ANTS: Adaptive Negative Textual Space Shaping for OOD Detection via Test-Time MLLM Understanding and Reasoning

Wenjie Zhu[1,2*] Yabin Zhang[3*] Xin Jin[2,4] Wenjun Zeng[2†] Lei Zhang[1†]
[1]The Hong Kong Polytechnic University [2]Eastern Institute of Technology, Ningbo
[3]Harbin Institute of Technology (Shenzhen) [4]Zhongguancun Academy
22040319r@connect.polyu.hk, wzeng-vp@eitech.edu.cn, cslzhang@comp.polyu.edu.hk

Abstract

The introduction of negative labels (NLs) has proven effective in enhancing Out-of-Distribution (OOD) detection. However, existing methods often lack an understanding of OOD images, making it difficult to construct an accurate negative space. Furthermore, the absence of negative labels semantically similar to ID labels constrains their capability in near-OOD detection. To address these issues, we propose shaping an Adaptive Negative Textual Space (ANTS) by leveraging the understanding and reasoning capabilities of multimodal large language models (MLLMs). Specifically, we cache images likely to be OOD samples from the historical test images and prompt the MLLM to describe these images, generating expressive negative sentences that precisely characterize the OOD distribution and enhance far-OOD detection. For the near-OOD setting, where OOD samples resemble the in-distribution (ID) subset, we cache the subset of ID classes that are visually similar to historical test images and then leverage MLLM reasoning to generate visually similar negative labels tailored to this subset, effectively reducing false negatives and improving near-OOD detection. To balance these two types of negative textual spaces, we design an adaptive weighted score that enables the method to handle different OOD task settings (near-OOD and far-OOD), making it highly adaptable in open environments. On the ImageNet benchmark, our ANTS significantly reduces the FPR95 by 3.1%, establishing a new state-of-the-art. Furthermore, our method is training-free and zero-shot, enabling high scalability. Codes are available at `https://github.com/ZhuWenjie98/ANTS`.

1. Introduction

Deep neural networks (DNNs) have achieved remarkable performance in classifying test samples that fall into the training distribution [10, 12]. However, it is well-known that DNNs tend to misclassify test samples from unknown classes, which are often called out-of-distribution (OOD) data [13]. Unfortunately, OOD data are inevitably encountered in open environments. Therefore, how to effectively identify OOD data is crucial for the reliable deployment of DNN models in open-world scenarios.

Traditional OOD detection methods in the image domain primarily rely on visual modality information [16, 42, 45, 48]. For example, MSP [13] utilizes the maximum softmax probability of a pre-trained vision model to detect OOD images. Recently, multimodal knowledge has attracted increasing attention in OOD detection [2, 19, 25, 30, 33, 34, 52, 57, 58]. In particular, NegLabel [19] introduces negative labels (NLs) by mining words that are semantically distant from in-distribution (ID) labels, and identifies OOD images

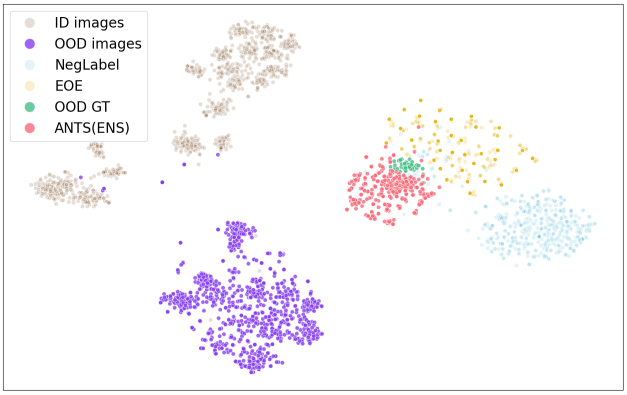


Figure 1. T-SNE visualization of the ID and OOD image features, the text features of NegLabel [19], EOE [4], OOD ground-truth, and the expressive negative sentences (ENS) of ANTS. We select ImageNet and SUN as the ID and OOD datasets, respectively. NegLabel and EOE lack a good understanding of OOD images, resulting in a greater distance between the OOD images and the text features. In contrast, our ANTS utilizes the MLLMs to understand OOD images during ENS generation, reducing the distance between ENS and OOD images and improving OOD detection performance.

*These authors contributed equally to this work.
†Corresponding authors.

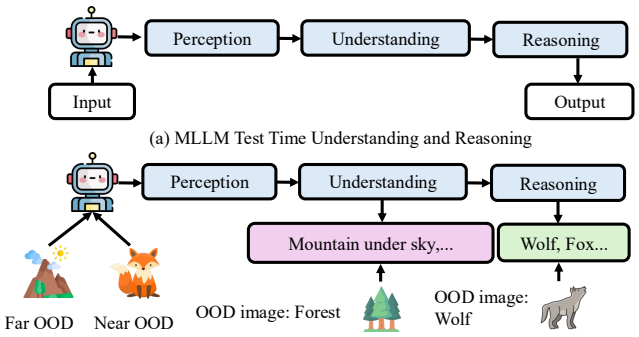


Figure 2. (a) Current MLLM improve their reasoning abilities by test time understanding and reasoning through chain-of-thought (CoT) prompting. (b) In our work, we leverage the test time understanding and reasoning capabilities of MLLM during inference to help visual-language models perform better on OOD detection.

by comparing their similarities to NLs and ID labels. Similar approaches generate NLs by prompting LLMs [4] or modifying superclass names [5]. Although these methods have achieved promising performance, they suffer from three key limitations. First, due to the lack of understanding of OOD images, the NLs are positioned far from the OOD image, as illustrated in Fig. 1. Second, these methods struggle with the challenging near-OOD setting, where OOD samples are semantically close to ID labels. NegLabel focuses on generating NLs that are semantically distant from ID classes, inherently overlooking such cases. While EOE [4] introduces visually similar labels for all ID classes to address this problem, it neglects the fact that OOD samples are typically similar to only a subset of ID classes, resulting in many false negative labels (see Fig. 6a). Third, these methods rely on the strong assumption that the target task setting (*e.g.*, near OOD or far OOD) is known in advance, allowing for the tailored design of NL generation rules. However, this assumption limits their applicability in complex, unknown, and dynamically changing open environments.

To address these challenges, we propose to shape an Adaptive Negative Textual Space (ANTS) by harnessing the understanding and reasoning capabilities of multimodal large language models (MLLMs) [1, 23, 28], as shown in Fig. 2. Specifically, we introduce expressive negative sentences (ENS), which effectively capture fine-grained details of OOD images. These negative sentences are generated by prompting MLLMs to describe online-mined negative images, leveraging their multimodal understanding capabilities and significantly enhancing the traditional far-OOD detection. While ENS shows greater expressive power in identifying far-OOD samples, it faces challenges in handling the near-OOD setting, where OOD samples are semantically close to certain ID classes. To address this, we dynamically identify the subset of ID classes most similar to the negative images and utilize the reasoning capabilities of MLLMs to construct visually similar negative labels (VSNL) tailored for this subset. This targeted approach reduces false negative labels and improves performance in the near-OOD setting (see Fig. 6a). Finally, to ensure adaptability across diverse task settings in open environments, we introduce an adaptive weighted score function to balance the two distinct negative textual spaces. This dynamic mechanism enables the model to seamlessly handle both near-OOD and far-OOD scenarios without prior knowledge of the task settings. The overall framework is presented in Fig. 3.

We conduct extensive experiments to validate the advantages of our ANTS method. On the large-scale ImageNet dataset, our approach significantly reduces FPR95 by 3.1% and 3.25% in the far-OOD and challenging near-OOD detection settings, respectively. Moreover, our method operates in a zero-shot and training-free manner, demonstrating strong scalability across different MLLMs. We summarize our contributions as follows:

- We identify three limitations of existing NLs-based methods: (1) lack understanding of OOD images; (2) struggle to address the challenging near-OOD setting, where OOD samples are semantically close to ID labels; (3) rely on the strong assumption that the target task setting (*e.g.*, near-OOD or far-OOD) is known in advance.
- To overcome these limitations, we propose the ANTS approach by leveraging the understanding and reasoning capabilities of MLLMs. Specifically, we (1) introduce two strategies including Negative Images Mining and Visually Similar ID-Classes Mining to avoid interference from ID noise and generate false negative labels; (2) design two types of prompt for MLLMs to generate expressive negative sentences and visually similar negative labels; and (3) design an adaptive weighted score to dynamically balance these two text spaces in open environments.
- Extensive experiments are conducted to validate the proposed components. Our method demonstrates new state-of-the-art performance on both near-OOD and far-OOD detection tasks. Our method is training-free, zero-shot, and does not require any auxiliary outlier images.

2. Related Work

Traditional OOD Detection. Traditional OOD detection methods can be categorized into the following groups: (1) classification-based methods [9, 13–15, 18, 22, 26, 27, 29, 31, 35, 36, 39, 41, 42, 49, 53, 56] that distinguish ID and OOD samples by designing a score function; (2) density-based methods [17, 37, 46, 61] that detect OOD samples by evaluating the likelihood or density of test data derived from probabilistic models; (3) distance-based methods [32, 54] that detect OOD samples by measuring their deviation from in-distribution class prototypes.

OOD Detection with Vision Language Model. VLM-

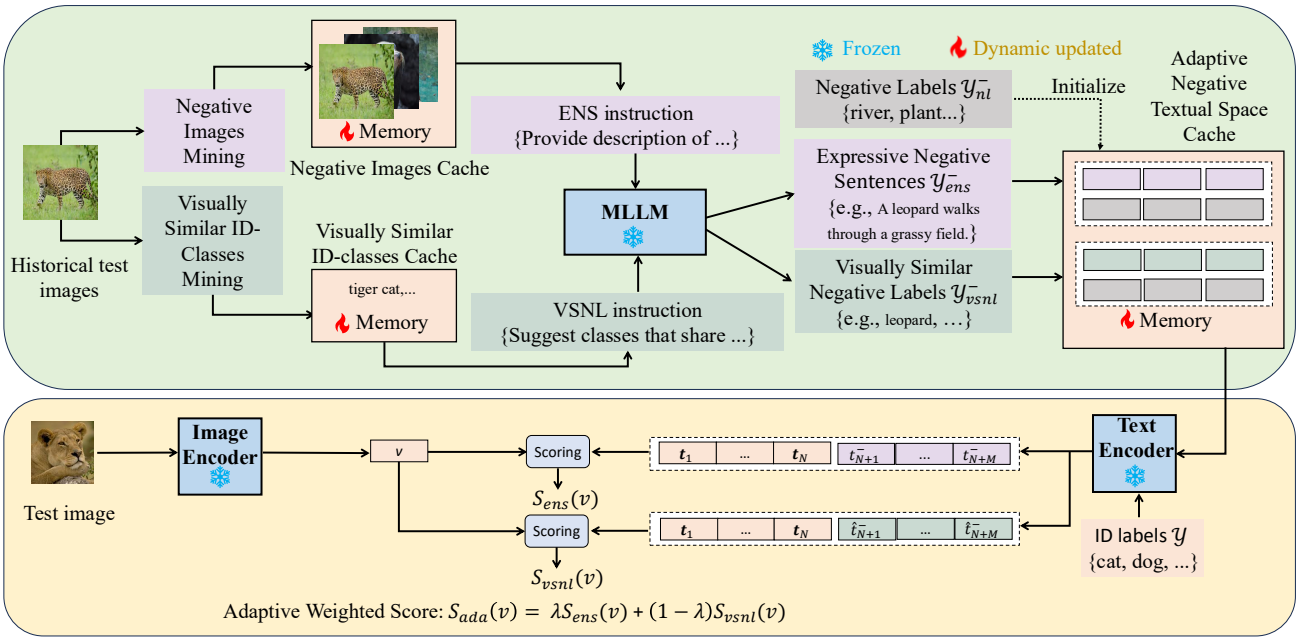


Figure 3. The overall framework of our ANTS. ANTS framework consists of in three stages: (1) caching negative images and visually similar ID classes mined from historical test images; (2) shaping two negative textual spaces by prompting an MLLM with the cached data to generate expressive negative sentences and visually similar labels; and (3) performing online evaluation of the test image using an adaptively weighted combination of these textual spaces.

based OOD detection methods can be broadly categorized into two settings: few-shot and zero-shot. Few-shot methods enhance OOD detection by using negative prompts to define boundaries between ID and OOD images [2, 25, 34], or by integrating non-ID or local ID regions for regularization [21, 33]. For zero-shot OOD detection, some works [20, 30, 51, 57, 60] design post-hoc strategies that utilize softmax scores or image feature information during testing. Some methods [47, 58] leverage auxiliary datasets to strengthen the detection of OOD samples. Other approaches [4, 5, 7, 11, 19, 36] retrieve negative labels from corpus databases or generating them using LLMs. However, these NLs methods lack understanding of actual OOD images, the semantic gap with OOD images limits their OOD detection capabilities.

3. Preliminary

OOD Detection Setup. Denote by $\mathcal{X}$ the image space and $\mathcal{Y} = \{y_1, \ldots, y_N\}$ the ID label space, with examples $\mathcal{Y} = \{cat, dog, \ldots, bird\}$ and N denoting the total number of classes. Given $\boldsymbol{x}_{in} \in \mathcal{X}$ as the ID random variable and $\boldsymbol{x}_{ood} \in \mathcal{X}$ as the OOD random variable, we denote their respective distributions as $\mathcal{P}_{\boldsymbol{x}_{in}}$ and $\mathcal{P}_{\boldsymbol{x}_{ood}}$. In closed-set scenarios, a test image $\boldsymbol{x}$ is expected to belong to one ID class, *i.e.*, $\boldsymbol{x} \in \mathcal{P}_{\boldsymbol{x}_{in}}$ and $y \in \mathcal{Y}$, where y is the label of $\boldsymbol{x}$. However, in real-world scenarios, AI systems may encounter samples that do not match any known class, *i.e.*, $\boldsymbol{x} \in \mathcal{P}_{\boldsymbol{x}_{ood}}$ and $y \notin \mathcal{Y}$, resulting in potential misclassifications and safety concerns [40]. To tackle these issues, OOD detection aims to distinguish ID and OOD samples using a scoring function S:

$$G_\gamma(\boldsymbol{x}) = \begin{cases} \text{ID} & S(\boldsymbol{x}) \geq \gamma, \\ \text{OOD} & S(\boldsymbol{x}) < \gamma, \end{cases} \tag{1}$$

where G_γ is the OOD detector with threshold γ.

OOD Detection with NLs. Enhancing OOD detection with textual knowledge has recently garnered increasing attention [30, 47, 57], while a representative type of approach introduces NLs [4, 19]. Specifically, in addition to the ID labels $\mathcal{Y}$, these methods introduce a disjoint set of NLs $\mathcal{Y}^-$ and classify a test sample as OOD if it exhibits high similarity to NLs and low similarity to ID labels. In this process, the quality of NLs is crucial. The pioneering method, NegLabel [19], selects words with large cosine distance to ID labels in a large corpus dataset $\mathcal{Y}^c = \{\widetilde{y}_1, \widetilde{y}_2, \ldots, \widetilde{y}_K\}$ as NLs:

$$\mathcal{Y}_{nl}^- = \mathcal{G}_{dis}(\mathcal{Y}, \mathcal{Y}^c, f_{clip}, M), \tag{2}$$

where the CLIP-like model f_{clip} defines the label similarity space. K and M represent the numbers of candidate labels in $\mathcal{Y}^c$ and the selected NLs in $\mathcal{Y}_{nl}^-$, where $M \leq K$. Another representative work, EOE [4], uses prompts to guide an LLM to generate NLs:

$$\mathcal{Y}_{eoe}^- = \mathcal{G}_{llm}(\mathcal{Y}, f_{llm}, \rho_{neg}, M), \tag{3}$$

where ρ_{neg} is a carefully designed textual prompt for the LLM f_{llm}. Given the generated NLs (*e.g.*, $\mathcal{Y}_{nl}^-$ [19]), the

score function for OOD detection can be formulated as:

$$S_{nl}(\boldsymbol{v}) = \frac{\sum_{y\in\mathcal{Y}} e^{\cos(\boldsymbol{v},\boldsymbol{t})/\tau}}{\sum_{y\in\mathcal{Y}} e^{\cos(\boldsymbol{v},\boldsymbol{t})/\tau} + \sum_{y^-\in\mathcal{Y}^-_{nl}} e^{\cos(\boldsymbol{v},\boldsymbol{t}^-)/\tau}}, \tag{4}$$

where $\tau > 0$ is the temperature scaling parameter. $\boldsymbol{v} \in \mathcal{R}^D$ represents the test image feature, while $\boldsymbol{t} \in \mathcal{R}^D$ and $\boldsymbol{t}^- \in \mathcal{R}^D$ denote the text features of ID labels $y \in \mathcal{Y}$ and NLs $y^- \in \mathcal{Y}^-_{nl}$, respectively, where D is the feature dimension.

4. Methodology

4.1. Motivation

Although NegLabel [19] and EOE [4] have advanced OOD detection using NLs, they face three key limitations: (1) lacking of understanding of OOD images, as shown in Fig.1; (2) poor performance in near-OOD settings due to false negatives by neglecting visually similar classes; and (3) reliance on prior task knowledge, limiting adaptability in open environments. This motivates us to raise the following question:

Can we leverage the test time understanding and reasoning capabilities of MLLMs to shape a more accurate and comprehensive negative textual space?

In this work, we attempt to answer this question by designing different prompts for MLLMs to leverage their test-time understanding and reasoning capabilities for OOD detection, as shown in Fig.2. The overall pipeline of our method is illustrated in Fig. 3.

4.2. Expressive Negative Sentences

Negative Images Mining. We leverage the image understanding capabilities of MLLMs to generate expressive negative sentences by describing negative images, which are historical test images likely to be OOD samples. We identify these negative images using the OOD detector of NegLabel, where historical test images with $\mathcal{S}_{nl}(\boldsymbol{x}) < \gamma$ are selected as negative images:

$$\mathcal{X}_{neg} = \{\boldsymbol{x} \mid S_{nl}(\boldsymbol{x}) < \gamma,\ \boldsymbol{x} \in \mathcal{X}^{\text{his}}_{test}\}, \tag{5}$$

where $\mathcal{X}^{\text{his}}_{test}$ denotes the historical test data. We find that manually defining a fixed γ is challenging for handling different testing scenarios, as the optimal threshold varies between different OOD datasets, as analyzed in Fig. 6b. To address this issue, we develop an adaptive threshold determination strategy based on the characteristics of the historical test data. Specifically, we filter out historical test samples with high $\mathcal{S}_{nl}$ scores using Eq. 5, as these samples are highly likely to be ID samples. For the remaining negative images $\widehat{\mathcal{X}}_{neg}$, which fall into a mixed set of ID and OOD samples, we select a proportion η of images with the lowest S_{nl} scores, and the adaptive threshold γ^* can be formulated as:

$$\mathcal{X}_{neg} = \text{Top}(\widehat{\mathcal{X}}_{neg}, \mathcal{O}_{nl}, \eta), \gamma^* \quad = \max_{\mathbf{x}\in\mathcal{X}_{neg}} S_{nl}(\mathbf{x}), \tag{6}$$

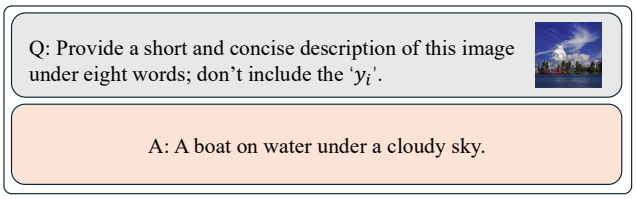


Figure 4. Expressive Negative Sentences, where y_i represents the predicted ID label of the negative image.

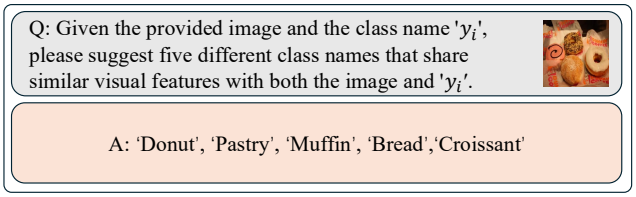


Figure 5. Visually Similar Negative Labels, where y_i represents the predicted ID label of the negative image.

where $\mathcal{O}_{nl} = \{-S_{nl}(\boldsymbol{x}) \mid \boldsymbol{x} \in \widehat{\mathcal{X}}_{neg}\}$, and $\eta \in (0, 1)$ determines the selection ratio. Here, function $\text{Top}(A, B, \eta)$ selects a proportion of η indices with the highest values in set B, and then retrieves the corresponding images from set A based on these indices. Since this approach relies on the distribution of $\mathcal{O}_{nl}$, it is equivalent to using an adaptive, data-dependent threshold γ, as illustrated in Fig. 6b.

ENS Generation and Score. With the mined negative images $\mathcal{X}_{neg}$, we introduce the expressive negative sentences as follows:

$$\mathcal{Y}^-_{ens} = \mathcal{G}_{ens}(\mathcal{Y}, \mathcal{X}_{neg}, f_{mllm}, M), \tag{7}$$

where $\mathcal{G}_{ens}$ is the negative sentence generation process detailed in Fig. 4. If $|\mathcal{X}_{neg}| \geq M$, we randomly select M sentences. Otherwise, we repeat the prompting process to generate M negative sentences. With the expressive negative sentences, we introduce the following negative score:

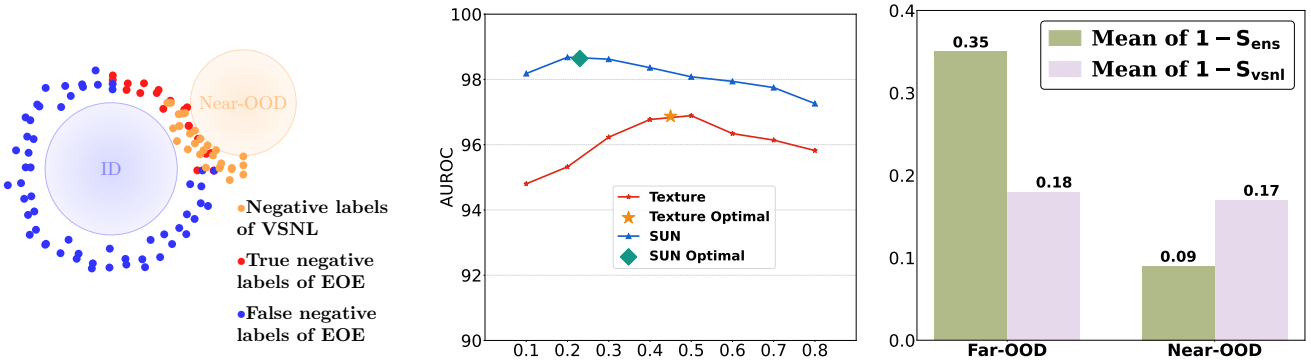
$$S_{ens}(\boldsymbol{v}) = \frac{\sum_{y\in\mathcal{Y}} e^{\cos(\boldsymbol{v},\boldsymbol{t})/\tau}}{\sum_{y\in\mathcal{Y}} e^{\cos(\boldsymbol{v},\boldsymbol{t})/\tau} + \sum_{y^-\in\mathcal{Y}^-_{ens}} e^{\cos(\boldsymbol{v},\boldsymbol{t}^-)/\tau}}. \tag{8}$$

4.3. Visually Similar Negative Labels

The expressive negative sentences introduced above can enhance the detection of far-OOD samples by describing negative images in detail. However, they struggle to distinguish ID samples from visually similar near-OOD data, as both conform to the sentence descriptions. To address this limitation, we prompt the MLLM to generate visually similar labels for the ID labels:

$$\mathcal{Y}^-_{vsl} = \mathcal{G}_{vsnl}(\mathcal{Y}, \mathcal{X}^{\text{his}}_{test}, f_{mllm}, M), \tag{9}$$

where $\mathcal{G}_{vsnl}$ represents the visually similar label generation process, as illustrated in Fig. 5.



(a) False Negative NLs (b) Different optimal thresholds γ (c) $1 - S_{ens}(\boldsymbol{v})$ Vs. $1 - S_{vsnl}(\boldsymbol{v})$

Figure 6. (a) Our VSNL generates visually similar labels only for the ID class subset, whose images are most similar to the near OOD samples, largely reducing false negative labels. (b) Different OOD datasets prefer different thresholds, and our proposed method can cache the historical test images and adaptively mine negative images, implicitly setting an dataset adaptive threshold. (c) S_{ens} and S_{vsnl} perform differently on far and near OOD, providing clues for designing an adaptive weight λ in Eq. 13.

Visually Similar ID-Classes Mining. While these visually similar labels cover the near-OOD regions, they typically include false NLs. Specifically, OOD data may only be similar to a subset of ID classes, while being distant from others. These visually similar NLs derived from OOD-unrelated ID classes are also far from OOD samples, thereby introducing false NLs and disturbing OOD detection, as intuitively shown in Fig. 6a. To address this issue, we first identify the subset of ID labels most similar to the OOD samples:

$$F(y_i) = \frac{|\{\boldsymbol{x} \in \mathcal{X}_{test}^{\text{his}} \mid H(\boldsymbol{x}) = y_i\}|}{|\mathcal{X}_{test}^{\text{his}}|}, \quad \forall y_i \in \mathcal{Y}, \\ \mathcal{Y}' = \text{Top}\left(\mathcal{Y}, F(\mathcal{Y}), \delta\right), \tag{10}$$

where $H(\boldsymbol{x})$ is the CLIP-based ID classifier with ID text features as weight, $|\cdot|$ measures the set size, $F(y_i)$ represents the proportion of historical test images in $\mathcal{X}_{test}^{\text{his}}$ being classified as $y_i \in \mathcal{Y}$, $F(\mathcal{Y})$ is the collection of $F(y_i)$, and $\delta \in (0, 1)$ serves as the selection ratio.

VSNL Generation and Score. After getting these filtered ID labels that share high similarity with negative images, we introduce the following visually similar negative labels:

$$\mathcal{Y}_{vsnl}^{-} = \mathcal{G}_{vsnl}(\mathcal{Y}', \mathcal{X}_{test}^{\text{his}}, f_{mllm}, M). \tag{11}$$

These visually similar negative labels adaptively capture the characteristics of the target OOD distribution, reducing false negative labels and resulting in the following score function:

$$S_{vsnl}(\boldsymbol{v}) = \frac{\sum_{y \in \mathcal{Y}} e^{\cos(\boldsymbol{v}, \boldsymbol{t})/\tau}}{\sum_{y \in \mathcal{Y}} e^{\cos(\boldsymbol{v}, \boldsymbol{t})/\tau} + \sum_{y^- \in \mathcal{Y}_{vsnl}^{-}} e^{\cos(\boldsymbol{v}, \widehat{\boldsymbol{t}^-})/\tau}}, \tag{12}$$

where $\widehat{\boldsymbol{t}^-}$ is the text feature of $y^- \in \mathcal{Y}_{vsnl}^{-}$.

4.4. Adaptive Weighted Score

Existing OOD detection methods rely on the assumption that the testing scenario (near-OOD or far-OOD) is human defined beforehand, but in real-world applications, this assumption often fails due to the dynamic nature of open environments. To address this, we propose an adaptive weighting strategy to balance these two scoring functions with an adaptive weight $\lambda \in [0, 1]$:

$$S_{ada}(\boldsymbol{v}) = \lambda S_{ens}(\boldsymbol{v}) + (1 - \lambda) S_{vsnl}(\boldsymbol{v}). \tag{13}$$

The weight λ adjusts dynamically based on the environment, approaching 1 in far-OOD scenarios to prioritize $S_{ens}(\boldsymbol{v})$, and 0 in near-OOD scenarios to emphasize $S_{vsnl}(\boldsymbol{v})$.

We design the adaptive weight λ by leveraging the performance differences of $S_{ens}(\boldsymbol{v})$ and $S_{vsnl}(\boldsymbol{v})$ on near and far OOD data. Specifically, ENS effectively characterizes far OOD samples, but its coarse-grained descriptions struggle to distinguish near OOD from ID samples, resulting in lower scores for far OOD samples and higher scores for near OOD samples. Conversely, VSNL better captures near OOD samples but, due to its ID-similarity, produces false negatives for far OOD samples, leading to higher scores for far OOD samples and lower scores for near OOD samples, as illustrated in Fig. 6c. Based on this observation, we define λ as:

$$\lambda = F\left(\frac{1}{|\mathcal{X}_{neg}|}\sum_{\boldsymbol{v} \in \mathcal{X}_{neg}} S_{ens}(\boldsymbol{v}), \frac{1}{|\mathcal{X}_{neg}|}\sum_{\boldsymbol{v} \in \mathcal{X}_{neg}} S_{vsnl}(\boldsymbol{v})\right), \tag{14}$$

where $F(a, b) = \frac{1-a}{(1-a)+(1-b)} \in (0, 1)$. One can see that when $\frac{1}{|\mathcal{X}_{neg}|}\sum_{\boldsymbol{v} \in \mathcal{X}_{neg}} S_{ens}(\boldsymbol{v}) > \frac{1}{|\mathcal{X}_{neg}|}\sum_{\boldsymbol{v} \in \mathcal{X}_{neg}} S_{vsnl}(\boldsymbol{v})$, λ approaches 0; otherwise, λ approaches 1. The algorithm is summarized in Alg. 1.

5. Experiments

5.1. Experiment Setup

Datasets and benchmarks. Following [16], we select ImageNet-1K [8] as the ID dataset and use iNaturalist [43],

Table 1. OOD detection results by using ImageNet-1k as ID dataset. ViTB/16 is used as the encoder. The results of traditional methods are available in the **supplementary materials**.

Methods	OOD datasets									
	iNaturalist		SUN		Places		Textures		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Training-required (or with Fine-tuning)										
MSP [13]	87.44	58.36	79.73	73.72	79.67	74.41	79.69	71.93	81.63	69.61
ZOC [11]	86.09	87.30	81.20	81.51	83.39	73.06	76.46	98.90	81.79	85.19
CLIPN [47]	95.27	23.94	93.93	26.17	92.28	33.45	90.93	40.83	93.10	31.10
LSN [34]	95.83	21.56	94.35	26.32	91.25	34.48	90.42	38.54	92.26	30.22
LoCoOp [33]	93.93	29.45	90.32	41.13	90.54	44.15	93.24	33.06	92.01	36.95
ID-Like [2]	98.19	8.98	91.64	42.03	90.57	44.00	94.32	25.27	93.68	30.07
NegPrompt [25]	90.49	37.79	92.25	32.11	91.16	35.52	88.38	43.93	90.57	37.34
SCT [52]	95.86	13.94	95.33	20.55	92.24	29.86	89.06	41.51	93.27	26.47
LAPT [58]	99.63	1.16	96.01	19.12	92.01	33.01	91.06	40.32	94.68	23.40
CMA [20]	99.62	1.65	96.36	16.84	93.11	27.65	91.64	33.58	95.13	19.93
SynOOD [24]	99.57	1.57	95.82	20.46	97.37	12.12	95.29	22.94	97.01	14.27
Zero Shot (No Training Required)										
MCM [30]	94.59	32.20	92.25	38.80	90.31	46.20	86.12	58.50	90.82	43.93
CoVer [55]	95.98	22.55	93.42	32.85	90.27	40.71	90.14	43.39	92.45	34.88
EOE [4]	97.52	12.29	95.73	20.40	92.95	30.16	85.64	57.63	92.96	30.09
NegLabel [19]	99.49	1.91	95.49	20.53	91.64	35.59	90.22	43.56	94.21	25.40
OODD [51]	99.36	2.22	95.01	21.49	87.10	44.76	93.27	30.69	93.69	24.79
AdaNeg [57]	99.71	0.59	97.44	9.50	94.55	34.34	94.93	31.27	96.66	18.92
CSP [5]	99.60	1.54	96.66	13.66	92.90	29.32	93.86	25.52	95.76	17.51
ANTS	**99.75**	**0.54**	**98.77**	**5.43**	96.10	20.21	**96.38**	**18.52**	**97.75**	**11.20**

Algorithm 1 Adaptive Negative Textual Space Shaping

Require: ID label space $\mathcal{Y}$, stream of testing batches $\{\mathcal{X}_t\}_{t=1}^{T}$.

Initialize ENS space $\mathcal{Y}_{ens}^{-}$ and VSNL space $\mathcal{Y}_{vsnl}^{-}$ with $\mathcal{Y}_{nl}^{-}$ of Eq. 2.
for each incoming batch $\mathcal{X}_t$ **do**
  Filter historical test samples with $\mathcal{S}_{nl}$ using Eq. 5;
  Collect negative images $\mathcal{X}_{neg}$ with implicit and adaptive threshold using Eq. 6;
  Generate expressive negative sentences $\mathcal{Y}_{ens}^{-}$ with the MLLM using Eq. 7; // ENS Generation
  Identify ID label subset similar to historical test images using Eq. 10;
  Generate visually similar negative labels $\mathcal{Y}_{vsnl}^{-}$ with MLLM using Eq. 11; // VSNL Generation
  Compute ENS scores S_{ens} and VSNL scores S_{vsnl} using Eq. 8 and Eq. 12, respectively;
  Calculate the dynamic weighting λ using Eq. 14;
  Get the final score S_{ada} by weighting the two scores using Eq. 13. //Adaptive Weighted Score
end for
Output Collected score S_{ada}.

SUN [50], Places [59], and Textures [6] as the OOD test datasets. We also validate our method on the OpenOOD benchmark, which contains SSB-hard [44] and NINCO [3] as near-OOD datasets, and iNaturalist [43], Texture [6], and OpenImage-O [46] as far-OOD datasets.

Implementation Details. We use the visual encoder of ViT-B/16 pretrained by CLIP [38]. We adopt the LLaVA-1.5-7B model as the default MLLM for our research, Following NegLabel [19], we adopt the text prompt of 'The nice <label>.', set temperature τ =0.01, and define the number M of negative labels as 10000 with group size 100. Additionally, we set the initial threshold $\gamma = 0.9$ and set η = 0.5. Our method follows a test-time adaptation setting as [57].

Evaluation Metrics. We employ two standard metrics to evaluate OOD detection: the false positive rate (FPR95), which measures the rate of OOD samples when the true positive rate for ID samples is at 95%, and the area under the receiver operating characteristic curve (AUROC).

5.2. Main Results

ImageNet-1k Benchmark. The results are shown in Tab. 1. We can see that other CLIP-based training-based methods [2, 25, 33, 34, 47, 52, 58] learn a negative branch or prompts by synthesizing negative samples, but they often fail to reflect the true OOD space. Other NL-based methods [4, 5, 19] directly retrieve negative labels from corpus datasets or generate them using LLMs, but they lack supervision from auxiliary OOD images. ANTS consistently

Table 2. OOD detection results of zero-shot methods on the OpenOOD benchmark. ImageNet-1k is adopted as ID dataset. Detailed results are available in the **supplementary materials**.

Methods	FPR95 ↓		AUROC ↑	
	Near-OOD	Far-OOD	Near-OOD	Far-OOD
MCM [30]	79.02	68.54	60.11	84.77
NegLabel [19]	68.18	27.34	76.92	93.30
EOE [4]	82.93	46.73	66.94	89.14
AdaNeg [56]	67.51	17.31	76.70	96.43
SynOOD [24]	71.68	17.11	77.55	96.21
ANTS	**60.98**	**15.38**	**82.15**	**96.50**

Table 3. OOD detection performance on other ID datasets.

ID Dataset	Method	AUROC↑	FPR95↓
CUB-200-2011	NegLabel [19]	99.93	0.13
	ANTS (Ours)	**99.95**	**0.01**
STANFORD-CARS	NegLabel [19]	99.99	0.01
	ANTS (Ours)	**99.99**	**0.00**
Food-101	NegLabel [19]	99.90	0.40
	ANTS (Ours)	**99.92**	**0.05**
Oxford-IIIT Pet	Neglabel [19]	99.62	1.70
	ANTS (Ours)	**99.99**	**0.02**

achieves remarkable improvements, which demonstrates the advantages of utilizing the understanding capabilities of MLLMs to shape a more accurate NL space. Compared with other test-time adaptation methods [51, 57], which store test images in memory to calculate the image proxy score and then combine the scores from both modalities, ANTS uses a text-only score to eliminate the modality gap when calculating the OOD score with ID classes, leading to better OOD detection results.

OpenOOD Benchmark. The results are shown in Tab. 2. Though the NL-based methods [4, 19] can handle small-scale near-OOD scenarios (*e.g.*, using ImageNet-10 and ImageNet-20 as ID and OOD data, respectively), these methods that selects semantically distant negative labels struggle to handle large-scale near-OOD scenarios such as using ImageNet-1k as ID. However, EOE [4], while using LLMs to generate visually similar labels, suffers from a growing number of false negatives with increasing ID classes. ANTS first identifies a subset of ID classes similar to OOD images, as shown in Fig. 6a, then it leverages the reasoning capabilities of MLLMs to generate visually similar labels. As a result, ANTS significantly outperforms its closest competitors [4, 19, 57] in both near-OOD and far-OOD scenarios, validating its scalability.

Results of other ID datasets. As shown in Tab. 3, our ANTS consistently surpasses existing methods in zero-shot OOD detection method NegLabel[19] across all in-distribution (ID) datasets. We also validate the robustness of ANTS to **Domain Shift** and **Adversarial Examples**, the detailed results are available in the supplementary materials.

Table 4. Ablation experiments. 'NIM' indicates the Negative Image Mining strategy in Eq. 6, and 'SIM' means the Visually Similar ID-Classes Mining strategy in Eq. 10.

	Components					FPR95 ↓	
	NIM	$\mathcal{Y}^-_{ens}$	SIM	$\mathcal{Y}^-_{vsnl}$	$S_{ada}(\boldsymbol{v})$	NearOOD	FarOOD
NegLabel[19]						68.18	27.34
A	✗	✓	✗	✗	✗	74.48	43.87
B	✓	✓	✗	✗	✗	73.70	19.22
C	✗	✗	✗	✓	✗	74.36	53.82
D	✗	✗	✓	✓	✗	63.11	23.44
E	✓	✓	✓	✓	✗	62.05	21.65
F	✓	✓	✓	✓	✓	60.98	15.38

5.3. Analyses and Discussions

Ablation Study. As illustrated in Tab. 4, it is necessary to introduce ENS $\mathcal{Y}^-_{ens}$ with mined negative images, as validated by the advantages of setting B over A in the far-OOD setting. Setting B significantly outperforms NegLabel, confirming the superiority of ENS over NLs. Generating visually similar labels to the mined ID class subset can significantly reduce false negative labels, as justified by the advantages of setting D over C. Combining ENS with VSNL by setting $\lambda = 0.5$ balances the results across different OOD sets, as shown in setting E, while using an adaptive λ leads to the best results in both OOD scenarios, as shown in setting F.

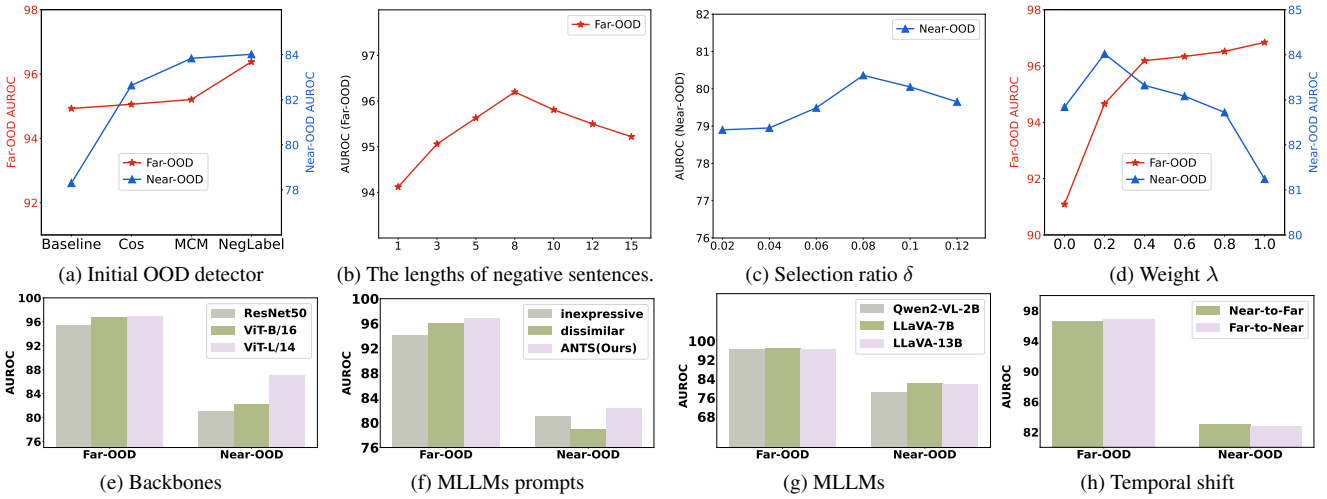
Analyses on initial OOD detectors. Besides NegLabel, we also tested two other variants: (1) a weak MCM detector, and (2) a cosine-distance filter that selects negative far from ID labels in the feature space. As shown in Fig. 7a, even with these weaker detectors, our method still outperforms previous SOTA baseline.

Analyses on the lengths of negative sentences. Due to the hallucination issue in MLLMs, we analyzed the length of generated negative sentences. As shown in Fig. 7b, appropriate increases in length enhance expressiveness and OOD detection, while excessive text introduces less discriminative words and hinders performance. Our ENS achieves an optimal balance at an average length of 8.4.

Ratio δ. As shown in Fig. 7c, generating visually similar labels for all ID classes ($\delta = 1$) performs poorly due to numerous false negative labels, as illustrated in Fig. 6a. However, using a too small δ will fail to adequately cover the OOD distribution. We set $\delta = 0.08$ in all experiments, although it is not optimal for specific datasets.

Weight λ. As shown in Fig. 7d, a larger λ emphasizes the $S_{ens}(x)$ score, improving far-OOD detection, while a smaller λ prioritizes the $S_{vsnl}(x)$ score, enhancing near-OOD detection. Our adaptive strategy automatically selects a suitable λ for various OOD settings.

Different Backbones. As illustrated in Fig. 7e, larger vi-



(a) Initial OOD detector (b) The lengths of negative sentences. (c) Selection ratio δ (d) Weight λ

(e) Backbones (f) MLLMs prompts (g) MLLMs (h) Temporal shift

Figure 7. Analysis on (a) different initial OOD detectors, (b) the lengths of negative sentences, (c) selection ratio δ, (d) weight λ, (e) CLIP image encoder backbones, (f) MLLMs prompts, and (g) different MLLMs. (h) Temporal shift. We use Texture [6] and NINCO [3] datasets as Far-OOD and Near-OOD, respectively.

sual backbones generally achieve improved OOD detection. Besides, our ANTS can generalize well to various VLM backbones, demonstrating its robustness.

Different Prompts for MLLMs. We designed two alternative prompts: an inexpressive prompt that limits negative image descriptions to under three words, and a dissimilar prompt that requests negative categories visually distinct from high-frequency ID classes. As shown in Fig. 7f, OOD detection performance declines with both alternatives, confirming our proposed prompts' efficacy.

Various MLLMs. When constructing the adaptive negative space, MLLMs of all sizes showed comparable far-OOD detection, but larger models excelled in near-OOD settings. As shown in Fig. 7g, LLaVA-7B performed best, owing to stronger reasoning.

Analysis of Temporal Shift. We evaluated ANTS under different temporal shifts. As shown in Fig. 7h, "Near-to-Far" and "Far-to-Near" indicate testing first on near (or far) OOD, then on the opposite, ANTS maintains strong performance, demonstrating robustness to temporal shift.

Complexity Analyses. As analyzed in Tab. 5 and Tab. 6, ANTS requires no learnable parameters. Although individual MLLM calls (ENS/VSNL) have higher latency, they are only selectively triggered for a small subset of samples. By amortizing these costs and utilizing a compact MLLM, ANTS maintains a competitive inference speed of 2.84 ms/image, as most samples are processed solely by the CLIP encoder.

Table 5. Latency (ms) breakdown (ImageNet).

Modules	ENS	VSNL	CLIP	Others
Per-call Latency	72.55	28.75	2.59	0.02
Avg. Latency/Img	0.18	0.06	2.59	0.02

Table 6. Complexity analyses. All results are obtained by using a GeForce RTX 3090 GPU.

Methods	Train Time	Latency (ms)	Param. (M)
ZOC	$>$ 24 h	5.38	336
CLIPN	$>$ 24 h	2.53	37.8
EOE	-	2.78	-
NegLabel	-	2.61	-
AdaNeg	-	2.70	-
ANTS	-	**2.84**	-

6. Conclusion and Future Work

This paper presents ANTS, a training-free, zero-shot framework for out-of-OOD detection. We first investigate three limitations of existing NLs methods. To address these issues, ANTS caches negative images and visually similar ID classes from historical test images, leveraging test-time MLLM understanding and reasoning through tailored prompts to construct a more accurate adaptive negative textual space. Two noise-filtering strategies are introduced to mitigate interference from ID noise and false negative labels. Finally, an adaptive scoring mechanism dynamically balances the two textual spaces, enhancing the framework's scalability across diverse OOD scenarios. Experimental results demonstrate that ANTS achieves state-of-the-art performance on zero-shot OOD detection benchmarks.

One minor limitation of our approach is that utilizing the MLLM model during testing necessitates GPU memory. More efficient utilization of MLLMs during the testing phase presents a meaningful direction for future work.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[2] Yichen Bai, Zongbo Han, Bing Cao, Xiaoheng Jiang, Qinghua Hu, and Changqing Zhang. Id-like prompt learning for few-shot out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17480–17489, 2024.

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

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

[5] 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.

[6] 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.

[7] Yi Dai, Hao Lang, Kaisheng Zeng, Fei Huang, and Yongbin Li. Exploring large language models for multi-modal out-of-distribution detection. *arXiv preprint arXiv:2310.08027*, 2023.

[8] 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.

[9] Xin Dong, Junfeng Guo, Ang Li, Wei-Te Ting, Cong Liu, and HT Kung. Neural mean discrepancy for efficient out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19217–19227, 2022.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] 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*, pages 6568–6576, 2022.

[12] 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.

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

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

[15] 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.

[16] 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.

[17] Dihong Jiang, Sun Sun, and Yaoliang Yu. Revisiting flow generative models for out-of-distribution detection. In *International Conference on Learning Representations*, 2021.

[18] Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Detecting out-of-distribution data through in-distribution class prior. In *International Conference on Machine Learning*, pages 15067–15088. PMLR, 2023.

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

[20] Jeonghyeon Kim and Sangheum Hwang. Enhanced ood detection through cross-modal alignment of multi-modal representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29979–29988, 2025.

[21] Marc Lafon, Elias Ramzi, Clément Rambour, Nicolas Audebert, and Nicolas Thome. Gallop: Learning global and local prompts for vision-language models. In *European Conference on Computer Vision*, pages 264–282. Springer, 2024.

[22] 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.

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[24] 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*, pages 4496–4506, 2025.

[25] 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.

[26] 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.

[27] Ziqian Lin, Sreya Dutta Roy, and Yixuan Li. Mood: Multi-level out-of-distribution detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 15313–15323, 2021.

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[29] 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.

[30] 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.

[31] Yifei Ming, Ying Fan, and Yixuan Li. Poem: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning*, pages 15650–15665. PMLR, 2022.

[32] 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.

[33] 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.

[34] 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.

[35] Aristotelis-Angelos Papadopoulos, Mohammad Reza Rajati, Nazim Shaikh, and Jiamian Wang. Outlier exposure with confidence control for out-of-distribution detection. *Neurocomputing*, 441:138–150, 2021.

[36] Jaewoo Park, Yoon Gyo Jung, and Andrew Beng Jin Teoh. Nearest neighbor guidance for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1695, 2023.

[37] Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. *Advances in neural information processing systems*, 31, 2018.

[38] 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.

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

[40] 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.

[41] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, pages 691–708. Springer, 2022.

[42] 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.

[43] 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.

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

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

[46] 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.

[47] 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.

[48] 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.

[49] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*, pages 23631–23644. PMLR, 2022.

[50] 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.

[51] Yifeng Yang, Lin Zhu, Zewen Sun, Hengyu Liu, Qinying Gu, and Nanyang Ye. Oodd: Test-time out-of-distribution detection with dynamic dictionary. *arXiv preprint arXiv:2503.10468*, 2025.

[52] 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.

[53] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9518–9526, 2019.

[54] Alireza Zaeemzadeh, Niccolo Bisagno, Zeno Sambugaro, Nicola Conci, Nazanin Rahnavard, and Mubarak Shah. Out-of-distribution detection using union of 1-dimensional subspaces. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 9452–9461, 2021.

[55] Boxuan Zhang, Jianing Zhu, Zengmao Wang, Tongliang Liu, Bo Du, and Bo Han. What if the input is expanded in ood detection? *Advances in Neural Information Processing Systems*, 37:21289–21329, 2024.

[56] Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Shi Han, Dongmei Zhang, et al. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *The Eleventh International Conference on Learning Representations*, 2022.

[57] 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.

[58] 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*, pages 271–288. Springer, 2024.

[59] 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.

[60] Wenjie Zhu, Yabin Zhang, Xin Jin, Wenjun Zeng, and Lei Zhang. Knowledge regularized negative feature tuning of vision-language models for out-of-distribution detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3565–3574, 2025.

[61] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018.