AdaNeg: Adaptive Negative Proxy Guided OOD Detection with Vision-Language Models

Yabin Zhang
The Hong Kong Polytechnic University
csybzhang@comp.polyu.edu.hk

Lei Zhang*
The Hong Kong Polytechnic University
cslzhang@comp.polyu.edu.hk

Abstract

Recent research has shown that pre-trained vision-language models are effective at identifying out-of-distribution (OOD) samples by using negative labels as guidance. However, employing consistent negative labels across different OOD datasets often results in semantic misalignments, as these text labels may not accurately reflect the actual space of OOD images. To overcome this issue, we introduce *adaptive negative proxies*, which are dynamically generated during testing by exploring actual OOD images, to align more closely with the underlying OOD label space and enhance the efficacy of negative proxy guidance. Specifically, our approach utilizes a feature memory bank to selectively cache discriminative features from test images, representing the targeted OOD distribution. This facilitates the creation of proxies that can better align with specific OOD datasets. While task-adaptive proxies average features to reflect the unique characteristics of each dataset, the sample-adaptive proxies weight features based on their similarity to individual test samples, exploring detailed sample-level nuances. The final score for identifying OOD samples integrates static negative labels with our proposed adaptive proxies, effectively combining textual and visual knowledge for enhanced performance. Our method is training-free and annotation-free, and it maintains fast testing speed. Extensive experiments across various benchmarks demonstrate the effectiveness of our approach, abbreviated as AdaNeg. Notably, on the large-scale ImageNet benchmark, our AdaNeg significantly outperforms existing methods, with a 2.45% increase in AUROC and a 6.48% reduction in FPR95. Codes are available at https://github.com/YBZh/OpenOOD-VLM.

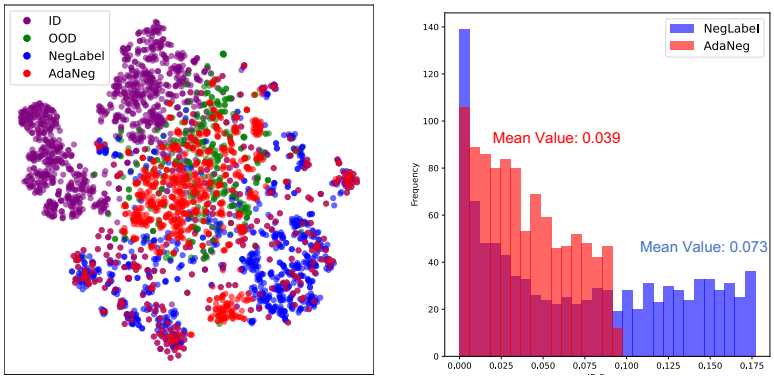
1 Introduction

In real applications, artificial intelligence (AI) systems often encounter test samples of unknown classes, termed out-of-distribution (OOD) data. These OOD data often result in overly confident errors [52, 44], posing security threats. Therefore, accurately identifying OOD data is essential for ensuring the reliability and security of AI systems in open-world environments.

Traditional OOD detection methods in image domain primarily rely on vision-only models [20, 33, 36]. Recent advancements in vision-language models (VLMs) have demonstrated remarkable OOD detection performance by leveraging multi-modal knowledge [17, 15, 39]. Recently, NegLabel [27] explores negative labels by identifying text labels that are semantically distant from the in-distribution (ID) labels. This method achieves state-of-the-art performance by detecting test images closer to negative labels as OOD. In other words, NegLabel regards these negative labels as proxies of OOD data. However, employing consistent negative labels across different OOD datasets often leads to semantic misalignment, where these text labels may not accurately reflect the actual label space

*Corresponding Author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



(a) t-SNE Visualization (b) Histogram of ISOR

Figure 1: Qualitative and quantitative analyses of semantic misalignment between OOD labels and negative proxies using ImageNet (ID) and SUN (OOD) datasets. (a) Visualization of ID labels, OOD labels, negative labels from NegLabel, and adaptive negative proxies (AdaNeg). (b) Quantitative analysis based on ID-Similarity to OOD Ratio (ISOR in short, see Appendix A.1). Lower ISOR indicates a higher similarity to OOD labels and reduced similarity to ID labels. AdaNeg consistently achieves lower ISOR, demonstrating enhanced alignment with OOD characteristics. Visualizations include the top 1,000 discriminative proxies from both NegLabel and AdaNeg.

of OOD images, as shown in Fig. 1. This misalignment between the proxies and targeted OOD distribution leads to sub-optimal performance.

To promote the alignment between the negative proxies and target OOD distribution, we introduce the **Ada**ptive **Neg**ative proxies (**AdaNeg**), which are dynamically generated during testing by exploring actual OOD images. Specifically, we start by initializing an empty category-split memory bank for each OOD dataset and selectively cache features of discriminative OOD images during testing. The OOD discrimination is assessed using mined negative labels, as detailed in [27]. With this feature memory, we develop task-adaptive proxies by simply averaging cached features within each category. These proxies, derived from actual OOD images, reflect the distinct characteristics of the target OOD dataset and align more closely with the underlying OOD label space.

The task-adaptive proxies mentioned previously provide unique proxies for different OOD datasets while maintaining consistency across various test samples within the same dataset. To delve into the fine-grained nuances at the sample level, we introduce the sample-adaptive proxies by weighting cached features based on their similarity to a particular test sample. This is achieved with an attention mechanism, where the feature memory serves as both keys and values, and the test feature acts as the query. The final score for detecting OOD samples integrates static negative labels with our adaptive proxies, effectively combining textual and visual knowledge for enhanced performance.

We conduct extensive experiments on standard benchmarks to validate the effectiveness of AdaNeg, where our proposed adaptive proxies outperform the negative-label-based one, enhancing performance with complementary multi-modal knowledge. Particularly, on the large-scale ImageNet dataset, our AdaNeg method outperforms existing methods by 2.45% AUROC and 6.48% FPR95. Notably, our method is training-free and annotation-free, and it maintains fast testing speed, as analyzed in Tab. 4. The ability to dynamically adjust to new OOD datasets without affecting testing speed or labor-intensive annotation/training makes our approach particularly valuable for real-world applications where adaptability and efficiency are crucial. We summarize our contribution as follows:

- We first identify the label space misalignment between existing negative-label-based proxies and the target OOD distributions. In response, we introduce adaptive negative proxies that are dynamically generated during testing by exploring actual OOD images, resulting in a more effective alignment with the OOD label space.

- Our adaptive negative proxies are constructed with a feature memory bank that selectively caches discriminative image features during testing. We instantiate this concept by developing task-adaptive proxies to reflect the unique characteristics of each OOD dataset and

sample-adaptive proxies to capture detailed sample-level nuances. The final OOD detection score combines these insights with complementary textual and visual knowledge.

- We conduct thorough analyses of the proposed components and perform extensive experiments on standard benchmarks. Our method is training-free and annotation-free, and it maintains fast testing speed and achieves state-of-the-art performance. Notably, our method significantly outperforms existing methods, with a 2.54% increase in AUROC and a 6.48% reduction in FPR95 on the large-scale ImageNet dataset.

2 Related Work

OOD Detection focuses on identifying OOD test samples with semantic shifts, thus distinguishing it from generalization studies which typically focus on covariate shifts [3–5, 75]. A variety of OOD detection techniques have been developed, which can be roughly categorized into score-based [20, 33, 36, 37, 66, 26, 64, 56], distance-based [58, 59, 57, 12, 41, 53], and generative-based [50, 29] methods. Among them, score-based methods are particularly notable by employing a variety of scoring functions to differentiate between ID and OOD samples. These functions include confidence-based [20, 36, 56, 64], discriminator-based [29], energy-based [37, 66], and gradient-based [25] scores. In contrast, distance-based methods determine OOD samples by evaluating the distance in the feature space between test data and the closest ID samples [58] or ID prototypes [59], using metrics such as KNN [57, 12, 41] or Mahalanobis distance [33, 53].

Despite their achievements, traditional OOD detection methods generally rely on manually annotated ID images and often overlook the integration of textual information. To leverage the textual knowledge, recent advancements have focused on employing VLMs [39, 40, 27, 77, 76, 15, 42, 65, 45]. Specifically, ZOC [15] 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. LoCoOp [42] adopts a novel approach by learning ID prompts from few-shot ID samples, and further enhances the robustness of these prompts by incorporating OOD features mined from the backgrounds of images. CLIPN [65], LSN [45] and LAPT [76] explore learning text prompts for expressing negative concepts. In specific, CLIPN initializes text prompts with the word 'no' combined with ID labels and refines them with large-scale multi-modal data; LSN starts with manually collected ID samples to learn negative prompts, offering a different approach to leveraging textual information in OOD detection; LAPT conducts automated prompt tuning with automatically collected training samples, boosting OOD detection without any manual effort. MCM [39] utilizes ID class names to facilitate effective zero-shot OOD detection. It is further refined by NegLabel [27], which incorporates additional negative class names mined from available data sources as negative proxies. However, as illustrated in Fig. 1, there is a mismatch between the negative-label-based proxies and the target OOD distribution, underscoring the limitations of this strategy. This observation has inspired us to construct adaptive proxies by exploring potential OOD test images during testing. This leads to an efficient method that aligns better with the target OOD distribution, resulting in enhanced OOD detection performance.

Furthermore, we clarify the relationship between our method and existing approaches on OOD exposure [17, 21, 73]. Most OOD exposure methods introduce manually collected negative images during training, where manual labor is necessary to ensure that the labels of negative images are different from ID ones. Moreover, involving negative images in training typically introduces additional computational overhead, impeding its practical deployment. Unlike these methods, NegLabel [27] is exposed to negative labels during the test phase in a training-free manner. However, given a fixed ID dataset, the exposed negative texts remain consistent for different OOD datasets, inevitably resulting in label misalignment, as shown in Fig. 1. To address this, we expose the VLMs to adaptive negative proxies, which explore actual OOD samples during testing and align more effectively with OOD distribution. Our method does not require manual annotations and works in a training-free manner, making it an appealing solution for real applications.

Test-time Adaptation. We adopt an online update of the negative proxies during testing, resembling test-time adaptation (TTA) methods [35, 63, 54]. Existing TTA methods primarily address covariate shifts between training and testing domains. In contrast, our approach mitigates the label shift between negative proxies and the target OOD distribution by exploring online test samples. Recently, TTA strategies have been considered in the field of OOD detection. However, these methods typically require test-time optimization [18, 72, 16], slowing down the testing process. In contrast, our method

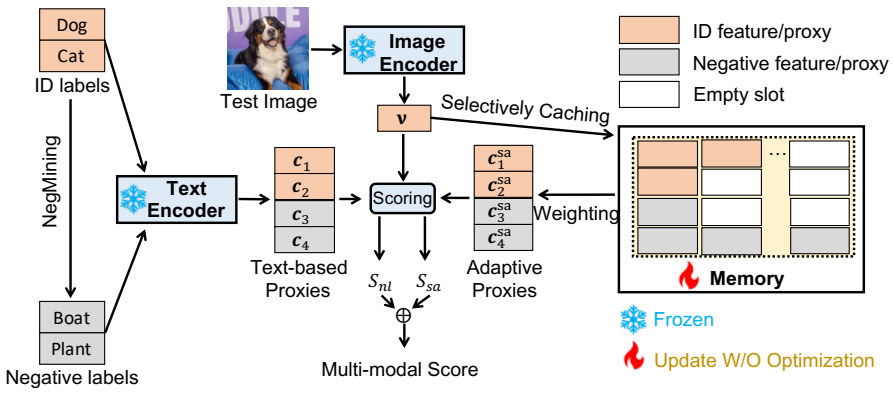


Figure 2: The overall framework of AdaNeg, where we selectively cache test images and generate adaptive proxies with an external feature memory bank. The final score combines textual and visual knowledge from static negative labels and our adaptive proxies, integrating multi-modal information.

is optimization-free and introduces only a lightweight memory interaction operation, enabling rapid and accurate testing, as analyzed in Tab. 4.

Memory Networks. The use of memory networks for storing and retrieving past knowledge [67, 55] has been extensively applied across various fields, including classification [28, 51, 77], segmentation [46, 69], detection [8, 6], and NLP [47]. To our best knowledge, our work is the first to apply memory networks to the field of OOD detection. By caching and retrieving test images with a feature memory, we propose adaptive proxies to more effectively align with the OOD distribution in a training-free manner. This innovative approach significantly enhances the OOD detection performance.

3 Methodology

3.1 Preliminaries

OOD Detection Setup. Consider $\mathcal{X}$ as the image domain and $\mathcal{Y} = \{y_1, \dots, y_C\}$ as the space of ID class labels, where $\mathcal{Y}$ comprises text elements such as $\mathcal{Y} = \{cat, dog, \dots, bird\}$, and C represents the total number of classes. Let $\boldsymbol{x}^{in}$ and $\boldsymbol{x}^{ood}$ be random variables representing ID and OOD samples from $\mathcal{X}$, respectively. We define $\mathcal{P}\boldsymbol{x}^{in}$ and $\mathcal{P}\boldsymbol{x}^{ood}$ as the marginal distributions for ID and OOD, respectively. In conventional classification scenarios, it is assumed that the test image $\boldsymbol{x}$ originates from the ID and is associated with a specific ID label, specifically, $\boldsymbol{x} \in \mathcal{P}\boldsymbol{x}^{in}$ and $y \in \mathcal{Y}$, with y being the label of $\boldsymbol{x}$. However, in real applications, AI systems often face data from unknown classes, denoted by $\boldsymbol{x} \in \mathcal{P}\boldsymbol{x}^{ood}$ and $y \notin \mathcal{Y}$. Such occurrences can make AI models incorrectly categorize these instances into familiar ID categories with substantial certainty [52, 44], resulting in security concerns. To address these challenges, OOD detection is proposed to accurately categorize ID samples into their respective classes and reject OOD samples as non-ID. Recognition within the ID categories is performed using a C-way classifier, following standard classification approaches [31, 19]. Concurrently, OOD detection typically employs a scoring mechanism S [33, 36, 37] to differentiate between ID and OOD inputs:

$$G_\gamma(\boldsymbol{x}) = \text{ID}, \text{ if } S(\boldsymbol{x}) \geq \gamma; \quad \text{otherwise, } G_\gamma(\boldsymbol{x}) = \text{OOD}, \tag{1}$$

where G_γ represents the OOD detector set at a threshold $\gamma \in \mathcal{R}$. The test sample $\boldsymbol{x}$ is identified as an ID sample if and only if $S(\boldsymbol{x}) \geq \gamma$.

CLIP and NegLabel. For an ID test image $\boldsymbol{x}$ within the label space $\mathcal{Y}$, we derive the image feature vector $\boldsymbol{v} = f_{img}(\boldsymbol{x}) \in \mathcal{R}^D$ and the text feature matrix $\boldsymbol{C}_{id} = f_{txt}(\rho(\mathcal{Y})) \in \mathcal{R}^{C \times D}$ using pre-trained CLIP encoders, where D represents the feature dimension. The functions $f_{img}(\cdot)$ and $f_{txt}(\cdot)$ are the encoders for images and text, respectively. The function $\rho(\cdot)$ is the text prompt mechanism, typically defined as 'a photo of a <label>.', where label is the actual class name, for example, 'cat' or 'dog'. Both $\boldsymbol{v}$ and $\boldsymbol{C}_{id}$ undergo L_2 normalization across the dimension D. Then, zero-shot classification probabilities are computed utilizing $\boldsymbol{C}$ as the classifier:

$$\boldsymbol{p}^{id} = \text{Softmax}(\boldsymbol{v}\boldsymbol{C}_{id}^T/\tau) \in \mathcal{R}^C, \tag{2}$$

where $\tau > 0$ is the scaling temperature.

The vanilla CLIP is proposed for zero-shot ID recognition and has recently been extended to OOD detection. Specifically, the NegLabel approach [27] introduces negative class names $\mathcal{Y}^- = \{y_{C+1}, \ldots, y_{C+M}\}$ sourced from broad text corpora, where M is the length of negative classes and $\mathcal{Y}^- \cap \mathcal{Y} = \emptyset$. Then, we can obtain the full text feature matrix $\boldsymbol{C} = f_{txt}(\rho(\mathcal{Y} \cup \mathcal{Y}^-)) \in \mathcal{R}^{(C+M)\times D}$ with the pre-trained CLIP text encoder, leading to the classification probability across $C + M$ classes:

$$\boldsymbol{p} = \text{Softmax}(\boldsymbol{v}\boldsymbol{C}^T/\tau) \in \mathcal{R}^{C+M}. \tag{3}$$

Assuming that ID samples exhibit greater similarity to ID labels and lesser similarity to negative labels compared to OOD samples, NegLabel introduces the following score for OOD detection:

$$S_{nl}(\boldsymbol{v}) = \sum_{i=1}^{C} \boldsymbol{p}_i, \tag{4}$$

where $\boldsymbol{p}_i$ is the i-th entry of $\boldsymbol{p}$, indicating the classification probability of the i-th class. Intuitively, the NegLabel method uses negative labels as proxies of the OOD distribution, detecting OOD images based on the similarity to these negative labels.

3.2 AdaNeg

Although NegLabel has successfully employed negative labels as the OOD proxies, there is a semantic misalignment between such OOD proxies and actual OOD labels, as illustrated in Fig. 1. We aim to obtain OOD proxies that align better to the targeted OOD distribution. However, acquiring such OOD proxies is challenging, as the OOD information is unknown prior to actual testing.

From another perspective, we can access real OOD information during testing, motivating us to refine OOD proxies in the testing stage. We can identify discriminative negative images during testing and then adjust the OOD proxies toward detected images. This is achieved by selectively caching features of test images into a task-aware memory bank, followed by memory reading operations to produce adaptive proxies. We detail our implementation as follows.

Task-aware Memory. We construct a task-aware memory as a category-split tensor $\boldsymbol{M} \in \mathbb{R}^{(C+M)\times L\times D}$, where L is the memory length for each category. $\boldsymbol{M}$ is initialized with zero values and gradually filled with features of selected images during testing. Specifically, for a test image with feature $\boldsymbol{v}$, we first calculate its score $S_{nl}(\boldsymbol{v})$ with Eq. (4). If $S_{nl}(\boldsymbol{v}) < \gamma$, we detect this test point as a negative image, and otherwise, it is identified as a positive sample. For negative and positive images, we respectively identify their closest labels as:

$$\text{Negative} : y = \arg\max_i \boldsymbol{p}_i^{ood} + C, \tag{5}$$

$$\text{Positive} : y = \arg\max_i \boldsymbol{p}_i^{id}, \tag{6}$$

where $\boldsymbol{p}^{ood} = \boldsymbol{p}[C : C + M]$ and $\boldsymbol{p}^{id} = \boldsymbol{p}[0 : C]$ are the classification probabilities corresponding to negative and ID class names, respectively. Then, we cache this feature $\boldsymbol{v}$ into the task-aware memory:

$$\boldsymbol{M}_{y,l} = \boldsymbol{v}, \tag{7}$$

where l indicates an empty slot of $\boldsymbol{M}_y \in \mathbb{R}^{L\times D}$. Once $\boldsymbol{M}_y$ is filled, we drop the image feature with the highest prediction entropy, as detailed in Appendix A.2. In one word, we keep confident image features with low prediction entropy in the memory.

The aforementioned strategy attempts to cache all test images, including those with high confusion between ID and OOD. However, we found that selectively retaining only those image features that exhibit strong ID/OOD distinguishing capabilities can effectively reduce this confusion. Specifically, we modify the selection criterion for memorization as follows:

$$\begin{aligned} \text{Negative} : S_{nl}(\boldsymbol{v}) < \gamma \rightarrow S_{nl}(\boldsymbol{v}) < \gamma - g\gamma, \\ \text{Positive} : S_{nl}(\boldsymbol{v}) \geq \gamma \rightarrow S_{nl}(\boldsymbol{v}) \geq \gamma + g(1-\gamma), \end{aligned} \tag{8}$$

where $g \in [0, 1]$ is a hyperparameter that introduces a gap in the score space. Consequently, image features falling within the gap $\gamma - g\gamma \leq S_{nl}(\boldsymbol{v}) < \gamma + g(1-\gamma)$ are considered to have low distinguishing confidence and are not cached.

Algorithm 1 Adaptive Negative Proxy Guided OOD Detection

Require: ID label space $\mathcal{Y}$ and test set $\mathcal{X}$
1: $\mathcal{Y}^- \leftarrow$ NegMine($\mathcal{Y}$) following [27]
2: Constructing an empty memory $\boldsymbol{M}$
3: **for** $\boldsymbol{x} \in \mathcal{X}$ **do**
4: Generating detection score with negative labels using Eq. 4
5: Determine whether $\boldsymbol{x}$ should be cached using Eq. 8
6: Caching $\boldsymbol{x}$ with Eq. 7 if needed
7: Generating adaptive proxies with memory bank using Eq. 9 or Eq. 11
8: Generating adaptive scores with Eq. 10 or Eq. 12
9: Generating final score S_{all} by merging multi-modal knowledge with Eq. 13
10: **end for**
11: **Return** Collected final scores $\{S_{all}\}$

Task-adaptive Proxies. Given the updated $\boldsymbol{M}$, we can easily get the task-adaptive proxy by averaging along the length dimension L:

$$\boldsymbol{C}^{ta} = \mathrm{L}_2\left(\frac{1}{L+1}\sum_{l=1}^{L+1}\widehat{\boldsymbol{M}}_{:,l,:}\right) \in \mathbb{R}^{(C+M)\times D}, \tag{9}$$

where $\mathrm{L}_2(\cdot)$ indicates the L_2 normalization along feature dimension D, and $\widehat{\boldsymbol{M}} = [\boldsymbol{M}, \boldsymbol{C}] \in \mathbb{R}^{(C+M)\times(L+1)\times D}$ is the extended memory with vanilla text proxies $\boldsymbol{C}$. Such a memory extension is necessary since $\boldsymbol{M}$ is initially empty and uninformative. Initially, this extension initializes the adaptive proxies $\boldsymbol{C}^{ta}$ with the basic text proxies $\boldsymbol{C}$. However, there are two key distinctions between the adaptive $\boldsymbol{C}^{ta}$ and the vanilla $\boldsymbol{C}$. First, unlike the NegLabel approach, which employs a fixed proxy $\boldsymbol{C}$ across various OOD datasets, our $\boldsymbol{C}^{ta}$ dynamically adjusts to the targeted OOD domain as the memory accumulates data, thereby providing dataset-specific adaptive proxies. Second, the $\boldsymbol{C}^{ta}$ primarily incorporates image features, offering modal knowledge that complements the text-based proxies $\boldsymbol{C}$. The benefits of this approach are further analyzed in Tab. 3.

The score function for OOD detection with the task-adaptive proxy $\boldsymbol{C}^{ta}$ is derived as:

$$S_{ta}(\boldsymbol{v}) = \frac{\sum_{i=1}^{C} e^{\cos(\boldsymbol{v}, \boldsymbol{c}_i^{ta})/\tau}}{\sum_{i=1}^{C} e^{\cos(\boldsymbol{v}, \boldsymbol{c}_i^{ta})/\tau} + \sum_{j=C+1}^{C+M} e^{\cos(\boldsymbol{v}, \boldsymbol{c}_j^{ta})/\tau}}, \tag{10}$$

where $\cos(\cdot,\cdot)$ measures the cosine similarity, and $\boldsymbol{c}_i^{ta}$ is the i-th entry of $\boldsymbol{C}^{ta}$.

Sample-adaptive Proxies. As evidenced in Table 3, our task-adaptive proxies significantly outperform the fixed text proxies by effectively adapting to the characteristics of target OOD dataset. Building on this success, we further refine our approach by leveraging finer-grained, sample-level nuances to introduce even more effective sample-adaptive proxies. Specifically, given the extended memory $\widehat{\boldsymbol{M}}$ and the test image feature $\boldsymbol{v}$, we introduce the sample-adaptive proxy $\boldsymbol{C}^{sa} \in \mathbb{R}^{(C+M)\times D}$ via the following cross-attention operation:

$$\boldsymbol{c}_i^{sa} = \mathrm{L}_2\left(\varphi\left(\boldsymbol{v}(\widehat{\boldsymbol{M}}_i)^\top\right)\widehat{\boldsymbol{M}}_i\right) \in \mathbb{R}^D, \tag{11}$$

where $\boldsymbol{v}(\widehat{\boldsymbol{M}}_i)^\top \in \mathbb{R}^{1\times(L+1)}$ measures the cosine similarities between normalized features of $\boldsymbol{v}$ and $\widehat{\boldsymbol{M}}_i$, and $\varphi(x) = \exp(-\beta(1-x))$ modulates the sharpness of x with hyper-parameter β. $\boldsymbol{c}_i^{sa}$ and $\widehat{\boldsymbol{M}}_i$ are the i-th entry of $\boldsymbol{C}^{sa}$ and $\widehat{\boldsymbol{M}}$, respectively.

Both the task-adaptive and the sample-adaptive proxies are derived from the memorized image features stored in $\widehat{\boldsymbol{M}}$. The primary distinction between them lies in their respective weighting strategies. For $\boldsymbol{C}^{ta}$, each feature $\widehat{\boldsymbol{M}}_{:,l,:}$ in the memory is assigned a uniform weighting coefficient of $\frac{1}{L+1}$. Conversely, in constructing $\boldsymbol{C}^{sa}$, the weighting coefficient for each cached feature is dynamically determined based on its cosine similarity to the test image feature, denoted as $\boldsymbol{v}(\widehat{\boldsymbol{M}}_i)^\top$. Consequently, while $\boldsymbol{C}^{ta}$ remains constant across different test samples, $\boldsymbol{C}^{sa}$ adapts to each individual test sample. This adaptability allows $\boldsymbol{C}^{sa}$ to tailor its response based on the specific characteristics

of each test image, thereby enhancing discrimination between ID and OOD samples, particularly in diverse and variable testing scenarios.

The score function for OOD detection with the sample-adaptive proxies $\boldsymbol{C}^{sa}$ is derived as:

$$S_{sa}(\boldsymbol{v}) = \frac{\sum_{i=1}^{C} e^{\cos(\boldsymbol{v},\boldsymbol{c}_i^{sa})/\tau}}{\sum_{i=1}^{C} e^{\cos(\boldsymbol{v},\boldsymbol{c}_i^{sa})/\tau} + \sum_{j=C+1}^{C+M} e^{\cos(\boldsymbol{v},\boldsymbol{c}_j^{sa})/\tau}}. \tag{12}$$

Multi-modal Score. As previously discussed, the score function $S_{nl}(\boldsymbol{v})$ relies primarily on text features, whereas the sample-adaptive score function $S_{sa}(\boldsymbol{v})$ utilizes cached image features. Given the complementary nature of text and image modalities, we derive the final score function by integrating knowledge from both modalities:

$$S_{all}(\boldsymbol{v}) = S_{nl}(\boldsymbol{v}) + \lambda S_{sa}(\boldsymbol{v}), \tag{13}$$

where $\lambda > 0$ is the hyperparameter balancing the two modalities. The overall pipeline of our method is illustrated in Fig. 2 and summarized in Algorithm 1.

4 Experiments

4.1 Setup

Datasets. We conduct extensive experiments with the large-scale ImageNet-1k [9] as ID data. Following prior practice [26, 39, 27], four OOD datasets of iNaturalist [60], SUN [68], Places [78], and Textures [7] are evaluated. We also validate our method on the OpenOOD benchmark [74, 71], where OOD datasets are grouped into near-OOD (*e.g.*, SSB-hard [62], NINCO [2]) and far-OOD (*e.g.*, iNaturalist [60], Textures [7], OpenImage-O [64]) according to their similarity to ImageNet dataset. Besides ImageNet, we also evaluate our method on smaller-sized CIFAR10/100 datasets [30] with the OpenOOD setup. Specifically, with the ID dataset of CIFAR10/100, we adopt near-OOD datasets of CIFAR100/10 and TIN [32], and far-ood datasets of MNIST [10], SVHN [43], Texture [7], and Plances365 [78]. These experiments with various ID and OOD datasets enable a comprehensive evaluation on various OOD settings.

Implementation Details. We adopt the visual encoder of VITB/16 pretrained by CLIP [48] and analyze more backbone architectures in Tab. A11. For hyper-parameters, we adopt the memory length L=10, threshold γ=0.5 with the gap g=0.5 in Eq. 8, β=5.5 in Eq. 11, and λ=0.1 in Eq. 13 in all experiments. These hyper-parameters are carefully analyzed in Sec. 4.3. Following NegLabel, we adopt the text prompt of 'The nice <label>.', set temperature τ=0.01, and define the number M of negative labels as $10,000$ for the ImageNet dataset. For the CIFAR datasets, we set the number M as $70,000$ since we find that a larger M leads to better results for CIFAR. Following common practice [26, 39, 27], we employ the evaluation metrics of FPR95, AUROC, and ID ACC, which are detailed in Appendix A.3. All experiments are conducted with a single Tesla V100 GPU.

4.2 Main Results

ImageNet Results with Four OOD Datasets. As illustrated in Tab. 1, our AdaNeg significantly outperforms existing training-free methods and even surpasses approaches requiring additional training. Specifically, we report traditional methods [20, 36, 37, 25, 64, 57, 13, 59] by fine-tuning CLIP-encoders with labeled training samples following [27], and report results of [45, 42, 27, 34, 1] from their original paper. Compared to the closest competitor [27], our method achieves consistent and notable improvements, validating the advantage of the proposed adaptive proxies over the negative-label-based ones.

ImageNet Results with OpenOOD Setup. Our method is extensively evaluated against a range of OOD datasets in Tab. 2 on the OpenOOD benchmark. Competing methods that require training are referred from OpenOOD. These methods utilize the full ImageNet training set and hold an unfair advantage over zero-shot, training-free methods like ours. Our AdaNeg consistently outperforms its closest competitor [27] in both near-OOD and far-OOD scenarios. Additionally, AdaNeg not only enhances OOD detection capabilities but also improves ID classification, presenting higher robustness under diverse conditions.

Results on CIFAR10/100 datasets are provided in Appendix A.5, where our advantages still hold.

Table 1: OOD detection results with ImageNet-1k, where a VITB/16 CLIP encoder is adopted.

Methods	OOD datasets									
	INaturalist		Sun		Places		Textures		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Methods requiring training (or fine-tuning)										
MSP [20]	87.44	58.36	79.73	73.72	79.67	74.41	79.69	71.93	81.63	69.61
ODIN [36]	94.65	30.22	87.17	54.04	85.54	55.06	87.85	51.67	88.80	47.75
Energy [37]	95.33	26.12	92.66	35.97	91.41	39.87	86.76	57.61	91.54	39.89
GradNorm [25]	72.56	81.50	72.86	82.00	73.70	80.41	70.26	79.36	72.35	80.82
ViM [64]	93.16	32.19	87.19	54.01	83.75	60.67	87.18	53.94	87.82	50.20
KNN [57]	94.52	29.17	92.67	35.62	91.02	39.61	85.67	64.35	90.97	42.19
VOS [13]	94.62	28.99	92.57	36.88	91.23	38.39	86.33	61.02	91.19	41.32
NPOS [59]	96.19	16.58	90.44	43.77	89.44	45.27	88.80	46.12	91.22	37.93
ZOC [15]	86.09	87.30	81.20	81.51	83.39	73.06	76.46	98.90	81.79	85.19
LSN [45]	95.83	21.56	94.35	26.32	91.25	34.48	90.42	38.54	92.96	30.22
CLIPN [65]	95.27	23.94	93.93	26.17	92.28	33.45	90.93	40.83	93.10	31.10
LoCoOp [42]	96.86	16.05	95.07	23.44	91.98	32.87	90.19	42.28	93.52	28.66
LAPT [76]	99.63	1.16	96.01	19.12	92.01	33.01	91.06	40.32	94.68	23.40
NegPrompt [34]	98.73	6.32	95.55	22.89	93.34	27.60	91.60	35.21	94.81	23.01
Zero-Shot Training-free Methods										
Mahalanobis [33]	55.89	99.33	59.94	99.41	65.96	98.54	64.23	98.46	61.50	98.94
Energy [37]	85.09	81.08	84.24	79.02	83.38	75.08	65.56	93.65	79.57	82.21
MCM [39]	94.59	32.20	92.25	38.80	90.31	46.20	86.12	58.50	90.82	43.93
NegLabel [27]	99.49	1.91	95.49	20.53	91.64	35.59	90.22	43.56	94.21	25.40
AdaNeg (Ours)	**99.71**	**0.59**	**97.44**	**9.50**	**94.55**	**34.34**	**94.93**	**31.27**	**96.66**	**18.92**

Table 2: OOD detection results on the OpenOOD benchmark, where ImageNet-1k is adopted as ID dataset. Full results are available in Tab. A7.

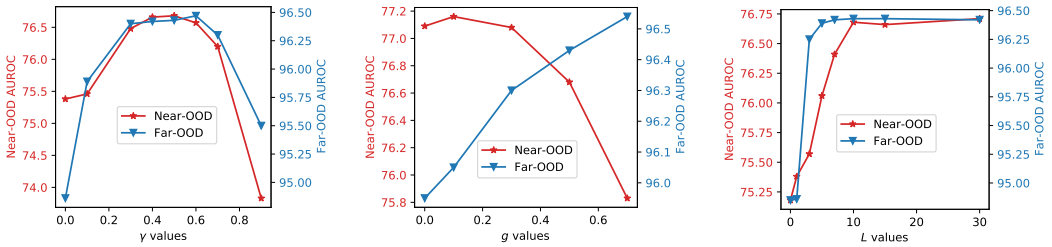
Methods	FPR95 ↓		AUROC ↑		ACC ↑
	Near-OOD	Far-OOD	Near-OOD	Far-OOD	ID
Methods requiring training (or fine-tuning)					
GEN [38]	–	–	78.97	90.98	81.59
AugMix [23] + ReAct [56]	–	–	79.94	93.70	77.63
RMDS [49]	–	–	80.09	92.60	81.14
SCALE [70]	–	–	81.36	96.53	76.18
AugMix [23] + ASH [11]	–	–	82.16	96.05	77.63
LAPT [76]	58.94	24.86	82.63	94.26	67.86
Zero-shot Training-free Methods					
MCM [39]	79.02	68.54	60.11	84.77	66.28
NegLabel [27]	69.45	23.73	75.18	94.85	66.82
AdaNeg (Ours)	**67.51**	**17.31**	**76.70**	**96.43**	**67.13**

4.3 Analyses and Discussions

Score Functions. As illustrated in Tab. 3, the adaptive score functions S_{ta} and S_{sa} consistently outperforms the fixed S_{nl}, validating the advantage of adaptive proxies over fixed label proxies. The sample-adaptive score S_{sa} slightly surpasses the task-adaptive one S_{ta}, verifying the usefulness of fine-grained sample characteristics. Combining adaptive image-based proxies and fixed text-based proxies leads to the best performance, proving their complementarity.

Table 3: OOD detection results with different score functions, where results are reported with ImageNet ID dataset under the OpenOOD setup.

S_{nl}	S_{ta}	S_{sa}	Near-OOD AUROC ↑	Far-OOD AUROC ↑
✓			75.18	94.85
	✓		75.76	96.20
		✓	76.03	96.35
✓	✓		76.49	96.45
✓		✓	76.70	96.63



(a) Threshold γ (b) Gap value g (c) Memory length L

Figure 3: Analyses on the hyper-parameters of (a) threshold γ in Eq. 8, (b) gap value g in Eq. 8, and (c) memory length L on the ImageNet dataset under OpenOOD setting.

Table 4: Analyses on the time complexity of our AdaNeg and its competitors. 'Training' measures the training time, and 'Param.' presents the number of learnable parameters. 'FPS' reflects the inference speed, measured with a batch size of 256. Results are achieved with a NVIDIA V100 GPU.

Methods	**Training**	**FPS**	**Param.**	**FPR95** ↓
ZOC [15]	>24h	287	336M	85.19
CLIPN [65]	>24h	605	37.8M	31.10
LoCoOp [42]	9h	625	8K	28.66
MCM [39]	–	625	–	43.93
NegLabel [27]	–	592	–	25.40
AdaNeg (Ours)	–	476	–	18.92

Threshold γ and gap g in Eq. 8. As illustrated in Fig. 3, employing a moderate threshold γ (*e.g.*, $0.4 < \gamma < 0.6$) proves effective in distinguishing OOD samples across various scenarios. However, the efficacy of the gap parameter g depends upon the specific characteristics of different OOD datasets. A larger g facilitates the identification of more confidently classified ID/OOD samples, thereby improving the detection of far-OOD samples. Conversely, a smaller g is advantageous for near-OOD detection as it accommodates low-confidence OOD samples, which are typically more prevalent in near-OOD scenarios.

In scenarios where the OOD distribution is entirely unknown, we adopt a conservative approach by setting g to 0.5 in all experiments. This balanced setting provides a robust baseline for performance across a variety of conditions. However, if prior knowledge regarding the difficulty levels of OOD datasets is accessible, tailoring the hyperparameters—such as opting for a smaller g in more challenging OOD contexts—can yield enhanced detection performance.

ID-OOD imbalanced Test Data. To investigate the stability of our method with imbalanced ID-OOD test data, we construct test sets with various mixture ratios of ID and OOD samples. Specifically, we adopted the 1.28M ImageNet training data as ID and randomly sampled 12.8K and 1.28K instances from the SUN OOD dataset to construct the ID:OOD ratios of 100:1 and 1000:1 settings, respectively. To construct the ID:OOD ratios of 1:100, 1:10, 1:1, and 10:1 settings, we randomly sampled 40K samples from the SUN dataset as OOD and randomly sampled 400, 4K, 40K, and 400K instances from the ImageNet training data. As shown in Tab. 5, our method outperforms NegLabel across a wide range of mixture ratios (from 1:100 to 100:1), validating the robustness and reliability of our approach. However, unbalanced mixture ratios do pose a challenge to our method. Our approach performs the best in scenarios with a balanced mixture of ID and OOD samples, reducing the FPR95 by 11.18%. As the mixture ratio becomes increasingly unbalanced, the improvement brought by our method gradually decreases. When the unbalanced ratio reaches 1000:1, our method shows some negative impact.

After a more detailed analysis, we discovered that the fundamental issue stems from the increased proportion of misclassified samples in the memory due to the growing ID-OOD imbalance. To effectively address this problem, we refine the selection criterion for memorizing OOD samples by adaptively adjusting the gap g. We refer to this adaptive gap strategy as AdaGap, which significantly improves the robustness of our method against ID-OOD imbalanced test data, as demonstrated in

Table 5: FPR95 (↓) with different mixture ratios of ID and OOD samples.

ID:OOD Ratio	1:100	1:10	1:1	10:1	100:1	1K:1
NegLabel	22.42	21.11	20.99	20.92	21.48	23.69
AdaNeg	21.00	12.49	9.81	15.61	20.71	26.28
AdaNeg (With AdaGap)	20.50	12.22	9.73	12.98	15.61	18.43

Table 6: OOD detection results with BIMCV-COVID19+ [61], where a VITB/16 CLIP encoder is adopted.

Methods	OOD datasets CT-SCAN		X-Ray-Bone		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
NegLabel	63.53	100	99.68	0.56	81.61	50.28
AdaNeg (Ours)	**93.48**	100	**99.99**	**0.11**	**96.74**	**50.06**

Table 5. To summarize, we first estimate the ratio of ID to OOD in the test data online. If there is a higher proportion of ID/OOD compared to OOD/ID, we adjust the standards for caching ID/OOD data into memory by dynamically modifying the gap in Eq. 8. For more detailed implementation, please refer to the Appendix A.6.

Generalization to Other Domains. Besides experiments with natural images, we further validate our method on the BIMCV-COVID19+ dataset [61], which includes medical images, following the OpenOOD setup. Specifically, we select BIMCV as the ID dataset, which includes chest X-ray images CXR (CR, DX) of COVID-19 patients and healthy individuals. For the OOD datasets, we follow the OpenOOD setup and use CT-SCAN and X-Ray-Bone datasets. The CT-SCAN dataset includes computed tomography (CT) images of COVID-19 patients and healthy individuals, while the X-Ray-Bone dataset contains X-ray images of hands. As illustrated in Tab. 6, our AdaNeg method consistently outperforms NegLabel on this medical image dataset.

Memory Length. As demonstrated in Fig. 3c, there is a positive correlation between the memory length L and the performance outcomes, affirming the efficacy of feature memorization from another perspective. In all our experiments, we have set L to 10.

Complexity Analyses. As analyzed in Table 4, our AdaNeg approach does not introduce any learnable parameters or require model training. Furthermore, it significantly enhances performance while maintaining a fast testing speed.

More analyses and discussions on the λ in Eq. 13, β in Eq. 11, various backbone architectures, the ordering of test samples, complementarity to training-based method, and the number of test data can be found in Appendix A.6.

5 Conclusion and Limitations

We proposed adaptive negative proxies that aligned more effectively with OOD distributions, thereby providing more effective guidance for OOD detection. These proxies were constructed using a task-aware feature memory that selectively cached discriminative image features during testing. Our approach facilitated the generation of both task-adaptive and sample-adaptive proxies through carefully designed memory reading operations. Notably, our method was training-free and annotation-free, and it maintained fast testing speed and achieved state-of-the-art results on various benchmarks. These results validated the effectiveness of the proposed adaptive proxies.

One minor limitation of our method is the introduction of an external memory requirement. For example, this memory occupies a storage space of 214.75MB when using the ImageNet dataset as the ID, which may pose challenges for storage-constrained applications.

References

[1] Yichen Bai, Zongbo Han, Changqing Zhang, Bing Cao, Xiaoheng Jiang, and Qinghua Hu. Id-like prompt learning for few-shot out-of-distribution detection. *arXiv preprint arXiv:2311.15243*, 2023.

[2] Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *ICML*, 2023.

[3] Chaoqi Chen, Jiongcheng Li, Hong-Yu Zhou, Xiaoguang Han, Yue Huang, Xinghao Ding, and Yizhou Yu. Relation matters: Foreground-aware graph-based relational reasoning for domain adaptive object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3677–3694, 2022.

[4] Chaoqi Chen, Luyao Tang, Yue Huang, Xiaoguang Han, and Yizhou Yu. Coda: generalizing to open and unseen domains with compaction and disambiguation. *Advances in Neural Information Processing Systems*, 36, 2023.

[5] Chaoqi Chen, Luyao Tang, Leitian Tao, Hong-Yu Zhou, Yue Huang, Xiaoguang Han, and Yizhou Yu. Activate and reject: towards safe domain generalization under category shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11552–11563, 2023.

[6] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10337–10346, 2020.

[7] 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*, pages 3606–3613, 2014.

[8] Hanming Deng, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan. Object guided external memory network for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6678–6687, 2019.

[9] 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*, pages 248–255. Ieee, 2009.

[10] 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.

[11] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. *arXiv preprint arXiv:2209.09858*, 2022.

[12] Xuefeng Du, Gabriel Gozum, Yifei Ming, and Yixuan Li. Siren: Shaping representations for detecting out-of-distribution objects. *Advances in Neural Information Processing Systems*, 35:20434–20449, 2022.

[13] Xuefeng Du, Xin Wang, Gabriel Gozum, and Yixuan Li. Unknown-aware object detection: Learning what you don't know from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13678–13688, 2022.

[14] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022.

[15] 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, pages 6568–6576, 2022.

[16] Ke Fan, Yikai Wang, Qian Yu, Da Li, and Yanwei Fu. A simple test-time method for out-of-distribution detection. *arXiv preprint arXiv:2207.08210*, 2022.

[17] 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.

[18] Zhitong Gao, Shipeng Yan, and Xuming He. Atta: Anomaly-aware test-time adaptation for out-of-distribution detection in segmentation. *Advances in Neural Information Processing Systems*, 36:45150–45171, 2023.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

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

[22] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019.

[23] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

[24] Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16783–16792, 2022.

[25] 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.

[26] 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*, pages 8710–8719, 2021.

[27] 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 *The Twelfth International Conference on Learning Representations*, 2024.

[28] Geethan Karunaratne, Manuel Schmuck, Manuel Le Gallo, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. Robust high-dimensional memory-augmented neural networks. *Nature communications*, 12(1):2468, 2021.

[29] Shu Kong and Deva Ramanan. Opengan: Open-set recognition via open data generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2021.

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

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

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

[33] 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.

[34] Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. Learning transferable negative prompts for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17584–17594, 2024.

[35] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *arXiv preprint arXiv:2303.15361*, 2023.

[36] 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.

[37] 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.

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

[39] 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.

[40] Yifei Ming and Yixuan Li. How does fine-tuning impact out-of-distribution detection for vision-language models? *International Journal of Computer Vision*, 132(2):596–609, 2024.

[41] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? *arXiv preprint arXiv:2203.04450*, 2022.

[42] 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.

[43] 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, page 7. Granada, Spain, 2011.

[44] 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*, pages 427–436, 2015.

[45] 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.

[46] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019.

[47] Charles Packer, Vivian Fang, Shishir G Patil, Kevin Lin, Sarah Wooders, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

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

[49] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

[50] Seonghan Ryu, Sangjun Koo, Hwanjo Yu, and Gary Geunbae Lee. Out-of-domain detection based on generative adversarial network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 714–718, 2018.

[51] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. PMLR, 2016.

[52] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012.

[53] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021.

[54] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

[55] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. *Advances in neural information processing systems*, 28, 2015.

[56] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021.

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

[58] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020.

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

[60] 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*, pages 8769–8778, 2018.

[61] Maria De La Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, et al. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. *arXiv preprint arXiv:2006.01174*, 2020.

[62] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? *arXiv preprint arXiv:2110.06207*, 2021.

[63] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

[64] 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*, pages 4921–4930, 2022.

[65] 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*, pages 1802–1812, 2023.

[66] Yezhen Wang, Bo Li, Tong Che, Kaiyang Zhou, Ziwei Liu, and Dongsheng Li. Energy-based open-world uncertainty modeling for confidence calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9302–9311, 2021.

[67] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

[68] 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*, pages 3485–3492. IEEE, 2010.

[69] Guo-Sen Xie, Huan Xiong, Jie Liu, Yazhou Yao, and Ling Shao. Few-shot semantic segmentation with cyclic memory network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7293–7302, 2021.

[70] Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for training time and post-hoc out-of-distribution detection enhancement. *arXiv preprint arXiv:2310.00227*, 2023.

[71] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. Openood: Benchmarking generalized out-of-distribution detection. 2022.

[72] Puning Yang, Jian Liang, Jie Cao, and Ran He. Auto: Adaptive outlier optimization for online test-time ood detection. *arXiv preprint arXiv:2303.12267*, 2023.

[73] Jingyang Zhang, Nathan Inkawhich, Yiran Chen, and Hai Li. Fine-grained out-of-distribution detection with mixup outlier exposure. *CoRR*, (abs/2106.03917), 2021.

[74] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023.

[75] Yabin Zhang, Bin Deng, Hui Tang, Lei Zhang, and Kui Jia. Unsupervised multi-class domain adaptation: Theory, algorithms, and practice. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2775–2792, 2020.

[76] 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*, 2024.

[77] Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. Dual memory networks: A versatile adaptation approach for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 28718–28728, 2024.

[78] 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.

Table A7: Detailed OOD detection results on the OpenOOD benchmark, where ImageNet-1k is adopted as ID dataset.

Near-/Far-OOD	Datasets	FPR95 ↓	AUROC ↑
Near-OOD	SSB-hard [62]	74.91	75.11
	NINCO [2]	60.10	78.30
	Mean	67.51	76.70
Far-OOD	iNaturalist [60]	0.72	99.72
	Textures [7]	21.40	95.71
	OpenImage-O [64]	29.81	93.87
	Mean	17.31	96.43

A Appendix

A.1 ID-Similarity to OOD Ratio with ground truth ID and OOD labels

We introduce the ID-Similarity to OOD Ratio (ISOR) to quantitatively measure the relative alignment of negative proxies with ground truth OOD labels, compared to their alignment with ID labels. In implementation, we adapt the score function of NegLabel (*i.e.*, Eq. 4), which originally measures the similarity of a test image to ID labels over negative labels. We modify this function by replacing the negative labels with ground truth OOD labels and changing the input from test images to negative proxies (*e.g.*, negative texts), while keeping other aspects consistent with Eq. 4. This modified score function allows us to assess the degree of similarity between the inputs and ground truth ID/OOD labels, thereby quantifying the relative alignment between negative proxies and OOD labels. Specifically, lower ISOR indicates a higher similarity to OOD labels and a reduced similarity to ID labels.

A.2 Entropy-based caching strategy for full memory

The memory we construct is of finite length; hence, as the number of cached images increases, it may become fully occupied. To address this, we have devised a simple strategy to retain only those images with low entropy, *e.g.*, high confidence. Specifically, when storing an image in memory, we also record its entropy pertinent to OOD detection:

$$\text{Entropy}(\boldsymbol{v}) = -S_{nl}(\boldsymbol{v}) \log(S_{nl}(\boldsymbol{v})) - (1 - S_{nl}(\boldsymbol{v})) \log(1 - S_{nl}(\boldsymbol{v})), \tag{A.1}$$

where $S_{nl}(\boldsymbol{v})$ represents the probability of belonging to the ID, as shown in Eq. (4). Given the entropy of the current test image and a full memory $\boldsymbol{M}_y$, we replace the item with the maximum entropy in $\boldsymbol{M}_y$ with the current image feature if the current test image exhibits lower entropy. Otherwise, we do not cache the current test sample.

A.3 Evaluation criteria

Following common practice [26, 39, 27], we employ the following three evaluation metrics: (1) FPR95, which measures the false positive rate for OOD samples when the detection accuracy for ID samples is at 95%; (2) AUROC, the area under the receiver operating characteristic curve; and (3) ID ACC, which quantifies the accuracy of correctly identifying and classifying ID samples.

A.4 Detailed results on ImageNet dataset under OpenOOD setting

The detailed OOD detection results on the OpenOOD benchmark are presented in Tab. A7.

A.5 Results on CIFAR10/100

As illustrated in Tab. A8, the advantage of our AdaNeg also holds on the CIFAR10/100 dataset. Notably, our method achieves new state-of-the-art results in the far-OOD setting under a zero-shot training-free manner, even outperforming its competitors training on the full labeled training set.

Table A8: OOD detection results with CIFAR10/100 on the OpenOOD benchmark. Full results are provided in Tables A10 and A9.

Methods	FPR95 ↓		AUROC ↑	
	Near-OOD	Far-OOD	Near-OOD	Far-OOD
Methods requiring training (or fine-tuning)				
PixMix [24] + KNN [57]	–	–	93.10	95.94
OE [21] + MSP [20]	–	–	94.82	96.00
PixMix [24] + RotPred [22]	–	–	94.86	98.18
Zero-shot Training-free Methods				
MCM [39]	30.86	17.99	91.92	95.54
NegLabel [27]	28.75	6.60	94.58	98.39
AdaNeg (Ours)	**20.40**	**2.79**	**94.78**	**99.26**

(a) CIFAR10 as ID dataset

Methods	FPR95 ↓		AUROC ↑	
	Near-OOD	Far-OOD	Near-OOD	Far-OOD
Methods requiring training (or fine-tuning)				
GEN [38]	–	–	81.31	79.68
VOS [14] + EBO [37]	–	–	80.93	81.32
SCALE [70]	–	–	80.99	81.42
OE [21] + MSP [20]	–	–	88.30	81.41
Zero-shot Training-free Methods				
MCM [39]	75.20	59.32	71.00	76.00
NegLabel [27]	71.44	40.92	70.58	89.68
AdaNeg (Ours)	**59.07**	**29.35**	**84.62**	**95.25**

(b) CIFAR100 as ID dataset

Table A9: Detailed OOD detection results of on the OpenOOD benchmark, where CIFAR100 is adopted as ID dataset.

Near-/Far-OOD	Datasets	FPR95 ↓	AUROC ↑
Near-OOD	CIFAR10 [30]	58.24	79.91
	TIN [32]	59.90	89.34
	Mean	59.07	84.62
Far-OOD	MNIST [10]	4.18	97.90
	SVHN [43]	6.03	97.60
	Texture [7]	30.00	95.14
	Places365 [78]	77.20	90.35
	Mean	29.35	95.25

A.6 More analyses

Detailed Implementation of AdaGap Module. We have implemented an adaptive gap (AdaGap) module to adjust the memorization selection criteria dynamically. This strategy builds on the observation that as the score S_{nl} increases/decreases, the probability that a sample is ID/OOD also increases accordingly. By enforcing a stringent selection criterion, we can effectively minimize the inclusion of misclassified samples in our memory. Specifically, we first online estimate the ratio of ID to OOD samples in the test data using a First-In-First-Out queue, which caches the ID/OOD estimation (cf. Eq. 8) of the most recent N samples:

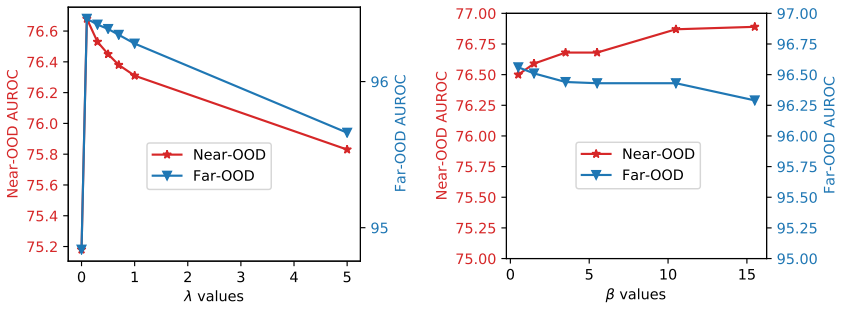
$$MR = \frac{\text{Estimated ID Number}}{\text{Estimated ID Number + Estimated OOD number}}, \tag{A.2}$$

where the ID and OOD numbers are acquired within the queue.

Leveraging the estimated mix ratio (MR), we can dynamically adjust the gap g in memory caching to avoid a majority of misclassified samples within the memory. For instance, if we find that ID

Table A10: Detailed OOD detection results of our method on the OpenOOD benchmark, where CIFAR10 is adopted as ID dataset.

Near-/Far-OOD	Datasets	FPR95 ↓	AUROC ↑
Near-OOD	CIFAR100 [30]	35.80	90.93
	TIN [32]	5.01	98.63
	Mean	20.40	94.78
Far-OOD	MNIST [10]	0.13	99.96
	SVHN [43]	0.04	99.97
	Texture [7]	0.04	99.82
	Places365 [78]	10.93	97.29
	Mean	2.79	99.26



(a) λ value (b) β value

Figure A4: Analyses on the hyper-parameters of (a) λ in Eq. 13 and (b) β in Eq. 11 on the ImageNet dataset under OpenOOD setting.

samples constitute the majority of the test samples (*i.e.*, $MR > 0.5$), this could lead to an increased proportion of ID samples in the OOD memory. In response, we can adjust the selection criterion for OOD memorization to only cache higher confidence OOD samples. This adjustment is achieved by modifying the selection criterion for memorization in Eq. 8 as follows:

$$\begin{aligned} \text{Negative}: & S_{nl}(\boldsymbol{v}) < \gamma - g\gamma \rightarrow S_{nl}(\boldsymbol{v}) < \gamma - \max(g, MR)\gamma, \\ \text{Positive}: & S_{nl}(\boldsymbol{v}) \geq \gamma + g(1-\gamma) \rightarrow S_{nl}(\boldsymbol{v}) \geq \gamma + \max(g, 1-MR)(1-\gamma), \end{aligned} \tag{A.3}$$

where $g = 0.5$ is the default gap analyzed in Figure 3b. In this way, under ID/OOD balanced conditions (*i.e.*, $MR = 0.5$), our method aligns with our original version. However, if the proportion of ID samples is higher in the test samples' estimation (*i.e.*, $MR > 0.5$), we raise the standard for storing negative samples in the memory. In an extreme case when $MR = 1$, we estimate that there might be no OOD samples among the test samples; thus, we stop storing test samples in the negative memory and only selectively cache test samples into the positive memory. We adjust our approach conversely when the MR value is lower than 0.5.

Please note that the selection criterion is dynamically adjusted online because the MR is estimated with the most recent N test samples. Here, we set $N = 10,000$ by default. This dynamic adjustment ensures that our memory caching strategy remains responsive to the evolving nature of the test sample distribution, thereby optimizing memory utilization and enhancing the accuracy of our domain distinction process.

λ in Eq. 13. As illustrated in Fig. A4a, the OOD detection performance remains robust across a wide range of λ values, *e.g.*, from 0.01 to 1. For all experiments, we have set λ to 0.1.

β in Eq. 11. Results with different β values are shown in Fig. A4b, where OOD detection performance is robust to the β values. We adopt β=5.5 in all experiments.

Ordering of Testing Data. Our AdaNeg selectively caches test data into memory, potentially causing variations in the results depending on the ordering of the test data. To rigorously test this

aspect, we randomly shuffled the order of the test data using three distinct seeds. We observed that our method exhibits robustness to changes in the ordering of test data. Specifically, across three experiments conducted on the ImageNet dataset, the AUROC scores were 96.65%, 96.69%, and 96.64%, respectively, demonstrating fluctuations of less than 0.1%. We report the average results from three random runs in our paper.

Various Backbone Architectures. Results with various VLMs architectures are illustrated in Tab. A11, where better results are typically achieved with stronger backbones.

Table A11: OOD detection results of our AdaNeg with different VLMs architectures, where ImageNet-1K is used as the ID dataset.

Backbone	OOD datasets									
	INaturalist		Sun		Places		Textures		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
ResNet50	99.58	1.18	97.37	10.56	93.84	43.19	94.18	35.00	96.24	22.48
VITL/32	99.59	1.02	97.53	9.63	93.99	38.45	94.21	37.92	96.33	21.76
VITB/16	99.71	0.59	97.44	9.50	94.55	34.34	94.93	31.27	96.66	18.92

Complementarity to Training-based Method. We validated the complementarity between our AdaNeg method and the latest works, NegPrompt [34] and LAPT [76], which use learned prompts. As shown in Table A12, our method significantly improves performance over these approaches, demonstrating its complementarity with training-based methods.

Table A12: FPR95 (↓) with the ID dataset of ImageNet.

Methods	INaturalist	SUN	Places	Textures	Average
NegPrompt [34]	6.76	23.41	28.32	34.57	23.27
+ **AdaNeg**	3.87	11.35	25.45	29.79	17.62
LAPT [76]	1.10	20.59	35.38	40.11	24.29
+ **AdaNeg**	**0.58**	**9.98**	**30.47**	**24.25**	**16.32**

Number of Testing Data. We examine the dependency of our approach on the number of test images by evaluating its performance across different scales of test samples. As the number of test samples increases (from 900 to 90K), the cached feature data also increases, leading to an improvement in our method's results, as shown in Tab. A13. Even with a small number of test samples (*e.g.*, 90 and 900), our method significantly reduces FPR95 compared to NegLabel, demonstrating its robustness across different numbers of test images.

Note that with only 90 test images, the task of distinguishing between ID and OOD samples degenerates into a simpler task since the number of test images is even smaller than the number of classes (*e.g.*, 1000 for ImageNet). Consequently, both NegLabel and our method achieve lower FPR95 in such an easier scenario.

Table A13: FPR95 (↓) with different numbers of test images, where test samples are randomly sampled from ImageNet (ID) and SUN (OOD) datasets, and we maintain a consistent ratio of ID to OOD samples at 5:4 throughout the experiments.

Num. of Test Images	90	900	9K	45K	90K
NegLabel	14.00	20.44	20.71	20.51	20.53
AdaNeg	**6.00**	**10.12**	**9.78**	**9.66**	**9.50**

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification:

   Guidelines: This paper contributes the adaptive negative proxies for OOD detection, which matches the main claims in the abstract and introduction.

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Please refer to Sec. 5 in the main paper.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

   Answer: [NA]

Justification: The main contributions are the algorithm design and experimental validation. No theory proofs are provided.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

   Answer: [Yes]

   Justification: We provide detailed algorithm designs in Sec. 3 and hyper-parameters in Sec 4.1 to facilitate reproducibility.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: The adopted datasets are publicly available. Codes are available at `https://github.com/YBZh/OpenOOD-VLM`.

Guidelines:

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

6. **Experimental Setting/Details**

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

Answer: [Yes]

Justification: The proposed method does not require training, and the testing details are illustrated in Sec 4.1.

Guidelines:

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

7. **Experiment Statistical Significance**

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

Answer: [Yes]

Justification: As analyzed in Sec. A.6 in the appendix, we examined the experimental results on the ImageNet dataset with different orderings of the testing data. We found that the order of the test data has a minimal impact on the final results, *e.g.*, the variation in three random experiments was less than 0.1%, which verified the statistical stability of our method.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: Our experiments can be reproduced with only one V100 GPU, as shown in Sec. 4.1.

   Guidelines:

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

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: I have read the NeurIPS Code of Ethics and confirm that this paper conforms.

   Guidelines:

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

10. **Broader Impacts**

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

    Answer: [Yes]

    Justification: **Positive Societal Impacts:** The proposed method for OOD detection can enhance the reliability and safety of AI systems, particularly in critical applications such as healthcare, autonomous driving, and cybersecurity. By improving the detection of out-of-distribution data, the system can better avoid potentially harmful decisions based on

anomalous inputs, thereby increasing trust in AI technologies and their deployment in various fields.

Negative Societal Impacts: One potential negative impact is the risk of privacy issues due to online caching and processing of test data, which may inadvertently store sensitive information. Additionally, the improved OOD detection might be misused in surveillance and monitoring applications, leading to ethical concerns regarding privacy and autonomy.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The used codes and datasets are well cited.

    Guidelines:

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

13. **New Assets**

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

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer:[NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

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