sum_{y\in\mathcal{Y}_{\text{I}}}\exp\left[\log\hat{\mathcal{W}}(\mathbf{x},y)\right]=\log\sum_{y\in\mathcal{Y}_{\text{I}}}\frac{\hat{p}_{Y|X}(y|\mathbf{x};\boldsymbol{\theta})}{p_Y(y)} \tag{36}
$$

Step 1:

Given that $h_{\boldsymbol{\theta}}(\mathbf{x},y)=\log\frac{p_{Y|X}(y|\mathbf{x})}{p_Y(y)}+c(\mathbf{x})$, when $N\to+\infty$, the Law of Large Numbers implies the following

$$
\begin{aligned}
&\lim_{N\to+\infty}\frac{1}{N}\left[\exp h_{\boldsymbol{\theta}}(\mathbf{x},y)+\sum_{j=1}^{N-1}\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y}_j)\right] \\
=&\lim_{N\to+\infty}\frac{1}{N}\exp h_{\boldsymbol{\theta}}(\mathbf{x},y)+\lim_{N\to+\infty}\frac{N-1}{N}\lim_{N\to+\infty}\frac{1}{N-1}\sum_{j=1}^{N-1}\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y}_j) \\
=&\mathbb{E}_{\hat{y}\sim\mathbb{P}_Y}\left[\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y})\right] \\
=&\mathbb{E}_{\hat{y}\sim\mathbb{P}_Y}\left[\frac{p_{Y|X}(\hat{y}|\mathbf{x})}{p_Y(\hat{y})}\exp c(\mathbf{x})\right]=\exp c(\mathbf{x})
\end{aligned}
\tag{37}
$$

Step 2:

This implies that

$$\begin{aligned}\lim_{N\to+\infty} E_{\boldsymbol{\theta}}(\mathbf{x}) &= \lim_{N\to+\infty} \log \sum_{y\in\mathcal{Y}_{\mathrm{I}}} \mathbb{E}_{\hat{\mathbf{y}}_{N-1}\sim\mathbb{P}_Y^{N-1}} \left[\frac{N\cdot \exp h_{\boldsymbol{\theta}}(\mathbf{x},y)}{\exp h_{\boldsymbol{\theta}}(\mathbf{x},y)+\sum_{j=1}^{N-1}\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y}_j)}\right] \\ &= \lim_{N\to+\infty} \log \sum_{y\in\mathcal{Y}_{\mathrm{I}}} \mathbb{E}_{\hat{\mathbf{y}}_{N-1}\sim\mathbb{P}_Y^{N-1}} \left[\frac{p_{Y|X}(y|\mathbf{x})}{p_Y(y)}\right] \\ &= \log \sum_{y\in\mathcal{Y}_{\mathrm{I}}} \frac{p_{Y|X}(y|\mathbf{x})}{p_Y(y)} = E(\mathbf{x})\end{aligned} \tag{38}$$

Step 3: Let $\hat{y}_i^{(t)}$ be the i-th sample drawn at the t-th round, the law of large numbers implies

$$\begin{aligned}\frac{\hat{p}_{Y|X}(y|\mathbf{x};\boldsymbol{\theta})}{p_Y(y)} &= \mathbb{E}_{\hat{\mathbf{y}}_{N-1}\sim\mathbb{P}_Y^{N-1}} \left[\frac{N\cdot \exp h_{\boldsymbol{\theta}}(\mathbf{x},y)}{\exp h_{\boldsymbol{\theta}}(\mathbf{x},y)+\sum_{j=1}^{N-1}\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y}_j)}\right] \\ &= \lim_{T\to+\infty} \frac{1}{T}\sum_{i=1}^{T} \frac{N\cdot \exp h_{\boldsymbol{\theta}}(\mathbf{x},y)}{\exp h_{\boldsymbol{\theta}}(\mathbf{x},y)+\sum_{j=1}^{N-1}\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y}_j^{(t)})} \\ &= \lim_{T\to+\infty} \frac{N}{T}\sum_{i=1}^{T} \hat{L}(x,y,t),\end{aligned} \tag{39}$$

where, for each $t=1,\ldots,T$,

$$\hat{L}(x,y,t) = \frac{\exp h_{\boldsymbol{\theta}}(\mathbf{x},y)}{\exp h_{\boldsymbol{\theta}}(\mathbf{x},y)+\sum_{j=1}^{N-1}\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y}_j^{(t)})} \tag{40}$$

As $\hat{\mathbf{y}}_{N-1}^{(t)} = (\hat{y}_i^{(t)})_{i=1}^{N-1}$ is sampled from $\mathbb{P}_Y^{N-1}$ i.i.d, we can build $\hat{L}(x,y,t)$ for any $\hat{y}_i^t \in \mathcal{Y}$. If we consider a valid case where each label in $\mathcal{G}_t \cup \mathcal{Y}_I \setminus \{y\}$ is exactly sampled to constitute $\hat{\mathbf{y}}_{N-1}^{(t)}$ when calculating $\hat{L}(x,y,t)$ in Eq. (39), we have $N = K + L/T$ such that $N\to+\infty$ requires $L/T\to+\infty$, which results in rewriting Eq. (39) as follows:

$$\begin{aligned}\lim_{N\to+\infty} E_{\boldsymbol{\theta}}(\mathbf{x}) &= \lim_{\substack{T\to+\infty\\ N\to+\infty}} \log \sum_{y\in\mathcal{Y}_{\mathrm{I}}} \frac{N}{T}\sum_{i=1}^{T} \hat{L}(x,y,t) \\ &= \lim_{\substack{T\to+\infty\\ N\to+\infty}} \log \sum_{y\in\mathcal{Y}_{\mathrm{I}}} \frac{K+L/T}{T}\sum_{i=1}^{T} \frac{\exp h_{\boldsymbol{\theta}}(\mathbf{x},y)}{\sum_{\hat{y}_j\in\mathcal{G}_t\cup\mathcal{Y}_I}\exp h_{\boldsymbol{\theta}}(\mathbf{x},\hat{y}_j)} \\ &= \lim_{\substack{T\to+\infty\\ L/T\to+\infty}} \log(K+L/T) S_{\text{NegLabel}}(\mathbf{x};\boldsymbol{\theta}).\end{aligned} \tag{41}$$

Step 2 and Step 3 implies this result. □

Remarks. We note that the same ideas of the derivation in Step 3 have been witnessed in contrastive learning that is known for minimizing InfoNCE [60, 51] of the form $-\mathbb{E}_{(x,y)\in\mathbb{P}_{XY}} R(x,y)$ where

$$R(x,y) = \mathbb{E}_{\hat{\mathbf{y}}_N\in\mathbb{P}_Y^N}\left[\log \frac{\exp h_\theta(x,y)}{\exp h_\theta(x,y)+\sum_{j=1}^{N}\exp h_\theta(x,\hat{y}_j)}\right]. \tag{42}$$

Recalling the batch-wise empirical loss of contrastive learners such as SimCLR [3] (x and y share a same modality) and CLIP (x and y are with different modalities), i.e.,

$$-\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp h_\theta(x_i,y_i)}{\exp h_\theta(x_i,y_i)+\sum_{j\neq i}\exp h_\theta(x_i,y_j)},$$

where $\mathcal{B} = \{(x_i,y_i)\}_{i=1}^B$ is the current training batch, one can check that each given $R(x,y)$ is estimated by exactly sampling $\hat{\mathbf{y}}_N$ as $\{y_j | (x_j,y_j)\in\mathcal{B} \text{ and } y_j\neq y\}$.

E Proof of Theorem 3

Definition 2 (Mutual Information (MI)). Given two random variables X and Y, the MI between X and Y is the Kullback-Leibler (KL) divergence between the joint distribution $\mathbb{P}_{XY}$ and the product of marginal distributions $\mathbb{P}_X\mathbb{P}_Y$, i.e.,

$$I(X;Y) \triangleq D_{KL}(\mathbb{P}_{XY}||\mathbb{P}_X\mathbb{P}_Y) = \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{XY}}\left[\log\frac{p_{XY}(\mathbf{x},y)}{p_X(\mathbf{x})p_Y(y)}\right] = \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{XY}}\left[\mathcal{W}(\mathbf{x},y)\right]. \tag{43}$$

where $\mathcal{W}(\mathbf{x},y)$ is the PMI defined in the main paper.

Lemma 1. *Given three random variables X, Y, and $\tilde{X}$, the mutual information $I(X;Y|V)$ can be decomposed into the following two ways:*

$$I(X,\tilde{X};Y) = I(X;Y) + I(\tilde{X};Y|X) = I(\tilde{X};Y) + I(X;Y|\tilde{X}), \tag{44}$$

where $I(X;Y|\tilde{X})$, i.e., the MI between X and Y conditioned on $\tilde{X}$, is defined as follows:

$$I(X;Y|\tilde{X}) \triangleq \mathbb{E}_{(\mathbf{x},y,\tilde{\mathbf{x}})\sim\mathbb{P}_{XY\tilde{X}}}\left[\log\frac{p_{XY|\tilde{X}}(\mathbf{x},y|\tilde{\mathbf{x}})}{p_{X|\tilde{X}}(\mathbf{x}|\tilde{\mathbf{x}})p_{Y|\tilde{X}}(y|\tilde{\mathbf{x}})}\right] = \mathbb{E}_{(\mathbf{x},y,\tilde{\mathbf{x}})\sim\mathbb{P}_{XY\tilde{X}}}\left[\mathcal{W}(\mathbf{x},y|\tilde{\mathbf{x}})\right]. \tag{45}$$

As a reminder, Theorem 3 is stated again as follows:

Theorem 3. *For any $\mathbf{x}\in\mathcal{X}$ and $y\in\mathcal{Y}$, let $\tilde{\mathbf{x}}=\mathcal{T}(\mathbf{x})$ be a sub-view of the input $\mathbf{x}$, $\mathcal{W}(\mathbf{x},y)$ can be decomposed into the following two terms:*

$$\mathcal{W}(\mathbf{x},y) = \mathcal{W}(\tilde{\mathbf{x}},y) + \mathcal{W}(\mathbf{x},y|\tilde{\mathbf{x}}), \tag{46}$$

where $\mathcal{W}(\mathbf{x},y|\tilde{\mathbf{x}})$, i.e., the PMI between $\mathbf{x}$ and y conditioned on $\tilde{\mathbf{x}}$, is defined as follows:

$$\begin{aligned}\mathcal{W}(\mathbf{x},y|\tilde{\mathbf{x}}) &\triangleq \log\frac{p_{XY|\tilde{X}}(\mathbf{x},y|\tilde{\mathbf{x}})}{p_{X|\tilde{X}}(\mathbf{x}|\tilde{\mathbf{x}})p_{Y|\tilde{X}}(y|\tilde{\mathbf{x}})}\\ &= \log\frac{p_{Y|X\tilde{X}}(y|\mathbf{x},\tilde{\mathbf{x}})}{p_{Y|\tilde{X}}(y|\tilde{\mathbf{x}})}.\end{aligned} \tag{47}$$

Proof. For any $\mathbf{x}\in\mathcal{X}$ and $y\in\mathcal{Y}$, let $\tilde{\mathbf{x}}=\mathcal{T}(\mathbf{x})$ be a sub-view of the input $\mathbf{x}$,the data processing inequality implies that $I(\tilde{\mathbf{x}};y|\mathbf{x})=0$, which means that $I(X;Y)=I(\tilde{X};Y)+I(X;Y|\tilde{X})$.

Given that

$$\begin{aligned}\mathbb{E}_{(\tilde{\mathbf{x}},y)\sim\mathbb{P}_{\tilde{X}Y}}\left[\log\frac{p_{\tilde{X}Y}(\tilde{\mathbf{x}},y)}{p_{\tilde{X}}(\tilde{\mathbf{x}})p_Y(y)}\right] &= \sum_{\tilde{\mathbf{x}}}\sum_{y}p_{\tilde{X}Y}(\tilde{\mathbf{x}},y)\left[\log\frac{p_{\tilde{X}Y}(\tilde{\mathbf{x}},y)}{p_{\tilde{X}}(\tilde{\mathbf{x}})p_Y(y)}\right]\\ &= \sum_{\tilde{\mathbf{x}}}\sum_{y}\left(\sum_{\mathbf{x}}p_{XY\tilde{X}}(\mathbf{x},y,\tilde{\mathbf{x}})\right)\left[\log\frac{p_{\tilde{X}Y}(\tilde{\mathbf{x}},y)}{p_{\tilde{X}}(\tilde{\mathbf{x}})p_Y(y)}\right]\\ &= \sum_{\tilde{\mathbf{x}}}\sum_{y}\sum_{\mathbf{x}}p_{XY\tilde{X}}(\mathbf{x},y,\tilde{\mathbf{x}})\left[\log\frac{p_{\tilde{X}Y}(\tilde{\mathbf{x}},y)}{p_{\tilde{X}}(\tilde{\mathbf{x}})p_Y(y)}\right]\\ &= \mathbb{E}_{(\mathbf{x},y,\tilde{\mathbf{x}})\sim\mathbb{P}_{XY\tilde{X}}}\left[\log\frac{p_{\tilde{X}Y}(\tilde{\mathbf{x}},y)}{p_{\tilde{X}}(\tilde{\mathbf{x}})p_Y(y)}\right]\end{aligned} \tag{48}$$

and

$$\begin{aligned}\mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{XY}}\left[\log\frac{p_{XY}(\mathbf{x},y)}{p_X(\mathbf{x})p_Y(y)}\right] &= \sum_{\mathbf{x}}\sum_{y}p_{XY}(\mathbf{x},y)\left[\log\frac{p_{XY}(\mathbf{x},y)}{p_X(\mathbf{x})p_Y(y)}\right]\\ &= \sum_{\tilde{\mathbf{x}}}\sum_{y}\left(\sum_{\mathbf{x}}p_{XY\tilde{X}}(\mathbf{x},y,\tilde{\mathbf{x}})\right)\left[\log\frac{p_{XY}(\mathbf{x},y)}{p_X(\mathbf{x})p_Y(y)}\right]\\ &= \sum_{\mathbf{x}}\sum_{y}\sum_{\tilde{\mathbf{x}}}p_{XY\tilde{X}}(\mathbf{x},y,\tilde{\mathbf{x}})\left[\log\frac{p_{XY}(\mathbf{x},y)}{p_X(\mathbf{x})p_Y(y)}\right]\\ &= \mathbb{E}_{(\mathbf{x},y,\tilde{\mathbf{x}})\sim\mathbb{P}_{XY\tilde{X}}}\left[\log\frac{p_{XY}(\mathbf{x},y)}{p_X(\mathbf{x})p_Y(y)}\right],\end{aligned} \tag{49}$$

we have the following:

$$
\begin{aligned}
&I(X;Y) = I(\tilde{X};Y) + I(X;Y|\tilde{X}) \\
\Leftrightarrow &\mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{XY}}\left[\log\frac{p_{XY}(\mathbf{x},y)}{p_X(\mathbf{x})p_Y(y)}\right] = \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{\tilde{X}Y}}\left[\log\frac{p_{\tilde{X}Y}(\tilde{\mathbf{x}},y)}{p_{\tilde{X}}(\tilde{\mathbf{x}})p_Y(y)}\right] + \mathbb{E}_{(\mathbf{x},y,\tilde{\mathbf{x}})\sim\mathbb{P}_{XY\tilde{X}}}\left[\log\frac{p_{XY|\tilde{X}}(\mathbf{x},y|\tilde{\mathbf{x}})}{p_{X|\tilde{X}}(\mathbf{x}|\tilde{\mathbf{x}})p_{Y|\tilde{X}}(y|\tilde{\mathbf{x}})}\right] \\
\Leftrightarrow &\mathbb{E}_{(\mathbf{x},y,\tilde{\mathbf{x}})\sim\mathbb{P}_{XY\tilde{X}}}\left[\log\frac{p_{XY}(\mathbf{x},y)}{p_X(\mathbf{x})p_Y(y)}\right] = \mathbb{E}_{(\mathbf{x},y,\tilde{\mathbf{x}})\sim\mathbb{P}_{XY\tilde{X}}}\left[\log\frac{p_{\tilde{X}Y}(\tilde{\mathbf{x}},y)}{p_{\tilde{X}}(\tilde{\mathbf{x}})p_Y(y)}\right] + \mathbb{E}_{(\mathbf{x},y,\tilde{\mathbf{x}})\sim\mathbb{P}_{XY\tilde{X}}}\left[\log\frac{p_{XY|\tilde{X}}(\mathbf{x},y|\tilde{\mathbf{x}})}{p_{X|\tilde{X}}(\mathbf{x}|\tilde{\mathbf{x}})p_{Y|\tilde{X}}(y|\tilde{\mathbf{x}})}\right] \\
\Leftrightarrow &\mathbb{E}_{(\mathbf{x},y,\tilde{\mathbf{x}})\sim\mathbb{P}_{XY\tilde{X}}}\left[\log\frac{p_{XY}(\mathbf{x},y)}{p_X(\mathbf{x})p_Y(y)} - \log\frac{p_{\tilde{X}Y}(\tilde{\mathbf{x}},y)}{p_{\tilde{X}}(\tilde{\mathbf{x}})p_Y(y)} - \log\frac{p_{XY|\tilde{X}}(\mathbf{x},y|\tilde{\mathbf{x}})}{p_{X|\tilde{X}}(\mathbf{x}|\tilde{\mathbf{x}})p_{Y|\tilde{X}}(y|\tilde{\mathbf{x}})}\right] = 0 \\
\Leftrightarrow &\mathbb{E}_{(\mathbf{x},y,\tilde{\mathbf{x}})\sim\mathbb{P}_{XY\tilde{X}}}\left[\mathcal{W}(\mathbf{x},y) - \mathcal{W}(\tilde{\mathbf{x}},y) - \mathcal{W}(\mathbf{x},y|\tilde{\mathbf{x}})\right] = 0 \\
\Leftrightarrow &\mathcal{W}(\mathbf{x},y) = \mathcal{W}(\tilde{\mathbf{x}},y) + \mathcal{W}(\mathbf{x},y|\tilde{\mathbf{x}})
\end{aligned}
\tag{50}
$$

□

F Proof of Theorem 4

As a reminder, Theorem 4 is stated as follows:

Theorem 4. *Let us define* $h_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{x}},y) \triangleq \mathbf{r}^\top[\beta\tilde{\mathbf{z}} + (1-\beta)\mathbf{z}]/\kappa$ *and* $h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}},y) = \tilde{\mathbf{z}}^\top\mathbf{r}/\tau$. *If we, following prior works [47, 36] respectively, assume that* $h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}},y) = \log\frac{p_{Y|\tilde{X}}(y|\tilde{\mathbf{x}})}{p_Y(y)} + c(\tilde{\mathbf{x}})$ *with* $c(\tilde{\mathbf{x}})$ *as a constant term depending on* $\tilde{\mathbf{x}}$, *and that* $h_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{x}},y) = \log\frac{p_{Y|X\tilde{X}}(y|\mathbf{x},\tilde{\mathbf{x}})}{p_{Y|\tilde{X}}(y|\tilde{\mathbf{x}})} + c(\mathbf{x},\tilde{\mathbf{x}})$ *with* $c(\mathbf{x},\tilde{\mathbf{x}})$ *as a constant term depending on* $\mathbf{x}$ *and* $\tilde{\mathbf{x}}$, *we then have the following for* $\alpha = 1$*:*

$$
E(\mathbf{x}) = \lim_{N\to+\infty} E_{\boldsymbol{\theta}}(\mathbf{x}) = \lim_{\substack{T\to+\infty \\ L/T\to+\infty}} \log(K+L)S_{ours}(\mathbf{x};\boldsymbol{\theta}). \tag{51}
$$

Proof. **Step 1:** Given that $\alpha = 1$, we have

$$
E(\mathbf{x}) = \frac{1}{\alpha}\log\sum_{y\in\mathcal{Y}_\mathrm{I}}\exp\left[\alpha\mathcal{W}(x,y)\right] = \log\sum_{y\in\mathcal{Y}_\mathrm{I}}\exp\left[\log\frac{p_{Y|X}(y|\mathbf{x})}{p_Y(y)}\right] = \log\sum_{y\in\mathcal{Y}_\mathrm{I}}\frac{p_{Y|X}(y|\mathbf{x})}{p_Y(y)}. \tag{52}
$$

Given that $h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}},y) = \log\frac{p_{Y|\tilde{X}}(y|\tilde{\mathbf{x}})}{p_Y(y)} + c(\tilde{\mathbf{x}})$, we have

$$
\begin{aligned}
\lim_{N\to+\infty}\hat{\mathcal{W}}(\tilde{\mathbf{x}},y;\boldsymbol{\theta}) &= \lim_{N\to+\infty}\log\frac{\hat{p}_{Y|\tilde{X}}(y|\tilde{\mathbf{x}};\boldsymbol{\theta})}{p_Y(y)} \\
&= \lim_{N\to+\infty}\log\mathbb{E}_{\hat{\mathbf{y}}_{N-1}\sim\mathbb{P}_Y^{N-1}}\left[\frac{N\cdot\exp h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}},y)}{\exp h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}},y) + \sum_{j=1}^{N-1}\exp h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}},\hat{y}_j)}\right] \\
&= \log\frac{p_{Y|\tilde{X}}(y|\tilde{\mathbf{x}})}{p_Y(y)} = \mathcal{W}(\tilde{\mathbf{x}},y)
\end{aligned}
\tag{53}
$$

Step 2:

Since $h_{\boldsymbol{\theta}}(\mathbf{x}, \tilde{\mathbf{x}}, y) = \log \frac{p_{Y|X\tilde{X}}(y|\mathbf{x},\tilde{\mathbf{x}})}{p_{Y|\tilde{X}}(y|\tilde{\mathbf{x}})} + c(\mathbf{x}, \tilde{\mathbf{x}})$, by closely following Eq. (37) and Eq. (38), Eq. (53) implies that

$$
\begin{aligned}
\lim_{N\to+\infty} E_{\boldsymbol{\theta}}(\mathbf{x}) &= \lim_{N\to+\infty} \log \sum_{y\in\mathcal{Y}_{\mathrm{I}}} \frac{\exp\left[\hat{\mathcal{W}}(\tilde{\mathbf{x}}, y; \boldsymbol{\theta}) + h_{\boldsymbol{\theta}}(\mathbf{x}, \tilde{\mathbf{x}}, y)\right]}{\mathbb{E}_{\hat{y}\sim\mathbb{P}_Y}\left[\exp\left[\hat{\mathcal{W}}(\tilde{\mathbf{x}}, \hat{y}; \boldsymbol{\theta}) + h_{\boldsymbol{\theta}}(\mathbf{x}, \tilde{\mathbf{x}}, \hat{y})\right]\right]} \\
&= \log \sum_{y\in\mathcal{Y}_{\mathrm{I}}} \frac{\exp\left[\mathcal{W}(\tilde{\mathbf{x}}, y) + h_{\boldsymbol{\theta}}(\mathbf{x}, \tilde{\mathbf{x}}, y)\right]}{\mathbb{E}_{\hat{y}\sim\mathbb{P}_Y}\left[\exp\left[\mathcal{W}(\tilde{\mathbf{x}}, \hat{y}) + h_{\boldsymbol{\theta}}(\mathbf{x}, \tilde{\mathbf{x}}, \hat{y})\right]\right]} \\
&= \log \sum_{y\in\mathcal{Y}_{\mathrm{I}}} \frac{\exp\left[\log \frac{p_{Y|\tilde{X}}(y|\tilde{\mathbf{x}})}{p_Y(y)} + \log \frac{p_{Y|X\tilde{X}}(y|\mathbf{x},\tilde{\mathbf{x}})}{p_{Y|\tilde{X}}(y|\tilde{\mathbf{x}})} + c(\mathbf{x}, \tilde{\mathbf{x}})\right]}{\mathbb{E}_{\hat{y}\sim\mathbb{P}_Y}\left[\exp\left[\log \frac{p_{Y|\tilde{X}}(\hat{y}|\tilde{\mathbf{x}})}{p_Y(\hat{y})} + \log \frac{p_{Y|X\tilde{X}}(\hat{y}|\mathbf{x},\tilde{\mathbf{x}})}{p_{Y|\tilde{X}}(\hat{y}|\tilde{\mathbf{x}})} + c(\mathbf{x}, \tilde{\mathbf{x}})\right]\right]} \\
&= \log \sum_{y\in\mathcal{Y}_{\mathrm{I}}} \frac{\frac{p_{Y|X\tilde{X}}(y|\mathbf{x},\tilde{\mathbf{x}})}{p_Y(y)}}{\mathbb{E}_{\hat{y}\sim\mathbb{P}_Y}\left[\frac{p_{Y|X\tilde{X}}(\hat{y}|\mathbf{x},\tilde{\mathbf{x}})}{p_Y(\hat{y})}\right]} \\
&= \log \sum_{y\in\mathcal{Y}_{\mathrm{I}}} \frac{p_{Y|X\tilde{X}}(y|\mathbf{x},\tilde{\mathbf{x}})}{p_Y(y)} = E(\mathbf{x})
\end{aligned}
\tag{54}
$$

Step 3:

Let $\hat{y}_i^{(t)}$ be the i-th sample drawn at the t-th round, the Law of Large Numbers implies

$$
\begin{aligned}
\mathcal{W}(\tilde{\mathbf{x}}, y) &= \lim_{N\to+\infty} \log \mathbb{E}_{\hat{\mathbf{y}}_{N-1}\sim\mathbb{P}_Y^{N-1}} \left[\frac{N \cdot \exp h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}, y)}{\exp h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}, y) + \sum_{j=1}^{N-1} \exp h_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}, \hat{y}_j)}\right] \\
&= \lim_{N\to+\infty} \lim_{T\to+\infty} \frac{1}{T} \sum_{i=1}^{T} \frac{N \cdot \exp h_{\boldsymbol{\theta}}(\mathbf{x}, y)}{\exp h_{\boldsymbol{\theta}}(\mathbf{x}, y) + \sum_{j=1}^{N-1} \exp h_{\boldsymbol{\theta}}(\mathbf{x}, \hat{y}_j^{(t)})}.
\end{aligned}
\tag{55}
$$

Similarly, by taking $\mathcal{G}_t \cup \mathcal{Y}_{\mathrm{I}} \setminus \{y\}$ as the examplar of $\left\{\hat{y}_i^{(t)}\right\}_{i=1}^{N-1}$ for $\mathcal{W}(\tilde{\mathbf{x}}, y)$ in Eq. (55), we have $N = K + L/T$ such that $N \to +\infty$ requires $L/T \to +\infty$, which results in rewriting Eq. (55) as follows:

$$
\begin{aligned}
\mathcal{W}(\tilde{\mathbf{x}}, y) &= \lim_{\substack{T\to+\infty \\ L/T\to+\infty}} \log \sum_{y\in\mathcal{Y}_{\mathrm{I}}} \frac{1}{T} \sum_{t=1}^{T} \frac{\exp(\tilde{\mathbf{z}}^\top \mathbf{r}/\tau)}{\sum\limits_{y_j\in\mathcal{G}_t\cup\mathcal{Y}_{\mathrm{I}}} \exp(\tilde{\mathbf{z}}^\top \mathbf{r}_j/\tau)} + \log(K + L/T) \\
&= \lim_{\substack{T\to+\infty \\ L/T\to+\infty}} \Lambda(\tilde{\mathbf{x}}, y)
\end{aligned}
\tag{56}
$$

Given that $N = T \cdot L/T \to +\infty$ when $L/T \to +\infty$ and $T \to +\infty$, as implied by the Law of Large Number, combining Eq. (54) with Eq. (56) arrives at the following:

$$
\begin{aligned}
&\lim_{\substack{T\to+\infty \\ L/T\to+\infty}} \log(K + L) S_{\text{ours}}(\mathbf{x}; \boldsymbol{\theta}) \\
&= \lim_{\substack{T\to+\infty \\ L/T\to+\infty}} \log \sum_{y\in\mathcal{Y}_{\mathrm{I}}} \frac{\exp\left[\Lambda(\tilde{\mathbf{x}}, y; \boldsymbol{\theta}) + h_{\boldsymbol{\theta}}(\mathbf{x}, \tilde{\mathbf{x}}, y)\right]}{\sum_{j=1}^{K+L} \exp\left[\Lambda(\tilde{\mathbf{x}}, y_j; \boldsymbol{\theta}) + h_{\boldsymbol{\theta}}(\mathbf{x}, \tilde{\mathbf{x}}, y_j)\right]} + \log(K + L) \\
&= \lim_{L\to+\infty} \log \sum_{y\in\mathcal{Y}_{\mathrm{I}}} \frac{\exp\left[\mathcal{W}(\tilde{\mathbf{x}}, y) + h_{\boldsymbol{\theta}}(\mathbf{x}, \tilde{\mathbf{x}}, y)\right]}{\sum_{j=1}^{K+L} \exp\left[\mathcal{W}(\tilde{\mathbf{x}}, y) + h_{\boldsymbol{\theta}}(\mathbf{x}, \tilde{\mathbf{x}}, y_j)\right]} + \log(K + L) \\
&= \log \sum_{y\in\mathcal{Y}_{\mathrm{I}}} \frac{\exp\left[\mathcal{W}(\tilde{\mathbf{x}}, y) + h_{\boldsymbol{\theta}}(\mathbf{x}, \tilde{\mathbf{x}}, y)\right]}{\mathbb{E}_{\hat{y}\sim\mathbb{P}_Y}\left[\exp\left[\mathcal{W}(\tilde{\mathbf{x}}, \hat{y}) + h_{\boldsymbol{\theta}}(\mathbf{x}, \tilde{\mathbf{x}}, \hat{y})\right]\right]} = E(\mathbf{x})
\end{aligned}
\tag{57}
$$

Step 2 and Step 3 imply this result. □

Table 4: Comparison with different VLM architectures on ImageNet-1K (ID).All values are percentages. ↑ indicates larger values are better and vice versa. The best results in the last two columns are shown in bold per ID dataset. Results are averaged over 5 independent runs.

ID Dataset	Method	iNaturalist		SUN		Places		Textures		Average	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
GroupViT [62]	NegLabel	98.07	8.60	91.52	35.12	88.85	41.63	88.45	47.06	91.72	33.10
	Ours	99.52	6.86	95.52	25.08	92.24	37.44	91.28	46.15	**94.70**	**29.08**

G Additional Experiments

G.1 More Backbones

G.2 Cropping vs Cutout

Table 5: Ablation study on ImageNet-1K w.r.t the choice of obtaining subviews. ↑ indicates larger values are better and vice versa. Results are averaged over 5 independent runs.

Method	iNaturalist		SUN		Places		Textures		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Cropping	99.64	1.04	96.32	18.45	95.81	31.15	92.15	38.79	96.00	22.36
Cutout	99.52	1.26	95.08	19.08	95.46	32.44	91.50	39.38	95.39	23.04

H Limitations

This paper directly uses the off-the-shelf negative labels mined by NegLabel for PMI estimation. It will be exciting to what makes good negative labels for OOD detection with pre-trained VLMs.

I Broader impacts

Our project aims to improve the reliability and safety of modern machine learning models, which leads to benefits and societal impacts, particularly for safety-critical applications such as autonomous driving. Our study does not involve any human subjects or violation of legal compliance. We do not anticipate any potentially harmful consequences to our work.

J Stability

To verify that our method consistently provides strong performance, we run with 10 independent seeds for ImageNet-1K and report the average and standard deviation of FPR95 and AUROC as follows.

Table 6: Ablation on stability. OOD detection performance based on CLIP-B/16 of our method on ImageNet-1K. Results are averaged over 5 independent runs.

iNaturalist		SUN		Places		Textures	
FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
0.0	0.1	1.2	0.5	3.0	1.0	1.6	0.6