Enhancing Zero-shot OOD Detection with Pre-trained Multimodal Foundation Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Out-of-distribution (OOD) detection is essential for reliable deployment of deep models in real-world scenarios. Advances in pre-trained multimodal foundation models have enabled **zero-shot OOD detection** using only in-distribution (ID) labels. Recent methods in this direction expand the label space with auxiliary labels to facilitate the discrimination between IDs and OODs. Inspired by the probabilistic formulation via Binomial distribution, we further discover the key factors that theoretically affect zero-shot OOD detection performance: the cardinality of the auxiliary label set, the similarity between labels and samples, and the uncertainty of the similarity scores. From the theoretical analysis, existing methods that construct fixed, single-modality auxiliary labels surely have limited effectiveness. To address these issues, we propose **Refer-OOD**, a framework that adaptively generates, filters, and retrieves multimodal references that explicitly account for these factors. It consists of three modules: reference acquisition, feature mapping, and decision module. Experiments across multiple benchmarks demonstrate that Refer-OOD consistently improves zero-shot OOD detection with both vision-language models (VLMs) and multimodal large language models (MLLMs).

1 Introduction

The rapid advancement of deep learning has led to significant progress in computer vision tasks such as image classification and object detection. However, despite the strong performance on in-distribution (ID) data, deep learning models still struggle with out-of-distribution (OOD) samples. Model predictions on OOD samples may be incorrect yet overconfident, undermining the reliability of these models in real-world applications [1, 2, 3]. Therefore, developing effective OOD detection methods is crucial for enhancing both model capability and safety.

Leveraging powerful feature representation and prior knowledge of pre-trained multimodal foundation models, **zero-shot OOD detection** [4] using only ID labels has garnered increasing attention. Recent methods along this direction distinguish OOD samples by expanding the label space with auxiliary OOD labels, either sampled from a semantic pool [5, 6] or generated via large language models (LLMs) [7, 8], and then classifying input images into ID/OOD groups based on CLIP [9]. Despite the great research progress, how to gather theoretically relevant auxiliary information for zero-shot OOD detection is still under-explored.

In this paper, inspired by recent works that model the OOD scores with Binomial distribution and infer the mathematical performance metric thereby [5, 6], we further discover that the performance of zero-shot OOD detection is closely related to the cardinality of label set, the similarity probabilities of ID and OOD samples within the constructed OOD label set, and the uncertainty of the similarity result. From this insight, previous methods that construct auxiliary (OOD) labels deviate from known

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

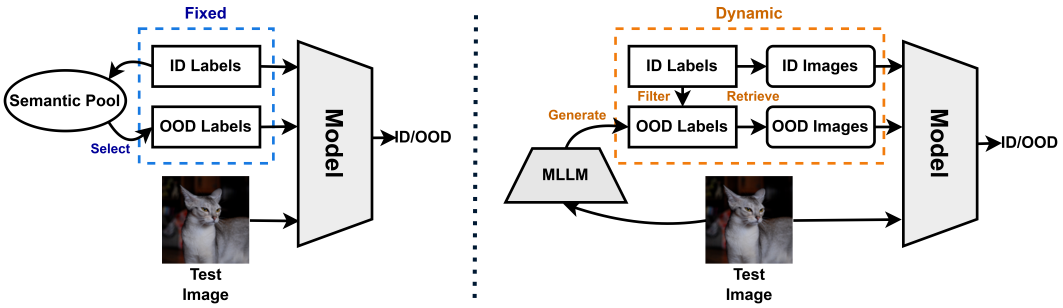


Figure 1: Comparison of two zero-shot OOD detection pipelines. Left: conventional methods mine a fixed set of OOD labels based on the ID label set from a semantic pool. Right: our proposed approach dynamically constructs reference sets by generating, filtering, and retrieving multimodal data related to the test sample at inference time.

ID labels within a single-modality framework are sub-optimal, ultimately resulting in degraded performance.

To address these issues and enhance zero-shot OOD detection, we propose **Refer-OOD**, which *adaptively* generates, filters, and retrieves *multimodal references* to increase the activation probability of OOD samples while maintaining the activation probability of ID samples. The comparison between conventional methods and Refer-OOD is illustrated in Figure 1. The entire detection process is implemented through three key modules (detailed in Figure 2): (1) a reference acquisition module for obtaining relevant references through generation, filtering and retrieval, (2) a feature mapping module that evaluates the relevance of the input image to the constructed references, and (3) a decision module that classifies samples as either ID or OOD. Theoretically, Refer-OOD can enhance OOD detection capability by dynamically integrating relevant references. Meanwhile, it is less sensitive to the reference set size.

We perform extensive experiments on coarse-grained and fine-grained OOD detection benchmarks using both traditional Vision-Language Models (VLMs) and Multimodal Large Language Models (MLLMs). The results show that Refer-OOD substantially improves model zero-shot performance in challenging OOD detection tasks. Moreover, our method consistently achieves state-of-the-art performance across various OOD detection benchmarks.

Our main contributions can be summarized as follows:

- We establish a theoretical framework for zero-shot OOD detection, based on which the key factors influencing detection performance and limitations in existing methods are identified.
- We propose Refer-OOD, a novel framework that comprehensively addresses all critical factors through adaptive label generation, similarity distribution regulation, and multimodal enhancement.
- We evaluate our method equipped with either VLMs or MLLMs and the results on both fine-grained and coarse-grained benchmarks verify the superiority of our method.

2 Related works

VLMs for Traditional Out-of-Distribution Detection. Pre-trained vision language models (VLMs) [9, 10] often require fine-tuning for effective adaptation to downstream tasks. For OOD detection, existing approaches either optimize visual or textual prompts [11, 12, 13, 14, 15, 16] or introduce OOD-specific regularization terms [14, 15, 16, 17, 18, 19]. However, these methods are computationally expensive and may undermine the generalization ability of pre-trained VLMs. Fine-tuning on ID data often leads to overfitting on seen categories, thereby reducing the model's ability to generalize to unseen ones and degrading OOD detection performance.

Zero-shot Out-of-Distribution Detection. Preserving VLMs' generalization ability while avoiding fine-tuning drawbacks, zero-shot OOD detection has emerged as a promising alternative. Leveraging the powerful representational capacity of pre-trained models, methods of this direction [20, 21, 22,

23, 24] bypass additional training by designing OOD detection scores to optimize the separability between IDs and OODs. Some approaches [25, 26, 27, 28] operate purely on the classification outputs of ID labels, while others [4, 6, 7, 8, 29, 30] introduce auxiliary OOD labels to recast the problem as a binary classification task distinguishing ID from OOD samples. However, constructing an appropriate OOD label set remains a non-trivial and open challenge.

Retrieval-Augmented Generation Methods. Retrieval-Augmented Generation (RAG) [31] combines generation with external knowledge retrieval to improve factual accuracy across various language tasks [32, 33, 34]. Recent works[35, 36, 37] extend RAG to multimodal settings by incorporating visual or auditory information, enabling richer context for generation. In this paper, we show that RAG can also enhance OOD detection by supporting retrieval-based reasoning over multimodal references.

3 Problem Analysis

3.1 Preliminaries

Zero-shot OOD detection. Zero-shot OOD detection aims to identify whether a test sample is in-distribution (ID) or out-of-distribution (OOD), using only ID class labels. Formally, given an ID label set $\mathcal{Y}^{\text{in}}$ of c classes, and a test image from either ID or OOD domains, i.e., $x \in \mathcal{X}^{\text{in}} \cup \mathcal{X}^{\text{out}}$ with $\mathcal{X}^{\text{in}} \cap \mathcal{X}^{\text{out}} = \emptyset$, the goal is to learn a detector $h(x; \mathcal{Y}^{\text{in}}) : x \to \{\text{ID}, \text{OOD}\}$.

OOD detection with auxiliary labels. To facilitate the identification of OOD samples, recent works propose to augment the label space with auxiliary OOD labels, either by sampling from a semantic pool [5, 6] or generating labels via LLMs [7]. Let $\mathcal{Y}^{\text{out}} = \{y_1^{\text{out}}, \dots, y_m^{\text{out}}\}$ denote the constructed OOD label set of size m. For a given image $x \in \mathcal{X}^{\text{in}} \cup \mathcal{X}^{\text{out}}$, its semantic similarity with an auxiliary label $y_i^{\text{out}} \in \mathcal{Y}^{\text{out}}$, can be computed as $s_i = \text{sim}(x, y_i^{\text{out}}) \in [0, 1]$. By applying a threshold ψ, this score can be converted into a binary label $b_i = \mathbb{1}_{s_i \geq \psi}$, which indicates the input is positive (OOD sample) with probability $p_i = P(s_i \geq \psi | y_i^{\text{out}}, x)$ according to the label y_i^{out}.

Probabilistic approximation. [5, 6] model the binary score b_i as a random variable following Bernoulli distribution with probability p_i. For $x \in \mathcal{X}^{\text{in}}$, the aggregated binary score $S^{\text{in}}(x) = \sum_{i=1}^{m} b_i^{\text{in}}$ is then a Poisson binomial variable with probabilities $\{p_i^{\text{in}}\}_{i=1}^{m}$. S^{out} can be defined similarly with probabilities $\{p_i^{\text{out}}\}_{i=1}^{m}$. According to the binomial approximation rules [38], as m increases, S^{in} and S^{out} can be approximated as normal distributions:

$$S^{\text{in}} \sim \mathcal{N}(mp^{\text{in}}, mp^{\text{in}}(1-p^{\text{in}}) - mv^{\text{in}}), \quad S^{\text{out}} \sim \mathcal{N}(mp^{\text{out}}, mp^{\text{out}}(1-p^{\text{out}}) - mv^{\text{out}}), \tag{1}$$

where $p^{\text{in}} = \mathbb{E}_i[p_i^{\text{in}}], v^{\text{in}} = \text{Var}_i[p_i^{\text{in}}], p^{\text{out}} = \mathbb{E}_i[p_i^{\text{out}}], v^{\text{out}} = \text{Var}_i[p_i^{\text{out}}]$. This leads to a closed-form approximation of the false positive rate (FPR) at a target true positive rate (TPR) $\lambda \in (0, 1]$:

$$\text{FPR}_\lambda = \frac{1}{2} + \frac{1}{2} \cdot \text{erf}\left(\sqrt{\frac{p^{\text{in}}(1-p^{\text{in}}) - v^{\text{in}}}{p^{\text{out}}(1-p^{\text{out}}) - v^{\text{out}}}} \text{erf}^{-1}(2\lambda - 1) + \frac{\sqrt{m(p^{\text{in}} - p^{\text{out}})}}{\sqrt{2p^{\text{out}}(1-p^{\text{out}}) - 2v^{\text{out}}}} \right), \tag{2}$$

where $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$ and a lower FPR_λ indicates better detection performance.

3.2 Theoretical Analysis for Performance Enhancement

How can we minimize the FPR?

According to Equation (2), FPR_λ is primarily influenced by four factors: p^{in}, p^{out}, m, and the similarity function $\text{sim}(\cdot)$. For simplicity, we fix the factor $\lambda = 0.5$ in our analysis.

Effect of $p_i^{\text{in}}, p_i^{\text{out}}, m$ on FPR. Let ζ denoting the formula input to function $\text{erf}(\cdot)$ in Equation (2), the partial derivatives of $\text{FPR}_{0.5}$ with respect to p^{in}, p^{out} and m are:

$$\begin{cases} \frac{\partial \text{FPR}_{0.5}}{\partial p^{\text{in}}} = \sqrt{\frac{m}{\pi}} \cdot e^{-\zeta^2} \cdot \frac{1}{(2p^{\text{out}}(1-p^{\text{out}}) - 2v^{\text{out}})^{\frac{1}{2}}} \geq 0, \\ \frac{\partial \text{FPR}_{0.5}}{\partial p^{\text{out}}} = -\sqrt{\frac{m}{\pi}} \cdot e^{-\zeta^2} \cdot \frac{p^{\text{out}} + p^{\text{in}} - 2p^{\text{in}}p^{\text{out}} - 2v^{\text{out}}}{(2p^{\text{out}}(1-p^{\text{out}}) - 2v^{\text{out}})^{\frac{3}{2}}} \leq 0, \\ \frac{\partial \text{FPR}_{0.5}}{\partial m} = \frac{1}{2\sqrt{\pi m}} \cdot e^{-\zeta^2} \cdot \frac{p^{\text{in}} - p^{\text{out}}}{\sqrt{2p^{\text{out}}(1-p^{\text{out}}) - 2v^{\text{out}}}} \leq 0, \text{when } p^{\text{in}} \leq p^{\text{out}}. \end{cases} \tag{3}$$

These results indicate that $\text{FPR}_{0.5}$ generally increases with p^{in} and decreases with p^{out} in most cases [5]. When $p^{\text{in}} \leq p^{\text{out}}$, increasing m further reduces $\text{FPR}_{0.5}$. Therefore, an ideal $\mathcal{Y}^{\text{out}}$ should have sufficiently large size m, and the labels $y_i^{\text{out}} \in \mathcal{Y}^{\text{out}}$ $(i = 1, \ldots, m)$ should make ID samples yield low p_i^{in} while OOD samples yield high p_i^{out}. Moreover, the relationship among m, p^{in} and p^{out} is more interdependent in practice. As m increases, some $y_i^{\text{out}} \in \mathcal{Y}^{\text{out}}$ may be irrelevant to most $x \in \mathcal{X}^{\text{out}}$, causing p^{out} to drop and eventually close to p^{in}.

Existing methods typically construct a fixed large-scale auxiliary label set $\mathcal{Y}^{\text{out}}$ by sampling or generating labels with the minimal similarity to $\mathcal{Y}^{\text{in}}$ [5, 6]. The constructed $\mathcal{Y}^{\text{out}}$ may suppress p^{in} for ID inputs. However, for OOD inputs, this does not necessarily guarantee that the expectation p^{out} is high enough. In practice, since $\mathcal{Y}^{\text{out}}$ is finite and fixed, there always $\exists x \in \mathcal{X}^{\text{out}}$ such that $\forall y_i^{\text{out}} \in \mathcal{Y}^{\text{out}}$, $\text{sim}(x, y_i^{\text{out}}) \ll 1$. To address the dependence between m, p^{in} and p^{out}, [5] proposes to filter uncommon or synonymous words from $\mathcal{Y}^{\text{out}}$ to increase the possibility of $\mathcal{Y}^{\text{out}}$ being activated by $\mathcal{X}^{\text{out}}$. However, such constructing strategy still faces the inherent limitation of uncontrollable p_i^{out} due to unknown $\mathcal{X}^{\text{out}}$. To address above issues, we propose to **(1) construct $\mathcal{Y}^{\text{out}}$ adaptively by generating relevant labels conditioned on the test input, rather than using a fixed set in contrast to $\mathcal{Y}^{\text{in}}$**.

Effect of sim$(\cdot)$ on FPR. Prior works [5, 6, 7, 25] typically compute similarity between a test image and a single reference label, which inevitably introduces high variance and limits the reliability of similarity-based decisions. We consider a more general setting where each class $y_i \in \{\mathcal{Y}^{\text{in}}, \mathcal{Y}^{\text{out}}\}$ is associated with n_i *diverse references* $\{r_{i_k}\}_{k=1}^{n_i}$ (e.g., diverse text or images). The average similarity score is defined as $\bar{s}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} s_{i_k} = \frac{1}{n_i}\sum_{k=1}^{n_i} \text{sim}(x, r_{i_k})$, where $\{s_{i_k}\}_{k=1}^{n_i}$ are assumed i.i.d. with mean μ_i^{in} (or μ_i^{out}) and variance σ_i^2. By the central limit theorem, $\bar{s}_i^{\text{in}} \sim \mathcal{N}(\mu_i^{\text{in}}, \sigma_i^2/n_i)$, and similarly for $\bar{s}_i^{\text{out}}$. Then, the probability p_i for ID or OOD inputs can be approximated by:

$$p_i^{\text{in}} = 1 - \text{erf}\left(\frac{\psi - \mu_i^{\text{in}}}{\sqrt{\sigma_i^2/n_i}}\right), \quad p_i^{\text{out}} = 1 - \text{erf}\left(\frac{\psi - \mu_i^{\text{out}}}{\sqrt{\sigma_i^2/n_i}}\right). \tag{4}$$

Reducing the variance to σ_i^2/n_i leads to sharper similarity distributions. When $\mu_i^{\text{out}} > \psi > \mu_i^{\text{in}}$, increasing n_i raises p_i^{out} while suppressing p_i^{in}, thereby enhancing separability and reducing FPR. In contrast, existing works using a single reference increase uncertainty and yield less discriminative similarity estimates. Therefore, we propose to **(2) use multiple diverse references per class rather than relying on a single reference label**.

4 Method

How can we construct a valid reference set?

Adaptive generation and filtering. According to theoretical analysis (1), we propose an adaptive reference generation strategy that dynamically constructs candidate labels conditioned on test samples, controllably improving the alignment with test data distribution and increasing p^{out}. To stabilize p^{in}, a filtering mechanism is performed to discard labels overly similar to known ID classes.

Multimodal retrieval. According to theoretical analysis (2), to overcome the limitation of a single reference label, we introduce a modality enhancement strategy that retrieves additional image representations via an online browser API. This increases modality diversity and improves OOD detection accuracy, especially for fine-grained samples.

Overall method. We propose Refer-OOD, a unified OOD detection framework based on adaptive generation and modality enhancement. Our method comprises three modules (Fig. 2): (1) the Reference Acquisition Module, which obtains multimodal reference samples; (2) the Feature Mapping Module, which evaluates the relevance of x to $\mathcal{Y}$; (3) the Decision Module, which determines whether the input x belongs to ID or OOD.

4.1 Reference Acquisition Module

The reference acquisition module consists of three sequential steps: generation, filtering, and retrieval, aiming to construct high-quality textural and visual references for zero-shot OOD detection.

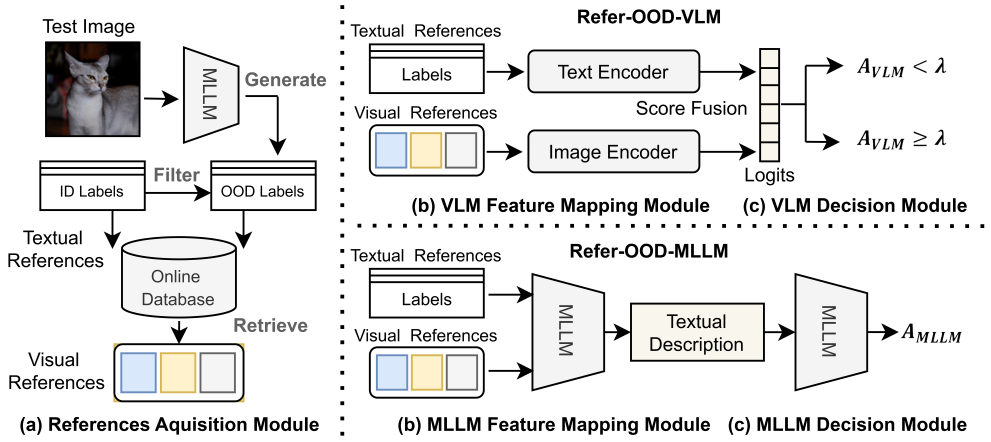


Figure 2: Detailed framework of Refer-OOD method. Refer-OOD comprises three modules: (1) Reference Acquisition Module, which obtains textual and visual references; (2) Feature Mapping Module, which evaluates the relevance of x to $\mathcal{Y}$ (OOD score for VLM-based and textual description for MLLM-based method); (3) Decision Module, which determines whether the input x belongs to ID or OOD.

Generation Phase. In the generation phase, we leverage a Multimodal Large Language Model (MLLM) $\mathcal{M}$ to produce textual labels that are semantically aligned with the input image x. With a prompt p_{per}, $\mathcal{M}$ generates a set of candidate OOD labels $\widetilde{\mathcal{Y}}^{\text{out}}$:

$$\widetilde{\mathcal{Y}}^{\text{out}} = \text{TOP}_{\mathcal{Y}}(p_{\mathcal{M}}(\mathcal{Y}|x, \mathrm{p}_{\text{per}}), m), \tag{5}$$

where $p_{\mathcal{M}}(\mathcal{Y}|x, \mathrm{p}_{\text{per}})$ represents the probability distribution of MLLM generated labels $\mathcal{Y}$ given x and prompt p_{per}. Unlike prior methods that rely on fixed or predefined OOD label set, our approach generates sample-specific labels on-the-fly. This adaptive generation explicitly increases the activation probability of OOD samples on p^{out}, as theoretically analyzed in Section 3.2.

Filtering Phase. To enhance semantic distinctiveness and avoid overlap between ID and OOD labels, we apply a filtering step that removes candidate OOD labels overly similar to known ID labels $\mathcal{Y}^{\text{in}}$. This results in a refined OOD label set:

$$\mathcal{Y}^{\text{out}} = \{y_i^{\text{out}} \in \widetilde{\mathcal{Y}}^{\text{out}} \mid \max_{y_j^{\text{in}} \in \mathcal{Y}^{\text{in}}} \text{sim}(y_i^{\text{out}}, y_j^{\text{in}}) < \tau\}, \tag{6}$$

where τ denotes a predefined threshold. This filtering ensures a low similarity between ID and OOD classes, thereby stabilizing p^{in}.

Retrieval Phase. Given the final label sets $\mathcal{Y}^{\text{out}}$ and $\mathcal{Y}^{\text{in}}$, we retrieve relevant visual references from external sources such as online search engines. The retrieval process returns the top-n_i^{I} images most semantically relevant to each label set:

$$\mathcal{I}^{\text{out}} = \text{TOP}_{\mathcal{I}}(p_{\text{retr}}(\mathcal{I}|\mathcal{Y}^{\text{out}}), n_i^{\text{I}}), \quad \mathcal{I}^{\text{in}} = \text{TOP}_{\mathcal{I}}(p_{\text{retr}}(\mathcal{I}|\mathcal{Y}^{\text{in}}), n_i^{\text{I}}), \tag{7}$$

where $p_{\text{retr}}(\mathcal{I} \mid \mathcal{Y})$ denotes the probability of retrieving image $\mathcal{I}$ given label set $\mathcal{Y}$.

Optimization Design. The reference acquisition module aims to construct reference sets that are relevant, informative and discriminative. In the generation phase, we optimize the prompt forms to guide the MLLM toward producing the highest semantically aligned labels. In the retrieval phase, we leverage the inherent ranking mechanism of the search engine to obtain top-ranked image references. This design makes the generation probability $p_{\mathcal{M}}$ and the retrieval probability p_{retr} feasible and optimal.

4.2 Feature Mapping Module

The Feature Mapping Module aims to assess the relevance of the input sample relative to the constructed references with Vision-Language Model (VLM) or Multi-Modal Large Model (MLLM).

With similar underlying targets, these two models are different in output formats: VLM quantifies the probability differences as OOD scores, whereas MLLM generates textual descriptions that emphasize these differences.

VLM-based Mapping Module. We define a unified similarity function between a test sample x and each (textual or visual) reference set $\{r_{i_k}\}_{k=1}^{n_i}$ corresponding to class $y_i \in \mathcal{Y}^{\text{in}} \cup \mathcal{Y}^{\text{out}}$, where $n_i = 1$ for textual reference and $n_i = n_i^{\text{I}}$ for visual references, as follows:

$$a_i(x) = \frac{|\langle I(x), \mathbb{E}_i[E(r_{i_k})]\rangle|}{|I(x)| \cdot |\mathbb{E}_i[E(r_{i_k})]|}, \tag{8}$$

where $E(\cdot)$ denotes either textual encoder $T(\cdot)$ or visual encoder $I(\cdot)$ depending on the modality of r_{i_k}. The OOD score for each modality follows the general form:

$$A(x) = \max_{v\in\{1,\ldots,c\}} \frac{e^{a_v(x)}}{\sum_{j=1}^{m+c} e^{a_j(x)}} - \beta \max_{v\in\{c+1,\ldots,m+c\}} \frac{e^{a_v(x)}}{\sum_{j=1}^{c+m} e^{a_j(x)}}, \tag{9}$$

where β is a balancing factor. The score for comparing the test image with textual references is denoted as A_{I2T}, and with visual references is A_{I2I}. Finally, the overall multimodal detection score is fused with weight coefficient α:

$$A_{\text{VLM}}(x) = \alpha A_{\text{I2I}}(x) + (1-\alpha) A_{\text{I2T}}(x). \tag{10}$$

MLLM-based Mapping Module. The MLLM generates reasoning descriptions based on textual reference ($\mathcal{Y}^{\text{in}}$, $\mathcal{Y}^{\text{out}}$) or visual references ($\mathcal{I}^{\text{in}}$, $\mathcal{I}^{\text{out}}$), prompted by p_{rea}, expressed as:

$$A_{\text{MLLM}}(x) = \begin{cases} A^{\text{T}}_{\text{MLLM}}(x) = \mathcal{M}(\text{p}_{\text{rea}}||x||\mathcal{Y}^{\text{out}}||\mathcal{Y}^{\text{in}}), \\ A^{\text{I}}_{\text{MLLM}}(x) = \mathcal{M}(\text{p}_{\text{rea}}||x||\mathcal{I}^{\text{out}}||\mathcal{I}^{\text{in}}), \end{cases} \tag{11}$$

where $A^{\text{T}}_{\text{MLLM}}$ and $A^{\text{I}}_{\text{MLLM}}$ are the reasoning texts generated by comparing the input sample x with the textual and visual references, respectively.

4.3 Decision Module

The Decision Module is responsible for determining whether a test sample x belongs to ID or OOD category based on the obtained mapping results.

VLM-based Decision Module. For VLM model, the decision process relies on the computed detection score and a predefined threshold λ:

$$h_{\text{VLM}}(x) = \begin{cases} \text{ID}, & A_{\text{VLM}} \geq \lambda \\ \text{OOD}, & A_{\text{VLM}} < \lambda. \end{cases} \tag{12}$$

where λ is typically set such that 95% of in-distribution (ID) data is correctly classified as ID.

MLLM-based Decision Module. For MLLM model, the decision process is directly based on the model-generated text A_{MLLM}, prompted by p_{ans} for final answer:

$$h_{\text{MLLM}}(x) = \mathcal{M}(\text{p}_{\text{ans}}||x||A_{\text{MLLM}}), \tag{13}$$

which ensures that the model makes the final textual judgment (ID/OOD) based on the input sample x and the relevance of corresponding references.

5 Experimental Analysis

5.1 Experimental Settings

Datasets. We classify OOD detection into coarse-grained and fine-grained tasks. Coarse-grained OOD detection follows the traditional setup [39], where ID and OOD belong to distinct datasets. Common ID datasets include CUB-200 [40], Stanford-Cars [41], Food-101 [42], Oxford-Pet [43], and ImageNet-1K [44], and OOD datasets include iNaturalist [45], SUN [46], Places [47], and Texture [48]. Fine-grained OOD detection is more challenging, with ID and OOD samples from the same dataset but differing at the subcategory level. Datasets are constructed by splitting categories

Table 1: Performance comparison for VLM-based methods on coarse-grained datasets.

Method	iNaturalist		SUN		Places		Texture		**Average**	
	FPR95↓	**AUROC**	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑
MCM	3.27	99.31	1.68	99.64	2.63	99.42	2.91	99.30	2.63	99.41
CLIPN	2.20	99.46	0.88	99.78	1.83	99.59	3.11	99.22	2.00	99.51
EOE	0.03	99.99	0.02	100.0	0.21	99.94	0.66	99.76	0.23	99.92
NegLabel	0.33	99.91	0.74	99.78	1.98	99.46	1.82	99.51	1.21	99.66
CSP	0.25	99.93	0.28	99.92	1.67	99.55	0.98	99.73	0.79	99.78
Refer-OOD-VLM	**0.01**	**100.0**	**0.01**	**100.0**	**0.12**	**99.97**	**0.06**	**99.99**	**0.03**	**99.99**

Table 2: Performance comparison for VLM-based methods on fine-grained datasets.

Method	CUB		Stanford-Cars		Food		Oxford-Pet		**Average**	
	FPR95↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑
MCM	83.72	67.50	84.02	68.76	44.10	91.37	64.03	84.88	68.97	78.20
CLIPN	83.89	67.36	82.92	69.37	42.46	91.28	68.88	85.03	69.54	78.39
EOE	74.13	73.18	77.60	70.98	39.66	**91.54**	55.17	90.30	61.64	81.50
NegLabel	81.23	71.92	79.62	70.18	42.85	90.88	64.56	87.45	67.06	80.10
CSP	80.48	69.88	78.05	70.58	52.00	89.54	62.97	89.25	68.37	79.81
Refer-OOD-VLM	**60.64**	**80.88**	**58.74**	**75.45**	**34.40**	91.50	**38.19**	**92.28**	**47.99**	**85.02**

within CUB-200, Oxford-Pet, Food-101, and Stanford-Cars. Note that OOD categories are disjoint from ID categories.

Experimental Setup. We use Bing image search and a Chrome retrieval plugin for retrieval. The Qwen-vl model [49] and the CLIP [9] with ViT-B/16 serves as the MLLM and VLM backbone, respectively. More comparison results are shown in the Appendix.

Evaluation Metrics. For VLM-based models, we report: (1) FPR95 (false positive rate at 95% TPR) and (2) AUROC (area under the ROC curve). For MLLM-based models, we evaluate: (1) F1 score (harmonic mean of precision and recall) and (2) ACC (accuracy of ID label predictions).

Comparative Methods. We compare two versions of our method (named **Refer-OOD-VLM** and **Refer-OOD-MLLM** according to different feature mapping and decision modules) with state-of-the-art zero-shot OOD detection methods. For Refer-OOD-VLM, we include MCM [25], CLIPN [19], NegLabel [6], CSP [5] and EOE [7] as comparative methods. Since MLLM-based OOD detection is unexplored, we define a straightforward baseline where the model is provided with the input data and the corresponding ID labels, and asked to determine whether the sample belongs to ID.

5.2 Main results

Results of VLM-based methods. Table 1 and Table 2 present the performance of the VLM-based methods in coarse-grained and fine-grained OOD detection tasks. In the coarse-grained task, VLM-based methods achieve strong performance across all datasets, with an average AUC above 99%, benefiting from CLIP's ability to differentiate datasets with large semantic gaps. Refer-OOD-VLM outperforms all other methods on these datasets. In the fine-grained task, where ID and OOD samples are more semantically similar, all methods face increased difficulty. Nevertheless, Refer-OOD-VLM achieves more superior performance, improving FPR95 by an average of 13.65% over EOE [7].

Results of MLLM-based methods. Table 3 and Table 4 evaluate MLLM-based models in coarse-grained and fine-grained OOD detection tasks. "Vanilla" refers to the constructed baseline. In the coarse-grained task, both MLLM-based methods excel at distinguishing OOD samples with large semantic gaps, achieving near-perfect accuracy. Refer-OOD-MLLM further enhances performance. In the fine-grained task, Refer-OOD-MLLM outperforms the vanilla approach across most datasets and metrics, especially showing great precision gains on Stanford-Cars and Food datasets.

5.3 Component Analysis

Analysis for Refer-OOD-VLM. As shown in Table 5, both textual and visual references have a positive influence on the model's performance. The dynamic label generation strategy in Refer-OOD-VLM outperforms the fixed label approach used in the EOE method, with an AUROC increase of

Table 3: Performance comparison for MLLM-based methods on coarse-grained datasets.

Method	iNaturalist		SUN		Places		Texture		Average		
	Precision↑	**F1↑**	**Precision↑**	**F1↑**	**Precision↑**	**F1↑**	**Precision↑**	**F1↑**	**Recall↑**	**Precision↑**	**F1↑**
Vanilla	100.0	83.76	99.67	83.66	99.05	83.45	99.06	83.56	72.49	99.44	83.60
Refer-OOD-MLLM	**100.0**	**90.38**	**100.0**	**90.38**	**99.24**	**90.06**	**99.32**	**90.09**	**82.75**	**99.63**	**90.16**

Table 4: Performance comparison for MLLM-based methods on fine-grained datasets.

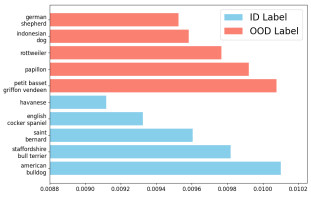
Method	CUB			Stanford-Cars			Food			Oxford-Pet			Average		
	Recall↑	**Precision↑**	**F1↑**	**Recall↑**	**Precision↑**	**F1↑**	**Recall↑**	**Precision↑**	**F1↑**	**Recall↑**	**Precision↑**	**F1↑**	**Recall↑**	**Precision↑**	**F1↑**
Vanilla	82.35	60.43	69.70	86.17	59.12	70.12	**91.83**	64.74	75.94	80.64	70.42	75.18	**85.24**	63.67	72.73
Refer-OOD-MLLM	**85.29**	**63.50**	**72.80**	**76.59**	**79.12**	**77.83**	79.59	**83.87**	**81.67**	**85.48**	**92.98**	**89.07**	81.73	**79.86**	**80.34**

11.14% on the CUB dataset. All modules effectively complement each other, boosting the average AUROC by 18.14% compared to the baseline.

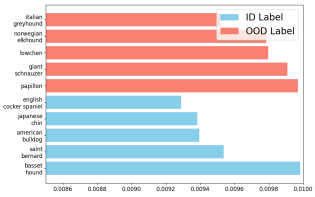
Analysis for Refer-OOD-MLLM. Table 6 presents the component effectiveness analysis for Refer-OOD-MLLM, reporting detection (F1) and prediction (ACC) performance. Again, both the textual and visual references enhance model performance.

5.4 Case Study

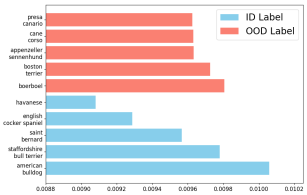
Case study for Refer-OOD-VLM. Figures 3a to 3d show the class probability scores. For the ID sample *American bulldog*, Refer-OOD-VLM achieves high confidence on the correct label while effectively suppressing the logits of OOD labels. For the OOD sample *Beagle*, EOE misclassifies it as a similar ID category due to the absence of appropriate OOD labels. In contrast, Refer-OOD-VLM assigns high confidence to a semantically correct OOD label, reducing confusion and improving detection accuracy.



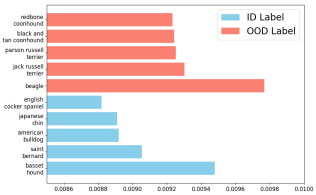
(a) EOE(American bulldog)



(b) EOE(Beagle)



(c) Refer-OOD-VLM (American bulldog)



(d) Refer-OOD-VLM (Beagle)

OOD label: Staffordshire Bull Terrier

Question: According to references, the image shows?
Answer: This conclusion is based on several features:
1. Build and Structure: …
2. Head Shape: …
3. Jaw and Snout: …
It aligns more closely with the American Pit Bull Terrier than the Staffordshire Bull Terrier.

Question: According to the analysis, is the image ID or OOD?
Answer: ID. American Pit Bull Terrier.

(e) Refer-OOD-MLLM (American bulldog)

Figure 3: Case study on Refer-OOD detection. (a)-(d) shows the class probability scores for VLM-based methods, and (e) shows Refer-OOD-MLLM result.

Case study for Refer-OOD-MLLM. Figure 3e illustrates one type of error made by the vanilla MLLM-based method. More illustrations in the Appendix include: (1) misclassifying ID samples as OOD, (2) predicting incorrect labels for ID samples, and (3) misclassifying OOD samples as ID. Refer-OOD-MLLM addresses these issues by incorporating valid multi-modality sets. For ID misdetection, Refer-OOD-MLLM enhances semantic understanding for accurate identification. For ID misclassification, it refines predictions using ID and OOD labels. For OOD misdetection, Refer-OOD-MLLM generates precise OOD labels to correctly classify OOD samples.

Table 5: Effectiveness of each module in Refer-OOD-VLM.

Baseline	Fixed Textual	Dynamic Textual	Visual	CUB		Stanford-Cars		Food		Oxford-Pet	
				FPR95↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑
✓				80.39	67.34	63.83	72.07	55.12	91.79	83.58	82.93
✓			✓	73.53	68.16	64.89	76.02	34.35	92.23	70.15	84.06
✓	✓			67.65	71.88	62.77	77.24	50.06	86.94	47.76	92.27
✓		✓		**49.02**	83.02	53.19	75.08	39.44	92.22	34.33	92.37
✓	✓		✓	65.69	74.41	60.64	**79.62**	49.18	87.66	41.79	91.91
✓		✓	✓	52.94	**85.48**	**51.06**	77.27	**32.83**	**92.31**	**29.85**	**93.81**

Table 6: Effectiveness of each module in Refer-OOD-MLLM.

Baseline	Textual	Visual	CUB		Standford-Cars		Food		Oxford-Pet	
			F1↑	**ACC**↑	**F1**↑	**ACC**↑	**F1**↑	**ACC**↑	**F1**↑	**ACC**↑
✓			69.70	49.01	70.12	68.08	75.94	**82.65**	75.18	77.41
✓	✓		71.65	**57.84**	75.93	71.27	**86.17**	76.53	88.13	82.25
✓		✓	**72.80**	55.88	**77.83**	**75.53**	81.67	70.40	**89.07**	**85.48**

5.5 Ablations

We further conduct ablation studies on the parameters, score functions, and foundation models. Please refer to the Appendix for all supporting figures and tables.

Effect of n_i^{I}. Figure 4 examines visual retrieval quantity n_i^{I}. Increasing n_i^{I} improves model performance by reducing similarity variance and enhancing separability.

Effect of m. Figure 5 presents the performance of different methods as the number of generated labels increases. In contrast, Refer-OOD consistently achieves stronger results, even with fewer generated labels. Moreover, while other methods are sensitive to the value of m, Refer-OOD demonstrates higher robustness.

Effect of $\beta\&\alpha$. Figure 6 evaluates β on fine-grained datasets. Performance improves as β increases, with optimal results near $\beta = 1$, balancing ID and OOD labels' contributions. Figure 7 explores α, which balances textual labels and image features in Equation (10). Introducing visual modality with well-balanced α outperforms single-modality approaches.

Effect of τ. Table 11 investigates the effect of the filtering threshold τ. The optimal τ typically correlates with the semantic gap between the ID and OOD datasets.

On score functions. Table 12 compares Refer-OOD-VLM's performance using standard scoring functions, including MSP [50], Energy [51], and MaxLogits [52]. Table 13 compares scoring function variants in Eq. 9, using max vs. sum over class logits.

On VLMs, MLLMs and APIs. We conduct comparative experiments across different VLMs, MLLMs and retrieval APIs. Table 14 evaluates VLM backbones including CLIP [9], ALIGN [53], and AltCLIP [54], showing that Refer-OOD consistently outperforms the comparative model across all architectures. Table 15 and Table 16 demonstrate that the results with GPT-4o are consistent with those with Qwen. Table 17 shows Refer-OOD's performance across different online retrieval APIs including Baidu and Google, which aligns with the results using Bing.

6 Conclusion

In this paper, we present a theoretical analysis on the zero-shot OOD detection paradigm, identifying key factors that influence detection performance, including label set size, similarity distributions, and metric uncertainty. Based on these insights, we propose Refer-OOD, a novel framework that systematically optimizes these factors through multimodal relevant references integration. Extensive experiments on both fine-grained and coarse-grained benchmarks validate the effectiveness of Refer-OOD, showing consistent improvements over prior methods.

Broader Impacts and Limitation. Our work promotes the reliable deployment of deep models in wide real-world scenarios, specifically on zero-shot OOD detection with pre-trained multimodal models. While our method outperforms existing approaches and maintains stable performance even with a reduced number of references, it incurs extra inference cost and possible security risks due to reference generation and retrieval.

References

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

[2] Angelos Filos, Panagiotis Tigkas, Rowan McAllister, Nicholas Rhinehart, Sergey Levine, and Yarin Gal. Can autonomous vehicles identify, recover from, and adapt to distribution shifts? In *International Conference on Machine Learning*, pages 3145–3153, 2020.

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

[4] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, pages 7068–7081, 2021.

[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] 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.

[7] 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. In *Proceedings of the 41st International Conference on Machine Learning*, pages 5629–5659, 2024.

[8] Yuxiao Lee, Xiaofeng Cao, Jingcai Guo, Wei Ye, Qing Guo, and Yi Chang. Concept matching with agent for out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4562–4570, 2025.

[9] 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, 2021.

[10] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

[11] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, pages 2337–2348, 2022.

[12] 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*, 2024.

[13] Marc Lafon, Elias Ramzi, Clément Rambour, Nicolas Audebert, and Nicolas Thome. Gallop: Learning global and local prompts for vision-language models. *arXiv preprint arXiv:2407.01400*, 2024.

[14] 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.

[15] 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.

[16] Yabin Zhang, Wenjie Zhu, Chenhang He, and Lei Zhang. Lapt: Label-driven automated prompt tuning for ood detection with vision-language models. *arXiv preprint arXiv:2407.08966*, 2024.

[17] 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.

[18] Kai Liu, Zhihang Fu, Chao Chen, Sheng Jin, Ze Chen, Mingyuan Tao, Rongxin Jiang, and Jieping Ye. Category-extensible out-of-distribution detection via hierarchical context descriptions. *Advances in Neural Information Processing Systems*, 2024.

[19] 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.

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

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

[22] Rundong He, Yue Yuan, Zhongyi Han, Fan Wang, Wan Su, Yilong Yin, Tongliang Liu, and Yongshun Gong. Exploring channel-aware typical features for out-of-distribution detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 12402–12410, 2024.

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

[24] Keke Tang, Chao Hou, Weilong Peng, Runnan Chen, Peican Zhu, Wenping Wang, and Zhihong Tian. Cores: Convolutional response-based score for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10916–10925, 2024.

[25] 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*, pages 35087–35102, 2022.

[26] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Zero-shot in-distribution detection in multi-object settings using vision-language foundation models. *arXiv preprint arXiv:2304.04521*, 2023.

[27] Yixia Li, Boya Xiong, Guanhua Chen, and Yun Chen. Setar: Out-of-distribution detection with selective low-rank approximation. *arXiv preprint arXiv:2406.12629*, 2024.

[28] Bin Zhang, Xiaoyang Qu, Guokuan Li, Jiguang Wan, and Jianzong Wang. Vista: Visual-contextual and text-augmented zero-shot object-level ood detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[29] 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.

[30] Haoyan Xu, Zhengtao Yao, Xuzhi Zhang, Ziyi Wang, Langzhou He, Yushun Dong, Philip S Yu, Mengyuan Li, and Yue Zhao. Glip-ood: Zero-shot graph ood detection with foundation model. *arXiv preprint arXiv:2504.21186*, 2025.

[31] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[32] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1, 2023.

[33] Karen Ka Yan Ng, Izuki Matsuba, and Peter Chengming Zhang. Rag in health care: a novel framework for improving communication and decision-making by addressing llm limitations. *NEJM AI*, 2(1):AIra2400380, 2025.

[34] Hasan Md Tusfiqur Alam, Devansh Srivastav, Md Abdul Kadir, and Daniel Sonntag. Towards interpretable radiology report generation via concept bottlenecks using a multi-agentic rag. In *European Conference on Information Retrieval*, pages 201–209. Springer, 2025.

[35] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training

with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23369–23379, 2023.

[36] Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. Videorag: Retrieval-augmented generation with extreme long-context videos. *arXiv preprint arXiv:2502.01549*, 2025.

[37] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571*, 2024.

[38] Ravi Parameswaran. Book review: Statistics for experimenters: An introduction to design, data analysis, and model building, 1979.

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

[40] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[41] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[42] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014.

[43] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[44] 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.

[45] 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.

[46] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

[47] 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*, pages 1452–1464, 2017.

[48] 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.

[49] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

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

[51] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, pages 21464–21475, 2020.

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

[53] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[54] Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. Altclip: Altering the language encoder in clip for extended language capabilities. *arXiv preprint arXiv:2211.06679*, 2022.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer:[Yes]

   Justification: A summary of the paper's contribution is provided in conclusion.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: See Section 6.

Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: See Section 3.

   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: See Section 5 and supplemental material for implementation details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

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

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

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

   Answer: [No]

   Justification: We include training and implementation details, but not code. Our code will be available if the paper is accepted.

   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: See Section 5 and supplemental material for implementation details.

   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: [NA]

   Justification: Not Applicable.

   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: See Section 5 and supplemental material for implementation details.

   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: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   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: See Section 6.

    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: [Yes]

    Justification: See Section 6.

    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: See References.

    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: Not Applicable.

    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: Not Applicable.

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: Not Applicable.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [NA]

Justification: Not Applicable.

Guidelines:

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