OPTLABEL: LABEL OPTIMIZATION FOR NEGATIVE-LABEL-BASED OOD DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Negative-label-based methods have emerged as a strong paradigm for CLIP-based out-of-distribution (OOD) detection. However, existing methods are largely text-driven, defining positive labels directly from in-distribution (ID) class names and selecting negative labels based on textual dissimilarity from ID classes. Consequently, their positive and negative labels may poorly represent the ID distribution, and OOD samples close to the ID boundary remain underrepresented. To address these limitations, we propose *OptLabel* that optimizes the representation of positive and negative labels under the ID-supervised setting. OptLabel treats the positive and negative labels from existing zero-shot negative-label-based methods as a semantic prior and refines their representation using the available ID data and the near-boundary OOD samples synthesized by applying Gaussian perturbation to ID samples in the CLIP embedding space. We show that, in the high-dimensional CLIP space, the placement of these synthesized OOD samples has an intuitive angular interpretation, and is thus easily controllable. OptLabel is a lightweight solution by keeping the CLIP frozen and requiring no image generation, external models, or auxiliary datasets. On the ImageNet-1K benchmark, OptLabel outperforms SynOOD, the strongest ID-supervised baseline, by 2.41 AUROC and 6.51 FPR95 points. The source code is available in the Supplementary Material.

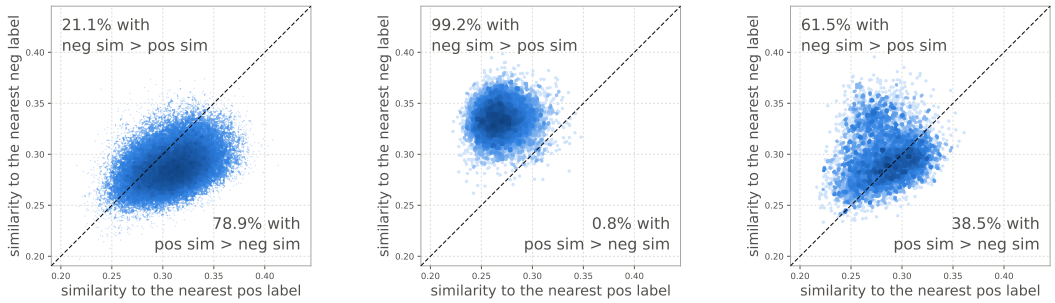
1 INTRODUCTION

In real-world environments, machine learning models inevitably encounter inputs that do not belong to any of the categories defined during training. Unfortunately, predictions on these unknown out-of-distribution (OOD) samples can be overconfident and unreliable, leading to critical risks and security vulnerabilities. OOD detection addresses this issue by identifying such inputs so that they can be flagged rather than incorrectly classified as one of the in-distribution (ID) classes (Hendrycks & Gimpel, 2017). Recent vision-language models such as CLIP (Radford et al., 2021) provide a promising foundation for OOD detection by leveraging not only visual representations of input images, but also semantic information from textual class labels, going beyond traditional methods that rely solely on the visual modality. Among CLIP-based OOD detection approaches, negative-label-based methods such as NegLabel (Jiang et al., 2024), CSP (Chen et al., 2024), NegRefine (Ansari et al., 2025), and ANTS (Zhu et al., 2026) have emerged as a particularly effective paradigm.

The main idea of negative-label-based methods is to treat the ID classes as positive labels and construct a set of negative labels from a large vocabulary, such as WordNet, by selecting concepts that are semantically distant from the ID classes. An input image is then classified as ID or OOD based on its CLIP similarity to the positive and negative labels: higher similarity to the positive labels relative to the negative labels indicates ID. By leveraging CLIP's powerful vision-language representations and the diverse semantics available in WordNet, these methods have outperformed previous few-shot and supervised approaches while operating in the zero-shot setting. Despite their strong performance, most existing negative-label-based methods remain entirely *text-driven*: the positive labels are derived directly from the textual ID class names, while the negative labels are selected based on their textual semantic similarity to the ID classes. This text-driven construction can limit ID/OOD separability in two important ways, as illustrated in Figure 1:

- **Limitation (i): ID samples may not be well represented by the positive and negative labels.** The positive labels, which are the ID class names, are often general and may not align sufficiently

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



(a) ImageNet-1K (ID) (b) iNaturalist (far-OOD) (c) NINCO (near-OOD)

Figure 1: Comparison of each image's similarity to its nearest positive and negative labels. Points above the dashed diagonal are more strongly matched by a negative label than by a positive label. The positive and negative labels are from NegRefine (Ansari et al., 2025).

with actual ID images. Also, the broad and noisy terms in WordNet can produce negative labels that align strongly with ID images, leading to misdetection of ID as OOD. For example, as shown in Figure 1a, 21.1% of ID images from ImageNet-1K are more strongly matched by a negative label than by any positive label.

- **Limitation (ii): Challenging near-boundary OOD samples are underrepresented by negative labels.** Since negative labels are deliberately selected to be semantically distant from ID class labels, they tend to represent OOD samples far from the ID distribution (far-boundary OOD samples) rather than those near its boundary (near-boundary OOD samples). As shown in Figure 1b, 99.2% of far-boundary OOD samples from iNaturalist are more strongly matched by a negative label than by any positive label, but this fraction drops substantially to 61.5% for near-boundary OOD samples from NINCO as shown in Figure 1c.

These mismatches translate into poor ID/OOD separability: for example, NegRefine achieves an FPR95 of 1.47% on far-boundary OOD dataset iNaturalist, but this increases sharply to 61.96% on near-boundary OOD dataset NINCO.

One way to address these limitations is to leverage ID training images that are readily available in many practical settings and provide richer supervision than class labels alone. This motivates extending the negative-label-based OOD detection in the zero-shot setting to the underexplored ID-supervised setting (Politowicz et al., 2026). However, ID supervision does not provide information for what should be rejected as OOD, and unlike ID data, the OOD space is open-ended and obtaining representative OOD samples is challenging. SynOOD (Li et al., 2025) addresses this challenge by synthesizing OOD images from non-ID background concepts present in the training ID images. This strategy provides only limited OOD coverage and depends on the diversity of non-ID concepts present in the ID images. Our experiments show that SynOOD's performance degrades when the ID dataset contains fewer classes.

Our Contributions. In this paper, we study negative-label-based OOD detection in the ID-supervised setting, shifting from a text-driven label paradigm toward a *data-driven* one. To this end, we introduce *OptLabel*, a framework that leverages the positive and negative labels from previous zero-shot negative-label-based methods as a semantic prior and optimizes them using the available ID training data and carefully synthesized OOD samples. Our key contributions are:

- At the core of OptLabel is a novel optimization procedure that treats the positive and negative label embeddings from existing zero-shot negative-label-based methods as learnable variables and further optimizes them in two stages. The first stage optimizes the representation of positive labels to align with the ID data, addressing Limitation (i). The second stage, with the aligned positive labels frozen, optimizes the representation of negative labels using both ID data and synthesized OOD samples, moving the negative labels closer to the ID boundary to better capture challenging near-boundary OOD samples. This addresses Limitation (ii).
- We synthesize near-boundary OOD samples directly from the ID data in CLIP's embedding space using Gaussian perturbations. Our method can probe the ID boundary in arbitrary directions and is

not dependent on non-ID concepts in training ID images. We show that, in the high-dimensional CLIP space, the synthesized samples exhibit a surprisingly simple and controllable geometric structure where a single angular parameter is sufficient without additional filtering. Moreover, one fixed parameter value performs well across popular OOD benchmarks.

- OptLabel is lightweight and efficient. It optimizes the label embeddings while keeping the CLIP encoders frozen, and its OOD synthesis requires no image generation, external models, or auxiliary image datasets.
- On the large-scale ImageNet-1K benchmark, OptLabel achieves state-of-the-art performance, outperforming SynOOD (Li et al., 2025), the strongest ID-supervised baseline, by 2.41 AUROC points and 6.51 FPR95 points. Our method also outperforms all previous zero-shot, few-shot, and test-time adaptation methods by a large margin.

In summary, OptLabel demonstrates its value through a complete design: strong semantic label initialization, simple and easily controllable OOD synthesis, separate optimization objectives, frozen positive labels during OOD training, and consistent gains at lightweight training cost.

2 Related Work

Zero-Shot Methods. In this setting, the ID class labels are given as input, and methods mainly differ in how they go beyond simply matching images against the ID labels. MCM (Ming et al., 2022a) uses the maximum softmax probability over ID-label similarities as the score. ZOC (Esmaeilpour et al., 2022) develops a captioning model to generate unseen labels. CLIPN (Wang et al., 2023) extends CLIP to understand negative prompts "not containing". GL-MCM (Miyai et al., 2025) extends MCM by using CLIP's local features. NegLabel (Jiang et al., 2024) introduces the negative-label-based paradigm by selecting a large number of negative labels from WordNet (Fellbaum, 1998) distant from ID classes. Proven effective, this paradigm was refined by subsequent works: CSP (Chen et al., 2024) expands the semantic pool using superclass labels, LAPT (Zhang et al., 2024) learns prompt tokens using images generated or retrieved from an external dataset, and NegRefine (Ansari et al., 2025) filters negative labels with an LLM and introduces a multi-matching score. Our work considers the ID-supervised setting.

Few-Shot Methods. These methods adapt to few-shot ID data through prompt learning: optimizing continuous context tokens preceding the ID class name in prompts such as "a photo of a [class]" given as input to CLIP's text encoder. LoCoOp (Miyai et al., 2023) and SCT (Yu et al., 2024) tune prompts to classify ID images while maximizing entropy on irrelevant regions of ID images. FA (Lu et al., 2025) focuses only on ID to learn prompts with richer ID representations. LSN (Nie et al., 2024) and NegPrompt (Li et al., 2024) additionally learn negative prompts by using other ID classes as OOD or encouraging negative prompts to be dissimilar to ID images. Local-Prompt (Zeng et al., 2025) learns local positive and negative prompts from selected ID image crops. These works optimize the context tokens in a prompt, instead of the classes themselves. Our method treats the labels from negative-label-based methods as the prior and directly optimizes the label embeddings, and this optimization takes place in CLIP's output space without further invoking the text encoder. For OOD supervision, we synthesize virtual OOD samples in the CLIP space rather than relying on local ID image features or crops, whose coverage is limited to content present in ID images.

ID-Supervised Methods. This setting assumes the availability of ID training data and is the setting in our work. CASOD (Politowicz et al., 2026), built on a pretrained ViT rather than CLIP, uses a segmenter to extract foreground and background views for learning better ID representations. NPOS (Tao et al., 2023) fine-tunes the last two blocks of CLIP's image encoder using OOD samples synthesized by applying Gaussian perturbations to selected boundary ID samples and retaining OOD samples far from ID data. We develop a geometric analysis showing that boundary ID sample selection and OOD sample filtering are unnecessary in the high-dimensional CLIP space, enabling more effective parameterization and synthesis. SynOOD (Li et al., 2025) optimizes a subset of negative labels by synthesizing OOD images using a separate model to identify non-ID elements in ID images and training a projection layer on top of the image encoder. We optimize the full sets of both positive and negative labels, and our OOD synthesis probes the ID boundary in arbitrary directions, and thus is not limited to concepts present in ID image backgrounds and does not require an external model. Our optimization is lightweight and efficient because it is performed in the output space of CLIP without further invoking CLIP.

Test-Time Adaptation Methods. These methods dynamically adapt the prediction function during inference using unlabeled test samples to improve predictions on subsequent samples. Several recent works extend the negative-label-based paradigm to this setting (Zhang & Zhang, 2024; Xu et al., 2026; Zhu et al., 2026; Zhang et al., 2026). Such test-time methods are orthogonal to our method.

3 PRELIMINARIES

OOD Detection Problem. OOD detection is a binary decision problem that aims to determine whether an input belongs to one of the ID classes Y or is OOD, i.e., does not belong to any ID class. An OOD detector defines an ID score $S(x) \in [0, 1]$. At test time, a threshold γ is applied to $S(x)$: x is predicted as ID if $S(x) \geq \gamma$, and as OOD otherwise. This threshold is usually set such that a large percentage, e.g., 95%, of ID data are detected as ID. Section 2 reviewed different settings.

Negative-Label-Based OOD Detection. The core idea of zero-shot negative-label-based methods (Jiang et al., 2024) is to treat the textual ID class labels Y as positive labels and create negative labels from WordNet concepts distant from the ID classes. Let E_{img} and E_{txt} denote the CLIP image and text encoders, which map inputs into a shared d-dimensional embedding space. Let $Y_{\text{pos}} = \{y_1^+, \ldots, y_K^+\}$ and $Y_{\text{neg}} = \{y_1^-, \ldots, y_M^-\}$ denote the text embeddings of the positive and negative labels, respectively. Given an input image x, let $z = E_{\text{img}}(x)/\|E_{\text{img}}(x)\|_2$ denote its normalized image embedding. The ID score is then defined as:

$$S(x) = \frac{\sum_{k=1}^{K} e^{\text{sim}(z, y_k^+)/\tau}}{\sum_{k=1}^{K} e^{\text{sim}(z, y_k^+)/\tau} + \sum_{m=1}^{M} e^{\text{sim}(z, y_m^-)/\tau}}, \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and τ is a temperature parameter. Higher similarity to the positive labels relative to the negative labels yields a higher ID score. Since S depends on x only through z, we write $S(z)$ when the score is computed directly from an embedding.

4 METHOD

We consider OOD detection in the ID-supervised setting, where the ID class labels Y and labeled ID training data $\mathcal{D}_{\text{ID}} = \{(x_i, c_i)\}_{i=1}^N$ are available, with c_i denoting the ground-truth class index of sample x_i. OptLabel treats the text-derived positive and negative label embeddings Y_{pos} and Y_{neg}, obtained from existing zero-shot negative-label-based methods, as learnable parameters and optimizes them to address the two limitations (i) and (ii) discussed in the Introduction. OptLabel uses the given $\mathcal{D}_{\text{ID}}$ and a carefully synthesized near-boundary OOD set $\mathcal{D}_{\text{OOD}} = \{\tilde{z}_i\}_{i=1}^N$ to optimize Y_{pos} and Y_{neg} through a two-stage procedure. The following subsections describe the two-stage label optimization and the near-boundary OOD synthesis method.

4.1 TWO-STAGE LABEL OPTIMIZATION

The goal is to optimize the label embeddings $Y_{\text{pos}} = \{y_1^+, \ldots, y_K^+\}$ and $Y_{\text{neg}} = \{y_1^-, \ldots, y_M^-\}$, initialized from existing zero-shot negative-label-based methods, directly in CLIP's embedding space while keeping the image and text encoders frozen. This optimization is done in two stages. In the first stage, we optimize the positive labels using only the ID training data $\mathcal{D}_{\text{ID}}$, addressing the limitation (i). In the second stage, we freeze the optimized positive labels while optimizing the negative labels using both $\mathcal{D}_{\text{ID}}$ and the synthesized OOD data $\mathcal{D}_{\text{OOD}}$, addressing the limitation (ii). This two-stage design finds a better representation of positive labels using high quality actual ID data and ensures that synthesized OOD samples, which provide only surrogate OOD supervision, influence only the negative labels. Each stage uses a single objective aligned with the intended role of the corresponding label set:

Stage 1: Positive-Label Optimization. We optimize Y_{pos} using $\mathcal{D}_{\text{ID}} = \{(x_i, c_i)\}_{i=1}^N$. Let z_i denote the normalized CLIP embedding of an ID sample x_i with ground-truth class index c_i, such that $y_{c_i}^+$ is its corresponding positive-label embedding. Treating the positive-label embeddings in Y_{pos} as learnable parameters, we minimize the standard cross-entropy loss to align each positive

label with the ID samples from its corresponding class:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\mathrm{sim}(z_i, y_{c_i}^{+})/\tau}}{\sum_{k=1}^{K} e^{\mathrm{sim}(z_i, y_k^{+})/\tau}}. \tag{2}$$

Stage 2: Negative-Label Optimization. In this stage, we freeze the optimized Y_{pos} and update only Y_{neg} using both $\mathcal{D}_{\mathrm{ID}} = \{(x_i, c_i)\}_{i=1}^{N}$ and $\mathcal{D}_{\mathrm{OOD}} = \{\tilde{z}_i\}_{i=1}^{N}$. We formulate this stage as a binary ID/OOD classification task and combine $\mathcal{D}_{\mathrm{ID}}$ and $\mathcal{D}_{\mathrm{OOD}}$ into $\mathcal{D} = \{(h_i, l_i)\}_{i=1}^{2N}$, where each h_i is either an ID sample embedding z_i with label $l_i = 1$ or a synthesized OOD $\tilde{z}_i$ with label $l_i = 0$. We minimize the binary cross-entropy loss over the ID score $S(\cdot)$ in Equation 1, computed from the frozen positive labels Y_{pos} and learnable negative labels Y_{neg}:

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{2N}\sum_{i=1}^{2N}[\, l_i \log S(h_i) + (1 - l_i)\log\big(1 - S(h_i)\big)\,]. \tag{3}$$

The initial negative labels are far from the ID boundary. This optimization tends to move the negative labels closer to the boundary to better capture the challenging near-boundary OOD samples.

Our optimizations are lightweight and efficient: in both stages, we update only the label embeddings, CLIP encoders remain frozen, and all ID and synthesized OOD representations can be precomputed and cached once, avoiding repeated encoder evaluations during optimization. At inference time, with our optimized positive and negative labels, the detection score $S(\cdot)$ in Equation 1 provides improved ID/OOD separability.

4.2 Near-Boundary OOD Synthesis

We focus on synthesizing challenging near-boundary OOD samples that are particularly relevant for improving ID/OOD separability (Du et al., 2022). Following NPOS (Tao et al., 2023), we synthesize OOD samples directly in the CLIP embedding space by applying Gaussian perturbation to individual ID image representations. This allows us to probe the neighborhood of each ID sample in arbitrary directions using only the available ID data, without requiring an external image generator or auxiliary dataset. Unlike NPOS, which restricts synthesis to boundary ID samples and applies k-NN filtering to retain samples in low-density regions, we perturb every ID representation and retain all synthesized samples. We justify this simpler design through a geometric analysis showing that, in the high-dimensional CLIP embedding space, the synthesized samples concentrate at a controllable angular distance from their ID anchors while moving away from their local ID neighborhoods.

Gaussian OOD Synthesis. Let $z_i \in \mathbb{R}^d$ denote the normalized CLIP image embedding of an ID training sample $x_i \in \mathcal{D}_{\mathrm{ID}}$, where d is the CLIP embedding dimension. For each z_i, we sample an isotropic Gaussian perturbation and re-normalize the perturbed feature onto the unit hypersphere:

$$\tilde{z}_i = \frac{z_i + \epsilon_i}{\|z_i + \epsilon_i\|_2}, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2 I_d). \tag{4}$$

Applying this procedure to all N ID samples yields the surrogate OOD set $\mathcal{D}_{\mathrm{OOD}} = \{\tilde{z}_i\}_{i=1}^{N}$. We re-normalize each perturbed feature because CLIP compares normalized image and text features on the unit hypersphere, where cosine similarity reduces to their dot product. The parameter σ controls how far the synthesized samples $\tilde{z}_i$ are displaced from their ID anchors z_i.

Concentration on an Angular Shell. In a high dimensional space, the norm of a Gaussian perturbation concentrates sharply such that most perturbations lie on a thin shell at approximately the same distance from the origin. In particular, for $\epsilon_i \sim \mathcal{N}(0, \sigma^2 I_d)$, we have $\|\epsilon_i\|_2 \approx \sigma\sqrt{d}$; see (Blum et al., 2020, Theorem 2.9) or (Vershynin, 2026, Theorem 3.1.1). The embedding dimension for CLIP ViT-B/16 is $d = 512$, which is commonly regarded as high dimensional (Yu et al., 2026). This result, however, does not directly describe the displacement of the final synthesized representation $\tilde{z}_i$, which is further normalized.

Since CLIP compares normalized representations using cosine similarity, this displacement is more naturally characterized by the angle θ between $\tilde{z}_i$ and its anchor z_i, i.e., $\mathrm{sim}(\tilde{z}_i, z_i) = \cos\theta$. As shown in Appendix A.1,

$$\theta \approx \arctan\big(\sigma\sqrt{d-1}\big). \tag{5}$$

Consequently, the synthesized samples $\tilde{z}_i$ concentrate on an angular shell at angle θ around their respective ID anchors z_i, as illustrated in Figure 2.

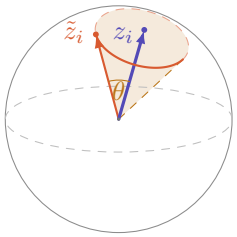


Figure 2: Angular view.

For a given embedding dimension size d, σ and θ are equivalent perturbation parameterizations, related by Equation 5 or $\sigma \approx \tan\theta/\sqrt{d-1}$. Accordingly, we use θ as the primary synthesis parameter because it directly determines the cosine similarity between each synthesized sample and its ID anchor. In all experiments, we set $\theta = 30°$, corresponding to $\sigma = 0.02554$ for $d = 512$ in CLIP ViT-B/16.

Relative to ID Neighborhood. A desirable property of a synthesized OOD sample $\tilde{z}_i$ is that it moves away not only from its ID anchor z_i but also from any training ID image z_j in the surrounding ID neighborhood of z_i. Let ϕ_{ij} denote the angle between z_i and z_j, i.e., $\mathrm{sim}(z_i, z_j) = \cos\phi_{ij}$. Recall that θ is the angle between $\tilde{z}_i$ and z_i. We show in Appendix A.2 that concentration in high dimensions gives

$$\mathrm{sim}(\tilde{z}_i, z_j) \approx \cos\theta\,\mathrm{sim}(z_i, z_j), \tag{6}$$

where the approximation error scales as $O\left(\sin\theta\sin\phi_{ij}/\sqrt{d-1}\right)$ with high probability. Consequently, with $\mathrm{sim}(z_i, z_j) > 0$ and $0 < \cos\theta < 1$, we typically have $\mathrm{sim}(\tilde{z}_i, z_j) < \mathrm{sim}(z_i, z_j)$, suggesting that $\tilde{z}_i$ is less similar to z_j than its anchor z_i is: their similarity is reduced to $\cos\theta$ times the local ID similarity $\mathrm{sim}(z_i, z_j)$. *In other words, $\tilde{z}_i$ moves away not only from z_i but also from z_i's neighboring ID image z_j, with the extent of this movement controlled by θ.* Therefore, once θ is chosen to provide the desired displacement, additional k-NN filtering to retain OOD samples farther from the ID data, as applied in NPOS (Tao et al., 2023), is unnecessary because the synthesized samples already move away from their local neighborhoods. Our empirical evaluation also shows that filtering worsens performance, since it discards useful synthesized OOD samples (see Figure 4).

It is worth noting that Equation 6 for high dimensional spaces is different from the general worst-case bound $|\mathrm{sim}(\tilde{z}_i, z_j) - \cos\theta\,\mathrm{sim}(z_i, z_j)| \leq \sin\theta\sin\phi_{ij}$ in Schubert (2021, Section 3.1) because the latter does not factor in the dimension size d.

5 Experiments

ImageNet-1K Benchmark. The main evaluation setting in most prior work (Huang & Li, 2021; Chen et al., 2024; Li et al., 2025) uses ImageNet-1K (Deng et al., 2009) as the ID dataset and iNaturalist (Van Horn et al., 2018), SUN (Xiao et al., 2010), Places (Zhou et al., 2017), and Textures (Cimpoi et al., 2014) as OOD test sets. Through manual inspection of 400 images randomly sampled from each commonly used OOD dataset, Bitterwolf et al. (2023) found that many images from these OOD datasets can in fact be assigned to one of the ImageNet-1K ID classes: 59.5% for Places and 25.6% for Textures, and Ansari et al. (2025) reported 26.2% for SUN. Because the semantic content of these images lies within the ID space, they are invalid as true OOD: a strong detector that correctly recognizes them as ID is penalized for disagreeing with the erroneous benchmark labels. For a more reliable evaluation, we follow (Ansari et al., 2025) and restrict our OOD test sets to datasets with little or no ID contamination: iNaturalist (Van Horn et al., 2018) (2.5% contamination), OpenImage-O (Wang et al., 2022) (4.9% contamination), together with Clean (Bitterwolf et al., 2023) and NINCO (Bitterwolf et al., 2023), both of which have 0% contamination.

ImageNet Subset and Spurious OOD Benchmarks. Following (Ming et al., 2022a; Jiang et al., 2024), we conduct additional evaluations on subsets of ImageNet-1K and the Spurious OOD benchmark. In the ImageNet-subset benchmark, ImageNet-10, ImageNet-20, and ImageNet-99, containing 10, 20, and 99 classes, respectively, are used as ID and OOD datasets under different pairwise settings established in prior work (Ming et al., 2022a; Jiang et al., 2024). ImageNet-99 is ImageNet-100 after removing the *race car* class that contains samples identical to those in the *sports car* class of ImageNet-10, which creates an overlap in the ImageNet-10 vs. ImageNet-99 setting (Ansari et al., 2025). For the Spurious OOD benchmark (Ming et al., 2022b), WaterBirds (Sagawa et al., 2019), containing bird objects over land or water backgrounds, is used as the ID dataset, while the corresponding background images, Placesbg, without bird objects, are treated as OOD samples.

Evaluation Protocol. For all benchmarks, we use the training split from the ID dataset as training data $\mathcal{D}_{\text{ID}}$, while evaluation is performed on the test split from the ID dataset together with the

Table 1: OOD detection on the ImageNet-1K ID benchmark; methods use CLIP ViT-B/16. The average is over the four OOD datasets. The metrics are percentages, with the best result shown in **bold** and the second-best underlined. Negative-label-based methods are highlighted in gray.

Methods	OOD Datasets								Average	
	iNaturalist		OpenImage-O		Clean		NINCO			
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
Test-Time Adaptation Methods										
InterNeg (Xu et al., 2026)	99.79	**0.40**	93.69	32.30	84.19	68.01	84.16	60.90	90.46	40.40
ANTS (Zhu et al., 2026)	99.75	0.54	93.13	28.89	84.24	64.73	82.32	59.23	89.86	38.35
TANL (Zhang et al., 2026)	**99.84**	0.42	91.92	39.95	86.15	51.53	85.10	57.61	90.75	37.38
Zero-Shot Methods										
ZOC (Esmaeilpour et al., 2022)	79.53	90.98	80.61	70.04	72.88	81.07	67.96	78.67	75.25	80.19
MCM (Ming et al., 2022a)	94.59	32.20	92.00	41.03	83.24	57.79	74.34	79.33	86.04	52.59
GL-MCM (Miyai et al., 2025)	96.44	17.42	92.91	34.52	84.78	51.49	76.03	74.40	87.54	44.46
CLIPN (Wang et al., 2023)	96.20	19.11	92.22	30.36	87.31	41.56	78.72	66.51	88.61	39.39
NegLabel (Jiang et al., 2024)	99.49	1.91	93.74	28.62	86.79	41.22	77.30	68.70	89.33	35.11
CSP (Chen et al., 2024)	99.60	1.54	94.09	28.94	88.32	38.75	77.88	68.65	89.97	34.47
LAPT (Zhang et al., 2024)	99.63	1.16	93.45	35.00	89.74	52.54	85.16	52.40	92.00	35.28
NegRefine (Ansari et al., 2025)	99.57	1.47	95.00	23.85	90.65	33.18	81.92	61.96	91.79	30.12
Few-Shot Methods (16-Shot)										
LoCoOp (Miyai et al., 2023)	96.86	16.05	92.64	32.60	85.18	47.51	71.76	75.26	86.61	42.86
SCT (Yu et al., 2024)	95.86	13.94	93.03	30.29	85.66	45.52	72.59	73.09	86.79	40.71
Local-Prompt (Zeng et al., 2025)	98.07	8.63	94.18	29.28	86.30	46.48	78.06	71.69	89.15	39.02
FA (Lu et al., 2025)	96.48	14.49	92.39	32.10	88.13	39.85	79.01	65.61	89.00	38.01
ID-Supervised Methods										
NPOS (Tao et al., 2023)	96.19	16.58	85.34	63.77	79.38	68.55	73.25	81.10	83.54	57.50
CASOD (Politowicz et al., 2026)	98.50	5.93	90.38	50.16	-	-	86.35	59.43	-	-
SynOOD (Li et al., 2025)	99.57	1.57	**97.06**	15.18	88.48	37.12	82.65	59.03	91.94	28.23
OptLabel (Ours)	99.72	1.14	96.96	**14.14**	**93.27**	**25.54**	**87.45**	**46.04**	**94.35**	**21.72**

benchmark's OOD test sets. For evaluation metrics, we use the Area Under the Receiver Operating Characteristic Curve (AUROC) and the False Positive Rate at 95% True Positive Rate (FPR95), treating ID samples as the positive class.

Implementation Details. For OptLabel, we set the size of $\mathcal{D}_{\text{OOD}}$ equal to that of $\mathcal{D}_{\text{ID}}$ and use $\theta = 30^\circ$ across all experiments. We optimize the positive labels for 50 and the negative labels for 25 epochs. Both stages use SGD with a learning rate of 10^{-3} and a batch size of 256, together taking less than 10 minutes on a single RTX 4090 GPU, excluding CLIP feature extraction. We initialize the negative labels from NegRefine (Ansari et al., 2025). Randomness in our method arises mostly from the OOD synthesis procedure. We therefore repeat all experiments over 10 seeds, reporting mean results in the main paper and their standard deviations in Appendix B. For the baselines, we report results from the original papers whenever available; otherwise, we use the authors' official implementations and released model checkpoints.

5.1 Main Results

ImageNet-1K Benchmark. Table 1 compares our method, OptLabel, with previous CLIP-based methods on the ImageNet-1K benchmark. OptLabel achieves the best average AUROC of 94.35 and average FPR95 of 21.72 across all method categories. It improves upon the previous best average AUROC of 92.00 (LAPT) by 2.35 points and the previous best average FPR95 of 28.23 (SynOOD) by 6.51 points. The FPR95 gains are particularly pronounced on Clean, where OptLabel reduces the previous best of 33.18 (NegRefine) to 25.54, and on the challenging near-boundary OOD NINCO, where it reduces the previous best of 52.40 (LAPT) to 46.04.

Further observations: First, compared with zero-shot methods, OptLabel reduces the best average FPR95 from 30.12 (NegRefine) to 21.72, illustrating the benefit of refining text-derived labels with ID data. Second, the negative-label-based methods (highlighted in gray) generally outperform the other baselines, underscoring the strength of this paradigm. Finally, test-time adaptation methods perform well on iNaturalist but substantially worse on Clean. These methods rely on cached representations of previously seen test samples; since Clean is small and diverse, the cache contains few relevant examples and may even hurt performance, illustrating a limitation of test-time adaptation.

ImageNet Subsets and Spurious OOD. Table 2 reports the results on the ImageNet-subset and Spurious OOD benchmarks, comparing OptLabel with related baselines that were also previously evaluated on these benchmarks. OptLabel shows the best or second-best performance across set-

Table 2: OOD detection on the ImageNet subsets and the Spurious OOD benchmark.

ID Dataset	ImageNet-10		ImageNet-10		ImageNet-20		ImageNet-99		WaterBirds		Average	
OOD Dataset	ImageNet-20		ImageNet-99		ImageNet-10		ImageNet-10		Placesbg			
Methods	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
NegLabel	98.94	4.09	**99.46**	1.94	97.71	15.21	88.62	55.56	89.22	25.54	94.79	20.47
CSP	99.11	2.96	99.43	1.83	98.65	6.81	90.70	46.50	93.23	11.06	96.22	13.83
NegRefine	98.79	4.30	99.13	2.45	**99.08**	**3.21**	91.43	41.90	93.21	11.10	96.33	12.59
SynOOD	98.26	7.35	97.83	8.11	97.57	14.21	86.61	44.37	**95.05**	**7.17**	95.06	16.24
OptLabel (Ours)	**99.16**	**1.43**	99.43	**1.07**	98.81	3.32	**93.61**	**28.28**	94.81	8.68	**97.16**	**8.56**

Table 3: Analysis of the positive and negative label optimization on the ImageNet-1K benchmark. ✕ indicates that the label set is not optimized, whereas ✓ indicates that the label set is optimized.

Label Optimization		iNaturalist		OpenImage-O		Clean		NINCO		Average	
Pos	Neg	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
✕	✕	99.59	1.50	94.82	25.90	89.42	35.36	79.77	66.43	90.90	32.30
✕	✓	99.58	1.59	94.88	21.99	90.33	32.45	82.93	54.70	91.93	27.68
✓	✕	**99.84**	**0.56**	96.84	16.45	92.70	29.15	86.85	56.35	94.06	25.63
✓	✓	99.72	1.14	**96.96**	**14.14**	**93.27**	**25.54**	**87.45**	**46.04**	**94.35**	**21.72**

tings, achieving the best average AUROC of 97.16 and FPR95 of 8.56. Specifically, on the hardest ImageNet-99 (ID) vs. ImageNet-10 (OOD), OptLabel reduces FPR95 from the previous best of 41.90 (NegRefine) to 28.28, an improvement of 13.62. This benchmark also reveals a key limitation of SynOOD that synthesizes OOD samples from non-ID concepts in the background of ID images: SynOOD performs strongly on ImageNet-1K but degrades on the ImageNet subsets because fewer ID classes provide less diverse background concepts for its OOD synthesis. SynOOD only slightly outperforms OptLabel on WaterBirds vs. Placesbg where test OOD samples are backgrounds of ID images (e.g., forests, oceans), matching SynOOD's synthesis strategy. However, in general, the OOD space is open-ended and OOD samples are not necessarily in the backgrounds of ID images.

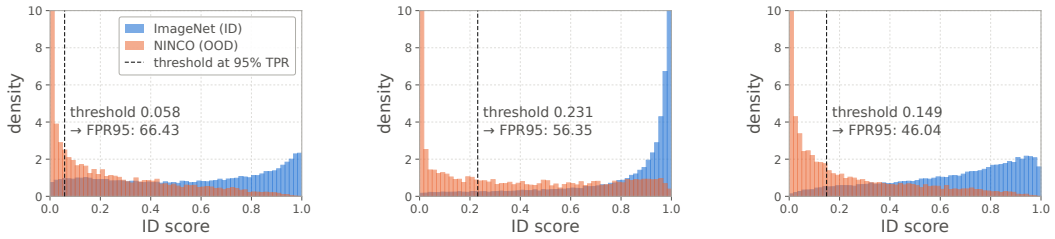
5.2 Analysis

We provide additional analysis using ImageNet-1K benchmark (ID). "Average" refers to the average result over the four OOD test sets.

Ablation of Positive- and Negative-Label Optimizations. In Table 3, we analyze the effect of each of the positive- and negative-label optimizations. The first row denotes the original, unoptimized embeddings; results differ from NegRefine because we use its labels without its multi-matching score. As shown in the Average column, optimizing only the negative labels (second row) reduces the average FPR95 from 32.30 to 27.68, while optimizing only the positive labels (third row) yields a larger improvement, reducing it to 25.63. This indicates that aligning the positive labels with the actual ID data is quite important. Applying negative-label optimization after positive-label optimization (last row) further reduces the average FPR95 from 25.63 to 21.72, achieving the best overall performance. This improvement is most pronounced on the near-boundary OOD NINCO, where FPR95 drops from 56.35 to 46.04, showing the importance of negative-label optimization for capturing challenging near-boundary OOD samples.

Figure 3 further illustrates how positive and negative label optimization reshape the ID-score distributions on the NINCO test set. With the original embeddings (Figure 3a), the ID distribution (blue) is spread across the full score range; thus, the two distributions overlap heavily, forcing a low 95%-TPR threshold and resulting in an FPR95 of 66.43. Positive-label optimization (Figure 3b) shifts the scores of most ID samples toward 1.0. This increases the 95%-TPR threshold and reduces FPR95 to 56.35, but it also shifts some OOD samples toward higher ID scores, leaving a long tail of OOD samples with scores closer to one. Negative-label optimization (Figure 3c) then shifts both ID and OOD distributions toward lower scores, but the shift is stronger for OOD samples, because the optimization moves negative labels closer to the ID boundary. Thus, despite the threshold decreasing, considerably fewer OOD samples remain above it, reducing FPR95 to 46.04.

Optimization Orders and Initializations. Table 4 examines other design choices of optimizations. Pos→Neg (random init.) represents our optimization order (i.e., first positive labels, then negative



(a) Original embeddings (b) Pos optimized (c) Pos and Neg optimized

Figure 3: ID-score distributions for ImageNet-1K (ID) and NINCO (OOD) across optimization.

Table 4: Different optimization orders and initializations in our method on ImageNet-1K benchmark. Train AUROC is computed on the train ID and synthesized OOD data.

Setting	Train	Average	
	AUROC ↑	AUROC ↑	FPR95 ↓
Pos→Neg (random init.)	70.98	92.77	28.51
Neg→Pos	77.94	94.29	23.44
Joint	84.37	93.91	23.40
Pos→Neg (Ours)	74.61	**94.35**	**21.72**

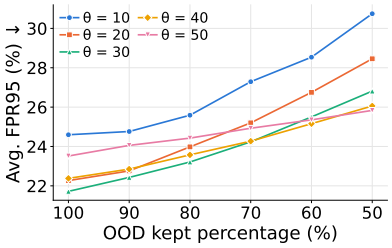


Figure 4: Ablation on θ and effect of filtering.

labels) but negative labels are initialized with random vectors instead of NegRefine's WordNet-derived labels. This increases average FPR95 from 21.72 for our method to 28.51. For the remaining rows, negative labels are initialized by NegRefine's labels: Neg→Pos represents the reverse optimization order, Joint represents optimization of both labels with a single summed objective, and Pos→Neg represents our optimization order. Pos→Neg achieves the best average AUROC and FPR95, supporting the choice to align positive labels with ID data before optimizing negative labels. Joint achieves the highest training AUROC but performs worse on real OOD data, suggesting that updating positive labels with OOD supervision encourages overfitting to synthesized samples and freezing positive labels during OOD training could act as a regularizer, further supporting our sequential design.

Ablation on θ and OOD Filtering. Figure 4 shows the average FPR95 for different perturbation angles θ. For each angle, we rank synthesized OOD samples by their distance to their 20th nearest ID neighbor and retain a specified percentage of samples that are farther away from ID. 100% retains all synthesized samples, which is the setting of OptLabel, and any smaller percentage will filter the remaining samples. Consistent with our theoretical analysis, any filtering worsens the performance across all angles. Among the θ values, $\theta = 30°$ yields the best results. A smaller angle (e.g., $10°$ or $20°$) produces samples too similar to ID, whereas a larger angle (e.g., $40°$ or $50°$) produces samples farther from ID. Both lead to worse performance. A similar study with $2N$ synthesized OOD samples shows no further improvement over N OOD samples; see Appendix Figure 7.

CLIP Architectures. Results with other CLIP architectures are provided in Appendix Table 7.

6 Conclusion

We proposed a label optimization framework, OptLabel, for negative-label-based OOD detection in the ID-supervised setting, motivated by the observation that the text-derived positive and negative labels of existing negative-label-based methods do not well align with actual ID and OOD distributions. OptLabel addresses this limitation through the demonstrated value of its complete design: semantic initialization of labels, simple and easily controllable OOD synthesis, separate objectives of positive and negative label optimizations, frozen positive labels during OOD training, and consistent gains at lightweight training cost.

AI USE STATEMENT

We used ChatGPT to improve the writing and provide feedback on research ideas and mathematical arguments, and Claude Code to assist with code development. We reviewed the AI-assisted text and code. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The source code of OptLabel is provided in the Supplementary Material; all datasets used in our experiments are publicly available from the references cited; parameters are clearly described in Implementation Details in Section 5; the detailed proofs of Equations 5 and 6 are given in Appendix A.

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.

Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *ICML*, 2023. URL `https://proceedings.mlr.press/v202/bitterwolf23a.html`.

Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of data science*. Cambridge University Press, 2020.

Mengyuan Chen, Junyu Gao, and Changsheng Xu. Conjugated semantic pool improves ood detection with pre-trained vision-language models. *arXiv preprint arXiv:2410.08611*, 2024.

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.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *Proceedings of the International Conference on Learning Representations*, 2022.

Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 6568–6576, 2022.

Christiane Fellbaum. *WordNet: An electronic lexical database*. MIT press, 1998.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017.

Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8710–8719, 2021.

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. In *International Conference on Learning Representations (ICLR)*, 2024.

Jinglun Li, Kaixun Jiang, Zhaoyu Chen, Bo Lin, Yao Tang, Weifeng Ge, and Wenqiang Zhang. Synthesizing near-boundary ood samples for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4496–4506, 2025.

Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. Learning transferable negative prompts for out-of-distribution detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17584–17594. IEEE, 2024.

Xinhua Lu, Runhe Lai, Yanqi Wu, Kanghao Chen, Wei-Shi Zheng, and Ruixuan Wang. Fa: Forced prompt learning of vision-language models for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1152–1161, 2025.

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

Yifei Ming, Hang Yin, and Yixuan Li. On the impact of spurious correlation for out-of-distribution detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 10051–10059, 2022b.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. *Advances in Neural Information Processing Systems*, 36:76298–76310, 2023.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Gl-mcm: Global and local maximum concept matching for zero-shot out-of-distribution detection. *International Journal of Computer Vision*, pp. 1–11, 2025.

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.

Alexander Politowicz, Sahisnu Mazumder, and Bing Liu. Improving out-of-distribution detection using segmented images and cross-view attention fusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5269–5279, 2026.

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.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Erich Schubert. A triangle inequality for cosine similarity. In *International Conference on Similarity Search and Applications*, pp. 32–44. Springer, 2021.

Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JHklpEZqduQ.

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.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2 edition, 2026.

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, 2023.

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.

Zhikang Xu, Qianqian Xu, Zitai Wang, Cong Hua, Sicong Li, Zhiyong Yang, and Qingming Huang. Mind the way you select negative texts: Pursuing the distance consistency in ood detection with vlms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34650–34661, 2026.

Geng Yu, Jianing Zhu, Jiangchao Yao, and Bo Han. Self-calibrated tuning of vision-language models for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 37:56322–56348, 2024.

Zijie Yu, Gaowen Liu, Ramana Rao Kompella, Philip S Yu, and Yue Song. The hyperspherical geometry of clip latent space: A semantic mixture model. *arXiv preprint arXiv:2607.13660*, 2026.

Fanhu Zeng, Zhen Cheng, Fei Zhu, Hongxin Wei, and Xu-Yao Zhang. Local-prompt: Extensible local prompts for few-shot out-of-distribution detection. In *International Conference on Learning Representations*, volume 2025, pp. 86715–86732, 2025.

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, 2024.

Yabin Zhang, Maya Varma, Yunhe Gao, Jean-Benoit Delbrouck, Jiaming Liu, Chong Wang, and Curtis Langlotz. Activation matters: Test-time activated negative labels for ood detection with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17462–17473, June 2026.

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.

Wenjie Zhu, Yabin Zhang, Xin Jin, Wenjun Zeng, and Lei Zhang. Ants: Adaptive negative textual space shaping for ood detection via test-time mllm understanding and reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20–30, 2026.

A Mathematical Analysis of Gaussian OOD Synthesis

Here, we provide the mathematical derivations for the results discussed in Section 4.2.

A.1 Angular Concentration after Normalization

In this subsection, we show that the angle θ between a synthesized sample $\tilde{z}_i$ and its ID anchor z_i concentrates around $\theta \approx \arctan\big(\sigma\sqrt{d-1}\big)$. For notational simplicity, we omit the sample index i only in this subsection and write z, ϵ, and $\tilde{z}$ in place of z_i, ϵ_i, and $\tilde{z}_i$, respectively. Recall that the ID anchor z is a normalized CLIP image embedding, i.e., $z \in \mathbb{R}^d$ and $\|z\|_2 = 1$.

Our synthesis first adds a Gaussian perturbation $\epsilon \sim \mathcal{N}(0, \sigma^2 I_d)$ to z and then normalizes the perturbed embedding to obtain $\tilde{z} = (z+\epsilon)/\|z+\epsilon\|_2$. Since normalization does not change the direction, the angle between z and $\tilde{z}$ is the same as that between z and $z + \epsilon$. We therefore characterize the angle between z and $z + \epsilon$.

We decompose ϵ into components parallel and orthogonal to the anchor z: $\epsilon = \epsilon_{\parallel} + \epsilon_{\perp}$. Since z is a unit vector, the parallel component can be written as $\epsilon_{\parallel} = az$, where $a = z^{\top}\epsilon$ is the scalar projection of ϵ onto z. Therefore,

$$z + \epsilon = z + \epsilon_{\parallel} + \epsilon_{\perp} = z + az + \epsilon_{\perp} = (1+a)z + \epsilon_{\perp}. \tag{7}$$

The above expresses $z + \epsilon$ as the sum of two orthogonal components: $(1 + a)z$, which lies along z, and $\epsilon_{\perp}$, which is perpendicular to z, as shown in Figure 5. Hence,

$$\tan\theta = \frac{\|\epsilon_{\perp}\|_2}{1+a}. \tag{8}$$

For the numerator, since ϵ is an isotropic Gaussian in d dimensions, its projection $\epsilon_{\perp}$ onto the subspace orthogonal to z is an isotropic Gaussian in $(d-1)$-dimensional subspace. By the standard concentration of the Euclidean norm of high-dimensional Gaussian vectors, its norm concentrates as $\|\epsilon_{\perp}\|_2 \approx \sigma\sqrt{d-1}$ (Vershynin, 2026, Theorem 3.1.1).

For the denominator, $a = z^{\top}\epsilon$ is Gaussian with mean zero and variance $\sigma^2\|z\|_2^2 = \sigma^2$; thus, $a \sim \mathcal{N}(0, \sigma^2)$. Since σ is generally small in our setting, a is close to zero, allowing us to approximate $1 + a \approx 1$ ($\sigma = 0.02554$ in our setting, $|a| \le 3\sigma \approx 0.0766$ with probability 99.7%, which justifies the approximation $1 + a \approx 1$).

Substituting these two approximations into Equation 8 gives $\tan\theta \approx \sigma\sqrt{d-1}$, and so $\theta \approx \arctan\big(\sigma\sqrt{d-1}\big)$.

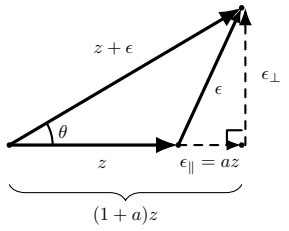


Figure 5: $z + \epsilon$ decomposition relative to z.

A.2 Similarity to Other ID Representations

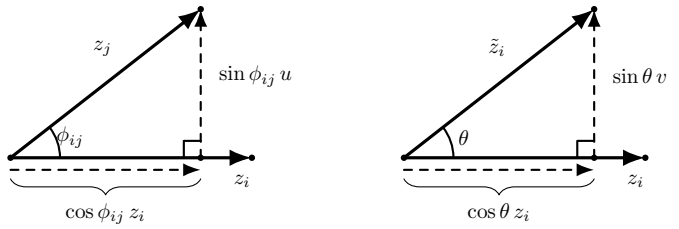
In this subsection, we show that, due to concentration in high dimensions, $\mathrm{sim}(\tilde{z}_i, z_j) \approx \cos\theta\,\mathrm{sim}(z_i, z_j)$, where the approximation error scales as $O\left(\sin\theta\sin\phi_{ij}/\sqrt{d-1}\right)$ with high probability. Recall that z_i, z_j, and $\tilde{z}_i$ are all normalized, and that ϕ_{ij} denotes the angle between the ID representations z_i and z_j, and θ denotes the angle between synthesized sample $\tilde{z}_i$ and z_i.

As illustrated in Figure 6, we decompose both z_j and $\tilde{z}_i$ into components parallel and orthogonal to z_i. Let u denote the unit vector orthogonal to z_i that points toward z_j. Similarly, let v denote the unit

vector orthogonal to z_i that points toward $\tilde{z}_i$. Note that $\|u\|_2 = \|v\|_2 = 1$ and $u^\top z_i = v^\top z_i = 0$. The two decompositions are

$$z_j = \cos\phi_{ij}\, z_i + \sin\phi_{ij}\, u, \qquad \tilde{z}_i = \cos\theta\, z_i + \sin\theta\, v. \tag{9}$$

In each decomposition, the first term is the component parallel to z_i, while the second term is the component orthogonal to z_i. Importantly, u and v generally represent different directions within the $(d-1)$-dimensional subspace orthogonal to z_i.



(a) Decomposition of z_j. (b) Decomposition of $\tilde{z}_i$.

Figure 6: Decompositions of z_j and $\tilde{z}_i$ relative to the vector z_i.

We now compute the cosine similarity between the synthesized sample $\tilde{z}_i$ and the neighboring ID representation z_j. Since both vectors are unit-normalized, their cosine similarity equals their dot product. Substituting the decompositions from Equation 9 gives

$$\begin{aligned}
\mathrm{sim}(\tilde{z}_i, z_j) &= \tilde{z}_i^\top z_j \\
&= (\cos\theta\, z_i + \sin\theta\, v)^\top (\cos\phi_{ij}\, z_i + \sin\phi_{ij}\, u) \\
&= \cos\theta\cos\phi_{ij}\, z_i^\top z_i + \cos\theta\sin\phi_{ij}\, z_i^\top u \\
&\quad + \sin\theta\cos\phi_{ij}\, v^\top z_i + \sin\theta\sin\phi_{ij}\, v^\top u \\
&= \cos\theta\cos\phi_{ij} + \sin\theta\sin\phi_{ij}\, v^\top u,
\end{aligned} \tag{10}$$

where the final equality follows from $z_i^\top z_i = 1$, $z_i^\top u = 0$, and $v^\top z_i = 0$.

The second term in Equation 10 depends on the inner product $v^\top u$, which measures the alignment between two directions in the $(d-1)$-dimensional subspace orthogonal to z_i. Here, u is determined by the data point z_j, whereas v is induced by the isotropic Gaussian perturbation ϵ_i and is uniformly distributed over the unit sphere in this subspace. By the concentration of uniform random vectors on a high-dimensional sphere, this inner product concentrates around zero, and, at any fixed high confidence level, its magnitude is $O\left(1/\sqrt{d-1}\right)$ (Vershynin, 2026, Theorem 3.4.5); see also (Blum et al., 2020, Section 2.4.2). Therefore, in high dimensions, with high probability we have

$$\mathrm{sim}(\tilde{z}_i, z_j) = \cos\theta\cos\phi_{ij} + O\left(\frac{\sin\theta\sin\phi_{ij}}{\sqrt{d-1}}\right). \tag{11}$$

For example, when $d = 512$, (Vershynin, 2026, Theorem 3.4.5) gives $|v^\top u| < 0.087$ with probability at least 95%. Thus, for $\theta = 30°$, as used in our experiments, the resulting approximation error is less than $0.0435 \sin\phi_{ij} \leq 0.0435$ with probability at least 95%.

B Extended Experimental Results

Tables 5 and 6 report means and standard deviations over 10 seeds for the ImageNet-1K and ImageNet-subset and Spurious OOD benchmarks, respectively. The small standard deviations indicate robustness to randomness in our method.

Figure 7 extends the angular and filtering ablation to $2N$ synthesized OOD samples. These results also show that filtering degrades performance. The $2N$ synthesized OOD samples yield an average FPR95 of 21.82 for $\theta = 30°$ and keeping 100% of synthesized OOD samples, providing almost the same result as 21.72 with N samples, our default synthesis size.

Methods	iNaturalist		OpenImage-O		Clean		NINCO		Average	
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
OptLabel (Ours)	99.72 ± 0.00	1.14 ± 0.02	96.96 ± 0.01	14.14 ± 0.08	93.27 ± 0.02	25.54 ± 0.11	87.45 ± 0.04	46.04 ± 0.13	94.35	21.72

Table 5: Mean and standard deviation over 10 seeds for OptLabel on the ImageNet-1K benchmark, corresponding to Table 1 in the main paper.

ID Dataset OOD Dataset	ImageNet-10 ImageNet-20		ImageNet-10 ImageNet-99		ImageNet-20 ImageNet-10		ImageNet-99 ImageNet-10		WaterBirds Placesbg		Average	
Methods	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
OptLabel (Ours)	99.16 ± 0.01	1.43 ± 0.02	99.43 ± 0.01	1.07 ± 0.06	98.81 ± 0.02	3.32 ± 0.14	93.61 ± 0.18	28.28 ± 0.42	94.81 ± 0.10	8.68 ± 0.05	97.16	8.56

Table 6: Mean and standard deviation over 10 seeds for OptLabel on the ImageNet-subset and Spurious OOD benchmarks, corresponding to Table 2 in the main paper.

Table 7 reports results across six additional CLIP architectures. Our method, OptLabel, consistently outperforms the compared baselines in average AUROC and FPR95 by a large margin.

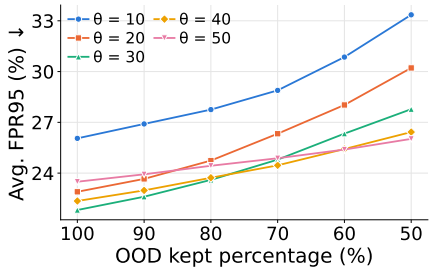


Figure 7: Ablation on θ and effect of filtering using $2 \times N$ synthesized OOD samples.

Table 7: OOD detection on the ImageNet-1K ID benchmark on different CLIP architectures.

Architecture	Methods	OOD Datasets								Average	
		iNaturalist		OpenImage-O		Clean		NINCO			
		AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
ResNet50	NegLabel	99.24	2.88	92.00	33.35	84.45	45.52	74.20	73.53	87.47	38.82
	CSP	**99.46**	**1.95**	91.83	33.91	86.08	42.06	75.53	72.86	88.22	37.70
	NegRefine	99.38	2.30	92.57	31.05	88.26	38.97	80.73	65.62	90.23	34.48
	OptLabel (Ours)	99.36	2.51	**94.71**	**22.88**	**89.23**	**36.98**	**83.69**	**56.48**	**91.75**	**29.71**
ResNet50x64	NegLabel	99.63	1.46	93.68	30.58	90.15	37.42	86.98	54.41	92.61	30.97
	CSP	**99.69**	**1.19**	93.95	30.21	91.48	33.92	87.18	52.94	93.08	29.57
	NegRefine	99.60	1.41	94.52	26.45	92.14	30.90	88.26	47.33	93.63	26.52
	OptLabel (Ours)	99.63	1.64	**96.36**	**16.45**	**93.98**	**23.13**	**90.11**	**37.14**	**95.02**	**19.59**
ResNet101	NegLabel	99.27	3.11	90.92	37.41	84.83	46.34	76.80	72.93	87.95	39.95
	CSP	99.47	2.04	91.90	35.58	85.81	43.43	75.92	72.25	88.28	38.33
	NegRefine	99.59	1.49	93.75	27.52	88.43	38.23	80.17	66.06	90.48	33.33
	OptLabel (Ours)	**99.64**	**1.45**	**95.32**	**21.26**	**90.62**	**33.55**	**86.36**	**53.03**	**92.98**	**27.32**
ViT-B/32	NegLabel	99.11	3.73	92.87	31.26	85.20	44.16	75.30	71.45	88.12	37.65
	CSP	99.46	2.37	93.37	31.01	87.42	40.15	76.96	70.62	89.30	36.04
	NegRefine	99.45	2.20	94.22	26.11	89.90	35.47	82.38	60.70	91.49	31.12
	OptLabel (Ours)	**99.59**	**1.64**	**96.31**	**16.83**	**91.63**	**29.61**	**85.98**	**51.94**	**93.38**	**25.01**
ViT-L/14	NegLabel	99.53	1.77	94.26	27.55	89.09	38.38	81.98	62.59	91.22	32.57
	CSP	99.72	1.21	95.02	25.59	91.52	34.22	84.94	57.20	92.80	29.55
	NegRefine	99.67	1.25	95.96	19.92	92.74	28.43	87.62	47.40	94.00	24.25
	OptLabel (Ours)	**99.76**	**0.95**	**97.76**	**10.07**	**95.17**	**20.04**	**91.31**	**33.40**	**96.00**	**16.11**
ViT-L/14-336px	NegLabel	99.67	1.31	94.26	27.28	89.64	36.87	83.70	61.06	91.82	31.63
	CSP	**99.79**	0.86	94.90	26.06	91.99	32.19	86.38	55.90	93.27	28.75
	NegRefine	99.71	1.01	95.73	20.46	93.16	26.92	88.81	44.68	94.35	23.27
	OptLabel (Ours)	**99.79**	**0.84**	**97.63**	**10.80**	**95.40**	**18.75**	**91.76**	**32.17**	**96.14**	**15.64**