A CLOSE LOOK AT NEGATIVE LABEL GUIDED OUT-OF-DISTRIBUTION DETECTION IN PRE-TRAINED VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Advances in pre-trained vision-language models have enabled zero-shot out-of-distribution (OOD) detection using only in-distribution (ID) labels. Recent methods in this direction expand the label space with negative labels to enhance the discrimination between ID and OOD inputs. Despite their promising progress, there remains a limited understanding of their empirical effectiveness in open-world scenarios, where negative labels can arbitrarily diverge from real OOD ones. This paper bridges this research gap with the helm of a novel energy-based framework, where the energy function is built upon the margin between the similarity of an input to ID labels and that to negative labels. Guided by this framework, we prove that the inherent tolerance of such methods to the sampling bias essentially stems from estimating the worst-case energy function over a KL-constrained set of potential distributions centered on the negative label distribution. Furthermore, our theoretical analysis reveals that existing methods suffer from over-pessimism and consequently high sensitivity to outliers. Provably, we can alleviate these problems by leveraging Rényi divergence to refine potential distributions. Extensive experiments empirically manifest that our method establishes a new state-of-the-art across a variety of OOD detection settings.

1 INTRODUCTION

Despite the significant progress in machine learning that has facilitated a broad spectrum of classification tasks (Masana et al., 2022; Zhao et al., 2019; Caruana & Niculescu-Mizil, 2006), 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 data during training, giving rise to what is known as out-of-distribution (OOD) data. These OOD data 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) data but also flag any OOD data as unknown. This directly motivates OOD detection (Lang et al., 2023; Salehi et al., 2021; Yang et al., 2021), which makes significant differences in ensuring the safety of decision-critical applications, e.g., autonomous driving (Huang et al., 2020), medical diagnosis (Zimmerer et al., 2022), and cyber-security (Nguyen et al., 2022).

This paper focuses on post-hoc OOD detection, which is more practical than learning-based methods that require resource-intensive retraining. Earlier studies (Liang et al., 2017; Liu et al., 2020; Huang et al., 2021; Sun et al., 2022; Peng et al., 2025; Zhang et al., 2024b) primarily utilized the single modality of pre-trained models, but the success of contrastive language-image pre-training (CLIP) (Radford et al., 2021a) has recently shifted research toward expanding post-hoc OOD detection from single-modal to multi-modal methods. The pioneering work, MCM (Ming et al., 2022), defines textual features as the concept for each ID class and uses the scaled distance between visual features and the closest ID prototype to measure OOD uncertainty. This 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 the text interpretation capabilities of VLMs underutilized. To address this, NegLabel (Jiang et al., 2024) selects negative labels from large-scale lexical databases, such as WordNet (Miller, 1995), based on their similarities to

the ID label space, which equips the model with stronger ability to distinguish OOD data. Despite promising potential of negative labels, there remains a limited theoretical understanding of their effectiveness in open-world scenarios, where real OOD data, due to its open-ended nature, can be arbitrarily different from the observed negative labels (Wang et al., 2023b;c).

To mitigate this research gap, this paper delivers a close look at CLIP-based post-hoc OOD detection with negative labels from the perspective of density estimation. We argue that this standpoint is well-suited for studying OOD detection, since OOD data, by definition, diverges from ID data in terms of their underlying density distributions. Following prior works (Liu et al., 2020), our analytical framework models ID data by resorting to the energy-based model (LeCun et al., 2006). However, we find that it is non-trivial to extend the energy function from a uni-modal to a multi-modal setting. Drawing inspiration from triplet-based metric learning (Sohn, 2016; Hermans et al., 2017), we propose to build the energy function upon the margin between the similarity of a given test-time input to ID labels and to negative labels. Guided by this framework, we theoretically show that NegLabel essentially augments the negative label distribution by constructing a distribution set contained within a Kullback–Leibler (KL) ball centered on it. Estimating the energy function against the worst-case distribution in this set ensures performance guarantees under all possible (or constrained) distribution shifts. This provides a theoretical explanation for why NegLabel remains effective when faced with unseen OOD data.

In addition, our theoretical analysis reveals that the paradigm of NegLabel is prone to induce an overly conservative worst-case distribution, as it assigns disproportionately large weights (governed by an exponential function) to negative labels that exhibit high similarity to the test-time input. In response, we transcend the boundary of KL divergence, but exploring a broader family of distribution divergence metrics — Rényi divergence (Rényi, 1961). As a generalization of KL divergence, Rényi divergence introduces an additional parameter, the order, which offers flexible control over the weighting distribution. We show the use of Rényi divergence enables to retain the aforementioned strengths while mitigating the conservativeness by shaping a milder, polynomial-bounded worst-case distribution with adaptively tunable order. Extensive experiments empirically manifest that our method establishes a new state-of-the-art in a variety of OOD detection setups.

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 (Ming et al., 2022), 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. Instead of relying on textual information from only ID label space, ZOC (Esmaeilpour et al., 2022) applies VLMs to discern OOD instances by training a captioner that generates potential OOD labels. Nevertheless, this captioner often fails to produce effective OOD labels, particularly for ID datasets containing many classes. Differently, NegLabel (Jiang et al., 2024) incorporates additional negative class names mined from available data sources as negative proxies. Considering the nonalignment between target visual OOD distribution and the generated negative textual OOD distribution, AdaNeg (Zhang & Zhang, 2024) leverages the benefits of test-time adaptation to generate adaptive proxies by exploring potential OOD images during testing. More recently, Peng et al. understand CLIP-based post-hoc OOD detection from an information-theoretical perspective. On the other hand, CLIP-based OOD detection can also be improved by prompt representation learning. In particular, LoCoOp (Miyai et al., 2024) learns ID text prompts by pushing them away from the portions of CLIP local features that have ID-irrelevant nuisances (e.g., backgrounds). CLIPN (Wang et al., 2023a) and LSN (Nie et al., 2024) design a learnable "no" prompt and a "no" text encoder to capture negation semantics within images. Differently, LAPT (Zhang et al., 2025) initializes prompts with negative labels (Jiang et al., 2024), followed by tuning prompts with cross-modal and cross-distribution mixing. *Due to space limitation, more related works are discussed in Appendix A.*

3 Preliminary

Notation. Let $\mathcal{X}$ and $\mathcal{Y}$ be the input space and the label space, respectively. Given a random variable $Y \in \mathcal{Y}$, we write $\mathbb{P}_Y$ as the marginal distribution defined over $\mathcal{Y}$, and use $y \sim \mathbb{P}_Y$ to

indicate a sample y drawn from $\mathbb{P}_Y$. Considering K-way classification as a case study, we write $\mathcal{Y}_{\text{I}} \triangleq \{y_1, \ldots, y_K\} \subset \mathcal{Y}$ as the *known* ID label space. The joint ID distribution, represented as $\mathbb{P}_{X_{\text{I}}Y_{\text{I}}}$, is a joint distribution defined over $\mathcal{X} \times \mathcal{Y}_{\text{I}}$. 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 $\mathcal{Y}_{\text{O}} \subseteq \mathcal{Y} \setminus \mathcal{Y}_{\text{I}}$ is the *unknown* OOD label space.

Post-hoc OOD Scoring. Existing methods (Hendrycks & Gimpel, 2016; Liang et al., 2017; Liu et al., 2020; Huang et al., 2021; Sun et al., 2022) tend to adopt a post-hoc strategy to detect OOD data, *i.e.,* given a pre-trained ID classification model f and a scoring function $S(\cdot; f) : \mathcal{X} \to \mathbb{R}$, then $\mathbf{x}$ is detected as ID data if and only if $S(\mathbf{x}; f) \geq \lambda$, for some given threshold λ:

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

Typically, λ is chosen to ensure a high fraction (e.g., 95%) of ID data to be correctly classified.

CLIP-based Models adopt a dual-stream architecture (Radford et al., 2021b) with one text encoder $f_{\mathcal{T}}$ and one image encoder $f_{\mathcal{X}}$ to map inputs of two modalities into a uni-modal hyper-spherical space $\mathbb{S}^{d-1} \triangleq \left\{\mathbf{z} \in \mathbb{R}^d | \, \|\mathbf{z}\|_2 = 1\right\}$. Zero-shot image classification based on a pre-trained CLIP-like model is to classify images into one of known ID classes by computing $\arg\max_{j=1,\ldots K} h(\mathbf{x}, y_j)$ where $h(\mathbf{x}, y_j) \triangleq f_{\mathcal{X}}(\mathbf{x})^\top f_{\mathcal{T}}\big(\Delta(y_j)\big)$ with $\Delta(\cdot)$ producing the text prompt for the input label.

CLIP-based OOD Detection with Negative Labels. CLIP-based models, thanks to their remarkable effectiveness (Radford et al., 2021b) and provable guarantees (Chen et al., 2023), are recently extended to the task of zero-shot OOD detection where there is no need to train on ID data. A popular pipeline is to leverage a L-sized set of negatives labels[1] $\hat{\mathcal{Y}} \triangleq \{\hat{y}_1, \ldots, \hat{y}_L\}$ to formulate the OOD scoring function of $\mathbf{x}$ as the model's prediction confidence that $\mathbf{x}$ belongs to $\mathcal{Y}_{\text{I}}$, i.e.,

$$S_{\text{NegLabel}}(\mathbf{x}; f) \triangleq \sum_{i=1}^{K} \frac{\exp\left[h(\mathbf{x}, y_i)/T\right]}{\sum_{j=1}^{K} \exp\left[h(\mathbf{x}, y_j)/T\right] + \sum_{j=1}^{L} \exp\left[h(\mathbf{x}, \hat{y}_j)/T\right]}, \tag{2}$$

where $T > 0$ is a temperature hyper-parameter.

Due to space limitation, detailed proofs of theorems in this paper are provided in Appendix B.

4 A Close Look at CLIP-based Out-of-distribution Detection with Negative Labels

While NegLabel (Jiang et al., 2024) has empirically emerged to be an effective post-hoc OOD detector, there is limited prior work providing a comprehensive explanation for its efficacy from a rigorous mathematical point of view. This paper fills this research gap from the perspective of *distributionally-augmented density estimation*. Inspired by Liu et al. (2020), 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 (LeCun et al., 2006). In particular, let $\hat{p}_{X_{\text{I}}}(\mathbf{x}; \boldsymbol{\theta})$ be an estimator of the modeled ID data density $\hat{p}_{X_{\text{I}}}(\mathbf{x})$ using the pre-trained CLIP-based model parameters $\boldsymbol{\theta}$, we have:

$$\hat{p}_{X_{\text{I}}}(\mathbf{x}; \boldsymbol{\theta}) = \frac{\exp\left[E(\mathbf{x}; \boldsymbol{\theta})\right]}{Z(\boldsymbol{\theta})} \propto \exp\left[E(\mathbf{x}; \boldsymbol{\theta})\right], \tag{3}$$

where $Z(\boldsymbol{\theta}) = \int \exp\left[E_{\boldsymbol{\theta}}(\mathbf{x}; \boldsymbol{\theta})\right] d\mathbf{x}$ is an *input-independent* normalization function with

$$E(\mathbf{x}; \boldsymbol{\theta}) = T \log \sum_{i=1}^{K} \exp\left[E(\mathbf{x}, y_i; \boldsymbol{\theta})/T\right].$$

The behavior of $E(\mathbf{x}; \boldsymbol{\theta})$ is largely determined by the formulation of $E(\mathbf{x}, y_i; \boldsymbol{\theta})$. A naive choice of $E(\mathbf{x}, y_i; \boldsymbol{\theta})$ is the CLIP-based zero-shot classifier logit $h(\mathbf{x}, y_j)$, which aligns with traditional energy-based OOD detection (Liu et al., 2020). However, Table 1 shows that Energy (zero-shot) achieves considerably far-from-satisfactory performance (79.57% AUROC and 82.21% FPR95 in average), which implies that it is non-trivial to extend energy function $E_{\boldsymbol{\theta}}(\mathbf{x})$ from single-modal to

[1] In accordance to Jiang et al. (2024), *negative* labels are defined as those semantically *irrelevant/dissimilar* to *all* ID labels.

multi-modal settings. Let $\mathbb{P}_{\hat{Y}}$ be the sampling distribution of negative labels, drawing inspiration from triplet-based metric learning (Sohn, 2016; Hermans et al., 2017), we define $E(\mathbf{x}, y_i; \boldsymbol{\theta})$ as:

$$E(\mathbf{x}, y_i; \boldsymbol{\theta}) \triangleq \mathbb{E}_{\hat{y}\in\mathbb{P}_{\hat{Y}}}\left[h(\mathbf{x}, y_i) - h(\mathbf{x}, \hat{y})\right] = h(\mathbf{x}, y_i) - \mathbb{E}_{\hat{y}\in\mathbb{P}_{\hat{Y}}}\left[h(\mathbf{x}, \hat{y})\right], \tag{4}$$

where the expectation can be effectively estimated using the observed negative labels $\hat{\mathcal{Y}}$. Table 1 shows that Eq. (4) achieves significantly better performance (93.65% AUROC and 28.82% FPR95 in average) than Energy (zero-shot), which empirically validates our design.

Intuitively, if negative labels are semantically similar to unseen ground-truth OOD labels, Eq. (4) will perform well when facing real OOD data. However, the two kinds of labels could be arbitrarily distinct from each other in practice (Wang et al., 2023b;c), posing us to suspect that the power of negative labels in Eq. (4) has yet to be fully unleashed. To verify this, we consider the situation where ground-truth OOD labels are accessible and considered as negative labels to estimate the the expectation in Eq. (4). We find in Table 1 that this oracle case contributes to a more satisfactory results, with AUROC of 97.11% and FPR95 of 15.38% in average.

In view of this, we further extend the formulation of $E(\mathbf{x}, y_i; \boldsymbol{\theta})$ in Eq. (4) beyond the given distribution $\mathbb{P}_{\hat{Y}}$ to a broader family of potential distributions with perturbations. To be specific, we are interested in the worst case of $E(\mathbf{x}, y_i; \boldsymbol{\theta})$ in Eq. (4) over a set of potential distributions $\mathbb{Q}_{\hat{Y}}$, which are centered on $\mathbb{P}_{\hat{Y}}$ and constrained by a metric function $D(\mathbb{Q}_{\hat{Y}}\|\mathbb{P}_{\hat{Y}})$ within a radius $\eta > 0$, i.e.,

$$\hat{E}(\mathbf{x}, y_i; \boldsymbol{\theta}) = h(\mathbf{x}, y_i) - \max_{\mathbb{Q}_{\hat{Y}}} \mathbb{E}_{\hat{y}\in\mathbb{Q}_{\hat{Y}}}\left[h(\mathbf{x}, \hat{y})\right] \quad \textit{s.t. } D(\mathbb{Q}_{\hat{Y}}\|\mathbb{P}_{\hat{Y}}) \leq \eta, \tag{5}$$

where $D(\mathbb{Q}\|\mathbb{P})$ measures the distribution discrepancy between $\mathbb{Q}$ and $\mathbb{P}$. Intuitively, $\mathbb{Q}_{\hat{Y}}$ acts as an "adversary", probing the hardest possible negative distribution. This makes $\hat{E}(\mathbf{x}, y_i; \boldsymbol{\theta})$ inherently more conservative and thus more reliable when real OOD labels differ from negatives labels.

Theorem 1. *By choosing $D(\cdot\|\cdot)$ as KL divergence, i.e., $D(\mathbb{Q}_{\hat{Y}}\|\mathbb{P}_{\hat{Y}}) = \int q_{\hat{Y}}(y)\log\frac{q_{\hat{Y}}(y)}{p_{\hat{Y}}(y)}dy$, we can rewrite $\hat{E}(\mathbf{x}, y_i; \boldsymbol{\theta})$ in Eq. (5) as follows:*

$$\hat{E}(\mathbf{x}, y_i; \boldsymbol{\theta}) = \alpha^*(\mathbf{x}, \mathbb{P}_{\hat{Y}})\log\frac{e^{h(\mathbf{x}, y_i)/\alpha^*(\mathbf{x}, \mathbb{P}_{\hat{Y}})}}{\mathbb{E}_{\hat{y}\sim\mathbb{P}_{\hat{Y}}}\left[e^{h(\mathbf{x}, \hat{y})/\alpha^*(\mathbf{x}, \mathbb{P}_{\hat{Y}})}\right]} - \alpha^*(\mathbf{x}, \mathbb{P}_{\hat{Y}})\cdot\eta, \tag{6}$$

where $\alpha^*(\mathbf{x}, \mathbb{P}_{\hat{Y}}) = \arg\min_{\alpha\geq 0}\left\{\alpha\eta + \alpha\log\mathbb{E}_{\hat{y}\sim\mathbb{P}_{\hat{Y}}}\left[e^{h(\mathbf{x}, \hat{y})/\alpha}\right]\right\}$.

Theorem 2 (Lemma 5 of Faury et al. (2020))**.** *The optimal $\alpha^*(\mathbf{x}, \mathbb{P}_{\hat{Y}})$ can be approximated as*

$$\alpha^*(\mathbf{x}, \mathbb{P}_{\hat{Y}}) \approx \sqrt{\mathbb{V}_{\hat{y}\sim\mathbb{P}_{\hat{Y}}}\left[h(\mathbf{x}, \hat{y})\right]/2\eta}, \tag{7}$$

where $\mathbb{V}_{\hat{y}\sim\mathbb{P}_{\hat{Y}}}\left[h(\mathbf{x}, \hat{y})\right]$ denotes the variance of $h(\mathbf{x}, \hat{y})$ over the distribution $\mathbb{P}_{\hat{Y}}$.

If the assumption of homoscedasticity (uniform variance) holds for each input $\mathbf{x} \in \mathcal{X}$ given a fixed $\mathbb{P}_{\hat{Y}}$, Theorem 2 implies that we can find a $\eta > 0$ to have $\alpha^*(\mathbf{x}, \mathbb{P}_{\hat{Y}}) \approx T$. Combining this with Theorem 1 implies that we can approximate the distributionally augmented energy function $\hat{E}(\mathbf{x}; \boldsymbol{\theta}) = T\log\sum_{i=1}^{K}\exp\left[\hat{E}(\mathbf{x}, y_i; \boldsymbol{\theta})/T\right]$ as follows:

$$\begin{aligned}\hat{E}(\mathbf{x}; \boldsymbol{\theta}) &\approx T\log\sum_{i=1}^{K}\frac{e^{h(\mathbf{x}, y_i)/T}}{\mathbb{E}_{\hat{y}\sim\mathbb{P}_{\hat{Y}}}\left[e^{h(\mathbf{x}, \hat{y})/T}\right]} - T\eta \\ &\approx \underbrace{T\log\sum_{i=1}^{K}\frac{e^{h(\mathbf{x}, y_i)/T}}{\sum_{j=1}^{L}e^{h(\mathbf{x}, \hat{y}_j)/T}}}_{\hat{S}_{\text{NegLabel}}(\mathbf{x}; f)} + \underbrace{T\log L - T\eta}_{\text{constant}}.\end{aligned} \tag{8}$$

Discussion. While $\hat{S}_{\text{NegLabel}}(\mathbf{x}; f)$ differs from $S_{\text{NegLabel}}(\mathbf{x}; f)$ in that the term $\sum_{j=1}^{K}e^{h(\mathbf{x}, y_j)/T}$ is excluded in its denominator, Table 1 shows that $\hat{S}_{\text{NegLabel}}(\mathbf{x}; f)$ performs on par with $S_{\text{NegLabel}}(\mathbf{x}; f)$, which implies the functional equivalence between $\hat{S}_{\text{NegLabel}}(\mathbf{x}; f)$ and $S_{\text{NegLabel}}(\mathbf{x}; f)$. The theoretical connection between $\hat{S}_{\text{NegLabel}}(\mathbf{x}; f)$ and $S_{\text{NegLabel}}(\mathbf{x}; f)$ can be found in Appendix J.1. Since the R#GPbo

Table 1: OOD detection results on ImageNet-1K with VIT B/16 CLIP as encoder. ↑ indicates larger values are better and vice versa. The best results in the last two columns are shown in bold. †: the baseline operates under the oracle setting where ground-truth OOD labels are known.

Dataset	iNaturalist		Sun		Places		Textures		Average	
Metric	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Energy (zero-shot)	85.09	81.08	84.24	79.02	83.38	75.08	65.56	93.65	79.57	82.21
Energy (Eq. (4))	98.49	7.11	94.72	25.93	90.35	41.35	91.02	40.89	93.65	28.82
Energy (Eq. (4))†	99.62	2.35	98.23	8.65	95.16	25.97	95.46	24.55	97.11	15.38
$S_{\text{NegLabel}}(\mathbf{x}; f)$ (Eq. (2))	99.30	2.65	95.06	23.11	90.90	40.35	89.76	46.63	93.76	28.19
$\hat{S}_{\text{NegLabel}}(\mathbf{x}; f)$ (Eq. (8))	99.29	2.67	95.02	23.22	90.93	40.30	89.85	46.31	93.77	28.13
$S_{\text{ours}}(\mathbf{x}; f)$ (Eq. (13))	99.64	1.29	95.71	17.94	91.90	33.98	90.79	39.45	**94.51**	**23.17**

function $\log(\cdot)$ is monotonically increasing, Eq. (8) admits the merit of $S_{\text{NegLabel}}(\mathbf{x}; f)$ as it is equivalent to augmenting the negative label distribution $\mathbb{P}_{\hat{Y}}$ by crafting a distribution set containing all the distributions in a KL ball centered on $\mathbb{P}_{\hat{Y}}$. This allows the energy-based density estimation via Eq. (4) to perform uniformly across various potential distributions of negative labels, thereby conferring inherent tolerance to distribution discrepancy between negative labels and real OOD labels.

Despite theoretical and empirical advantages of $\hat{E}(\mathbf{x}; \boldsymbol{\theta})$ over $E(\mathbf{x}; \boldsymbol{\theta})$, it is worth noting that the use of KL divergence to measure distribution discrepancy suffers from being overly pessimistic. To illustrate this, let us start from exploring the worst case of negative label distribution as follows.

Theorem 3. *Let us define*

$$\mathbb{Q}^*_{\hat{Y}} = \arg\max_{\mathbb{Q}_{\hat{Y}}} \mathbb{E}_{\hat{y}\in\mathbb{Q}_{\hat{Y}}}\left[h(\mathbf{x}, \hat{y})\right] \quad \text{s.t. } D(\mathbb{Q}_{\hat{Y}} \| \mathbb{P}_{\hat{Y}}) \leq \eta.$$

If we choose $D(\cdot\|\cdot)$ as KL divergence, then we have $q^*_{\hat{Y}}(\hat{y}) = \omega_{KL}(\mathbf{x}, \hat{y}) p_{\hat{Y}}(\hat{y})$ where

$$\omega_{KL}(\mathbf{x}, \hat{y}) \triangleq \frac{e^{h(\mathbf{x},\hat{y})/\alpha^*(\mathbf{x},\mathbb{P}_{\hat{Y}})}}{\mathbb{E}_{\tilde{y}\sim\mathbb{P}_{\hat{Y}}}\left[e^{h(\mathbf{x},\tilde{y})/\alpha^*(\mathbf{x},\mathbb{P}_{\hat{Y}})}\right]} \propto e^{h(\mathbf{x},\hat{y})/\alpha^*(\mathbf{x},\mathbb{P}_{\hat{Y}})}. \tag{9}$$

Theorem 3 implies that the use of KL divergence leads to assigning a weight $\omega_{KL}(\mathbf{x}, \hat{y})$ to each negative label $\hat{y} \sim \mathbb{P}_{\hat{Y}}$, with the weight $\omega_{KL}(\mathbf{x}, \hat{y})$ proportional to the exponential of the scaled cosine similarity. However, the explosive nature of the exponential function would make the resulting weight distribution tend to be highly skewed so that the worst-case expectation in Eq. (5) can be dominated by outliers, i.e., those exhibiting excessively high cosine simiarity to the input $\mathbf{x}$, which could greatly degrade the ability of $\hat{E}(\mathbf{x}; \boldsymbol{\theta})$ in Eq. (3) to detect OOD inputs especially when the outliers contain false negative labels[2]. We note that our theoretical analysis is consistent with empirical observations in Table 1: $\hat{S}_{\text{NegLabel}}(\mathbf{x}; f)$ performs marginally better than Energy (Eq. (4)) in average and even worse than Energy (Eq. (4)) on Textures[3].

5 Methodology

Our goal is to refine the worst-case distribution, aiming to assign more reasonable weights to negative labels. To this end, we propose to use Rényi divergence, a generalization of KL divergence that is defined with an additional parameter called an order, to measure distribution discrepancy. In particular, we focus on the Cressie-Read family of Rényi divergence (Duchi & Namkoong, 2021; Rényi, 1961) due to its analytical benefits that can be reflected by the following theorem:

Theorem 4. *By choosing $D(\cdot\|\cdot)$ as the Cressie-Read family of Rényi divergence, i.e.,*

$$D(\mathbb{Q}_{\hat{Y}} \| \mathbb{P}_{\hat{Y}}) = \int q_{\hat{Y}}(y) \phi_\gamma \left(\frac{q_{\hat{Y}}(y)}{p_{\hat{Y}}(y)} \right) dy, \tag{10}$$

[2]Prior works filter negative labels from a unlabeled wild corpus database with a cosine similarity-based strategy. However, there is no theoretical guarantees that cosine similarity could correctly capture semantic relationships so that the observed negative labels are inevitably contaminated by false negative labels.

[3]While $\hat{S}_{\text{NegLabel}}(\mathbf{x}; f)$ and $S_{\text{NegLabel}}(\mathbf{x}; f)$ can be enhanced by the grouping strategy as described in Jiang et al. (2024), existing works are fall short in providing theoretical justification for this heuristic trick.

where $\phi_\gamma(t) = \frac{1}{\gamma(\gamma-1)}(t^\gamma - \gamma t + \gamma - 1)$ *with* $\gamma > 1$*, we can rewrite* $\hat{E}(\mathbf{x}, y_i; \boldsymbol{\theta})$ *in Eq. (5) as:*

$$\hat{E}(\mathbf{x}, y_i; \boldsymbol{\theta}) = h(\mathbf{x}, y_i) - \left\{ c_\gamma(\eta) \mathbb{E}_{\hat{y} \sim \mathbb{P}_{\hat{Y}}} \left[\left(h(\mathbf{x}, \hat{y}) - \beta^*_{\mathbf{x}} \right)_+^{\gamma^*} \right]^{\frac{1}{\gamma^*}} + \beta^*_{\mathbf{x}} \right\}, \tag{11}$$

where $\gamma^* = \gamma/(\gamma-1)$*,* $c_\gamma(\eta) = (1 + \gamma(\gamma-1)\eta)^{\frac{1}{\gamma}}$*,* $(a)_+ = \max\{a, 0\}$*, and*

$$\beta^*_{\mathbf{x}} = \arg\min_{\beta} \left\{ c_\gamma(\eta) \mathbb{E}_{\hat{y} \sim \mathbb{P}_{\hat{Y}}} \left[\left(h(\mathbf{x}, \hat{y}) - \beta \right)_+^{\gamma^*} \right]^{\frac{1}{\gamma^*}} + \beta \right\}. \tag{12}$$

Note that Rényi divergence in Eq. (10) introduces an order parameter γ to adjust the polynomial relationships of the probability distance measure with the probability ratio. This provides enhanced flexibility in measuring distribution discrepancy by re-framing the metric function design as a search for the optimal γ within a narrow range. A similar spirit is also witnessed in Peng et al. (2024). Since Rényi divergence recovers KL divergence as $\gamma \to 1$ (Van Erven & Harremos, 2014), one can intuitively believe that Eq. (11) should perform at least not worse than Eq. (8).

Based on Theorem 4, we can formulate the distributionally augmented energy function $\hat{E}(\mathbf{x}; \boldsymbol{\theta}) = T \log \sum_{i=1}^K \exp\left[\hat{E}(\mathbf{x}, y_i; \boldsymbol{\theta})/T\right]$ under Rényi divergence as follows:

$$\begin{aligned} \hat{E}(\mathbf{x}; \boldsymbol{\theta}) &= T \log \sum_{i=1}^K \frac{\exp\left\{ \frac{1}{T} \cdot [h(\mathbf{x}, y_i) - \beta^*_{\mathbf{x}}] \right\}}{\exp\left\{ \frac{c_\gamma(\eta)}{T} \cdot \mathbb{E}_{\hat{y} \sim \mathbb{P}_{\hat{Y}}} \left[\left(h(\mathbf{x}, \hat{y}) - \beta^*_{\mathbf{x}} \right)_+^{\gamma^*} \right]^{\frac{1}{\gamma^*}} \right\}} \\ &\approx S_{\text{ours}}(\mathbf{x}; \boldsymbol{\theta}) \\ &\triangleq T \log \sum_{i=1}^K \frac{\exp\left\{ \frac{1}{T} \cdot [h(\mathbf{x}, y_i) - \beta^*_{\mathbf{x}}] \right\}}{\exp\left\{ \frac{c_\gamma(\eta)}{T} \cdot \left[\frac{1}{L} \sum_{j=1}^L \left(h(\mathbf{x}, \hat{y}_j) - \beta^*_{\mathbf{x}} \right)_+^{\gamma^*} \right]^{\frac{1}{\gamma^*}} \right\}}, \end{aligned} \tag{13}$$

In realization of $S_{\text{ours}}(\mathbf{x}; \boldsymbol{\theta})$ in Eq. (13), one requires to obtain $\beta^*_{\mathbf{x}}$ via solving the optimization problem in Eq. (12). While Eq. (12) does not have a closed-form solution, the convexity of Eq. (12) with regard to $\beta_{\mathbf{x}}$ (as proved in Appendix C) motivates us to find $\beta^*_{\mathbf{x}}$ via stochastic gradient descent (SGD) with a given learning rate lr, i.e.,

$$\beta_{\mathbf{x}} \leftarrow \beta_{\mathbf{x}} - lr \cdot \frac{\partial}{\partial \beta_{\mathbf{x}}} \left\{ c_\gamma(\eta) \left[\frac{1}{L} \sum_{j=1}^L \left(h(\mathbf{x}, \hat{y}_j) - \beta_{\mathbf{x}} \right)_+^{\gamma^*} \right]^{\frac{1}{\gamma^*}} + \beta_{\mathbf{x}} \right\}. \tag{14}$$

In the following, we disclose why Eq. (13) can be less vulnerable to the over-pessimism issue.

Theorem 5. *Let us define*

$$\mathbb{Q}^*_{\hat{Y}} = \arg\max_{\mathbb{Q}_{\hat{Y}}} \mathbb{E}_{\hat{y} \in \mathbb{Q}_{\hat{Y}}} [h(\mathbf{x}, \hat{y})] \quad \text{s.t. } D(\mathbb{Q}_{\hat{Y}} \| \mathbb{P}_{\hat{Y}}) \leq \eta.$$

If we choose $D(\cdot\|\cdot)$ *as the Cressie-Read family of Rényi divergence defined in Eq. (10), then we have* $q^*_{\hat{Y}}(\hat{y}) = \omega_\gamma(\mathbf{x}, \hat{y}) p_{\hat{Y}}(\hat{y})$*, where*

$$\omega_\gamma(\mathbf{x}, \hat{y}) \triangleq c_\gamma(\eta) \frac{\left(h(\mathbf{x}, \hat{y}) - \beta^*_{\mathbf{x}} \right)_+^{\frac{1}{\gamma-1}}}{\mathbb{E}_{\tilde{y} \sim \mathbb{P}_{\hat{Y}}} [\left(h(\mathbf{x}, \tilde{y}) - \beta^*_{\mathbf{x}} \right)_+^{\gamma^*}]^{\frac{1}{\gamma}}} \propto \left(h(\mathbf{x}, \hat{y}) - \beta^*_{\mathbf{x}} \right)_+^{\frac{1}{\gamma-1}}. \tag{15}$$

It can be found that the weight $\omega_\gamma(\mathbf{x}, \hat{y})$ in Eq. (15) acts as a polynomial function, therefore being relatively milder than $\omega_{KL}(\mathbf{x}, \hat{y})$ in Eq. (9). This tempers pessimism by flattening the effect of outliers: those with high cosine similarity to the input $\mathbf{x}$ still matter, but not disproportionately. Table 1 shows that the theoretical superiority (c.f. Theorem 5) indeed translates into strong empirical performance, where $S_{\text{ours}}(\mathbf{x}; \boldsymbol{\theta})$ significantly outperforms $\hat{S}_{\text{NegLabel}}(\mathbf{x}; f)$ and Energy (Eq. (4)).

Table 2: OOD detection results on ImageNet-1K with VIT B/16 CLIP as encoder. ↑ indicates larger values are better and vice versa. The best results in the last two columns are shown in bold.

Dataset	iNaturalist		Sun		Places		Textures		Average	
Metric	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Methods requiring training (or fine-tuning)										
MSP	87.44	58.36	79.73	73.72	79.67	74.41	79.69	71.93	81.63	69.61
ODIN	94.65	30.22	87.17	54.04	85.54	55.06	87.85	51.67	88.80	47.75
Energy	95.33	26.12	92.66	35.97	91.41	39.87	86.76	57.61	91.54	39.89
GradNorm	72.56	81.50	72.86	82.00	73.70	80.41	70.26	79.36	72.35	80.82
ViM	93.16	32.19	87.19	54.01	83.75	60.67	87.18	53.94	87.82	50.20
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	88.80	46.12	91.22	37.93
LSN	95.83	21.56	94.35	26.32	91.25	34.48	90.42	38.54	92.96	30.22
CLIPN	95.27	23.94	93.93	26.17	92.28	33.45	90.93	40.83	93.10	31.10
LoCoOp	96.86	16.05	95.07	23.44	91.98	32.87	90.19	42.28	93.52	28.66
LAPT	99.63	1.16	96.01	19.12	92.01	33.01	91.06	40.32	94.68	23.40
Zero-Shot Training-free Methods										
Mahalanobis	55.89	99.33	59.94	99.41	65.96	98.54	64.23	98.46	61.50	98.94
Energy	85.09	81.08	84.24	79.02	83.38	75.08	65.56	93.65	79.57	82.21
ZOC	86.09	87.30	81.20	81.51	83.39	73.06	76.46	98.90	81.79	85.19
MCM	94.59	32.20	92.25	38.80	90.31	46.20	86.12	58.50	90.82	43.93
NegLabel	99.49	1.91	95.49	20.53	91.64	35.59	90.22	43.56	94.21	25.40
Ours	99.64	1.29	95.71	17.94	91.90	33.98	90.79	39.45	**94.51**	**23.17**
NegLabel+AdaNeg	99.71	0.59	97.44	9.50	94.55	34.34	94.93	31.27	96.66	18.92
Ours+AdaNeg	99.75	0.47	98.01	8.69	95.63	30.24	95.86	27.28	**97.31**	**16.67**
NegLabel+CSP	99.60	1.54	96.66	13.66	92.90	29.32	93.86	25.52	95.76	17.52
Ours+CSP	99.70	1.32	97.52	11.35	94.89	24.98	94.16	23.65	**96.57**	**15.33**

6 EXPERIMENTS

Implementation. Unless otherwise specified, we employ CLIP-B/16 for zero-shot OOD detection. Following prior works (Jiang et al., 2024; Zhang & Zhang, 2024), we adopt the text prompt of 'The nice <label>.' and select $L = 10000$ negative labels from WordNet (Miller, 1995) using the same NegMining algorithm as NegLabel (Jiang et al., 2024). Notably, we show in Section 6.3 that our method can generalize well to various CLIP architectures and corpus sources. Regarding hyper-parameters, we set $T = 0.01$, $\gamma = 1.05$ and $c_\gamma(\eta) = 1.2$. We learn each input-specific constant $\beta^*_{\mathbf{x}}$ by performing SGD for only 15 steps with learning rate $lr = 1e-2$, which results in negligible computational overhead. Notably, we do not leverage the heuristic grouping strategy as described in Jiang et al. (2024). The reported results of our method are averaged over 5 independent runs.

Baselines. We compare our method with MSP (Hendrycks & Gimpel, 2016), ODIN (Liang et al., 2017), Energy (Liu et al., 2020), Gradnorm (Huang et al., 2021), Vim (Du et al., 2022), KNN (Sun et al., 2022), VOS (Tao et al., 2023), NPOS (Wang et al., 2023a), ZOC (Esmaeilpour et al., 2022), CLIPN (Wang et al., 2023a), LoCoOp (Miyai et al., 2024), LSN (Nie et al., 2024), LAPT (Zhang et al., 2025), Mahalanobis (Lee et al., 2018), MCM (Ming et al., 2022), NegLabel (Jiang et al., 2024), AdaNeg (Zhang & Zhang, 2024) and CSP (Chen et al., 2024).

6.1 MAIN RESULTS

Following prior work (Ming et al., 2022; Jiang et al., 2024; Chen et al., 2024; Zhang & Zhang, 2024), We evaluate our method on the popular ImageNet-1K benchmark (Deng et al., 2009), where the validation set of ImageNet-1K is designated as the ID dataset while iNaturalist (Van Horn et al., 2018), SUN (Xiao et al., 2010), Places365 (Zhou et al., 2017), and Textures (Cimpoi et al., 2014) are considered as OOD datasets. The methods listed in the upper section of Table 2, ranging from MSP (Hendrycks & Gimpel, 2016) to VOS (Tao et al., 2023), represent traditional visual OOD detection methods. Conversely, the methods in the lower section, extending from ZOC (Esmaeilpour et al., 2022) to NegLabel (Jiang et al., 2024), employ pre-trained VLMs like CLIP. Our method chieves the state-of-the-art on the ImageNet-1k benchmark, which highlights its superior performance in the zero-shot setting. Furthermore, our method can surpass traditional methods with a finetuned CLIP, demonstrating CLIP's strong OOD detection capabilities in zero-shot scenarios. This is because CLIP can parse images in a fine-grained manner, which is achieved through its pre-training on a large-scale image-text dataset.

Table 3: Evaluation on domain-generalizable OOD detection with VIT B/16 as encoder. ↑ indicates larger values are better and vice versa. The best results in the last two columns are shown in bold.

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	99.15	2.93	95.58	16.82	92.69	26.92	89.94	38.33	**94.34**	**24.25**
ImageNet-A	MCM	79.50	76.85	76.19	79.78	70.95	80.51	61.98	86.37	72.16	80.88
	NegLabel	98.80	4.09	89.83	44.38	82.88	60.10	80.25	64.34	87.94	43.23
	Ours	99.03	3.09	90.04	40.83	82.79	58.42	80.25	63.83	**88.03**	**41.54**
ImageNet-R	MCM	83.22	71.51	80.31	74.98	75.53	76.67	67.66	83.72	76.68	76.72
	NegLabel	99.58	1.60	96.03	15.77	91.97	29.48	90.60	35.67	94.54	20.63
	Ours	99.74	1.01	96.63	12.5.	92.90	27.25	92.06	32.42	**95.33**	**18.30**
ImageNetV2	MCM	91.79	45.90	89.88	50.73	86.52	56.25	81.51	69.57	87.43	55.61
	NegLabel	99.40	2.47	94.46	25.69	90.00	42.03	88.46	48.90	93.08	29.77
	Ours	99.64	1.35	94.72	23.20	89.93	42.05	44.77	89.43	**93.44**	**27.84**

Table 4: Evaluation on hard OOD detection, where a VIT B/16 CLIP encoder is adopted. ↑ indicates larger values are better and vice versa. The best results are shown in bold.

ID dataset	ImageNet-10		ImageNet-20		ImageNet-100		ImageNet-10	
OOD dataset	ImageNet-20		ImageNet-10		ImageNet-10		ImageNet-100	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
MCM	98.60	6.00	98.09	13.04	87.20	60.00	98.39	2.50
NegLabel	98.86	5.10	98.81	4.60	90.19	40.20	99.51	1.68
Ours	**99.12**	**3.75**	**99.27**	**2.35**	**91.64**	**34.36**	**99.65**	**1.17**

6.2 Extensions

Domain-generalizable OOD Detection. With ImageNet-1K as a case study, we, following Jiang et al. (2024), consider ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a) and ImageNetV2 (Recht et al., 2019) as ID data respectively. The experiment results on four OOD datasets are shown in Table 3. It is apparent that the performance of MCM significantly deteriorates across diverse domain shifts, indicating the difficulty of OOD detection under such conditions. NegLabel achieves remarkably better performances than MCM, thus demonstrating the significance of introducing negative labels for OOD detection. Our method consistently outperform NegLabel cross diverse ID datasets, which implies stronger robustness of our method against domain shifts.

Hard OOD Detection. Following prior works (Ming et al., 2022; Jiang et al., 2024), we alternate using ImageNet-10 and ImageNet-20 as ID and OOD data, as well as using ImageNet-10 and ImageNet-100 to mimic the setting in Fort et al. (2021) with high-resolution images. The results in Table 4 show that our method consistently outperforms MCM and NegLabel in all settings, demonstrating that our method has strong discriminative power for semantically hard OOD data.

Table 5: OOD detection results with different CLIP architectures on ImageNet-1k as ID. ↑ indicates larger values are better and vice versa. The best results in the last two columns are shown in bold.

Backbone	Method	iNaturalist		SUN		Places		Textures		Average	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
ViT-B/32	MCM	92.68	40.49	89.95	47.83	88.10	51.47	85.98	60.04	89.96	49.96
	NegLabel	99.11	3.73	95.27	22.48	91.72	34.94	88.57	50.51	93.67	27.92
	Ours	99.47	2.10	95.60	19.36	91.94	32.83	89.62	44.87	**94.16**	**24.79**
ViT-L/14	MCM	93.58	36.80	92.80	36.77	90.90	41.35	85.05	61.70	90.58	44.16
	NegLabel	99.53	1.77	95.63	22.33	93.01	32.22	89.71	42.92	94.47	24.81
	Ours	99.67	1.22	96.06	19.15	93.16	30.57	90.42	38.32	**94.83**	**22.31**
ResNet50	MCM	91.88	42.97	89.31	52.84	84.12	65.75	85.55	62.15	87.71	55.93
	NegLabel	99.24	2.88	94.54	26.51	89.72	42.60	88.40	50.80	92.97	30.70
	Ours	99.54	1.48	94.61	24.58	89.69	41.64	90.23	42.70	**93.52**	**27.60**

Table 6: OOD detection results with different input resolution on ImageNet-1k as ID, where a VIT L/14 CLIP encoder is adopted. ↑ indicates larger values are better and vice versa. The best results in the last two columns are shown in bold.

Resolution	Method	iNaturalist		SUN		Places		Textures		Average	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
336×336	NegLabel	99.71	1.12	95.68	21.84	93.15	31.79	90.55	40.46	94.77	23.80
	Ours	99.72	1.09	96.17	18.24	93.39	29.62	90.67	37.68	**94.99**	**21.66**

Table 7: OOD detection results with different corpus sources on ImageNet-1k as ID, where a VIT B/16 CLIP encoder is adopted. ↑ indicates larger values are better and vice versa. The best results in the last two columns are shown in bold.

Corpus	Method	iNaturalist		SUN		Places		Textures		Average	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Common	NegLabel	86.91	65.43	95.03	24.22	91.52	34.83	83.69	67.75	89.29	48.06
	Ours	88.77	57.76	94.46	25.57	91.70	34.18	85.15	60.04	**90.02**	**44.64**
Part-of-Speech	NegLabel	99.23	3.25	94.20	25.93	90.17	43.09	87.77	50.11	92.84	30.59
	Ours	99.42	2.46	94.82	23.29	91.75	39.48	91.59	41.86	**94.40**	**26.77**

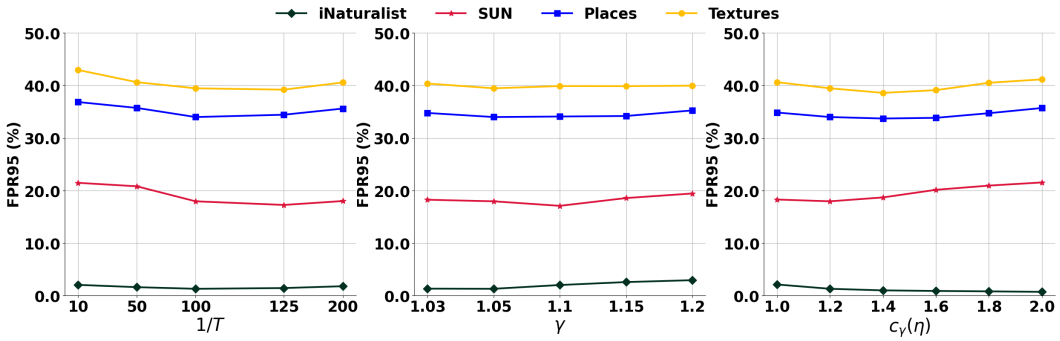

Figure 1: Hyper-parameter analysis on ImageNet-1K w.r.t T (left), γ (middle), and $c_\gamma(\eta)$ (right).

6.3 Ablation Study

Architectures. In principle, our method is generic to the choice of visual encoder. We evaluate our method with different visual encoder architectures, including ViT-B/32, ViT-L/14 and ResNet-50, and report the corresponding OOD detection results in Table 5. On the one hand, the performance of OOD detection can be enhanced by more powerful visual encoders. On the other hand, our method consistently outperforms the most recent NegLabel regardless of the backbone architecture used, which implies the better generalization of our method over NegLabel.

Input Size. In principle, our method is generic to the input resolution. We evaluate our method with a larger input size, i.e., 336×336, and report the corresponding OOD detection results in Table 6. On the one hand, the performance of OOD detection can be enhanced by a larger input size. On the other hand, our method consistently outperforms the most recent NegLabel regardless of the input resolution, which implies the better generalization of our method over NegLabel.

Corpus Sources. The role of the corpus is to provide a larger and more comprehensive semantic space. While our method is generic to the input resolution, we also conduct ablative analysis with different corpus sources, including Part-of-Speech Tags and Common-20K. As for Part-of-Speech Tags, we, following NegLabel (Jiang et al., 2024), randomly sample 70000 words to constitute the corpus source. It can be found that Table 7 that our method consistently outperforms NegLabel on multiple corpora, which implies that the flexibility of our method.

Hyper-parameter Analysis. We evaluate the hyper-parameters most essential to our design, including the temperature T, the order γ, and $c_\gamma(\eta)$. The corresponding results are plotted in Figure 1.

7 CONCLUSION

This work presents a distributionally-augmented energy-based framework to provide a novel perspective on CLIP-based OOD detection with Negative labels. We show that existing methods in this direction essentially estimate the energy function against a worst-case distribution within a KL-divergence ball, thereby tolerating sampling bias between observed negative labels and real OOD labels. We also identify the inherent over-pessimism of KL-based formulations. In response, we propose a Rényi-divergence–based refinement for a more flexible and balanced worst-case distribution, achieving state-of-the-art results in various setups of OOD detection.

ETHICS STATEMENT

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

REPRODUCIBILITY STATEMENT

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

REFERENCES

Amirhossein Ansari, Ke Wang, and Pulei Xiong. Negrefine: Refining negative label-based zero-shot ood detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 573–582, 2025.

Aharon Ben-Tal and Marc Teboulle. An old-new concept of convex risk measures: the optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.

Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. *arXiv preprint arXiv:2306.00826*, 2023.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

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.

Mengyuan Chen, Junyu Gao, and Changsheng Xu. Conjugated semantic pool improves ood detection with pre-trained vision-language models. *Advances in Neural Information Processing Systems*, 37:82560–82593, 2024.

Zixiang Chen, Yihe Deng, Yuanzhi Li, and Quanquan Gu. Understanding transferable representation learning and zero-shot transfer in clip. *arXiv preprint arXiv:2310.00927*, 2023.

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

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

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

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

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.

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*, 2024a.

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

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

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.

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

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.

Louis Faury, Ugo Tanielian, Elvis Dohmatob, Elena Smirnova, and Flavian Vasile. Distributionally robust counterfactual risk minimization. In *AAAI*, pp. 3850–3857. AAAI Press, 2020.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in neural information processing systems*, 34:7068–7081, 2021.

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

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

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

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021a.

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, 2021b.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

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.

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.

Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided ood detection with pretrained vision-language models. *arXiv preprint arXiv:2403.20078*, 2024.

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

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

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

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.

Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

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

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

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

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.

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.

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

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

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

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

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

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 7. Granada, 2011.

Andre T Nguyen, Fred Lu, Gary Lopez Munoz, Edward Raff, Charles Nicholas, and James Holt. Out of distribution data detection using dropout bayesian neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7877–7885, 2022.

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

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

Bo Peng, Jie Lu, Guangquan Zhang, and Zhen Fang. An information-theoretical framework for understanding out-of-distribution detection with pretrained vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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

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

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, 2021a.

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, 2021b.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pp. 547–562. University of California Press, 1961.

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.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

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.

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

Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

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.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? 2021.

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.

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, 2023a.

Qizhou Wang, Zhen Fang, Yonggang Zhang, Feng Liu, Yixuan Li, and Bo Han. Learning to augment distributions for out-of-distribution detection. *Advances in neural information processing systems*, 36:73274–73286, 2023b.

Qizhou Wang, Junjie Ye, Feng Liu, Quanyu Dai, Marcus Kalander, Tongliang Liu, Jianye Hao, and Bo Han. Out-of-distribution detection with implicit outlier transformation. *arXiv preprint arXiv:2303.05033*, 2023c.

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.

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

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. *Journal of Data-centric Machine Learning Research*, 2024a. ISSN XXXX-XXXX. URL https://openreview.net/forum?id=cnnTnJQigs. Dataset Certification.

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

Yabin Zhang, Wenjie Zhu, Chenhang He, and Lei Zhang. Lapt: Label-driven automated prompt tuning for ood detection with vision-language models. In *European Conference on Computer Vision*, pp. 271–288. Springer, 2025.

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

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

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

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

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.

A Related Work on Traditional OOD Detection

The popularity of OOD detection is motivated by the empirical observation (Nguyen et al., 2015) that neural networks tend to be over-confident in OOD data.

One line of work performs OOD detection by devising post-hoc scoring functions, including confidence-based methods (Hendrycks et al., 2019; Ming et al., 2022; Zhang & Xiang, 2023), energy-based methods (Liu et al., 2020), distance-based approaches (Lee et al., 2018; Sun et al., 2022; Sohn, 2016; Morteza & Li, 2022; Peng et al., 2024), gradient-based approaches (Huang et al., 2021), and Bayesian approaches (Kristiadi et al., 2020; Malinin & Gales, 2019). Another line of work addresses OOD detection by fine-tuning a pre-trained discrimination model with training-time regularizations that help the model learn ID/OOD discrepancy following the guideline of outlier exposure (Hendrycks et al., 2018). For instance, the discriminative model is regularized to produce lower confidence (Lee et al., 2017; Malinin & Gales, 2018), smaller feature magnitudes (Liu et al., 2020) or higher energy (Dhamija et al., 2018) for outlier points. More recently, some works have considered a practical scenario where the auxiliary outliers can be arbitrarily different from the real OOD data, therefore distributionally augmenting the observed OOD data. Besides, the given OOD samples tend to include unlabelled ID counterparts (Katz-Samuels et al., 2022). Because of this, WOOD (Katz-Samuels et al., 2022) formulates learning with noisy OOD samples as a constrained optimization problem while SAL (Du et al., 2024a) separates candidate outliers from the unlabeled data and then trains a binary classifier using the candidate outliers and the labelled ID data.

On the theoretical side, there are various attempts to explore the theoretical understanding of OOD detection. Fang et al. (2022; 2024) study the generalization of OOD detection by PAC learning and find a necessary condition for the learnability of OOD detection. Du et al. (2024a) provides a provable understanding of the OOD detection result by modeling the feature space as a mixture of multivariate Gaussian distributions. Peng et al. (2024) weakens distributional assumption from Gaussian distribution to exponential family distribution. Du et al. (2024b) studies the impact of ID labels on OOD detection.

B Proofs of Main Theorems

B.1 Proof of Theorem 3

Proof. We consider the following optimization problem

$$\mathbb{Q}^*_{\hat{Y}} = \arg\max_{\mathbb{Q}_{\hat{Y}}} \mathbb{E}_{\hat{y}\in\mathbb{Q}_{\hat{Y}}}\left[h(\mathbf{x},\hat{y})\right] \quad \textit{s.t. } D_{KL}(\mathbb{Q}_{\hat{Y}},\mathbb{P}_{\hat{Y}}) = \int q_{\hat{Y}}(\hat{y})\log\frac{q_{\hat{Y}}(\hat{y})}{p_{\hat{Y}}(\hat{y})}d\hat{y} \le \eta.$$

Introducing multipliers $\alpha \ge 0$ for the KL constraint and δ for normalization $\int q_{\hat{Y}}(\hat{y})d\hat{y} = 1$:

$$\mathcal{L} = \int q_{\hat{Y}}(\hat{y})h(\mathbf{x},\hat{y})\,d\hat{y} + \alpha\Big(\eta - \int q_{\hat{Y}}(\hat{y})\log\frac{q_{\hat{Y}}(\hat{y})}{p_{\hat{Y}}(\hat{y})}\,d\hat{y}\Big) + \delta\Big(1 - \int q_{\hat{Y}}(\hat{y})\,d\hat{y}\Big). \tag{16}$$

Note that $\mathcal{L}$ both depend on $\mathbb{Q}_{\hat{Y}}, \mathbf{x}, \alpha, \delta$, but we suppress the dependence from the notation for simplicity.

Taking the functional derivative with respect to $q_{\hat{Y}}(\hat{y})$ gives

$$\frac{\partial\mathcal{L}}{\partial q_{\hat{Y}}(\hat{y})} = h(\mathbf{x},\hat{y}) - \alpha\Big(\log\frac{q_{\hat{Y}}(\hat{y})}{p_{\hat{Y}}(\hat{y})} + 1\Big) - \delta.$$

Stationarity requires $\frac{\partial\mathcal{L}}{\partial q_{\hat{Y}}(\hat{y})} = 0$, hence

$$\log\frac{q^*_{\hat{Y}}(\hat{y})}{p_{\hat{Y}}(\hat{y})} = \frac{h(\mathbf{x},\hat{y}) - \delta - \alpha}{\alpha}.$$

Exponentiating yields

$$q^*_{\hat{Y}}(\hat{y}) = p_{\hat{Y}}(\hat{y})\exp\left(\frac{h(\mathbf{x},\hat{y}) - \delta - \alpha}{\alpha}\right) \propto p_{\hat{Y}}(\hat{y})\exp\left(\frac{h(\mathbf{x},\hat{y})}{\alpha}\right).$$

Replacing α with the optimal $\alpha^*(\mathbf{x},\mathbb{P}_{\hat{Y}}) = \arg\min_{\alpha\ge0}\left\{\alpha\eta + \alpha\log\mathbb{E}_{\hat{y}\sim\mathbb{P}_{\hat{Y}}}[e^{h(\mathbf{x},\hat{y})/\alpha}]\right\}$ yields

$$q^*_{\hat{Y}}(\hat{y}) \propto p_{\hat{Y}}(\hat{y})\exp\left(\frac{h(\mathbf{x},\hat{y})}{\alpha^*(\mathbf{x},\mathbb{P}_{\hat{Y}})}\right).$$

□

B.2 Proof of Theorem 1

Proof. Let $\ell(\hat{y}) = q_{\hat{Y}}(\hat{y})/p_{\hat{Y}}(\hat{y})$ and $\varphi(a) = a\log a - a + 1$, then we have

$$\int q_{\hat{Y}}(\hat{y}) h(\mathbf{x}, \hat{y})\, d\hat{y} = \mathbb{E}_{\mathbb{P}_{\hat{Y}}}[h(\mathbf{x}, \hat{y})\ell(\hat{y})]$$

$$\int q_{\hat{Y}}(\hat{y}) \log \frac{q_{\hat{Y}}(\hat{y})}{p_{\hat{Y}}(\hat{y})}\, d\hat{y} = \mathbb{E}_{\mathbb{P}_{\hat{Y}}}[\varphi(\ell(\hat{y}))]$$

$$\int q_{\hat{Y}}(\hat{y})\, d\hat{y} = \mathbb{E}_{\mathbb{P}_{\hat{Y}}}[\ell(\hat{y})]$$

According to Eq. (16), we can rewrite $\hat{E}(\mathbf{x}, y_j; \boldsymbol{\theta})$ in Eq. (5) as follows:

$$\begin{aligned}
\hat{E}(\mathbf{x}, y_j; \boldsymbol{\theta}) &= h(\mathbf{x}, y_j) - \min_{\alpha\geq 0,\delta} \max_{\mathbb{Q}_{\hat{Y}}} \mathcal{L} \\
&= h(\mathbf{x}, y_j) - \min_{\alpha\geq 0,\delta} \max_{\mathbb{Q}_{\hat{Y}}} \{\mathbb{E}_{\mathbb{P}_{\hat{Y}}}[h(\mathbf{x}, \hat{y})\ell(\hat{y})] - \alpha[\mathbb{E}_{\mathbb{P}_{\hat{Y}}}[\varphi(\ell(\hat{y}))] - \eta] + \delta(\mathbb{E}_{\mathbb{P}_{\hat{Y}}}[\ell(\hat{y})] - 1)\} \\
&= h(\mathbf{x}, y_j) - \min_{\alpha\geq 0,\delta} \{\alpha\eta - \delta + \alpha \max_{\mathbb{Q}_{\hat{Y}}} \{\mathbb{E}_{\mathbb{P}_{\hat{Y}}}[\frac{h(\mathbf{x}, \hat{y}) + \delta}{\alpha}\ell(\hat{y}) - \varphi(\ell(\hat{y}))]\}\} \qquad (17) \\
&= h(\mathbf{x}, y_j) - \min_{\alpha\geq 0,\delta} \{\alpha\eta - \delta + \alpha\mathbb{E}_{\mathbb{P}_{\hat{Y}}}[\max_{\ell(\hat{y})} \{\frac{h(\mathbf{x}, \hat{y}) + \delta}{\alpha}\ell(\hat{y}) - \varphi(\ell(\hat{y}))\}]\} \qquad (18) \\
&= h(\mathbf{x}, y_j) - \min_{\alpha\geq 0,\delta} \{\alpha\eta - \delta + \alpha\mathbb{E}_{\mathbb{P}_{\hat{Y}}}[\varphi^*(\frac{h(\mathbf{x}, \hat{y}) + \delta}{\alpha})]\} \qquad (19) \\
&= h(\mathbf{x}, y_j) - \min_{\alpha\geq 0,\delta} \{\alpha\eta - \delta + \alpha\mathbb{E}_{\mathbb{P}_{\hat{Y}}}[e^{\frac{h(\mathbf{x}, \hat{y}) + \delta}{\alpha}} - 1]\} \qquad (20) \\
&= h(\mathbf{x}, y_j) - \min_{\alpha\geq 0} \{\alpha\eta + \alpha\log \mathbb{E}_{\mathbb{P}_{\hat{Y}}}[e^{\frac{h(\mathbf{x}, \hat{y})}{\alpha}}]\} \qquad (21) \\
&= \alpha^*(\mathbf{x}, \mathbb{P}_{\hat{Y}}) \log \frac{e^{h(\mathbf{x}, \hat{y})/\alpha^*(\mathbf{x}, \mathbb{P}_{\hat{Y}})}}{\mathbb{E}_{\mathbb{P}_{\hat{Y}}}[e^{h(\mathbf{x}, \hat{y})/\alpha^*(\mathbf{x}, \mathbb{P}_{\hat{Y}})}]} - \alpha^*(\mathbf{x}, \mathbb{P}_{\hat{Y}})\eta,
\end{aligned}$$

where $\alpha^*(\mathbf{x}, \mathbb{P}_{\hat{Y}}) = \arg\min_{\alpha\geq 0} \{\alpha\eta + \alpha\log \mathbb{E}_{\hat{y}\sim\mathbb{P}_{\hat{Y}}}[e^{h(\mathbf{x}, \hat{y})/\alpha}]\}$.

We kindly note that 1) Eq. (17) holds due to the strong duality (Boyd & Vandenberghe, 2004); 2) Eq. (18) is derived via a re-arrangement for optimizing over $\mathbb{P}_{\hat{Y}}$; 3) the derivation of Eq. (19) follows by Ben-Tal & Teboulle (2007); 4) Eq. (20) is established based on the definition of convex conjugate (Hiriart-Urruty & Lemaréchal, 2004), i.e., $\varphi^*(a) = e^a - 1$.

To prove Eq. (21), Fix $\alpha > 0$ and minimize Eq. (20) over δ gives R#gfTd

$$\min_{\delta} \left[\alpha\eta - \delta + \alpha\mathbb{E}_{P_{\hat{Y}}}[e^{(h(\mathbf{x}, \hat{y}) + \delta)/\alpha} - 1]\right].$$

Expand and separate δ gives

$$\alpha\eta - \delta + \alpha\left(e^{\delta/\alpha}\mathbb{E}_{P_{\hat{Y}}}[e^{h(\mathbf{x}, \hat{y})/\alpha}] - 1\right) = \alpha\eta - \alpha + \underbrace{(\alpha R e^{\delta/\alpha} - \delta)}_{=:g(\delta)},$$

where $R := \mathbb{E}_{P_{\hat{Y}}}[e^{h(\mathbf{x}, \hat{y})/\alpha}] > 0$ (for simplicity, we omit the dependence on $\mathbf{x}$, $P_{\hat{Y}}$ and α).

Compute the derivative of $g(\delta)$ w.r.t. δ and set it to zero:

$$g'(\delta) = Re^{\delta/\alpha} - 1 = 0 \to \delta^* = -\alpha\log R.$$

Since $g''(\delta) = \frac{1}{\alpha}Re^{\delta/\alpha} > 0$, δ^* gives the minimum, i.e.,

$$\min_{\delta} g(\delta) = g(\delta^*) = \alpha R \cdot \frac{1}{R} - (-\alpha\log R) = \alpha + \alpha\log R.$$

Therefore, for fixed α,

$$\min_{\delta} \left[\alpha\eta - \delta + \alpha\mathbb{E}_{P_{\hat{Y}}}[e^{(h(\mathbf{x}, \hat{y}) + \delta)/\alpha} - 1]\right] = \alpha\eta - \alpha + (\alpha + \alpha\log R) = \alpha\eta + \alpha\log \mathbb{E}_{P_{\hat{Y}}}[e^{h(\mathbf{x}, \hat{y})/\alpha}],$$

such that

$$h(\mathbf{x}, y_j) - \min_{\alpha\geq 0,\delta} \left[\alpha\eta - \delta + \alpha\mathbb{E}_{P_{\hat{Y}}}[e^{\frac{h(\mathbf{x}, \hat{y}) + \delta}{\alpha}} - 1]\right] = h(\mathbf{x}, y_j) - \min_{\alpha\geq 0} \left[\alpha\eta + \alpha\log \mathbb{E}_{P_{\hat{Y}}}[e^{h(\mathbf{x}, \hat{y})/\alpha}]\right]$$

□

B.3 Proof of Theorem 4

Proof. We consider the following optimization problem

$$\mathbb{Q}^*_{\hat{Y}} = \arg\max_{\mathbb{Q}_{\hat{Y}}} \mathbb{E}_{\hat{y}\in\mathbb{Q}_{\hat{Y}}}\left[h(\mathbf{x},\hat{y})\right] \quad s.t.\ D_\gamma(\mathbb{Q}_{\hat{Y}},\mathbb{P}_{\hat{Y}}) = \int q_{\hat{Y}}(\hat{y})\phi_\gamma\left(\frac{q_{\hat{Y}}(\hat{y})}{p_{\hat{Y}}(\hat{y})}\right)d\hat{y} \le \eta, \tag{22}$$

where $\phi_\gamma(t) = \frac{1}{\gamma(\gamma-1)}(t^\gamma - \gamma t + \gamma - 1)$ with $\gamma > 1$.

Introducing multipliers $\alpha \ge 0$ for the Rényi constraint and δ for normalization $\int q_{\hat{Y}}(\hat{y})d\hat{y} = 1$:

$$\begin{aligned}\mathcal{L}_\gamma &= \int q_{\hat{Y}}(\hat{y})h(\mathbf{x},\hat{y})\,d\hat{y} + \alpha\Big(\eta - \int q_{\hat{Y}}(\hat{y})\phi_\gamma\left(\frac{q_{\hat{Y}}(\hat{y})}{p_{\hat{Y}}(\hat{y})}\right)d\hat{y}\Big) + \delta\Big(1 - \int q_{\hat{Y}}(\hat{y})\,d\hat{y}\Big)\\ &= \mathbb{E}_{\mathbb{P}_{\hat{Y}}}[h(\mathbf{x},\hat{y})\ell(\hat{y})] + \alpha\Big(\eta - \mathbb{E}_{\mathbb{P}_{\hat{Y}}}[\phi_\gamma(\ell(\hat{y}))]\Big) + \delta\Big(1 - \mathbb{E}_{\mathbb{P}_{\hat{Y}}}[\ell(\hat{y})]\Big),\end{aligned} \tag{23}$$

where $\ell(\hat{y}) = q_{\hat{Y}}(\hat{y})/p_{\hat{Y}}(\hat{y})$. Note that $\mathcal{L}_\gamma$ both depend on $\mathbb{Q}_{\hat{Y}}, \mathbf{x}, \alpha, \delta$, but we suppress the dependence from the notation for simplicity.

Then solving Eq. (22) is equivalent to solving the following problem:

$$\begin{aligned}&\min_{\alpha\ge0,\delta}\max_{\mathbb{Q}_{\hat{Y}}}\mathcal{L}_\gamma\\ &= \min_{\alpha\ge0,\delta}\max_{\mathbb{Q}_{\hat{Y}}}\mathbb{E}_{\mathbb{P}_{\hat{Y}}}[h(\mathbf{x},\hat{y})\ell(\hat{y})] + \alpha\Big(\eta - \mathbb{E}_{\mathbb{P}_{\hat{Y}}}[\phi_\gamma(\ell(\hat{y}))]\Big) + \delta\Big(1 - \mathbb{E}_{\mathbb{P}_{\hat{Y}}}[\ell(\hat{y})]\Big)\\ &= \min_{\alpha\ge0,\delta}\left\{\alpha\eta + \delta + \alpha\max_{\mathbb{Q}_{\hat{Y}}}\mathbb{E}_{\mathbb{P}_{\hat{Y}}}\left[\frac{h(\mathbf{x},\hat{y})-\delta}{\alpha}\ell(\hat{y}) - \phi_\gamma(\ell(\hat{y}))\right]\right\}\\ &= \min_{\alpha\ge0,\delta}\left\{\alpha\eta + \delta + \alpha\mathbb{E}_{\mathbb{P}_{\hat{Y}}}\left[\max_{\ell(\hat{y})}\left(\frac{h(\mathbf{x},\hat{y})-\delta}{\alpha}\ell(\hat{y}) - \phi_\gamma(\ell(\hat{y}))\right)\right]\right\}\end{aligned} \tag{24}$$

Note that $\max_{\ell(\hat{y})}\left\{\frac{h(\mathbf{x},\hat{y})-\delta}{\alpha}\ell(\hat{y}) - \phi_\gamma(\ell(\hat{y}))\right\} = \phi^*_\gamma\left(\frac{h(\mathbf{x},\hat{y})-\delta}{\alpha}\right)$ is the Fenchel Conjugate Function of $\phi_\gamma(\frac{h(\mathbf{x},\hat{y})-\delta}{\alpha})$, then we have $\phi^*_\gamma(a) = \frac{1}{\gamma}\left((\gamma-1)a+1\right)^{\gamma^*}_+ - \frac{1}{\gamma}$ with $\gamma^* = \frac{\gamma}{\gamma-1}$. Please refer to Duchi & Namkoong (2021) for more details. Then Eq. (24) can be rewritten as follows:

$$\begin{aligned}&\min_{\alpha\ge0,\delta}\left\{\alpha\eta + \delta + \alpha\mathbb{E}_{\mathbb{P}_{\hat{Y}}}\left[\max_{\ell(\hat{y})}\left(\frac{h(\mathbf{x},\hat{y})-\delta}{\alpha}\ell(\hat{y}) - \phi_\gamma(\ell(\hat{y}))\right)\right]\right\}\\ &= \min_{\alpha\ge0,\delta}\left\{\alpha\eta + \delta + \alpha\mathbb{E}_{\mathbb{P}_{\hat{Y}}}\left[\phi^*_\gamma\left(\frac{h(\mathbf{x},\hat{y})-\delta}{\alpha}\right)\right]\right\}.\end{aligned} \tag{25}$$

Following Duchi & Namkoong (2021), with $\beta = \delta - \frac{\alpha}{\gamma-1}$, we can arrive at the closed-form formulation of the optimal α^* that minimizes Eq. (25) as follows:

$$\alpha^* = (\gamma-1)(\gamma(\gamma-1)\eta+1)^{-\frac{1}{\gamma^*}}\mathbb{E}_{\mathbb{P}_{\hat{Y}}}\left[(h(\mathbf{x},\hat{y})-\beta)^{\gamma^*}_+\right]^{\frac{1}{\gamma^*}}, \tag{26}$$

By substituting α^*, β and $\phi^*_\gamma\left(\frac{h(\mathbf{x},\hat{y})-\delta}{\alpha}\right)$ into Eq. (25), we have:

$$\max_{\mathbb{Q}_{\hat{Y}}:D_\gamma(\mathbb{Q}_{\hat{Y}}\|\mathbb{P}_{\hat{Y}})\le\eta}\mathbb{E}_{\hat{y}\in\mathbb{Q}_{\hat{Y}}}\left[h(\mathbf{x},\hat{y})\right] = \min_{\alpha\ge0,\delta}\max_{\mathbb{Q}_{\hat{Y}}}\mathcal{L}_\gamma = \min_\beta\left\{c_\gamma(\eta)\mathbb{E}_{\hat{y}\sim\mathbb{P}_{\hat{Y}}}\left[(h(\mathbf{x},\hat{y})-\beta)^{\gamma^*}_+\right]^{\frac{1}{\gamma^*}} + \beta\right\}, \tag{27}$$

where $c_\gamma(\eta) = (\gamma(\gamma-1)\eta+1)^{\frac{1}{\gamma}}$, such that

$$\begin{aligned}\hat{E}(\mathbf{x},y_j;\boldsymbol{\theta}) =& h(\mathbf{x},y_j) - \max_{\mathbb{Q}_{\hat{Y}}:D_\gamma(\mathbb{Q}_{\hat{Y}}\|\mathbb{P}_{\hat{Y}})\le\eta}\mathbb{E}_{\hat{y}\in\mathbb{Q}_{\hat{Y}}}\left[h(\mathbf{x},\hat{y})\right]\\ =& h(\mathbf{x},y_j) - \left\{c_\gamma(\eta)\mathbb{E}_{\hat{y}\sim\mathbb{P}_{\hat{Y}}}\left[\left(h(\mathbf{x},\hat{y})-\beta^*_{\mathbf{x}}\right)^{\gamma^*}_+\right]^{\frac{1}{\gamma^*}} + \beta^*_{\mathbf{x}}\right\},\end{aligned} \tag{28}$$

where

$$\beta^*_{\mathbf{x}} = \arg\min_\beta\left\{c_\gamma(\eta)\mathbb{E}_{\hat{y}\sim\mathbb{P}_{\hat{Y}}}\left[(h(\mathbf{x},\hat{y})-\beta)^{\gamma^*}_+\right]^{\frac{1}{\gamma^*}} + \beta\right\}. \tag{29}$$

□

B.4 Proof of Theorem 5

Proof. Taking the functional derivative of $\frac{h(\mathbf{x},\hat{y})-\delta}{\alpha}\ell(\hat{y}) - \phi_\gamma\left(\ell(\hat{y})\right)$ in Eq. (24) w.r.t. $q_{\hat{Y}}(\hat{y})$ gives

$$\begin{aligned}
&\frac{\partial}{\partial \ell(\hat{y})}\left\{\frac{h(\mathbf{x},\hat{y})-\delta}{\alpha}\ell(\hat{y}) - \phi_\gamma\left(\ell(\hat{y})\right)\right\} \\
=&\frac{h(\mathbf{x},\hat{y})-\delta}{\alpha} - \frac{\partial \phi_\gamma\left(\ell(\hat{y})\right)}{\partial \ell(\hat{y})} \\
=&\frac{h(\mathbf{x},\hat{y})-\delta}{\alpha} - \frac{1}{\gamma-1}\left[\ell(\hat{y})^{\gamma-1}-1\right].
\end{aligned} \tag{30}$$

Stationarity requires

$$\frac{\partial}{\partial \ell(\hat{y})}\left\{\frac{h(\mathbf{x},\hat{y})-\delta}{\alpha}\ell(\hat{y}) - \phi_\gamma\left(\ell(\hat{y})\right)\right\} = 0,$$

hence

$$h(\mathbf{x},\hat{y}) - \delta = \frac{\alpha}{\gamma-1}\left[\ell^*(\hat{y})\right]^{\gamma-1},$$

where $\ell^*(\hat{y}) = q^*_{\hat{Y}}(\hat{y})/p_{\hat{Y}}(\hat{y})$. Replacing α with the optimal α^*, then we have:

$$\ell^*(\hat{y}) = c_\gamma(\eta)\frac{\left(h(\mathbf{x},\hat{y})-\beta^*_{\mathbf{x}}\right)_+^{\frac{1}{\gamma-1}}}{\mathbb{E}_{\tilde{y}\sim\mathbb{P}_{\hat{Y}}}\left[\left(h(\mathbf{x},\tilde{y})-\beta^*_{\mathbf{x}}\right)_+^{\gamma^*}\right]^{\frac{1}{\gamma}}},$$

such that

$$q^*_{\hat{Y}}(\hat{y}) = c_\gamma(\eta)\frac{\left(h(\mathbf{x},\hat{y})-\beta^*_{\mathbf{x}}\right)_+^{\frac{1}{\gamma-1}}}{\mathbb{E}_{\tilde{y}\sim\mathbb{P}_{\hat{Y}}}\left[\left(h(\mathbf{x},\tilde{y})-\beta^*_{\mathbf{x}}\right)_+^{\gamma^*}\right]^{\frac{1}{\gamma}}} p_{\hat{Y}}(\hat{y}) \propto \left(h(\mathbf{x},\hat{y})-\beta^*_{\mathbf{x}}\right)_+^{\frac{1}{\gamma-1}} p_{\hat{Y}}(\hat{y}).$$

□

C Convexity with regard to β

Let $\zeta(\beta) = c_\gamma(\eta)\mathbb{E}_{\hat{y}\sim\mathbb{P}_{\hat{Y}}}\left[\left(h(\mathbf{x},\hat{y})-\beta\right)_+^{\gamma^*}\right]^{\frac{1}{\gamma^*}} + \beta$, then we have:

$$\begin{aligned}
&\zeta(\delta\beta_1 + (1-\delta)\beta_2) \\
=&\delta\beta_1 + (1-\delta)\beta_2 + c_\gamma(\eta)\mathbb{E}_{\hat{y}\sim\mathbb{P}_{\hat{Y}}}\left[\left(\delta(h(\mathbf{x},\hat{y})-\beta_1) + (1-\delta)(h(\mathbf{x},\hat{y})-\beta_2)\right)_+^{\gamma^*}\right]^{\frac{1}{\gamma^*}}
\end{aligned} \tag{31}$$

According to Minkowski's inequality, we have:

$$\begin{aligned}
&\mathbb{E}_{\hat{y}\sim\mathbb{P}_{\hat{Y}}}\left[\left(\delta\cdot(h(\mathbf{x},\hat{y})-\beta_1) + (1-\delta)\cdot(h(\mathbf{x},\hat{y})-\beta_2)\right)_+^{\gamma^*}\right]^{\frac{1}{\gamma^*}} \\
\leq&\delta\cdot\mathbb{E}_{\hat{y}\sim\mathbb{P}_{\hat{Y}}}\left[\left(h(\mathbf{x},\hat{y})-\beta_1\right)_+^{\gamma^*}\right]^{\frac{1}{\gamma^*}} + (1-\delta)\cdot\mathbb{E}_{\hat{y}\sim\mathbb{P}_{\hat{Y}}}\left[\left(h(\mathbf{x},\hat{y})-\beta_2\right)_+^{\gamma^*}\right]^{\frac{1}{\gamma^*}},
\end{aligned} \tag{32}$$

such that

$$\zeta(\delta\beta_1 + (1-\delta)\beta_2) \leq \delta\zeta(\beta_1) + (1-\delta)\zeta(\beta_2). \tag{33}$$

D 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.

E LIMITATION

Our framework only provides a theoretical explanation for CLIP-based OOD detection with negative labels, leaving a crucial gap in the success of methods in other directions.

F EVALUATION METRIC

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. The reported results of our method are averaged over 5 independent runs.

G USAGE CLAIM OF LLMS

We use ChatGPT for grammar and spelling checks only, with the prompt "Proofread the sentences".

H ALGORITHMIC SUMMARY

R#gfTd

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

Algorithm 1:

Input : Test-time input $\mathbf{x}$, ID labels $\mathcal{Y}_{\text{I}} = \{y_1, \ldots, y_K\}$, Negative labels $\hat{\mathcal{Y}} = \{\hat{y}_1, \ldots, \hat{y}_L\}$, learning rate lr, critic $h(\cdot, \cdot)$, maximum iteration M, hyper-parameters $\gamma^*, c_\gamma(\eta)$

Output: OOD scoring function $S_{\text{ours}}(\mathbf{x}; \boldsymbol{\theta})$

// Step 1: Optimizing $\beta_{\mathbf{x}}$

for $iter = 1$ **to** M **do**

$$\beta_{\mathbf{x}} \leftarrow \beta_{\mathbf{x}} - lr \cdot \frac{\partial}{\partial \beta_{\mathbf{x}}} \left\{ c_\gamma(\eta) \left[\frac{1}{L} \sum_{j=1}^{L} \left(h(\mathbf{x}, \hat{y}_j) - \beta_{\mathbf{x}} \right)_+^{\gamma^*} \right]^{\frac{1}{\gamma^*}} + \beta_{\mathbf{x}} \right\}.$$

// Step 2: OOD Scoring

$$S_{\text{ours}}(\mathbf{x}; \boldsymbol{\theta}) = T \log \sum_{i=1}^{K} \frac{\exp\left\{ \frac{1}{T} \cdot [h(\mathbf{x}, y_i) - \beta_{\mathbf{x}}] \right\}}{\exp\left\{ \frac{c_\gamma(\eta)}{T} \cdot \left[\frac{1}{L} \sum_{j=1}^{L} \left(h(\mathbf{x}, \hat{y}_j) - \beta_{\mathbf{x}} \right)_+^{\gamma^*} \right]^{\frac{1}{\gamma^*}} \right\}}$$

I QUANTITATIVE ANALYSIS ON COMPUTATION TIME

R#gfTd, R#GPbo, R#44Et, R#KbNX

As shown in Algorithm 1, our method consists of two stages: 1) Optimizing $\beta_{\mathbf{x}}$ and 2) OOD scoring.

As for the first stage, we start with deriving the specific form of the gradient w.r.t $\beta_{\mathbf{x}}$ as follows:

$$1 - c_\gamma(\eta) \phi(\beta)^{\frac{1}{\gamma^*} - 1} \frac{1}{L} \sum_{j=1}^{L} \mathbf{1}\left(h(\mathbf{x}, \hat{y}_j) > \beta \right) \left(h(\mathbf{x}, \hat{y}_j) - \beta \right)_+^{\gamma^* - 1},$$

where $\phi(\beta) = \frac{1}{L} \sum_{j=1}^{L} (h(\mathbf{x}, \hat{y}_j) - \beta)_+^{\gamma^*}$.

Clearly, the time complexity of gradient computation is $O(L)$, which, same as NegLabel, linearly grows with the number of negative labels. Here we omit the complexity introduced by the dot-product computation, as it is orthogonal to our algorithmic design. Therefore, the time complexity of the first stage which involve M-step SGD is $O(M \cdot L)$, where the maximum iteration M is usually set to a relative small value (e.g., $M = 15$ in this paper)

Table 8: Detailed OOD detection results of our method on the OpenOOD benchmark, where ImageNet-1K is adopted as ID dataset.

Ours (Stage 1)	Ours (Stage 2)	Ours (Stage 1+2)	NegLabel (with grouping strategy)
1.51ms	0.12ms	1.63ms	0.14ms

As for the second step, one can easily check that the time complexity of our proposed scoring function $S_{\text{ours}}(\mathbf{x};\boldsymbol{\theta})$ is $O(K+L)$, which is same as that of $S_{\text{NegLabel}}(\mathbf{x};\boldsymbol{\theta})$ in Eq. (2).

Table 8 reports the average computation time of our method and NegLabel per test-time input on a single NVIDIA A100. Note that we omit the computation cost introduced by extracting features from pre-trained CLIP, as our method keeps the same feature extraction procedure as NegLabel.

Although the optimization process adds roughly 1.5 ms per test-time input, this cost is small in absolute terms. Therefore, both theoretically and empirically, the additional computation is limited and does not contradict our claim that the extra cost is negligible in practice.

J More Discussions

J.1 Connection between $S_{\text{NegLabel}}(\mathbf{x};f)$ and $\hat{S}_{\text{NegLabel}}(\mathbf{x};f)$

R#GPbo

For any $\alpha \in [0,1]$, the α-skew negative label distribution $\hat{\mathbb{P}}_{\hat{Y}}$ is defined as

$$\hat{\mathbb{P}}_{\hat{Y}} = \alpha \mathbb{P}_{Y_{\text{I}}} + (1-\alpha)\mathbb{P}_{\hat{Y}},$$

where $\mathbb{P}_{Y_{\text{I}}}$ is the ID label distribution. Replacing $\hat{\mathbb{P}}_{\hat{Y}}$ with $\mathbb{P}_{\hat{Y}}$ in Eq. (8) gives

$$\begin{aligned}
& T\log\sum_{i=1}^{K}\frac{e^{h(\mathbf{x},y_j)/T}}{\mathbb{E}_{\hat{y}\sim\hat{\mathbb{P}}_{\hat{Y}}}[e^{h(\mathbf{x},\hat{y})/T}]} - T\eta \\
=& T\log\sum_{i=1}^{K}\frac{e^{h(\mathbf{x},y_j)/T}}{\alpha\mathbb{E}_{\hat{y}\sim\mathbb{P}_{Y_{\text{I}}}}[e^{h(\mathbf{x},\hat{y})/T}] + (1-\alpha)\mathbb{E}_{\hat{y}\sim\mathbb{P}_{\hat{Y}}}[e^{h(\mathbf{x},\hat{y})/T}]} - T\eta \\
\approx& T\log\sum_{i=1}^{K}\frac{\exp\left[h(\mathbf{x},y_i)/T\right]}{\frac{\alpha}{K}\sum_{j=1}^{K}\exp\left[h(\mathbf{x},y_j)/T\right] + \frac{1-\alpha}{L}\sum_{j=1}^{L}\exp\left[h(\mathbf{x},\hat{y}_j)/T\right]} - T\eta.
\end{aligned}$$

With $\alpha = \frac{K}{K+L}$, we have

$$\begin{aligned}
& \log\sum_{i=1}^{K}\frac{\exp\left[h(\mathbf{x},y_i)/T\right]}{\frac{\alpha}{K}\sum_{j=1}^{K}\exp\left[h(\mathbf{x},y_j)/T\right] + \frac{1-\alpha}{L}\sum_{j=1}^{L}\exp\left[h(\mathbf{x},\hat{y}_j)/T\right]} \\
=& \underbrace{\log\sum_{i=1}^{K}\frac{\exp\left[h(\mathbf{x},y_i)/T\right]}{\sum_{j=1}^{K}\exp\left[h(\mathbf{x},y_j)/T\right] + \sum_j=1^{L}\exp\left[h(\mathbf{x},\hat{y}_j)/T\right]}}_{S_{\text{NegLabel}}(\mathbf{x};f)} + \log(K+L)
\end{aligned} \tag{34}$$

The above implies that $S_{\text{NegLabel}}(\mathbf{x};f)$ essentially estimates a worst-case energy function over a KL-divergence-constrained set, therefore being functionally equivalent to $\hat{S}_{\text{NegLabel}}(\mathbf{x};f)$ in a broader sense. In a narrow sense, $S_{\text{NegLabel}}(\mathbf{x};f)$ can be considered as a slightly noisy version of $\hat{S}_{\text{NegLabel}}(\mathbf{x};f)$ (up to a constant) with the noise level $\alpha = \frac{1000}{1000+10000} \approx 0.09$.

J.2 Comparison with Test-time Adaptation OOD Detection

R#GPbo

We notice that, superficially, the use of 15-step SGD in test-time somewhat resembles the idea of test-time adaptation (TTA). However, we draw a conceptual distinction based on what is being updated and what the goal is:

Table 9: OOD detection results on the OpenOOD benchmark, where CIFAR-100 is adopted as ID dataset. Full results are provided in Table 10.

Methods	FPR95 ↓		AUROC ↑	
	Near-OOD	Far-OOD	Near-OOD	Far-OOD
MCM	75.20	59.32	71.00	76.00
NegLabel	71.44	40.92	70.58	89.68
Ours	**68.12**	**34.81**	**72.84**	**92.95**

Table 10: Detailed OOD detection results of our method on the OpenOOD benchmark, where CIFAR-100 is adopted as ID dataset.

Settings	Datasets	FPR95 ↓	AUROC ↑
Near-OOD	CIFAR-10	69.59	64.76
	TIN	66.65	80.92
	Average	68.12	72.84
Far-OOD	MNIST	10.05	96.84
	SVHN	13.24	96.97
	Texture	35.32	93.74
	Places	80.63	84.25
	Average	34.81	92.95

TTA OOD DETECTION (E.G., ADANEG)

- **What is updated:** Negative proxies that are shared by test-time inputs.
- **Scope of update:** Updation is based on test-time inputs and can **accumulate over time**.
- **Goal:** Explicitly **adapt the negative proxies** to the test distribution.
- **Statefulness:** The negative proxies and/or their internal state after processing earlier test-time inputs **influence predictions on future test-time inputs**.

OUR METHOD

- **What is updated:** We **only optimize a scalar variable $\beta_{\mathbf{x}}$ per input**, which is *not* shared across test-time inputs.
- **Scope of update:** The optimization is **strictly local and input-dependent**. For each $\mathbf{x}$, $\beta_{\mathbf{x}}$ is reinitialized and optimized from scratch. There is no cross-sample sharing.
- **Goal:** We are **not** adapting negative labels to a new distribution. Instead, we are computing an **input-specific optimum** of a fixed energy function that was fully specified by the pre-trained CLIP model and the pre-computed negative labels.
- **Statefulness:** There is **no persistent state** that changes over time. OOD scoring for $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent since optimizing $\beta_{\mathbf{x}_1}$ does not influence $\beta_{\mathbf{x}_2}$ or any future $\beta_{\mathbf{x}}$.

K EXPERIMENT ON CIFAR-100 BENCHMARK

R#gfTd, R#GPbo

Besides ImageNet, we also assess our method on the smaller CIFAR-100 dataset (Krizhevsky et al., 2009) under the OpenOOD setup (Zhang et al., 2024a). Specifically, we utilize Near-OOD datasets including CIFAR-10 (Krizhevsky et al., 2009) and Tiny-ImageNet (TIN) (Le & Yang, 2015),and Far-OOD datasets including MNIST (Deng, 2012), SVHN (Netzer et al., 2011), Texture (Cimpoi et al., 2014), and Places (Zhou et al., 2017). As illustrated in Table 9, our advantage still holds.

L MORE EXPERIMENT ON IMAGENET-1K BENCHMARK

R#GPbo, R#44Et

Table 11: OOD detection results on the OpenOOD benchmark, where ImageNet-1K is adopted as ID dataset. Full results are provided in Table 12.

Methods	FPR95 ↓		AUROC ↑	
	Near-OOD	Far-OOD	Near-OOD	Far-OOD
MCM	79.02	68.54	60.11	84.77
NegLabel	69.45	23.73	75.18	94.85
Ours	**68.09**	**21.50**	**75.65**	**95.35**

Table 12: Detailed OOD detection results of our method on the OpenOOD benchmark, where ImageNet-1K is adopted as ID dataset.

Settings	Datasets	FPR95 ↓	AUROC ↑
Near-OOD	SSB-hard	70.11	76.04
	NINCO	66.07	75.26
	Average	68.09	75.65
Far-OOD	iNaturalist	1.29	99.64
	Texture	39.45	90.79
	OpenImage-O	23.75	95.62
	Average	21.50	95.35

Our method is extensively evaluated against a range of OOD datasets on the OpenOOD benchmark. Specifically, with ImageNet-1K as the ID dataset, we utilize Near-OOD datasets including SSB-hard (Vaze et al., 2021) and NINCO (Bitterwolf et al., 2023),and Far-OOD datasets including iNaturalist (Van Horn et al., 2018), Textures (Cimpoi et al., 2014), and OpenImage-O (Wang et al., 2022). As illustrated in Table 11, our advantage still holds. Moreover, as shown in Table 13, our method not only demonstrates high distinguish ability against semantic shifts but also exhibits strong robustness to covariate shifts.

M Ablation Study on Text Prompts

R#44Et

We evaluate our method with different prompt templates on the ImageNet-1k benchmark, as shown in Table 15.

N Ablation Study on Negative Labels

R#44Et

We explore the impact of the number of selected negative labels on OOD detection. The results shown in Table 16 demonstrate that our method consistently outperforms NegLabel with different number of negative labels and therefore is not sensitive with the number of negative labels.

We also investigate the impact the strategy of mining negative labels on OOD detection. The results shown in Table 17 demonstrate that our method consistently outperforms NegLabel with mining negative and therefore is not sensitive with the the quality of negative-label mining.

Table 13: Full-spectrum OOD detection results on the OpenOOD benchmark, where ImageNet-1K is adopted as ID dataset. Full results are provided in Table 14.

Methods	FPR95 ↓		AUROC ↑	
	Near-OOD	Far-OOD	Near-OOD	Far-OOD
MCM	85.37	69.87	58.97	77.11
NegLabel	76.25	33.30	72.77	92.02
Ours	**73.78**	**29.65**	**74.18**	**93.61**

Table 14: Detailed full-spectrum OOD detection results of our method on the OpenOOD benchmark, where ImageNet-1K is adopted as ID dataset.

Settings	Datasets	FPR95 ↓	AUROC ↑
Near-OOD	SSB-hard	75.29	73.92
	NINCO	72.27	74.44
	Average	73.78	74.18
Far-OOD	iNaturalist	2.32	99.35
	Texture	48.36	88.40
	OpenImage-O	38.27	93.08
	Average	29.65	93.61

Table 15: OOD detection results with different prompt templates on ImageNet-1k as ID. ↑ indicates larger values are better and vice versa. The best results in the last two columns are shown in bold.

Prompt	**Method**	**iNaturalist**		**SUN**		**Places**		**Textures**		**Average**	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
A photo of a <label>.	NegLabel	99.59	1.74	94.83	26.35	90.17	46.92	80.79	72.11	91.34	36.78
	Ours	99.59	1.62	95.18	22.94	91.09	41.87	83.70	65.76	**92.39**	**33.05**
<label>.	NegLabel	99.52	1.91	95.77	19.32	92.43	32.79	86.89	59.34	93.65	28.34
	Ours	99.60	1.51	96.02	17.17	93.59	29.65	89.35	55.90	**94.71**	**26.06**

O MORE ABLATION STUDY ON BACKBONES

R#44Et

Table 18 assesses the performance of our method using ResNet50x16 as the backbone to examine whether performance gains diminish when the base model already separates ID/OOD well.

Table 16: Impact of the number of negative labels L selected by NegMining (Jiang et al., 2024) from WordNet on ImageNet-1k benchmark. ↑ indicates larger values are better and vice versa. The best results in the last two columns are shown in bold.

L	Method	iNaturalist		SUN		Places		Textures		Average	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
5000	NegLabel	99.64	1.39	94.83	23.28	91.28	37.58	89.97	44.26	93.93	26.63
	Ours	99.54	1.74	95.34	19.51	91.82	34.05	90.47	41.76	**94.30**	**24.26**
20000	NegLabel	99.18	3.16	95.24	21.47	91.16	37.23	89.83	44.40	93.85	26.56
	Ours	99.47	2.78	95.55	19.23	92.19	34.62	90.59	42.07	**94.45**	**24.68**

Table 17: Impact of mining strategy of $L = 10000$ negative labels from WordNet on ImageNet-1k benchmark. ↑ indicates larger values are better and vice versa. The best results in the last two columns are shown in bold.

Strategy	Method	iNaturalist		SUN		Places		Textures		Average	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Random Selection	NegLabel	97.96	9.11	93.93	28.67	89.54	45.10	86.62	55.87	92.01	34.69
	Ours	98.40	7.60	94.91	24.78	92.23	38.59	89.42	50.26	**93.74**	**30.31**
NegRefine (Ansari et al., 2025)	NegLabel	99.57	1.51	94.64	22.93	90.42	39.10	94.69	21.15	94.83	21.17
	Ours	99.70	1.03	95.28	16.13	91.78	36.53	95.09	20.39	**95.46**	**18.52**

Table 18: OOD detection results with ResNet50x16 or ViT-G/14 as the backbone on ImageNet-1k. ↑ indicates larger values are better and vice versa. The best results in the last two columns are shown in bold.

Backbone	Method	iNaturalist		SUN		Places		Textures		Average	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
ResNet50x16	NegLabel	99.48	2.00	94.18	29.11	88.85	48.14	91.23	38.74	93.43	29.50
	Ours	99.63	1.56	95.06	25.35	90.75	44.56	91.84	36.21	**94.32**	**26.92**
ViT-G/14	NegLabel	99.70	1.15	96.17	18.01	93.21	29.37	90.26	39.96	94.84	22.12
	Ours	99.78	0.82	96.59	15.13	93.96	26.53	91.07	36.54	**95.35**	**19.75**