Envisioning Outlier Exposure by Large Language Models for Out-of-Distribution Detection

Chentao Cao [1] **Zhun Zhong** [2 3] **Zhanke Zhou** [1] **Yang Liu** [4] **Tongliang Liu** [5] **Bo Han** [1]

Abstract

Detecting out-of-distribution (OOD) samples is essential when deploying machine learning models in open-world scenarios. Zero-shot OOD detection, requiring no training on in-distribution (ID) data, has been possible with the advent of vision-language models like CLIP. Existing methods build a text-based classifier with only closed-set labels. However, this largely restricts the inherent capability of CLIP to recognize samples from large and open label space. In this paper, we propose to tackle this constraint by leveraging the expert knowledge and reasoning capability of large language models (LLM) to Envision potential Outlier Exposure, termed EOE, without access to any actual OOD data. Owing to better adaptation to open-world scenarios, EOE can be generalized to different tasks, including far, near, and fine-grained OOD detection. Technically, we design (1) LLM prompts based on visual similarity to generate potential outlier class labels specialized for OOD detection, as well as (2) a new score function based on potential outlier penalty to distinguish hard OOD samples effectively. Empirically, EOE achieves state-of-the-art performance across different OOD tasks and can be effectively scaled to the ImageNet-1K dataset. The code is publicly available at: https://github.com/tmlr-group/EOE.

[1]TMLR Group, Department of Computer Science, Hong Kong Baptist University [2]School of Computer Science and Information Engineering, Hefei University of Technology [3]School of Computer Science, University of Nottingham [4]Computer Science and Engineering, University of California, Santa Cruz [5]Sydney AI Centre, The University of Sydney. Correspondence to: Bo Han <bhanml@comp.hkbu.edu.hk>, Zhun Zhong <zhunzhong007@gmail.com>.

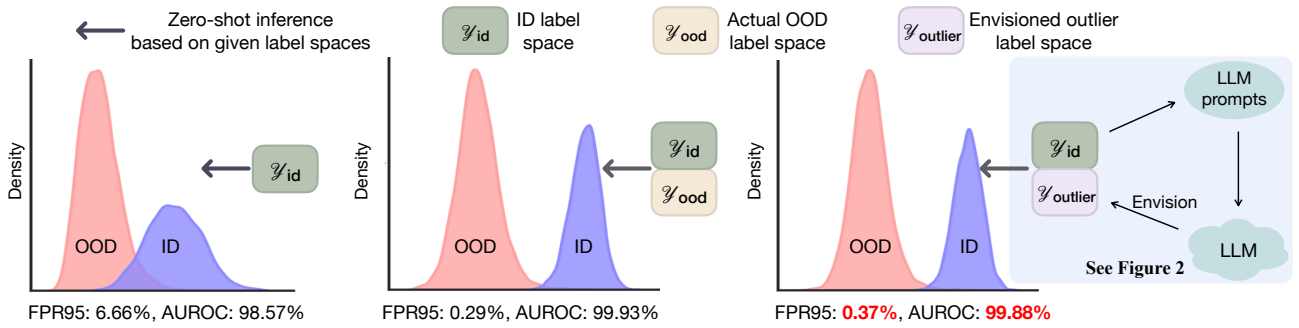
Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

Machine learning models excel in closed-set scenarios, where training and testing datasets share identical distribution. However, in open-world settings, especially in high-stakes scenarios like autonomous driving where the consequence of making an error can be fatal, these models often encounter out-of-distribution (OOD) samples that fall outside the label space of the training dataset, leading to unpredictable and frequently erroneous model behaviors. Consequently, there is a growing interest in OOD detection (Yang et al., 2021; Zhang et al., 2023; Salehi et al., 2021), aiming to distinguish OOD samples from test-time data while maintaining classification accuracy.

Most existing OOD detection methods (Hendrycks & Gimpel, 2017; Lee et al., 2018; Hendrycks et al., 2019b; Liu et al., 2020; Sehwag et al., 2021) can effectively detect OOD samples based on a well-trained in-distribution (ID) classifier. However, they are constrained to ID datasets with different label spaces. Besides, these methods solely depend on vision patterns, ignoring the connection between visual images and textual labels. Recently, MCM (Ming et al., 2022) introduced the setting of zero-shot OOD detection, which aims to leverage the capabilities of large-scale vision-language models (VLMs), *e.g.*, CLIP (Radford et al., 2021), to detect OOD samples across diverse ID datasets without training samples. By constructing a textual classifier with only ID class labels, MCM achieves impressive performance compared to previous OOD detection methods.

However, such an approach often fails when encountering hard OOD samples, as shown in Figure 1 (a). One might wonder 1) if this issue arises because the pre-trained models (*e.g.*, CLIP) are not strong enough or require further fine-tuning; or 2) if it is attributable to the usages of these pre-trained models, *e.g.*, *an exclusive reliance on closed-set ID classes*. Surprisingly, our findings suggest that CLIP can achieve superior OOD detection results by incorporating actual OOD class labels, as depicted in Figure 1 (b). This reinforces that relying solely on ID class labels is inadequate for distinguishing hard OOD samples. Unfortunately, we are unable to access the actual OOD label space in practical open-world scenarios. Therefore, we raise an open question:



(a) **Closed-Set:** Using only closed-set ID classes

(b) **Ground Truth:** Incorporating with actual OOD classes (unavailable)

(c) **EOE (Ours):** Incorporating with envisioned outlier classes

Figure 1: Comparison of zero-shot OOD detection score distribution. Compared to the model using (a) only closed-set ID classes, (b) adding actual OOD class labels can largely increase the OOD detection performance. (c) By adding the outlier classes generated by our method, the OOD detection results can also be significantly improved without using the actual OOD class labels. We use CUB-200-2011 (Wah et al., 2011) as ID classes and Places (Zhou et al., 2017) as OOD classes.

Is it possible to generate the potential outlier class labels for OOD detection without access to test-time data?

To answer this question, we take a step further in this work and ponder whether we can employ large language models (LLMs) to address this challenge. We propose a knowledge-enhanced approach that harnesses the expert knowledge and reasoning capabilities of LLMs to Envision potential Outlier Exposure, termed EOE, without relying on any actual or auxiliary OOD data, as shown in Figure 1 (c). Technically, we design LLM prompts to generate potential outlier class labels specialized for OOD detection, following a visual similarity rule. For example, "*Give three categories that visually resemble a horse*", in which "*horse*" is an ID class. Furthermore, we introduce a new score function based on potential outlier penalty to distinguish hard OOD samples effectively. Different from ZOC (Esmaeilpour et al., 2022) and CLIPN (Wang et al., 2023), which also attempt to generate "NOT ID" classes, ZOC requires additional training on a text-based image description generator, and CLIPN necessitates a large dataset to train the extra CLIP encoder. EOE only utilizes ID class labels to generate outlier classes.

The proposed EOE brings significant performance improvements and enjoys the advantages of: (1) *OOD-Agnostic*, which does not require any prior knowledge of the unknown OOD data; (2) *Zero-Shot*, which serves various task-specific ID datasets with a single pre-trained model; (3) *Scalability and Generalizability*, which effectively adapts to large-scale datasets such as ImageNet-1K (Deng et al., 2009) with 1000 ID classes, and is flexible and generalizable across far, near, and fine-grained OOD detection tasks.

Our contributions can be summarized as follows:

- We propose a new perspective that leverages expert knowledge from LLM to envision potential outlier class labels, facilitating OOD detection (Section 3).
- We propose EOE, providing LLM prompts based on the visual similarity rule to envision potential outlier class labels. A score function is further designed based on potential outlier penalty, helping the model effectively distinguish between ID samples and OOD samples (Section 3).
- Our EOE is superior, significantly outperforming existing methods. EOE achieves improvements of 2.47%, 2.13%, 3.59%, and 12.68% on the far OOD, near OOD, fine-grained OOD, and ImageNet-1K far OOD detection tasks in terms of FPR95 (Section 4).

2. Preliminaries

Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) is trained on 400 million (image, text) pairs collected from the internet using self-supervised contrastive representation learning (Tian et al., 2020). The model comprises an image encoder $\mathcal{I}(\cdot)$, adopting either ViT (Dosovitskiy et al., 2021) or ResNet (He et al., 2016) architecture, and a text encoder $\mathcal{T}(\cdot)$, utilizing the Transformer (Vaswani et al., 2017) architecture. During testing, the model evaluates the similarity of visual and textual features to choose the best-matching class label. CLIP enables open-set textual inputs as class labels without retraining or fine-tuning based on specific training data. This feature renders it widely applied to zero-shot downstream tasks, such as visual classification, detection, segmentation, and so on.

Large Language Models (LLMs) refer to natural language processing models trained on massive data, with more than hundreds of billions or more parameters, such as GPT-3 (Brown et al., 2020) GPT-4 (OpenAI, 2023), PaLM (Chowdhery et al., 2022) and LLaMA (Touvron et al., 2023). These models demonstrate the proficiency to understand and generate natural language text, thus enabling them to undertake a multitude of linguistic tasks. Considering the

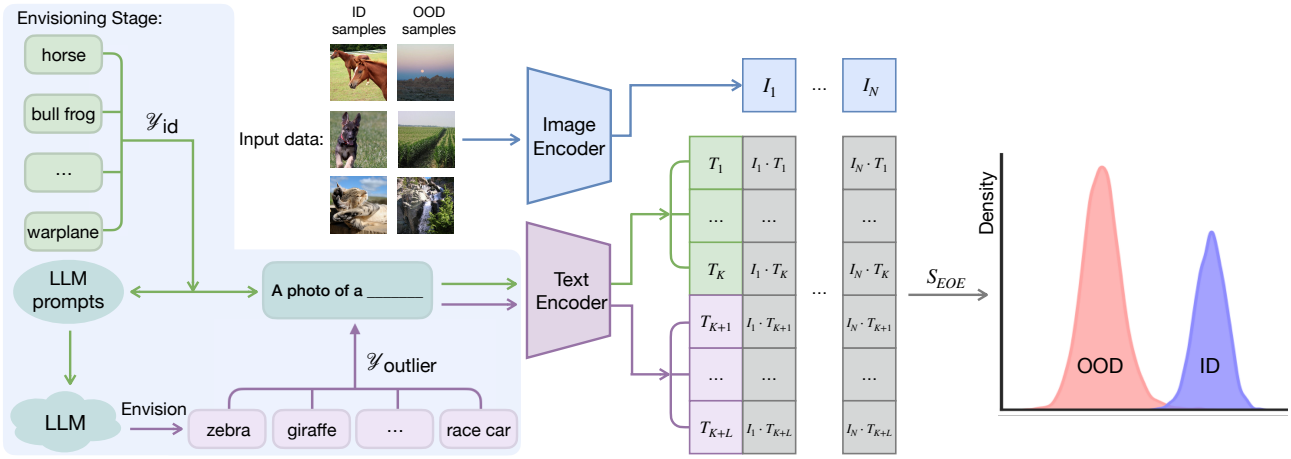


Figure 2: The framework of the proposed EOE. Given a set of ID class labels $\mathcal{Y}_{\text{id}}$, we first leverage the designed prompts to generate a set of outlier class labels, $\mathcal{Y}_{\text{outlier}}$, by using a LLM. Then, we input both the ID and generated OOD class labels into the text encoder for building the textual classifier. During the test stage, given an input image, we obtain the visual feature by the image encoder and calculate the similarities between the visual feature and the textual classifier. Finally, the OOD score is obtained by scaling the similarities with the proposed detection score function S_{EOE}.

comprehensive nature of their training datasets, which span a wide variety of knowledge domains, the expert knowledge embedded in LLMs can be employed to provide class labels visually similar to the ID class to meet our needs.

Zero-Shot OOD Detection (Esmaeilpour et al., 2022; Ming et al., 2022; Wang et al., 2023) enables detecting OOD samples across diverse ID datasets using the same pre-trained model, *e.g.*, CLIP, without re-training on any unseen ID data. It can be viewed as a binary classification problem:

$$G_\lambda(x; \mathcal{Y}_{\text{id}}, \mathcal{I}, \mathcal{T}) = \begin{cases} \text{ID} & S(x) \geq \lambda \\ \text{OOD} & S(x) < \lambda \end{cases}, \tag{1}$$

where $G(\cdot)$ is the OOD detector, x denotes the input image, $x \in \mathcal{X}, \mathcal{X} = \{\text{ID}, \text{OOD}\}$ and $\mathcal{Y}_{\text{id}}$ defines the space of ID class labels. The OOD detection score function S is derived from the similarity between the visual representation $\mathcal{I}(x)$ and textual representation $\mathcal{T}(t)$, t is the textual input to the text encoder, such as *"a photo of a <ID class>"*, and λ is the threshold to distinguish ID/OOD classes.

3. Envisioning Outlier Exposure for Zero-shot OOD Detection

In this paper, we aim to enhance zero-shot OOD detection performance by harnessing LLM to generate potential outlier class labels. However, several challenges require attention: (1) How to guide LLM to generate the desired outlier class label? (2) How can we sharpen the distinction between ID and OOD samples given the envisioned outlier class label? To address these issues, we propose LLM prompts specifically tailored for OOD detection and introduce a novel score function for better differentiation. The overall framework of our method is illustrated in Figure 2.

3.1. Acquiring Envisioned Outlier Class Labels

We first categorize OOD detection tasks into three types: far, near, and fine-grained OOD detection. Then, we elaborate the corresponding three LLM prompts anchored on *visual similarity* to generate outlier class labels as follows, which are general to different datasets for each OOD task.

Far OOD Detection refers to identifying OOD classes that are distant from the ID classes in the label space, with most being effectively discerned. Building upon the existing ID class labels, we first guide LLM to summarize these classes and determine their respective major categories. Subsequently, we ask LLM to provide outlier class labels that are visually similar to these major categories. Since LLM first summarizes the ID classes into major categories, our approach can be easily extended to large-scale datasets, such as ImageNet-1K. The LLM prompt for far OOD detection is illustrated in Figure 3. Wherein, the $\mathcal{Y}_{\text{id}}$ represents the set of ID class labels, $\mathcal{Y}_{\text{id}} = \{y_1, y_2, \cdots, y_K\}$, and the K signifies the total number of categories encompassed within the ID class labels. Similarly, $\mathcal{Y}_{\text{outlier}}$ indicates the set of envisioned outlier class labels generated by LLM, and $\mathcal{Y}_{\text{outlier}} = \{n_1, n_2, \cdots, n_L\}$.

Near OOD Detection pertains to identifying OOD classes that are relatively close to the ID class, *e.g.*, "*horse*" and "*zebra*", presenting an increased propensity to come across OOD samples that bear visual resemblances to ID classes. Consequently, for each ID class label, we instruct LLM to provide l outlier class labels that exhibit visual resemblance

Q: I have gathered images of K distinct categories: $\mathcal{Y}_{\text{id}}$. Summarize what broad categories these categories might fall into based on **visual features**. Now, I am looking to identify L classes that **visually resemble** these broad categories but have no direct relation to these broad categories. Please list these L categories for me.

Far OOD prompt **A:** These L categories are:

Figure 3: LLM prompt for far OOD detection, consisting of both the contents of **Q** and **A**.

Q: Given the image category y_i, please suggest **visually similar** categories that are not directly related or belong to the same primary group as y_i. Provide suggestions that **share visual characteristics** but are from broader and different domains than y_i.

Near OOD prompt **A:** There are l classes similar to y_i, and they are from broader and different domains than y_i:

Figure 4: LLM prompt for near OOD detection.

Q: I have a dataset containing K different species of *class-type*. I need a list of L distinct ***class-type* species** that are NOT present in my dataset, and ensure there are no repetitions in the list you provide. For context, the species in my dataset are: $\mathcal{Y}_{\text{id}}$.

Fine-grained OOD prompt **A:** The other L *class-type* species not in the dataset are:

Figure 5: LLM prompt for fine-grained OOD detection.

with ID class labels, *i.e.*, $l \times K = L$. Overlapping classes in $\mathcal{Y}_{\text{outlier}}$ with $\mathcal{Y}_{\text{id}}$ are removed by string matching. The LLM prompt for near OOD detection is shown in Figure 4.

Fine-grained OOD Detection, also known as open-set recognition (OSR) (Vaze et al., 2022), focuses on semantic shift instead of distributional shift primarily comprised in previous OOD detection. In fine-grained OOD detection, both ID and OOD samples fall under the same major category (*e.g.*, "*bird*"), and intrinsic visual similarities exist among subclasses (*e.g.*, "*Frigatebird*", "*Ovenbird*"). Therefore, it is more appropriate to instruct the LLM to provide different subclasses within the same major category directly. The LLM prompt for fine-grained OOD detection is presented below Figure 5, where *class-type* refers to the major category, such as "*bird*".[1]

3.2. A New OOD Detection Score

The purpose of the OOD detection score function is to better distinguish between ID and OOD score distributions. By having the outlier labels available, we can effectively enhance the score function's ability to differentiate between ID and OOD samples. We introduce a new score function S_{EOE} based on potential outlier $\mathcal{Y}_{\text{outlier}}$ penalty to distinguish hard OOD samples effectively. First, the label-wise matching score $s_i(x)$ is expressed as:

$$s_i(x) = \frac{\mathcal{I}(x) \cdot \mathcal{T}(t_i)}{\|\mathcal{I}(x)\| \cdot \|\mathcal{T}(t_i)\|}; \quad t_i \in \mathcal{Y}_{\text{id}} \cup \mathcal{Y}_{\text{outlier}}. \tag{2}$$

Subsequently, the proposed score $S_{\text{EOE}}(\cdot)$ is defined, taking into account the outlier class labels, as follows:

$$S_{\text{EOE}}(x; \mathcal{Y}_{\text{id}}, \mathcal{Y}_{\text{outlier}}, \mathcal{T}, \mathcal{I}) = \max_{i \in [1,K]} \frac{e^{s_i(x)}}{\sum_{j=1}^{K+L} e^{s_j(x)}} - \beta \cdot \max_{k \in (K,K+L]} \frac{e^{s_k(x)}}{\sum_{j=1}^{K+L} e^{s_j(x)}}, \tag{3}$$

where β is a hyperparameter, and we set $\beta = 0.25$ in the main results. For more details on the design of S_{EOE} and β, please refer to Appendix A.2. Based on S_{EOE}, the OOD detector $G(x; \mathcal{Y}_{\text{id}}, \mathcal{I}, \mathcal{T})$ can be viewed as the binary classification:

$$G_\lambda(x; \mathcal{Y}_{\text{id}}, \mathcal{Y}_{\text{outlier}}, \mathcal{T}, \mathcal{I}) = \begin{cases} \text{ID} & S_{\text{EOE}}(x) \geq \lambda \\ \text{OOD} & S_{\text{EOE}}(x) < \lambda \end{cases}, \tag{4}$$

where λ is a selected threshold such that a high fraction of ID data (typically 95%) exceeds this value. The zero-shot OOD detection procedure is shown in Algorithm 1.

Algorithm 1 Zero-shot OOD detection with envisioned outlier class labels

1: **Input:** ID class labels $\mathcal{Y}_{\text{id}}$, test sample x, text encoder $\mathcal{T}$, image encoder $\mathcal{I}$, LLM, β, threshold λ;
Envisioning Stage:
2: Given $\mathcal{Y}_{\text{id}}$, $\mathcal{Y}_{\text{outlier}} = \text{LLM}(\text{prompt}(\mathcal{Y}_{\text{id}}))$;
Testing Stage:
3: $K = \text{len}(\mathcal{Y}_{\text{id}}), L = \text{len}(\mathcal{Y}_{\text{outlier}})$;
// Compute label-wise matching score
4: $\{s_i(x) = \frac{\mathcal{I}(x) \cdot \mathcal{T}(t_i)}{\|\mathcal{I}(x)\| \cdot \|\mathcal{T}(t_i)\|}\}_{i=1}^{K+L}; \quad t_i \in \mathcal{Y}_{\text{id}} \cup \mathcal{Y}_{\text{outlier}}$;
// Compute OOD detection score
5: $S_{\text{EOE}}(x) = \max_{i \in [1,K]} \frac{e^{s_i(x)}}{\sum_{j=1}^{K+L} e^{s_j(x)}} - \beta \max_{k \in (K,K+L]} \frac{e^{s_k(x)}}{\sum_{j=1}^{K+L} e^{s_j(x)}}$;
6: **Output:** OOD detection decision $\mathbf{1}\{S_{\text{EOE}} \geq \lambda\}$.

We summarize the advantages of our approach as follows:

1. **OOD-Agnostic**: EOE does not rely on prior knowledge of unknown OOD data, making it particularly suitable and adaptable to open-world scenarios.
2. **Zero-Shot**: A single pre-trained model efficiently serves various task-specific ID datasets without the need for individual training on each specific ID dataset. EOE can achieve superior OOD detection performance by merely knowing the ID class labels.

[1]For detailed LLM prompts and the outlier class labels generated by LLM, please refer to Appendix F.

Table 1: Zero-shot **far** OOD detection results. The **bold** indicates the best performance on each dataset. The gray indicates that the comparative methods require an additional massive auxiliary dataset. Ground Truth is the result of incorporating the actual OOD class labels, which are unavailable in fact.

ID Dataset	Method	OOD Dataset								Average	
		iNaturalist		SUN		Places		Texture			
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CUB-200-2011	CLIPN	0.10	99.97	0.06	99.98	0.33	99.91	0.17	99.95	0.17	99.95
	Energy	0.46	99.89	0.03	99.99	0.30	99.92	0.02	100.00	0.20	99.95
	MaxLogit	0.35	99.92	0.06	99.99	0.35	99.91	0.00	100.00	0.19	99.95
	MCM	9.83	98.24	4.93	99.10	6.66	98.57	6.99	98.75	7.10	98.66
	EOE (Ours)	0.06	99.98	0.03	100.00	0.37	99.88	0.01	100.00	**0.12**	**99.96**
	Ground Truth	-	-	-	-	0.29	99.93	0.00	99.99	-	-
STANFORD-CARS	CLIPN	0.00	99.99	0.02	99.99	0.13	99.96	0.02	99.99	0.04	99.98
	Energy	0.01	100.00	0.04	99.99	0.42	99.90	0.04	99.99	0.13	99.97
	MaxLogit	0.00	100.00	0.02	99.99	0.26	99.94	0.00	100.00	0.07	99.98
	MCM	0.05	99.77	0.02	99.95	0.24	99.89	0.02	99.96	0.08	99.89
	EOE (Ours)	0.00	100.00	0.01	100.00	0.06	99.99	0.00	100.00	**0.02**	**100.00**
	Ground Truth	-	-	-	-	0.07	99.99	0.00	100.00	-	-
Food-101	CLIPN	0.70	99.83	0.10	99.96	0.26	99.94	5.35	98.19	1.60	99.48
	Energy	0.92	99.75	0.18	99.92	0.54	99.86	12.43	96.55	3.52	99.02
	MaxLogit	0.56	99.86	0.09	99.95	0.49	99.88	8.33	97.44	2.37	99.28
	MCM	0.64	99.78	0.90	99.75	1.86	99.58	4.04	98.62	1.86	99.43
	EOE (Ours)	0.06	99.99	0.00	100.00	0.10	99.98	2.45	99.04	**0.65**	**99.76**
	Ground Truth	-	-	-	-	0.02	99.99	0.59	99.83	-	-
Oxford-IIIT Pet	CLIPN	0.01	99.99	1.08	99.78	0.97	99.80	1.42	99.61	0.87	99.80
	Energy	0.06	99.97	0.05	99.98	0.23	99.94	0.35	99.88	0.17	99.94
	MaxLogit	0.02	99.98	0.05	99.97	0.20	99.94	0.27	99.91	0.14	99.95
	MCM	2.80	99.38	1.05	99.73	2.11	99.56	0.80	99.81	1.69	99.62
	EOE (Ours)	0.00	100.00	0.01	99.99	0.15	99.96	0.12	99.97	**0.07**	**99.98**
	Ground Truth	-	-	-	-	0.08	99.98	0.09	99.98	-	-
Average	CLIPN	0.20	99.95	0.32	99.93	0.42	99.90	1.74	99.44	0.67	99.80
	Energy	0.36	99.90	0.07	99.97	0.37	99.91	3.21	99.10	1.01	99.72
	MaxLogit	0.23	99.94	0.06	99.98	0.33	99.92	2.15	99.34	0.69	99.79
	MCM	3.33	99.29	1.73	99.63	2.71	99.40	2.96	99.28	2.68	99.40
	EOE (Ours)	0.03	99.99	0.01	100.00	0.17	99.95	0.64	99.75	**0.21**	**99.92**
	Ground Truth	-	-	-	-	0.12	99.97	0.17	99.95	-	-

3. **Scalability and Generalizability**: In contrast to the existing zero-shot OOD detection method (Esmaeilpour et al., 2022) that generates candidate OOD class labels, EOE can be easily applied to large-scale datasets like ImageNet-1K. Moreover, EOE exhibits generalizability across diverse tasks, including far, near, and fine-grained OOD detection.

4. Experiments

4.1. Setups

Far OOD Detection. The ID datasets for far OOD detection encompass CUB-200-2011 (Wah et al., 2011), STANFORD-CARS (Krause et al., 2013), Food-101 (Bossard et al., 2014), Oxford-IIIT Pet (Parkhi et al., 2012) and ImageNet-1K (Deng et al., 2009). As for the OOD datasets, we use the large-scale OOD datasets iNaturalist (Van Horn et al., 2018), SUN (Xiao et al., 2010), Places (Zhou et al., 2017), and Texture (Cimpoi et al., 2014) curated by MOS (Huang & Li, 2021). For more details about how each dataset is adapted to be ID-OOD, please refer to Appendix B.1.

Near OOD Detection. We adopt ImageNet-10 and ImageNet-20 alternately as ID and OOD datasets, proposed by MCM (Ming et al., 2022), both of which are subsets extracted from ImageNet-1K. The ImageNet-10 dataset curated by MCM mimics the class distribution of CIFAR-10 (Krizhevsky et al., 2009). The ImageNet-20 dataset consists of 20 classes semantically similar to ImageNet-10 (*e.g.*, "*horse*" (ID) *vs.* "*zebra*" (OOD)).

Fine-grained OOD Detection. We split CUB-200-2011, STANFORD-CARS, Food-101, and Oxford-IIIT Pet. Specifically, half of the classes from each dataset are randomly selected as ID data, while the remaining classes constitute OOD data. Importantly, there is no overlap between the above ID dataset and the corresponding OOD dataset.

Evaluation Metrics. We employ two widely-used metrics for evaluation: (1) FPR95, the false positive rate of OOD data when the true positive rate is at 95% for ID data, where a lower value indicates better performance; (2) AUROC, the area under the receiver operating characteristic curve, with a higher value signifying superior performance. In addition, we report results in terms of AUPR in Appendix B.2.2.

EOE Setups. We employ CLIP (Radford et al., 2021) as the backbone of our framework. Unless otherwise specified, we adopt ViT-B/16 as the image encoder and masked self-attention Transformer (Vaswani et al., 2017) as the text encoder in our experiments. The pre-trained weights of CLIP are sourced from the official weights provided by Ope-

Table 2: Zero-shot **far** OOD detection results for ImageNet-1K as ID dataset. The **black bold** indicates the best performance. The gray indicates that the comparative methods require training or an additional massive auxiliary dataset. Energy (FT) requires fine-tuning, while Energy is post-hoc.

Method	OOD Dataset								Average	
	iNaturalist		SUN		Places		Texture			
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MOS (BiT)	9.28	98.15	40.63	92.01	49.54	89.06	60.43	81.23	39.97	90.11
Fort et al.	15.07	96.64	54.12	86.37	57.99	85.24	53.32	84.77	45.12	88.25
Energy(FT)	21.59	95.99	34.28	93.15	36.64	91.82	51.18	88.09	35.92	92.26
MSP	40.89	88.63	65.81	81.24	67.90	80.14	64.96	78.16	59.89	82.04
CLIPN	19.13	96.20	25.69	94.18	32.14	92.26	44.60	88.93	30.39	92.89
Energy	81.08	85.09	79.02	84.24	75.08	83.38	93.65	65.56	82.21	79.57
MaxLogit	61.66	89.31	64.39	87.43	63.67	85.95	86.61	71.68	69.08	83.59
MCM	30.92	94.61	37.59	92.57	44.71	89.77	57.85	86.11	42.77	90.77
EOE (Ours)	12.29	97.52	20.40	95.73	30.16	92.95	57.53	85.64	**30.09**	**92.96**
Ground Truth	-	-	-	-	13.24	96.96	24.29	95.04	-	-

nAI. In addition, we adopt the GPT-3.5-turbo-16k model as the LLM for our research, with the temperature parameter setting to 0. To reduce the potential impact of randomness, we instruct LLM to envision the outlier class three times on each ID dataset independently, and the final reported results are the average of these three experiments.

Compared Methods. We compare our method with state-of-the-art OOD detection methods, including zero-shot and those requiring fine-tuning. For fair comparisons, all compared methods employ CLIP as their backbone, which is consistent with EOE. With respect to fine-tuning methods, we consider MSP (Hendrycks & Gimpel, 2017), Energy (Liu et al., 2020), MOS (Huang & Li, 2021), and the method proposed by Fort et al. (Fort et al., 2021). As for zero-shot methods, our comparisons are drawn towards MCM (Ming et al., 2022) and CLIPN (Wang et al., 2023). What's more, We implement post-hoc methods (Energy (Liu et al., 2020) and MaxLogit (Hendrycks et al., 2019a)) as additional baselines on CLIP backbone. It is worth noting that CLIPN relies on a large-scale auxiliary dataset (Sharma et al., 2018) to additionally pre-train an encoder. Instead, our EOE does not require any such dataset.

4.2. Main Results

Far OOD Detection. Table 1 presents the comparison with the recent state-of-the-art zero-shot OOD detection method across four ID datasets: CUB-200-2011, STANFORD-CARS, Food-50, and Oxford-IIIT Pet. Ground Truth is the result of incorporating the actual OOD class labels, which are in fact unavailable. For each dataset, we guide the LLM to envision 500 outlier classes, *i.e.*, $L = 500$. Clearly, EOE achieves superior results on these four ID datasets, with an average FPR95 of 0.21% and AUROC of 99.92%. This indicates substantial improvement over the strong baseline. Notably, our EOE outperforms the strong baseline by 6.98% when using CUB-200-2011 as the ID dataset.

We then conduct experiments on the large-scale dataset (ImageNet-1K) for far OOD detection. Results are reported in Table 2. We adopt the results of fine-tuning methods reported by MCM. EOE is comparable to fine-tuning methods and surpasses MCM. Although CLIPN uses an additional text encoder and large-scale datasets for training the "no" prompt, our EOE still outperforms CLIPN. Furthermore, CLIPN utilizes an ensemble strategy for the textual inputs, in which the ensemble and learnable textual inputs are effective in enhancing performance (Zhou et al., 2022b;a).

Near OOD Detection. The results for near OOD detection are presented in Table 3. For each ID class label, EOE instructs the LLM to return three outlier class labels. EOE outperforms the strong baseline MCM by achieving improvements of 2.13% in average FPR95 and 0.30% in AUROC. And our EOE also surpasses Energy and MaxLogit scores. Compared to CLIPN, which uses extra large datasets for re-training, our EOE clearly outperforms it. Note that when using ImageNet-20 as the ID dataset, we filtered out candidate class names with a cosine similarity above 85% to the ID class names to prevent excessive ID samples from being classified as OOD candidates. We also report results on large-scale benchmarks organized by OpenOOD in Table 12, specifically ImageNet-1K *vs.* SSB-hard (Vaze et al., 2022) and NINCO (Bitterwolf et al., 2023).

Fine-grained OOD Detection. Table 4 shows the performance of fine-grained OOD detection. EOE guides the LLM to generate 500 outlier class labels for each ID dataset. Compared to MCM, our EOE increases the average OOD performance by 3.59% in FPR95. Despite the unfair comparison, our EOE still outperforms CLIPN on the Food-50 and Oxford-Pet-18 datasets in terms of FPR95.

4.3. Ablation Study

Score Functions. To demonstrate the superiority of the proposed OOD detection score S_{EOE}, we compare it with the other score functions: $S_{\text{MAX}}, S_{\text{MSP}}, S_{\text{Energy}}$ and S_{MaxLogit}. These score functions are all designed based on potential outlier class labels. The comparison of these score functions is shown in Figure 6 (a). Please refer to Appendix D.2 for the specific forms and results on more datasets. Results show that our S_{EOE} achieves the best OOD performance.

Table 3: Zero-shot **near** OOD detection results. The **bold** indicates the best performance on each dataset, and the gray indicates methods requiring an additional massive auxiliary dataset.

Method	ID OOD	ImageNet-10 ImageNet-20		ImageNet-20 ImageNet-10		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CLIPN		7.80	98.07	13.67	97.47	10.74	97.77
Energy		10.30	97.94	16.40	97.37	13.35	97.66
MaxLogit		9.70	98.09	14.00	97.81	11.85	97.95
MCM		5.00	98.71	17.40	97.87	11.20	98.29
EOE (Ours)		4.20	99.09	13.93	98.10	**9.07**	**98.59**
Ground Truth		0.20	99.80	0.20	99.93	0.20	99.87

Table 4: Zero-shot **fine-grained** OOD detection results. The **bold** indicates the best performance on each dataset, and the gray indicates methods requiring an additional massive auxiliary dataset.

Method	ID OOD	CUB-100 CUB-100		Stanford-Cars-98 Stanford-Cars-98		Food-50 Food-51		Oxford-Pet-18 Oxford-Pet-19		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CLIPN		73.54	74.65	53.33	82.25	43.33	88.89	53.90	86.92	56.05	83.18
Energy		76.13	72.11	73.78	73.82	44.95	89.97	68.51	88.34	65.84	81.06
MaxLogit		76.89	73.00	72.18	74.80	41.73	90.79	65.66	88.49	64.11	81.77
MCM		83.58	67.51	83.99	68.71	43.38	91.75	63.92	84.88	68.72	78.21
EOE (Ours)		74.74	73.41	76.83	71.60	37.95	91.96	52.55	90.33	**60.52**	**81.82**
Ground Truth		61.23	81.42	58.31	83.71	11.34	97.79	29.17	95.58	40.01	89.63

This verifies the superiority and importance of the proposed OOD detection score.

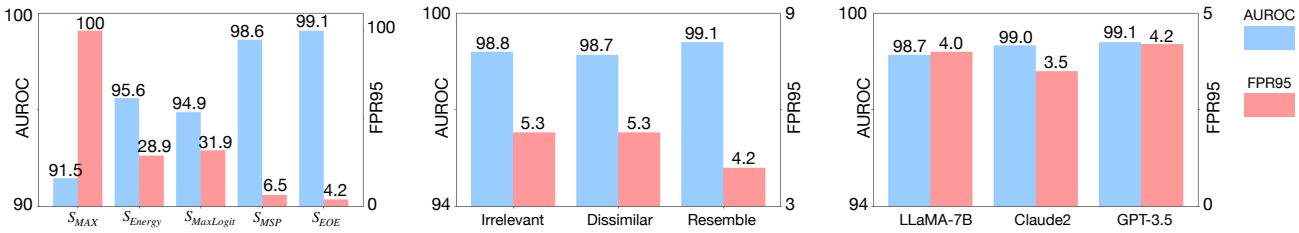
LLM Prompts. To investigate the effectiveness of the visual similarity rule, we design two additional types of LLM prompts, one termed *visually irrelevant* and the other *visually dissimilar*. The rest of the LLM prompt content remains unchanged. Specifically, the *'irrelevant'* LLM prompt instructs the LLM to generate arbitrary outlier class labels for the ID class without adhering to a *visually resemblance* constraint. Conversely, the *'dissimilar'* LLM prompt asks the LLM to envision outlier class labels for the ID class under a *visually dissimilar* constraint. As shown in Figure 6 (b), without the *visually resemble* constraint proposed in our EOE, the OOD performance degrades on both FPR95 and AUROC metrics, indicating the significance of the proposed constraint. For detailed LLM prompts and results on more datasets, please refer to Appendix D.3.

Various LLMs. We conduct experiments with various LLMs to provide a more comprehensive understanding of EOE's effectiveness. Specifically, we use LLaMA2-7B (Touvron et al., 2023) and Claude 2 (Anthropic, 2023) to envision outlier class labels. The results on ImageNet-10 (ID) are shown in Figure 6 (c). EOE achieves better results than the baseline MCM across different LLMs. Moreover, LLaMA2-7B and Claude 2 outperform GPT-3.5-turbo-16k in terms of FPR95. These results demonstrate the generalizability of our method. For further results on more datasets related to LLM prompts, refer to Appendix D.4.

Number of Outlier Class Labels. We conduct experiments to investigate the impact of the number of outlier class labels, *i.e.*, L, generated by LLM. We instruct the LLM to return 100, 300, and 500 outlier class labels for the far and fine-grained OOD detection for each ID dataset. For the near OOD detection, we ask the LLM to return outlier class labels in quantities of 1, 3, and 10 for each ID class label. Figure 7 presents the respective average metrics for varying numbers of outlier class labels across the three tasks. EOE consistently outperforms the baseline MCM and maintains relatively stable performance across different values of L. What's more, we conduct experiments on $\beta \in \{0, 0.25, 0.50, 0.75, 1\}$ with different outlier class number L. For detailed results, please refer to Appendix A.2.

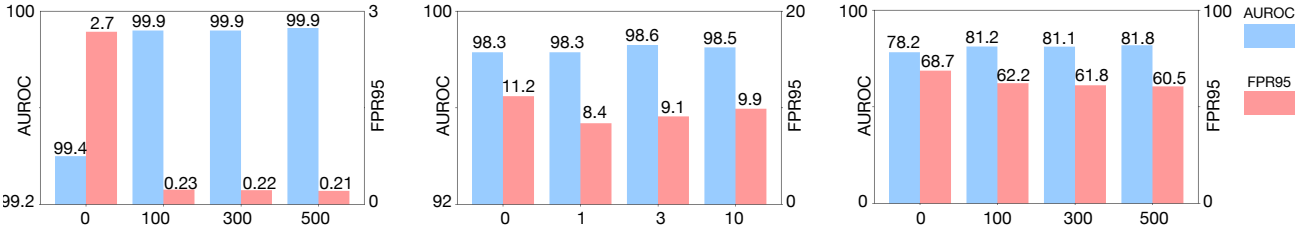
4.4. Further Analysis

More Experimental Results. We perform experiments with CIFAR-10/CIFAR-100 (Krizhevsky et al., 2009) benchmarks to support our method further. Please refer to Appendix C.1 for the results. We also explore the performance of using different CLIP vision encoders and different VLMs in Appendix D.1. When using different backbones, our EOE's performance is significantly better than the other methods. For example, when employing GroupViT (Xu et al., 2022) as the backbone, our EOE method outperforms MCM significantly, achieving a 23.22% improvement in FPR95 when using ImageNet-1K as the ID dataset. These results indicate that our EOE is more generalizable to different VLMs. To explore the robustness of EOE, we conduct experiments on more datasets, including ImageNet-Sketch (Wang et al., 2019) and ImageNet-C (Hendrycks & Dietterich, 2019). Compared to MCM, EOE demonstrates a 27.19% improvement in terms of FPR95 when using ImageNet-Sketch as the ID dataset. The detailed results are in Appendix C.2. What's more, We also perform prompt ensembling strategy for text input using different CLIP backbones. Interestingly, only our EOE achieves gains



(a) Score Function (b) LLM Prompts (c) Various LLMs

Figure 6: Ablation study on (a) score function, (b) LLM prompts, and (c) various LLMs. ID dataset: ImageNet-10; OOD dataset: ImageNet-20.

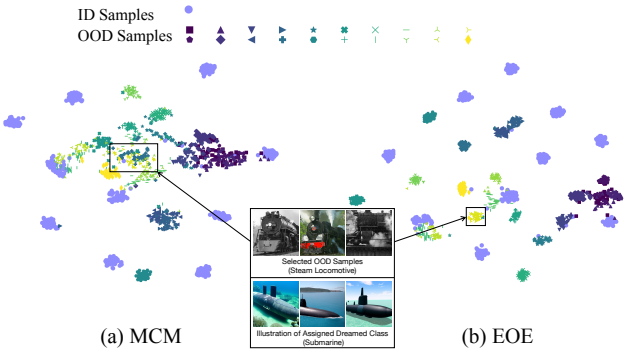


(a) Far OOD detection (b) Near OOD detection (c) Fine-grained OOD detection

Figure 7: Evaluation on the number of outlier class labels. When the number of outlier class labels is zero, the method reduces to the baseline MCM.

with the prompt ensembling strategy across different vision encoders. See Table 19 in Appendix C.3 for details.

Understanding of EOE. In fact, it is impractical for the generated outlier class labels to have a high probability of hitting the ground truth (GT) OOD class. This is because the OOD data encountered are diverse and unpredictable in the actual deployment of the model. However, it is for this reason that our EOE contributes to the OOD detection community. Based on the visual similarity rule, EOE asks the LLM to generate potential outlier class labels. *Even without hitting the GT OOD, these potential outlier classes can still enhance our performance in OOD detection.* To better understand this argument, we display the visualizations derived from the softmax output for the label-wise matching score via T-SNE (Van Der Maaten, 2014). Results compared between our EOE and the baseline MCM are shown in Figure 8. Based on the ID class labels of ImageNet-10, LLM generates the potential outlier label 'submarine' anchored on the visual similarity rule. When encountering the OOD class '*Steam Locomotive*'(class in ImageNet-20), '*Steam Locomotive*' exhibits the highest similarity with '*submarine*' among $\mathcal{Y}_{\text{id}}$ and $\mathcal{Y}_{\text{outlier}}$. Therefore, EOE will cluster it as '*submarine*', thus detecting it as an OOD class. Without the potential outlier class labels, we can find that MCM tends to cluster all the OOD class labels together. This could lead to identifying hard OOD samples as ID classes. We also present the visualization when using ImageNet-20 as the ID dataset, as shown in Figure 9. EOE will cluster actual OOD class '*antelope*' into potential outlier class '*Giraffe*'. In summary, within our EOE framework, 1) OOD samples



(a) MCM (b) EOE

Figure 8: T-SNE visualizations obtained by the classifier output. ID set: ImageNet-10; OOD set: ImageNet-20. We use distinct colors to represent different OOD classes. The illustrated envisioned OOD name is the class assigned with the corresponding cluster, and its examples are generated by Stable Diffusion (Rombach et al., 2022).

belonging to the same class tend to be clustered together, and 2) samples in the same group are classified into the envisioned outlier class that is visually similar to them (Steam Locomotive *vs* Submarine). These observations indicate that our EOE can enhance OOD detection without hitting actual OOD classes and is more semantically explainable.

Limitation I. Our method assumes that the LLM understands the ID categories. However, when the ID categories fall outside the expertise of the LLM, it may not provide appropriate candidate OOD categories. This situation can arise when the ID categories require extreme expertise or when the ID categories are entirely new, and the LLM has

never encountered them before. To handle this issue, we can adopt a preprocessing strategy. Specifically, before asking the LLM to provide candidate OOD categories, it should be queried whether it understands the given ID categories. If the LLM demonstrates understanding, our algorithm can proceed as designed; otherwise, it should fall back to the MCM approach. Furthermore, thanks to the generality of linguistic descriptions and the LLM's remarkable capacity for in-context learning, we can describe ID categories that are beyond the LLM's expert knowledge. By teaching the LLM about the visual features associated with these categories, we can then prompt the LLM to generate appropriate candidate OOD categories. This approach can also somewhat mitigate the impact of this failure case.

Limitation II. Another minor limitation of our approach is the use of different prompts, although our experiments demonstrate that the three proposed prompts are universal and can be applied across various datasets for each task. This implicitly assumes that we are aware of the specific type of OOD task at hand (far, near, or fine-grained). Despite this, the restriction imposed by this assumption is minimal and does not undermine the core principles of OOD detection. Additionally, we conducted experiments without this assumption, where we used the far OOD detection prompt (Figure 3) in all tasks. The results are shown in Table 25, our EOE method still surpasses MCM in most cases. In this paper, we hope to propose a new perspective using LLMs to explore potential OOD candidates for solving the zero-shot OOD detection task.

5. Related Works

OOD Detection. Methods adopted for previous OOD detection can be broadly categorized into classification-based (Hendrycks & Gimpel, 2017), density-based (Ren et al., 2019; Xiao et al., 2020), and reconstruction-based (Denouden et al., 2018; Zhou, 2022; Liu et al., 2023). Classification-based methods leverage a well-trained ID classifier and formulate a score function to recognize OOD samples. The score function can be formulated from input (Liang et al., 2017), hidden (Sun et al., 2022; Lee et al., 2018; Sun & Li, 2022; Zhu et al., 2023; Wang et al., 2022), output (Hendrycks & Gimpel, 2017; Hendrycks et al., 2019a; Liu et al., 2020), and gradient space (Huang et al., 2021). When the label space of the test ID data differs from the training data, the model needs to be re-trained from scratch or fine-tuned in previous OOD detection scenarios, which require significant computational overhead.

Zero-Shot OOD Detection. Owing to the powerful capabilities of the VLMs, zero-shot OOD detection methods have shown promising results. Typically, the similarity between the feature representations of input images and the textual input is measured to identify OOD samples. MCM (Ming et al., 2022) is entirely dependent on closed-set ID class labels and does not effectively harness the potent abilities of CLIP for handling open-world inputs. Although ZOC (Esmaeilpour et al., 2022) and CLIPN (Wang et al., 2023) take into account the open-world setting, ZOC requires an additional auxiliary dataset to train a text-based image description generator to generate candidate unknown class names for each input sample. This makes ZOC ill-suited for handling large-scale datasets. Similarly, CLIPN requires an extra auxiliary dataset for training the text encoder. By contrast, EOE not only considers the open-world scenario but also foregoes the need for any auxiliary datasets for extra training and can be easily scaled to large datasets.

LLM for Visual Classification. Drawing upon the expert knowledge embedded in LLMs has emerged as a novel trend in vision tasks and remains under-explored. (Menon & Vondrick, 2023; Maniparambil et al., 2023) employ the expertise within LLMs to extract information contained in ID class labels, thereby enhancing the performance of image classification (classification by description). Differing from this, EOE leverages the expert knowledge in LLMs to envision outlier OOD classes based on the visual similarity rule, harnessing the capabilities of VLMs more effectively, thus improving the performance of identifying OOD samples.

6. Conclusion

In this paper, we propose a new paradigm for zero-shot OOD detection, called EOE, by harnessing the expert knowledge embedded in LLMs to envision outlier exposure without relying on actual or auxiliary OOD data. Based on the designed visual similarity rule, the proposed three LLM prompts are applicable across various datasets for far, near, and fine-grained OOD detection tasks. We introduce a new score function based on potential outlier penalty, enabling us to recognize OOD samples effectively. Extensive experiments show that EOE achieves new state-of-the-art performance and can be effectively scaled to the ImageNet-1K dataset. We hope this work could open a new door in future research in the OOD detection field.

Acknowledgements

CTC, ZKZ, and BH were supported by the NSFC General Program No. 62376235, Guangdong Basic and Applied Basic Research Foundation Nos. 2024A1515012399 and 2022A1515011652, HKBU Faculty Niche Research Areas No. RC-FNRA-IG/22-23/SCI/04, ByteDance Faculty Research Award, and HKBU CSD Departmental Incentive Grant. YL was partially supported by the National Science Foundation (NSF) under grants IIS-2007951 and IIS-2143895. TLL was partially supported by the following Australian Research Council projects: FT220100318, DP220102121, LP220100527, LP220200949, and IC190100031.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Anthropic. Claude 2. Anthropic Blog, July 11 2023. URL https://www.anthropic.com/index/claude-2.

Anthropic, A. The claude 3 model family: Opus, sonnet, haiku. Claude-3 Model Card, 2024.

Bitterwolf, J., Mueller, M., and Hein, M. In or out? fixing imagenet out-of-distribution detection evaluation. arXiv preprint arXiv:2306.00826, 2023.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101 – mining discriminative components with random forests. In ECCV, 2014.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In NeurIPS, 2020.

Chen, Z., Liu, G., Zhang, B.-W., Ye, F., Yang, Q., and Wu, L. Altclip: Altering the language encoder in clip for extended language capabilities. arXiv preprint arXiv:2211.06679, 2022.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311, 2022.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In CVPR, 2014.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In CVPR, 2009.

Denouden, T., Salay, R., Czarnecki, K., Abdelzad, V., Phan, B., and Vernekar, S. Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. arXiv preprint arXiv:1812.02765, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021.

Esmaeilpour, S., Liu, B., Robertson, E., and Shu, L. Zero-shot out-of-distribution detection based on the pre-trained model clip. In AAAI, 2022.

Fort, S., Ren, J., and Lakshminarayanan, B. Exploring the limits of out-of-distribution detection. In NeurIPS, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In CVPR, 2016.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261, 2019.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In ICLR, 2017.

Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., and Song, D. Scaling out-of-distribution detection for real-world settings. arXiv preprint arXiv:1911.11132, 2019a.

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. In ICLR, 2019b.

Huang, R. and Li, Y. Mos: Towards scaling out-of-distribution detection for large semantic space. In CVPR, 2021.

Huang, R., Geng, A., and Li, Y. On the importance of gradients for detecting distributional shifts in the wild. In NeurIPS, 2021.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In ICML, 2021.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., et al. Mixtral of experts. arXiv preprint arXiv:2401.04088, 2024.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In ICCV Workshops, 2013.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In NeurIPS, 2018.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690, 2017.

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. In NeurIPS, 2020.

Liu, Z., Zhou, J. P., Wang, Y., and Weinberger, K. Q. Unsupervised out-of-distribution detection with diffusion inpainting. arXiv preprint arXiv:2302.10326, 2023.

Maniparambil, M., Vorster, C., Molloy, D., Murphy, N., McGuinness, K., and O'Connor, N. E. Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. arXiv preprint arXiv:2307.11661, 2023.

Menon, S. and Vondrick, C. Visual classification via description from large language models. In ICLR, 2023.

Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., and Li, Y. Delving into out-of-distribution detection with vision-language representations. In NeurIPS, 2022.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In NeurIPS Workshop, 2011.

OpenAI. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. Cats and dogs. pp. 3498–3505, 2012.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In NeurIPS, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In ICML, 2021.

Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.-b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530, 2024.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., and Lakshminarayanan, B. Likelihood ratios for out-of-distribution detection. In NeurIPS, 2019.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In CVPR, 2022.

Salehi, M., Mirzaei, H., Hendrycks, D., Li, Y., Rohban, M. H., and Sabokrou, M. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. arXiv preprint arXiv:2110.14051, 2021.

Sehwag, V., Chiang, M., and Mittal, P. Ssd: A unified framework for self-supervised outlier detection. In ICLR, 2021.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In ACL, 2018.

Sun, Y. and Li, Y. Dice: Leveraging sparsification for out-of-distribution detection. In European Conference on Computer Vision, pp. 691–708. Springer, 2022.

Sun, Y., Ming, Y., Zhu, X., and Li, Y. Out-of-distribution detection with deep nearest neighbors. In ICML, 2022.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In ECCV, 2020.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

Van Der Maaten, L. Accelerating t-sne using tree-based algorithms. JMLR, 15(1):3221–3245, 2014.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In CVPR, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In NeurIPS, 2017.

Vaze, S., Han, K., Vedaldi, A., and Zisserman, A. Open-set recognition: A good closed-set classifier is all you need? 2022.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. 2011.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. In NeurIPS, 2019.

Wang, H., Li, Y., Yao, H., and Li, X. Clipn for zero-shot ood detection: Teaching clip to say no. arXiv preprint arXiv:2308.12213, 2023.

Wang, Q., Liu, F., Zhang, Y., Zhang, J., Gong, C., Liu, T., and Han, B. Watermarking for out-of-distribution detection. In NeurIPS, 2022.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In CVPR, 2010.

Xiao, Z., Yan, Q., and Amit, Y. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. In NeurIPS, 2020.

Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., and Wang, X. Groupvit: Semantic segmentation emerges from text supervision. In CVPR, 2022.

Yang, J., Zhou, K., Li, Y., and Liu, Z. Generalized out-of-distribution detection: A survey. arXiv preprint arXiv:2110.11334, 2021.

Yang, Z., Xu, Q., Bao, S., He, Y., Cao, X., and Huang, Q. Optimizing two-way partial auc with an end-to-end framework. IEEE TPAMI, 2022.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. Lsun: construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365, 2015.

Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Zhou, K., Zhang, W., Li, Y., Liu, Z., Chen, Y., and Li, H. Openood v1.5: Enhanced benchmark for out-of-distribution detection. arXiv preprint arXiv:2306.09301, 2023.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. IEEE TPAMI, 40(6):1452–1464, 2017.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In CVPR, 2022a.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. IJCV, 130(9):2337–2348, 2022b.

Zhou, Y. Rethinking reconstruction autoencoder-based out-of-distribution detection. In CVPR, 2022.

Zhu, J., Li, H., Yao, J., Liu, T., Xu, J., and Han, B. Unleashing mask: Explore the intrinsic out-of-distribution detection capability. In ICML, 2023.

Appendix

A Further Analysis 13

A.1 Understanding EOE's Effectiveness: Without Hitting Actual OOD Classes . . . 13

A.2 Further Analysis on the Design of S_{EOE} . . . 13

B Experimental Details 14

B.1 Dataset Details . . . 14

B.2 Implement Details . . . 14

C Additional Empirical Results 20

C.1 Other OOD Detection Benchmarks . . . 20

C.2 Robustness . . . 20

C.3 Prompt Ensembling for Text Input . . . 20

D Additional Ablation Studies 23

D.1 VLM Backbones . . . 23

D.2 Score Functions . . . 26

D.3 LLM Prompts . . . 26

D.4 Various LLMs . . . 27

E Limitations 28

F Specific example of LLM prompts 28

A. Further Analysis

A.1. Understanding EOE's Effectiveness: Without Hitting Actual OOD Classes

Figure 9 presents the visualization when using ImageNet-20 as the ID dataset. This visualization indicates that OOD samples belonging to the same class tend to be clustered together under our EOE framework. OOD samples within the same group are classified into the envisioned outlier class that is visually similar to them. This demonstrates that our EOE method can improve OOD detection performance without hitting actual OOD class labels.

A.2. Further Analysis on the Design of S_{EOE}

How to utilize available outlier labels to enhance the ability of the score function to distinguish ID and OOD samples is crucial. The intuitive idea is to incorporate the envisioned class into the denominator of the MSP score:

$$S_{\text{MSP}}(x; \mathcal{Y}_{\text{id}}, \mathcal{Y}_{\text{outlier}}, \mathcal{T}, \mathcal{I}) = \max_{i \in [1,K]} \frac{e^{s_i(x)}}{\sum_{j=1}^{K+L} e^{s_j(x)}}, \tag{5}$$

However, in this case, the envisioned class only functions in the denominator, which doesn't significantly impact the final score distribution, implying that the envisioned class is not fully utilized. To amplify the role of the envisioned class, EOE further subtracts the second item ($-\beta \cdot \max_{k \in [K+1,K+L]} \frac{e^{s_k(x)}}{\sum_{j=1}^{K+L} e^{s_j(x)}}$), *i.e.*,

$$S_{\text{EOE}}(x; \mathcal{Y}_{\text{id}}, \mathcal{Y}_{\text{outlier}}, \mathcal{T}, \mathcal{I}) = \max_{i \in [1,K]} \frac{e^{s_i(x)}}{\sum_{j=1}^{K+L} e^{s_j(x)}} - \beta \cdot \max_{k \in (K,K+L]} \frac{e^{s_k(x)}}{\sum_{j=1}^{K+L} e^{s_j(x)}}, \tag{6}$$

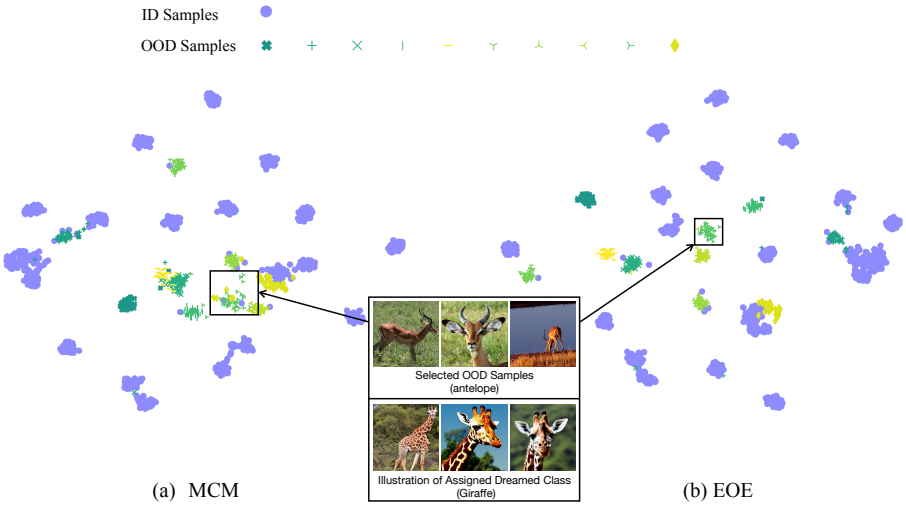


Figure 9: T-SNE visualizations obtained by the classifier output. ID set: ImageNet-20; OOD set: ImageNet-10. We use distinct colors to represent different OOD classes. The illustrated envisioned OOD name is the class assigned with the corresponding cluster, and its examples are generated by Stable Diffusion (Rombach et al., 2022).

It stems from an intuition: samples visually similar to the envisioned class should have lower scores, thus making it easier to distinguish between the ID and OOD score distribution. We conduct experiments on $\beta \in \{0, 0.25, 0.50, 0.75, 1\}$ with different outlier class number L, and the results are shown in Tables. 5 to 9. Empirically, $\beta \in [0.25, 0.5]$ is a suitable choice and EOE outperforms the strong baseline MCM even if L changes significantly.

B. Experimental Details

B.1. Dataset Details

In this section, we explain in detail how the ID dataset and OOD dataset are divided.

Far OOD Detection. Following the settings of MCM (Ming et al., 2022), we conduct our experiments on large-scale datasets[2]. The ID datasets include CUB-200-2011, STANFORD-CARS, Food-101, Oxford-IIIT Pet and ImageNet-1K. The OOD datasets are curated by MOS (Huang & Li, 2021), including subsets of iNaturalist, SUN, Places, and Textures. The OOD datasets are constructed based on the concept of the group, *i.e.*, decomposing a large semantic space into smaller groups with similar concepts, thereby simplifying the decision boundaries of OOD class labels and ID class labels. All OOD datasets exclude class labels that are present in the ID datasets. Based on the selected concepts, 10,000 images are sampled from each of the iNaturalist, SUN, and Places datasets. For the Textures dataset, the entire dataset is used.

Near OOD Detection. We use ImageNet-10 and ImageNet-20 curated by MCM, which are extracted from ImageNet-1K. Specifically, the class labels of ImageNet-10 and ImageNet-20 are shown in Figure 10

Fine-grained OOD Detection. We perform random partitioning of classes in the datasets CUB-200-2011, STANFORD-CARS, Food-101, and Oxford-IIIT Pet. We designate half of these classes as ID classes and the other half as OOD classes. This partitioning ensures that there is no overlap between the above ID dataset and the corresponding OOD dataset. After random partitioning, the ID/OOD datasets are CUB-100/CUB100, Stanford-Cars-98/Stanford-Cars-98, Food-50/Food-51, and Oxford-Pet-18/Oxford-Pet-19. In order to illustrate our random division more clearly, we take CUB-100(ID) and CUB-100(OOD) as examples, as shown in Figure 11.

B.2. Implement Details

B.2.1. Environment

EOE does not require any training. All experiments are conducted under the zero-shot setting without training samples. All experiments are performed using the PyTorch 1.13 library (Paszke et al., 2019) and Python 3.10.8, running on an NVIDIA

[2] https://github.com/deeplearning-wisc/large_scale_ood

Table 5: Zero-shot far OOD detection results with different outlier class label number L and β.

ID Dataset	Outlier Class Label Number	β	OOD Dataset iNaturalist FPR95↓	iNaturalist AUROC↑	SUN FPR95↓	SUN AUROC↑	Places FPR95↓	Places AUROC↑	Texture FPR95↓	Texture AUROC↑	Average FPR95↓	Average AUROC↑
CUB-200-2011	0 (MCM)		9.83	98.24	4.93	99.10	6.66	98.57	6.99	98.75	7.10	98.66
	$L = 100$	0	0.37	99.92	0.09	99.97	0.70	99.80	0.15	99.96	0.33	99.91
		0.25	0.06	99.98	0.04	99.99	0.37	99.88	0.01	99.99	0.12	99.96
		0.50	0.06	99.98	0.01	100.00	0.31	99.90	0.00	100.00	0.10	99.97
		0.75	0.07	99.98	0.01	100.00	0.28	99.92	0.00	100.00	**0.09**	**99.98**
		1.00	0.09	99.98	0.01	100.00	0.29	99.92	0.00	100.00	0.10	**99.98**
	$L = 300$	0	0.18	99.95	0.07	99.98	0.52	99.84	0.07	99.98	0.21	99.93
		0.25	0.06	99.98	0.03	99.99	0.39	99.88	0.01	99.99	0.12	**99.96**
		0.50	0.06	99.98	0.02	100.00	0.36	99.89	0.00	100.00	**0.11**	**99.96**
		0.75	0.07	99.97	0.02	100.00	0.38	99.89	0.01	99.98	0.12	**99.96**
		1.00	0.11	99.95	0.03	99.99	0.43	99.88	0.02	99.95	0.14	99.94
	$L = 500$	0	0.13	99.96	0.05	99.98	0.54	99.84	0.07	99.98	0.20	99.94
		0.25	0.06	99.98	0.03	100.00	0.37	99.88	0.01	100.00	0.12	99.96
		0.50	0.06	99.98	0.02	100.00	0.33	99.90	0.00	100.00	**0.10**	**99.97**
		0.75	0.08	99.97	0.02	100.00	0.32	99.90	0.00	99.99	0.11	**99.97**
		1.00	0.10	99.96	0.02	100.00	0.37	99.90	0.02	99.97	0.13	99.96
STANFORD-CARS	0 (MCM)		0.05	99.77	0.02	99.95	0.24	99.89	0.02	99.96	0.08	99.89
	$L = 100$	0	0.00	99.99	0.01	99.99	0.13	99.97	0.00	100.00	0.04	**99.99**
		0.25	0.00	99.99	0.01	99.99	0.11	99.97	0.00	100.00	**0.03**	**99.99**
		0.50	0.00	99.99	0.01	99.99	0.12	99.97	0.00	100.00	**0.03**	**99.99**
		0.75	0.01	99.97	0.07	99.97	0.14	99.95	0.01	100.00	0.06	99.97
		1.00	0.18	99.89	0.47	99.86	0.32	99.89	0.04	99.97	0.25	99.90
	$L = 300$	0	0.00	100.00	0.01	100.00	0.11	99.98	0.00	100.00	**0.03**	**100.00**
		0.25	0.00	100.00	0.01	100.00	0.09	99.98	0.00	100.00	**0.03**	**100.00**
		0.50	0.00	100.00	0.01	99.99	0.10	99.98	0.00	100.00	**0.03**	99.99
		0.75	0.00	99.99	0.03	99.99	0.12	99.96	0.01	100.00	0.04	99.98
		1.00	0.03	99.97	0.18	99.94	0.22	99.93	0.01	99.99	0.11	99.96
	$L = 500$	0	0.00	100.00	0.01	100.00	0.05	99.99	0.00	100.00	**0.02**	**100.00**
		0.25	0.00	100.00	0.01	100.00	0.06	99.99	0.00	100.00	**0.02**	**100.00**
		0.50	0.00	100.00	0.01	100.00	0.07	99.98	0.00	100.00	**0.02**	**100.00**
		0.75	0.00	100.00	0.01	100.00	0.11	99.98	0.00	100.00	0.03	**100.00**
		1.00	0.00	100.00	0.03	99.99	0.15	99.96	0.00	100.00	0.04	99.99
Average	0 (MCM)		4.94	99.01	2.48	99.53	3.45	99.23	3.51	99.35	3.59	99.28
	$L = 100$	0	0.19	99.95	0.05	99.98	0.42	99.88	0.07	99.98	0.19	99.95
		0.25	0.03	99.98	0.02	99.99	0.24	99.92	0.01	100.00	0.07	99.97
		0.50	0.03	99.98	0.01	99.99	0.21	99.94	0.00	100.00	**0.06**	**99.98**
		0.75	0.04	99.98	0.04	99.98	0.21	99.94	0.01	100.00	0.08	**99.98**
		1.00	0.13	99.94	0.24	99.93	0.31	99.90	0.02	99.98	0.17	99.94
	$L = 300$	0	0.09	99.97	0.04	99.99	0.31	99.91	0.03	99.99	0.12	99.96
		0.25	0.03	99.99	0.02	100.00	0.24	99.93	0.00	100.00	**0.07**	**99.98**
		0.50	0.03	99.99	0.01	100.00	0.23	99.93	0.00	100.00	**0.07**	**99.98**
		0.75	0.04	99.98	0.02	99.99	0.25	99.93	0.01	99.99	0.08	99.97
		1.00	0.07	99.96	0.10	99.97	0.32	99.90	0.02	99.97	0.12	99.95
	$L = 500$	0	0.07	99.98	0.03	99.99	0.29	99.92	0.03	99.99	0.11	99.97
		0.25	0.03	99.99	0.02	100.00	0.21	99.94	0.00	100.00	0.07	**99.98**
		0.50	0.03	99.99	0.01	100.00	0.20	99.94	0.00	100.00	**0.06**	**99.98**
		0.75	0.04	99.99	0.01	100.00	0.22	99.94	0.00	100.00	0.07	**99.98**
		1.00	0.05	99.98	0.02	100.00	0.26	99.93	0.01	99.99	0.09	99.97

ImageNet-10:
- "n04552348": "warplane"
- "n04285008": "sports car"
- "n01530575": "brambling bird"
- "n02123597": "Siamese cat"
- "n02422699": "antelope"
- "n02107574": "swiss mountain dog"
- "n01641577": "bull frog"
- "n03417042": "garbage truck"
- "n02389026": "horse"
- "n03095699": "container ship"

ImageNet-20:
- "n04147183": "sailboat"
- "n02782093": "balloon"
- "n03773504": "missile"
- "n02317335": "starfish"
- "n01630670": "common newt"
- "n02391049": "zebra"
- "n01697457": "African crocodile"
- "n02114367": "timber wolf"
- "n03785016": "moped"
- "n04266014": "space shuttle"
- "n02951358": "canoe"
- "n04389033": "tank"
- "n02917067": "bullet train"
- "n01632458": "spotted salamander"
- "n01631663": "eft"
- "n01693334": "green lizard"
- "n02120079": "Arctic fox"
- "n02132136": "brown bear"
- "n04310018": "steam locomotive"
- "n04252077": "snowmobile"

Figure 10: ImageNet-10 and ImageNet-20 class labels. To illustrate how each dataset is adapted to be ID-OOD.

Table 6: Zero-shot far OOD detection results with different outlier class label number L and β.

ID Dataset	Outlier Class Label Number	β	OOD Dataset iNaturalist		SUN		Places		Texture		Average	
			FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Food-101	0 (MCM)		0.64	99.78	0.90	99.75	1.86	99.58	4.04	98.62	1.86	99.43
	$L = 100$	0	0.12	99.97	0.02	99.99	0.29	99.94	2.73	98.97	0.79	99.72
		0.25	0.08	99.98	0.00	100.00	0.18	99.96	2.58	99.00	0.71	99.73
		0.50	0.06	99.98	0.00	100.00	0.14	99.97	2.61	99.01	**0.70**	**99.74**
		0.75	0.06	99.98	0.01	99.99	0.14	99.96	2.74	98.98	0.74	99.73
		1.00	0.07	99.98	0.01	99.98	0.18	99.94	3.06	98.89	0.83	99.70
	$L = 300$	0	0.09	99.98	0.00	100.00	0.11	99.98	2.55	99.04	0.69	**99.75**
		0.25	0.08	99.98	0.00	100.00	0.07	99.98	2.52	99.01	**0.67**	99.74
		0.50	0.07	99.98	0.00	100.00	0.08	99.98	2.64	98.96	0.70	99.73
		0.75	0.07	99.98	0.00	99.99	0.09	99.97	2.91	98.87	0.77	99.70
		1.00	0.08	99.97	0.01	99.99	0.12	99.95	3.23	98.71	0.86	99.65
	$L = 500$	0	0.07	99.99	0.00	100.00	0.10	99.98	2.43	99.08	**0.65**	**99.76**
		0.25	0.06	99.99	0.00	100.00	0.10	99.98	2.45	99.04	**0.65**	**99.76**
		0.50	0.04	99.99	0.00	100.00	0.09	99.98	2.58	98.98	0.68	99.74
		0.75	0.04	99.99	0.01	99.99	0.13	99.97	2.83	98.86	0.75	99.70
		1.00	0.08	99.98	0.02	99.98	0.22	99.94	3.39	98.64	0.93	99.63
Oxford-IIIT Pet	0 (MCM)		2.80	99.38	1.05	99.73	2.11	99.56	0.80	99.81	1.69	99.62
	$L = 100$	0	0.00	100.00	0.02	99.99	0.16	99.96	0.14	99.97	0.08	99.98
		0.25	0.00	100.00	0.01	99.99	0.14	99.97	0.11	99.97	**0.07**	**99.98**
		0.50	0.00	100.00	0.01	99.99	0.14	99.97	0.14	99.97	0.08	**99.98**
		0.75	0.00	100.00	0.01	99.99	0.14	99.97	0.14	99.96	0.08	**99.98**
		1.00	0.00	100.00	0.01	99.99	0.14	99.96	0.16	99.95	0.08	99.97
	$L = 300$	0	0.00	100.00	0.01	99.99	0.15	99.97	0.15	99.97	0.08	**99.98**
		0.25	0.00	100.00	0.01	99.99	0.13	99.97	0.14	99.97	**0.07**	**99.98**
		0.50	0.00	100.00	0.01	99.99	0.13	99.97	0.15	99.96	0.08	**99.98**
		0.75	0.00	100.00	0.02	99.99	0.13	99.97	0.17	99.96	0.08	**99.98**
		1.00	0.00	99.99	0.02	99.99	0.14	99.96	0.20	99.94	0.09	99.97
	$L = 500$	0	0.00	100.00	0.01	99.99	0.16	99.95	0.12	99.97	**0.07**	99.97
		0.25	0.00	100.00	0.01	99.99	0.15	99.96	0.12	99.97	**0.07**	**99.98**
		0.50	0.00	100.00	0.01	99.99	0.15	99.96	0.12	99.97	**0.07**	**99.98**
		0.75	0.00	100.00	0.02	99.99	0.15	99.95	0.13	99.96	0.08	99.97
		1.00	0.00	99.99	0.03	99.99	0.19	99.94	0.15	99.96	0.09	99.97
Average	0 (MCM)		1.72	99.58	0.98	99.74	1.98	99.57	2.42	99.21	1.78	99.53
	$L = 100$	0	0.06	99.99	0.02	99.99	0.23	99.95	1.43	99.47	0.44	99.85
		0.25	0.04	99.99	0.01	99.99	0.16	99.96	1.34	99.49	**0.39**	**99.86**
		0.50	0.03	99.99	0.01	100.00	0.14	99.97	1.38	99.49	**0.39**	**99.86**
		0.75	0.03	99.99	0.01	99.99	0.14	99.96	1.44	99.47	0.41	99.85
		1.00	0.04	99.99	0.01	99.99	0.16	99.95	1.61	99.42	0.46	99.83
	$L = 300$	0	0.05	99.99	0.01	100.00	0.13	99.98	1.35	99.51	0.38	99.87
		0.25	0.04	99.99	0.01	100.00	0.10	99.98	1.33	99.49	**0.37**	**99.86**
		0.50	0.04	99.99	0.01	100.00	0.11	99.98	1.40	99.46	0.39	**99.86**
		0.75	0.04	99.99	0.01	99.99	0.11	99.97	1.54	99.41	0.43	99.84
		1.00	0.04	99.98	0.01	99.99	0.13	99.95	1.72	99.33	0.47	99.81
	$L = 500$	0	0.03	99.99	0.01	100.00	0.13	99.96	1.27	99.52	**0.36**	**99.87**
		0.25	0.03	100.00	0.01	100.00	0.12	99.97	1.28	99.51	**0.36**	**99.87**
		0.50	0.02	99.99	0.01	99.99	0.12	99.97	1.35	99.47	0.38	99.86
		0.75	0.02	99.99	0.01	99.99	0.14	99.96	1.48	99.41	0.41	99.84
		1.00	0.04	99.98	0.02	99.98	0.20	99.94	1.77	99.30	0.51	99.80

Table 7: Zero-shot far OOD detection results for ImageNet-1K as the ID dataset with different outlier class label number L and β.

Outlier Class Label Number	β	**OOD Dataset** iNaturalist		SUN		Places		Texture		**Average**	
		FPR95↓	**AUROC↑**	**FPR95↓**	**AUROC↑**	**FPR95↓**	**AUROC↑**	**FPR95↓**	**AUROC↑**	**FPR95↓**	**AUROC↑**
0 (MCM)		30.92	94.61	37.59	92.57	44.71	89.77	57.85	86.11	42.77	90.77
$L = 100$	0	28.91	94.89	34.90	93.06	42.23	90.29	57.22	86.24	40.81	91.12
	0.25	20.40	96.23	22.65	95.27	31.82	92.54	58.54	85.49	33.36	92.39
	0.50	18.24	96.58	18.63	96.09	27.69	93.53	62.16	83.97	**31.68**	**92.54**
	0.75	19.93	96.38	18.01	96.28	27.18	93.82	67.60	81.96	33.18	92.11
	1.00	23.94	95.91	19.10	96.16	28.54	93.72	72.89	79.71	36.12	91.37
$L = 300$	0	29.42	94.78	33.83	93.28	41.11	90.52	57.03	86.26	40.35	91.21
	0.25	20.19	96.17	22.64	95.25	31.77	92.65	57.19	85.69	32.94	92.44
	0.50	18.17	96.52	19.57	95.90	28.50	93.50	60.84	84.16	**31.77**	**92.52**
	0.75	19.95	96.24	19.85	95.89	28.49	93.59	66.09	81.87	33.59	91.90
	1.00	24.25	95.56	22.03	95.53	30.77	93.22	73.04	79.15	37.52	90.86
$L = 500$	0	24.34	95.62	31.04	93.77	38.92	90.94	55.87	86.56	37.54	91.72
	0.25	12.29	97.52	20.40	95.73	30.16	92.95	57.53	85.64	30.09	92.96
	0.50	8.20	98.19	16.96	96.42	26.69	93.82	62.23	83.80	**28.52**	**93.06**
	0.75	7.14	98.33	16.55	96.52	26.31	94.01	68.18	81.27	29.55	92.53
	1.00	7.47	98.20	17.92	96.30	28.04	93.81	75.04	78.36	32.12	91.66

Table 8: Zero-shot near OOD detection results with different outlier class label number L and β. K is the number of classes in the corresponding ID dataset.

Outlier Class Label Number	β	**ID** ImageNet-10 **OOD** ImageNet-20		**ID** ImageNet-20 **OOD** ImageNet-10		**Average**	
		FPR95↓	**AUROC↑**	**FPR95↓**	**AUROC↑**	**FPR95↓**	**AUROC↑**
$L = 0$ (MCM)		5.00	98.71	17.40	97.87	11.20	98.29
$L = 1 \times K$	0	6.43	98.49	14.07	98.12	10.25	**98.30**
	0.25	7.40	97.95	9.40	98.63	**8.40**	98.29
	0.50	12.73	96.64	10.20	98.26	11.47	97.45
	0.75	24.40	94.10	19.27	97.04	21.83	95.57
	1.00	60.63	89.24	33.40	95.07	47.02	92.15
$L = 3 \times K$	0	6.50	98.62	16.40	97.95	11.45	98.28
	0.25	4.20	99.09	13.93	98.10	**9.07**	**98.59**
	0.50	4.00	98.84	15.47	97.83	9.73	98.33
	0.75	11.10	97.66	22.13	96.95	16.62	97.30
	1.00	32.60	94.79	30.00	95.45	31.30	95.12
$L = 10 \times K$	0	6.10	98.49	19.27	97.74	12.68	98.12
	0.25	3.63	99.11	16.20	97.92	9.92	**98.52**
	0.50	1.97	99.18	15.60	97.68	**8.78**	98.43
	0.75	4.30	98.72	19.07	96.81	11.68	97.77
	1.00	18.37	97.26	30.67	95.16	24.52	96.21

CUB-100 (ID):

- "Laysan Albatross"
- "Crested Auklet"
- "Rusty Blackbird"
- "Indigo Bunting"
- "Spotted Catbird"
- "Eastern Towhee"
- "Bobolink"
- "Bronzed Cowbird"
- "American Crow"
- "Acadian Flycatcher"
- "Gadwall"
- "Frigatebird"
- "American Goldfinch"
- "Eared Grebe"
- "Pied billed Grebe"
- "Evening Grosbeak"
- "Herring Gull"
- "Horned Grebe"
- "Ivory Gull"
- "Anna Hummingbird"
- "Blue Jay"
- "Gray Kingbird"
- "Horned Lark"
- "Western Meadowlark"
- "Ovenbird"
- "Orchard Oriole"
- "Sayornis"
- "Whip poor Will"
- "Fox Sparrow"
- "Chipping Sparrow"
- "House Sparrow"
- "Grasshopper Sparrow"
- "Le Conte Sparrow"
- "Vesper Sparrow"
- "Cliff Swallow"
- "Black Tern"
- "Brown Thrasher"
- "Black capped Vireo"
- "Philadelphia Vireo"
- "Caspian Tern"
- "Bewick Wren"
- "Purple Finch"
- "Magnolia Warbler"
- "Palm Warbler"
- "Pine Warbler"
- "Swainson Warbler"
- "Worm eating Warbler"
- "Cedar Waxwing"
- "Downy Woodpecker"
- "House Wren"
- "Groove billed Ani"
- "Parakeet Auklet"
- "Red faced Cormorant"
- "Painted Bunting"
- "Yellow breasted Chat"
- "Chuck will Widow"
- "Pelagic Cormorant"
- "Shiny Cowbird"
- "Golden winged Warbler"
- "Olive sided Flycatcher"
- "Yellow bellied Flycatcher"
- "Scissor tailed Flycatcher"
- "Boat tailed Grackle"
- "Glaucous winged Gull"
- "Blue Grosbeak"
- "Pine Grosbeak"
- "Pigeon Guillemot"
- "Rose breasted Grosbeak"
- "Ring billed Gull"
- "Pomarine Jaeger"
- "Green Jay"
- "Ringed Kingfisher"
- "Red legged Kittiwake"
- "Red breasted Merganser"
- "White breasted Nuthatch"
- "Hooded Oriole"
- "Western Wood Pewee"
- "White necked Raven"
- "Brewer Sparrow"
- "Clay colored Sparrow"
- "Great Grey Shrike"
- "Henslow Sparrow"
- "Savannah Sparrow"
- "Bank Swallow"
- "Scarlet Tanager"
- "Golden winged Warbler"
- "Sage Thrasher"
- "Blue headed Vireo"
- "Warbling Vireo"
- "Common Yellowthroat"
- "Cape May Warbler"
- "Hooded Warbler"
- "Myrtle Warbler"
- "Orange crowned Warbler"
- "Prothonotary Warbler"
- "Tennessee Warbler"
- "Northern Waterthrush"
- "Red bellied Woodpecker"
- "Blue winged Warbler"
- "Black throated" "Blue Warbler"

CUB-100 (OOD):

- "Sooty Albatross"
- "Least Auklet"
- "Brewer Blackbird"
- "Gray Catbird"
- "Cardinal"
- "Brandt Cormorant"
- "Fish Crow"
- "Mangrove Cuckoo"
- "Northern Flicker"
- "Northern Fulmar"
- "Heermann Gull"
- "Western Grebe"
- "California Gull"
- "Nighthawk"
- "Western Gull"
- "Green Violetear"
- "Florida Jay"
- "Common Raven"
- "Green Kingfisher"
- "Pacific Loon"
- "Mallard"
- "Mockingbird"
- "Brown Pelican"
- "Scott Oriole"
- "White Pelican"
- "Horned Puffin"
- "American Redstart"
- "Loggerhead Shrike"
- "Geococcyx"
- "Harris Sparrow"
- "Tree Sparrow"
- "Song Sparrow"
- "Seaside Sparrow"
- "Summer Tanager"
- "Tree Swallow"
- "Artic Tern"
- "Elegant Tern"
- "Least Tern"
- "Red eyed Vireo"
- "Common Tern"
- "Canada Warbler"
- "Forsters Tern"
- "Mourning Warbler"
- "Prairie Warbler"
- "Yellow Warbler"
- "Marsh Wren"
- "Winter Wren"
- "Cactus Wren"
- "Carolina Wren"
- "Rock Wren"
- "Black footed Albatross"
- "Rhinoceros Auklet"
- "Red winged Blackbird"
- "Lazuli Bunting"
- "Yellow headed Blackbird"
- "Brown Creeper"
- "Black billed Cuckoo"
- "Yellow billed Cuckoo"
- "Gray crowned Rosy Finch"
- "Least Flycatcher"
- "Great Crested Flycatcher"
- "European Goldfinch"
- "Vermilion Flycatcher"
- "Ruby throated Hummingbird"
- "Rufous Hummingbird"
- "Long tailed Jaeger"
- "Dark eyed Junco"
- "Belted Kingfisher"
- "Pied Kingfisher"
- "White breasted Kingfisher"
- "Hooded Merganser"
- "Slaty backed Gull"
- "Baltimore Oriole"
- "Clark Nutcracker"
- "American Pipit"
- "Tropical Kingbird"
- "Black throated Sparrow"
- "Baird Sparrow"
- "Field Sparrow"
- "Lincoln Sparrow"
- "White crowned Sparrow"
- "Nelson Sharp tailed Sparrow"
- "White throated Sparrow"
- "Barn Swallow"
- "Cape Glossy Starling"
- "Bay breasted Warbler"
- "Chestnut sided Warbler"
- "Green tailed Towhee"
- "White eyed Vireo"
- "Black and white Warbler"
- "Cerulean Warbler"
- "Kentucky Warbler"
- "Nashville Warbler"
- "Wilson Warbler"
- "Louisiana Waterthrush"
- "Pileated Woodpecker"
- "Red cockaded Woodpecker"
- "Red headed Woodpecker"
- "Bohemian Waxwing"
- "American Three toed" "Woodpecker"

Figure 11: CUB-100 (ID) and CUB-100 (OOD) class labels. To illustrate how each dataset is adapted to be ID-OOD.

Table 9: Zero-shot fine-grained OOD detection results with different outlier class label number L and β.

Outlier Class Label Number	β	ID / OOD	CUB-100 / CUB-100		Stanford-Cars-98 / Stanford-Cars-98		Food-50 / Food-51		Oxford-Pet-18 / Oxford-Pet-19		Average	
			FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
$L = 0$ (MCM)			83.58	67.51	83.99	68.71	43.38	91.75	63.92	84.88	68.72	78.21
$L = 100$	0		79.58	71.23	81.89	68.99	40.74	91.63	64.89	88.16	66.77	80.00
	0.25		76.79	72.31	77.73	70.92	38.36	91.83	56.03	89.63	62.23	**81.17**
	0.50		75.58	71.92	74.69	71.52	40.79	90.93	47.63	90.03	59.67	81.10
	0.75		75.54	70.14	72.89	70.85	45.39	89.29	44.52	89.46	**59.59**	79.93
	1.00		77.10	67.80	73.94	69.14	50.26	87.30	44.06	88.16	61.34	78.10
$L = 300$	0		80.55	70.72	80.90	69.26	46.90	90.03	61.89	87.52	67.56	79.38
	0.25		76.01	72.75	76.35	71.45	42.63	90.13	52.32	89.97	61.83	81.07
	0.50		72.62	73.51	72.64	72.71	42.30	89.13	44.60	91.51	58.04	**81.71**
	0.75		71.53	72.66	70.01	72.70	44.34	87.38	36.88	92.11	**55.69**	81.21
	1.00		71.99	70.63	70.90	71.35	47.51	85.29	35.84	91.80	56.56	79.77
$L = 500$	0		79.65	71.29	80.90	69.55	41.13	91.61	59.90	87.94	65.39	80.10
	0.25		74.74	73.41	76.83	71.60	37.95	91.96	52.55	90.33	60.52	81.82
	0.50		70.57	74.18	73.42	72.77	37.50	91.35	44.53	91.88	56.51	**82.54**
	0.75		69.35	73.27	70.12	72.77	40.60	89.92	38.03	92.48	**54.52**	82.11
	1.00		70.14	71.12	70.10	71.33	45.39	87.92	36.57	92.05	55.55	80.61

A100 80GB PCIe GPU and AMD EPYC 7H12 CPU.

B.2.2. Evaluation Metrics

We use the most commonly used metrics (FPR95 and AUROC) (Yang et al., 2022) in the OOD detection community on the main pages. To further explore the performance of EOE under other metrics, we report results in terms of AUPR on CLIP(ViT-B/16) backbone in Table 10. Our EOE achieves the best results on AUPR.

Table 10: Performance in terms of AUPR. ID dataset: The experiments are zero-shot OOD detection results with ImageNet-1K as the ID dataset. The **black bold** indicates the best performance. The gray indicates that the comparative methods require training or an additional massive auxiliary dataset.

Method	OOD Dataset				Average
	iNaturalist	SUN	Places	Texture	
CLIPN	99.15	98.59	98.22	98.38	98.59
Energy	96.84	96.50	96.16	94.66	96.04
MaxLogit	97.74	97.12	96.65	95.61	96.78
MCM	98.86	98.28	97.49	98.04	98.17
EOE (Ours)	99.47	99.01	98.27	97.97	**98.68**

Table 11: Additional empirical results with CIFAR-10 and CIFAR-100 as ID datasets. The **bold** indicates the best performance on each dataset. The gray indicates that the comparative methods require training or an additional massive auxiliary dataset.

ID Dataset	Method	OOD Dataset								Average	
		SVHN		LSUN		Texture		Places			
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CIFAR-10	CLIPN	53.28	74.20	27.89	92.72	3.58	98.93	9.82	96.98	23.64	90.63
	Energy	18.97	96.67	60.60	88.81	16.13	96.59	17.48	95.42	28.29	94.37
	MaxLogit	6.50	98.27	36.54	94.20	11.37	97.61	16.67	95.56	17.77	96.41
	MCM	3.98	99.03	5.12	98.72	16.35	96.44	36.55	90.79	15.50	96.25
	EOE (Ours)	5.78	98.20	4.69	98.64	5.61	98.66	11.94	96.36	**7.01**	**97.97**
CIFAR-100	CLIPN	71.72	68.20	84.42	80.90	37.74	90.92	51.06	87.25	61.24	81.82
	Energy	72.54	88.20	93.64	73.08	65.55	80.43	59.86	83.47	72.90	81.30
	MaxLogit	59.05	91.01	82.48	83.06	62.82	82.08	65.58	80.88	67.48	84.26
	MCM	64.45	89.96	47.26	91.69	90.30	73.61	98.42	61.37	75.11	79.16
	EOE (Ours)	68.47	88.78	36.85	93.29	66.89	82.64	77.60	78.06	**62.45**	**85.69**
Average	CLIPN	62.50	71.20	56.16	86.81	20.66	94.93	30.44	92.12	42.44	86.23
	Energy	45.76	92.44	77.12	80.95	40.84	88.51	38.67	89.45	50.60	87.84
	MaxLogit	32.78	94.64	59.51	88.63	37.10	89.85	41.13	88.22	42.63	90.34
	MCM	34.22	94.50	26.19	95.21	53.33	85.03	67.49	76.08	45.31	87.71
	EOE (Ours)	37.31	93.43	22.02	95.75	37.45	90.31	45.79	86.77	**35.64**	**91.56**

Table 12: Zero-shot **near** OOD detection results on large-scale datasets. The **bold** indicates the best performance on each dataset.

Method	ID OOD	ImageNet SSB-hard		ImageNet NINCO		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Energy		90.51	61.66	86.10	69.94	88.31	65.80
MaxLogit		88.25	62.68	80.35	72.81	84.30	67.75
MCM		88.71	64.69	79.40	74.27	84.06	69.48
EOE (Ours)		**85.99**	**66.36**	**73.63**	**76.93**	**79.81**	**71.65**

C. Additional Empirical Results

C.1. Other OOD Detection Benchmarks

We conduct experiments on the CIFAR-10/CIFAR-100 (as ID datasets) (Krizhevsky et al., 2009) benchmarks. The test OOD datasets include SVHN (Netzer et al., 2011), LSUN (Yu et al., 2015), Texture (Cimpoi et al., 2014), Places (Zhou et al., 2017). The results are shown in Table 11. Clearly, EOE achieves superior results on CIFAR-10/CIFAR-100. We additionally conduct experiments on the setting of large-scale near-OOD benchmarks organized by OpenOOD (Zhang et al., 2023). The ID dataset used is ImageNet (Deng et al., 2009), while the OOD datasets are SSB-hard (Vaze et al., 2022) and NINCO (Bitterwolf et al., 2023). Results in the Table 12 show that EOE achieves improvements of 4.25% on average FPR95 and 2.17% on average AUROC compared to MCM.

C.2. Robustness

To explore the OOD detection performance of data from different domains under the same ID class label, we conducted experiments on ImageNet-Sketch[3] (Wang et al., 2019) and ImageNet-C[4] (Hendrycks & Dietterich, 2019). Both ImageNet-Sketch and ImageNet-C contain the same 1000 classes as ImageNet. ImageNet-Sketch was constructed through Google Image queries 'sketch of *class_name*', and the images are all black and white. ImageNet-C includes five major types of corruption (Noise, Blur, Weather, Digital, Extra) with a total of 19 types, and it has five levels of severity. We report the results for level-1. Table 13 shows the experimental results with ImageNet-Sketch as the ID dataset.

Since ImageNet-Sketch shares the same ID class labels as ImageNet, EOE uses the same outlier class labels as ImageNet. Experiments show that the performance of MCM on ImageNet-Sketch deteriorates significantly, while our method still maintains good performance. *Compared to MCM, EOE achieves improvements of* 27.19% *in terms of FPR95.* Table 14 to 18 shows the performance of EOE under different types of corruption. Our EOE still significantly outperforms the compared methods. These results on ImageNet-Sketch and ImageNet-C indicate that our EOE exhibits stronger robustness against domain shifts compared to other methods.

Table 13: Robustness results on ImageNet-Sketch dataset. The ID class labels are the same as ImageNet. The **black bold** indicates the best performance.

Method	OOD Dataset iNaturalist		SUN		Places		Texture		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Energy	82.16	81.62	79.72	80.51	75.71	79.32	93.95	56.32	82.88	74.44
MaxLogit	70.90	84.44	72.90	81.91	70.69	79.91	89.80	60.77	76.07	76.76
MCM	62.88	88.08	68.55	85.01	71.25	81.01	77.84	75.35	70.13	82.36
EOE (Ours)	23.97	95.10	33.18	92.45	43.62	88.49	71.00	77.82	**42.94**	**88.47**

C.3. Prompt Ensembling for Text Input

To investigate the effect of prompt ensembling for CLIP text input on EOE, we conduct experiments using five types of prompts for CLIP text input from MCM under different vision encoders. The five prompts are shown in Figure 12. Table 19 presents the results of prompt ensembling under different vision encoders. **Interestingly, only our EOE achieves gains with the prompt ensembling strategy across different vision encoders.** In contrast, MCM exhibits varying degrees of

[3] https://github.com/HaohanWang/ImageNet-Sketch

[4] https://zenodo.org/records/2235448

Table 14: Robustness results on ImageNet-C dataset (corruption type: Noise). The ID class labels are the same as ImageNet. The **black bold** indicates the best performance.

ID Dataset	Method	OOD Dataset								Average	
		iNaturalist		SUN		Places		Texture			
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Gaussian	Energy	88.40	81.17	84.96	80.24	80.83	79.27	95.62	58.28	87.45	74.74
	MaxLogit	75.17	85.06	76.64	82.80	73.77	81.06	91.17	63.76	79.19	78.17
	MCM	64.26	88.04	69.88	85.12	72.31	81.42	78.74	76.11	71.30	82.67
	EOE (Ours)	27.16	94.91	35.87	92.27	46.16	88.30	73.17	77.63	**45.59**	**88.28**
Shot	Energy	87.40	81.18	84.14	80.20	80.03	79.18	95.32	57.70	86.72	74.56
	MaxLogit	74.53	85.00	76.08	82.70	73.32	80.89	90.98	63.22	78.73	77.95
	MCM	66.25	87.54	71.72	84.55	73.69	80.78	80.00	75.36	72.91	82.06
	EOE (Ours)	28.99	94.62	37.47	91.89	47.88	87.81	74.38	76.83	**47.18**	**87.79**
Impulse	Energy	87.47	78.07	84.18	76.95	80.10	75.86	95.35	51.57	86.78	70.61
	MaxLogit	78.82	81.37	79.54	78.74	76.36	76.79	92.02	56.79	81.69	73.42
	MCM	76.85	83.71	80.38	80.31	80.94	76.19	85.07	70.04	80.81	77.56
	EOE (Ours)	44.44	91.53	51.27	88.08	60.64	83.00	83.27	69.31	**59.90**	**82.98**
Average	Energy	87.76	80.14	84.43	79.13	80.32	78.10	95.43	55.85	86.98	73.30
	MaxLogit	76.17	83.81	77.42	81.41	74.48	79.58	91.39	61.26	79.87	76.51
	MCM	69.12	86.43	73.99	83.33	75.65	79.46	81.27	73.84	75.01	80.76
	EOE (Ours)	33.53	93.69	41.54	90.74	51.56	86.37	76.94	74.59	**50.89**	**86.35**

Table 15: Robustness results on ImageNet-C dataset (corruption type: Blur). The ID class labels are the same as ImageNet. The **black bold** indicates the best performance.

ID Dataset	Method	OOD Dataset								Average	
		iNaturalist		SUN		Places		Texture			
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Defocus	Energy	69.35	85.40	69.09	84.28	66.54	83.02	90.07	61.20	73.76	78.47
	MaxLogit	63.92	87.03	66.39	84.68	65.32	82.75	87.36	64.89	70.75	79.84
	MCM	84.88	81.53	87.26	78.01	87.12	73.73	89.13	67.32	87.10	75.15
	EOE (Ours)	43.67	91.96	50.61	88.58	60.05	83.60	82.97	70.16	**59.33**	**83.58**
Glass	Energy	66.72	86.25	67.13	85.13	64.68	83.80	89.38	61.95	71.98	79.28
	MaxLogit	61.03	87.72	63.81	85.35	63.19	83.36	86.40	65.39	68.61	80.45
	MCM	83.00	82.02	85.79	78.54	85.84	74.32	88.32	68.00	85.74	75.72
	EOE (Ours)	42.28	92.04	49.61	88.60	59.16	83.60	82.34	70.10	**58.35**	**83.58**
Motion	Energy	70.88	85.52	70.69	84.45	67.60	83.25	90.73	62.16	74.97	78.84
	MaxLogit	61.13	88.08	63.90	85.84	63.29	83.99	86.49	67.02	68.70	81.23
	MCM	70.65	86.67	75.45	83.52	76.53	79.52	82.20	73.77	76.21	80.87
	EOE (Ours)	31.66	94.18	39.82	91.30	50.20	86.98	76.34	75.35	**49.50**	**86.95**
Zoom	Energy	51.18	89.51	53.74	88.40	52.86	86.97	84.08	66.05	**60.47**	**82.73**
	MaxLogit	53.25	89.02	57.36	86.65	57.51	84.57	83.48	66.58	62.90	81.70
	MCM	92.87	75.13	93.61	71.21	93.20	66.70	93.33	59.63	93.25	68.17
	EOE (Ours)	57.37	88.59	61.96	84.58	70.10	78.76	88.85	62.96	69.57	78.72
Average	Energy	64.53	86.67	65.16	85.56	62.92	84.26	88.56	62.84	70.30	79.83
	MaxLogit	59.83	87.96	62.86	85.63	62.33	83.67	85.93	65.97	67.74	80.81
	MCM	82.85	81.34	85.53	77.82	85.67	73.57	88.24	67.18	85.57	74.98
	EOE (Ours)	43.74	91.69	50.50	88.27	59.88	83.23	82.63	69.64	**59.19**	**83.21**

Table 16: Robustness results on ImageNet-C dataset (corruption type: Weather). The ID class labels are the same as ImageNet. The **black bold** indicates the best performance.

ID Dataset	Method	OOD Dataset								Average	
		iNaturalist		SUN		Places		Texture			
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Snow	Energy	88.06	78.76	84.71	77.69	80.60	76.67	95.55	53.58	87.23	71.67
	MaxLogit	79.73	82.16	80.26	79.67	77.16	77.84	92.20	58.83	82.34	74.62
	MCM	77.45	84.17	80.91	80.87	81.42	76.80	85.46	70.81	81.31	78.16
	EOE (Ours)	52.85	90.16	58.33	86.76	66.92	81.63	86.93	67.72	**66.26**	**81.57**
Frost	Energy	75.03	85.72	74.20	84.75	70.57	83.69	92.02	64.13	77.95	79.57
	MaxLogit	63.27	88.10	65.84	85.97	64.76	84.26	87.13	68.16	70.25	81.62
	MCM	67.69	87.56	72.84	84.65	74.60	80.96	80.73	75.67	73.97	82.21
	EOE (Ours)	33.72	93.93	41.72	91.14	51.91	86.98	77.69	75.82	**51.26**	**86.97**
Fog	Energy	72.37	85.41	71.74	84.35	68.54	83.18	91.24	62.35	75.97	78.82
	MaxLogit	64.20	87.67	66.70	85.45	65.51	83.63	87.41	66.74	70.96	80.87
	MCM	71.89	86.20	76.22	83.03	77.37	79.01	82.73	73.22	77.05	80.37
	EOE (Ours)	36.54	93.31	44.19	90.31	54.23	85.80	79.26	73.64	**53.56**	**85.77**
Brightness	Energy	85.99	82.73	82.96	81.83	78.77	80.91	94.89	61.24	85.65	76.68
	MaxLogit	68.47	87.35	70.72	85.28	69.10	83.65	88.94	67.83	74.31	81.03
	MCM	41.83	92.69	47.88	90.34	54.15	87.18	65.48	82.93	52.34	88.29
	EOE (Ours)	16.72	96.68	25.48	94.61	35.70	91.42	63.57	82.96	**35.37**	**91.42**
Average	Energy	80.36	83.16	78.40	82.16	74.62	81.11	93.42	60.33	81.70	76.69
	MaxLogit	68.92	86.32	70.88	84.09	69.13	82.34	88.92	65.39	74.47	79.53
	MCM	64.72	87.66	69.46	84.72	71.89	80.99	78.60	75.66	71.17	82.26
	EOE (Ours)	34.96	93.52	42.43	90.71	52.19	86.46	76.86	75.03	**51.61**	**86.43**

Table 17: Robustness results on ImageNet-C dataset (corruption type: Digital). The ID class labels are the same as ImageNet-1K. The **black bold** indicates the best performance.

ID Dataset	Method	OOD Dataset								Average	
		iNaturalist		SUN		Places		Texture			
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Contrast	Energy	79.17	83.57	77.40	82.54	73.43	81.45	93.09	60.43	80.77	77.00
	MaxLogit	68.30	86.74	70.42	84.51	68.97	82.74	88.78	65.80	74.12	79.95
	MCM	71.34	86.64	75.87	83.57	76.97	79.67	82.52	74.06	76.67	80.98
	EOE (Ours)	34.05	93.83	42.05	90.94	52.17	86.60	77.87	74.93	**51.54**	**86.57**
Elastic Transformation	Energy	83.45	81.09	80.84	80.02	76.80	78.95	94.33	56.59	83.86	74.16
	MaxLogit	71.72	84.91	73.58	82.52	71.31	80.66	90.04	62.49	76.66	77.64
	MCM	68.27	87.02	73.42	83.94	75.05	80.04	81.06	74.44	74.45	81.36
	EOE (Ours)	30.94	94.24	39.27	91.36	49.68	87.06	75.80	75.51	**48.92**	**87.04**
Pixelate	Energy	82.32	83.51	79.97	82.54	75.90	81.50	94.04	61.07	83.06	77.16
	MaxLogit	68.89	86.86	70.93	84.67	69.30	82.93	89.06	66.23	74.55	80.17
	MCM	63.25	88.16	68.93	85.16	71.63	81.30	78.07	75.82	70.47	82.61
	EOE (Ours)	27.72	94.78	36.40	92.06	46.74	87.97	73.63	76.97	**46.12**	**87.94**
JPEG	Energy	87.33	82.47	84.09	81.61	79.97	80.76	95.27	61.64	86.67	76.62
	MaxLogit	73.30	86.47	74.87	84.42	72.17	82.86	90.55	67.18	77.72	80.23
	MCM	51.52	91.13	58.09	88.61	62.43	85.33	71.63	80.78	60.92	86.46
	EOE (Ours)	20.13	96.19	28.97	94.01	39.55	90.67	67.12	81.78	**38.94**	**90.66**
Average	Energy	83.07	82.66	80.58	81.68	76.53	80.67	94.18	59.93	83.59	76.23
	MaxLogit	70.55	86.25	72.45	84.03	70.44	82.30	89.61	65.42	75.76	79.50
	MCM	63.60	88.24	69.08	85.32	71.52	81.59	78.32	76.28	70.63	82.85
	EOE (Ours)	28.21	94.76	36.67	92.09	47.03	88.07	73.61	77.30	**46.38**	**88.06**

Table 18: Robustness results on ImageNet-C dataset (corruption type: Extra). The ID class labels are the same as ImageNet-1K. The **black bold** indicates the best performance.

ID Dataset	Method	OOD Dataset								Average	
		iNaturalist		SUN		Places		Texture			
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Speckle	Energy	83.33	82.83	80.70	81.85	76.56	80.83	94.27	60.08	83.72	76.40
	MaxLogit	69.52	86.36	71.49	84.13	69.68	82.37	89.38	65.36	75.02	79.56
	MCM	63.90	88.09	69.46	85.15	72.05	81.41	78.51	76.08	70.98	82.68
	EOE (Ours)	28.10	94.76	36.78	92.06	47.22	88.03	73.97	77.20	**46.51**	**88.02**
Gaussian	Energy	82.96	82.31	80.38	81.30	76.35	80.24	94.17	58.77	83.46	75.66
	MaxLogit	70.06	86.26	72.15	83.99	70.13	82.19	89.56	64.86	75.47	79.33
	MCM	63.56	88.30	69.20	85.32	71.85	81.48	78.28	76.05	70.72	82.79
	EOE (Ours)	27.06	94.88	35.85	92.18	46.19	88.12	73.19	77.22	**45.57**	**88.10**
Spatter	Energy	82.96	83.08	80.40	82.13	76.39	81.16	94.17	61.06	83.48	76.86
	MaxLogit	67.98	87.20	70.04	85.07	68.70	83.40	88.65	67.21	73.84	80.72
	MCM	48.49	91.36	55.08	88.77	59.96	85.33	70.07	80.61	58.40	86.52
	EOE (Ours)	22.03	95.77	30.79	93.40	41.33	89.77	68.85	80.07	**40.75**	**89.76**
Saturate	Energy	75.42	85.64	74.64	84.69	70.84	83.64	92.15	64.26	78.26	79.56
	MaxLogit	62.18	88.80	64.92	86.76	64.00	85.09	86.81	69.53	69.48	82.54
	MCM	52.30	90.68	59.00	88.00	63.21	84.45	72.16	79.57	61.67	85.67
	EOE (Ours)	22.27	95.75	31.26	93.36	41.81	89.70	69.21	79.96	**41.14**	**89.69**
Average	Energy	81.17	83.46	79.03	82.49	75.03	81.47	93.69	61.04	82.23	77.12
	MaxLogit	67.44	87.16	69.65	84.99	68.13	83.26	88.60	66.74	73.45	80.54
	MCM	57.06	89.61	63.18	86.81	66.77	83.17	74.75	78.08	65.44	84.42
	EOE (Ours)	24.86	95.29	33.67	92.75	44.14	88.91	71.30	78.61	**43.49**	**88.89**

performance degradation under ViT-B/16, ViT-B/32, RN50, and RN50x16.

Prompt set
- a photo of a <label>.
- a blurry photo of a <label>.
- a photo of many <label>.
- a photo of the large <label>.
- a photo of the small <label>.

Figure 12: Prompt set for text encoder input. <label> is the ID class name, such as "*horse*". The prompt set is extracted from CLIP (Radford et al., 2021) by MCM (Ming et al., 2022).

D. Additional Ablation Studies

D.1. VLM Backbones

In this section, we conduct experiments with different CLIP vision encoders to investigate the performance of EOE. Furthermore, we also report additional experimental results beyond CLIP models.

Table 20 shows the performance of ImageNet-1K(ID) with different CLIP vision encoders, including ViT-B/32[5], ViT-L/14[6], RN50[7], RN50x4, RN50x16 and RN101. Our EOE achieves the best results across all CLIP vision encoders. Compared with ViT-B/16, EOE yields an enhancement of 1.74% and 0.29% in FPR95 and AUROC based on ViT-L/14, respectively. Moreover, DOC achieves the best OOD detection performance compared to both zero-shot methods and fine-tuning methods in terms of FPR95 based on ViT-L/14. It should be noted that CLIPN achieves worse performance when using the VIT-L/14 as the backbone than when using VIT-B/16. Instead, our EOE is more generalizable to different backbones and produces clearly better OOD performance than CLIPN when using VIT-L/14 as the backbone. Compared to other zero-shot methods, our EOE also achieved the best results on the ResNet backbone. The results of fine-tuning methods are reported by MCM (Ming et al., 2022). Due to limitations in computational resources, we do not report the results of other fine-tuning methods on the ResNet backbone here.

[5] https://huggingface.co/openai/clip-vit-base-patch32
[6] https://huggingface.co/openai/clip-vit-large-patch14
[7] https://github.com/openai/CLIP

Table 19: Prompt ensembling for text input using different backbones. The ID dataset is ImageNet-1K. The "(+/- $\times\times$)" in the Average column indicates the difference compared to the text prompt of "*a photo of a <label>*.". Enhancements are marked in green, while performance degradation is indicated in red. Only our EOE achieves gains using prompt ensembling strategies across different backbones.

Method	**OOD Dataset**								**Average**	
	iNaturalist		SUN		Places		Texture			
	FPR95↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑	**FPR95**↓	**AUROC**↑
Energy (ViT-B/16)	79.75	83.75	79.81	83.21	70.28	83.95	88.23	71.51	79.52 (+2.69)	80.60 (+1.03)
MaxLogit (ViT-B/16)	67.24	87.31	66.14	86.36	61.09	85.96	80.83	76.01	68.83 (+0.25)	83.91 (+0.32)
MCM (ViT-B/16)	40.33	92.75	35.43	92.78	44.08	89.60	54.41	87.10	43.56 (-0.79)	90.56 (-0.21)
EOE (Ours) (ViT-B/16)	15.24	96.86	18.62	95.99	28.75	93.03	53.95	87.15	**29.14** (+0.95)	**93.26** (+0.30)
Energy (ViT-B/32)	89.22	79.15	81.01	81.62	61.22	87.20	87.64	71.36	79.77 (-1.87)	79.83 (-0.31)
MaxLogit (ViT-B/32)	79.45	83.75	68.89	84.85	52.30	88.60	79.88	75.29	70.13 (-3.17)	83.12 (-0.42)
MCM (ViT-B/32)	49.81	91.37	40.31	91.80	42.94	90.08	59.33	85.32	48.10 (-2.33)	89.64 (-0.28)
EOE (Ours) (ViT-B/32)	20.33	96.16	20.52	95.64	28.25	93.31	57.71	85.49	**31.71** (+0.01)	**92.65** (+0.15)
Energy (ViT-L/14)	79.20	85.29	76.83	84.68	65.62	87.59	87.23	70.14	77.22 (+3.26)	81.93 (+2.04)
MaxLogit (ViT-L/14)	63.06	89.02	60.26	88.29	52.51	89.65	80.66	73.96	64.12 (+3.08)	85.23 (+1.15)
MCM (ViT-L/14)	31.63	94.43	23.64	94.99	30.99	92.79	57.77	85.19	36.01(+2.09)	91.85 (+0.36)
EOE (Ours) (ViT-L/14)	13.26	97.43	14.51	97.01	22.35	94.79	56.77	85.27	**26.72** (+1.63)	**93.62** (+0.37)
Energy (RN50)	94.75	75.56	86.24	81.39	86.42	78.68	92.98	69.87	90.10(-1.26)	76.38 (-0.40)
MaxLogit (RN50)	86.45	81.21	74.56	84.31	78.15	81.10	86.45	74.61	81.40 (-2.22)	80.31 (-0.66)
MCM (RN50)	45.42	91.50	43.33	91.40	55.92	86.73	55.92	86.68	50.15 (-0.43)	89.08 (+0.09)
EOE (Ours) (RN50)	20.45	96.05	22.58	95.31	35.32	91.33	54.82	86.81	**33.30** (+1.18)	**92.38** (+0.39)
Energy (RN50x4)	85.55	81.25	80.13	84.81	68.84	85.40	92.09	69.28	81.65 (+1.23)	80.19 (+1.35)
MaxLogit (RN50x4)	74.51	85.14	65.51	87.61	58.86	87.26	84.47	74.81	70.84 (+0.03)	83.70 (+0.49)
MCM (RN50x4)	48.00	90.86	33.81	93.14	42.90	89.93	52.16	87.44	44.22 (+1.12)	90.34 (+0.35)
EOE (Ours) (RN50x4)	25.65	95.19	20.13	95.90	30.18	92.77	52.03	87.55	**32.00** (+2.56)	**92.85** (+0.60)
Energy (RN50x16)	73.44	86.95	65.15	88.97	73.74	83.97	84.43	76.11	74.19(-0.50)	84.00 (+0.64)
MaxLogit (RN50x16)	62.10	89.05	52.35	90.45	64.74	85.69	75.66	79.37	63.71 (-2.06)	86.14 (-0.05)
MCM (RN50x16)	43.02	91.69	34.24	93.27	46.96	89.27	51.93	87.94	44.04 (+1.14)	90.54 (+0.53)
EOE (Ours) (RN50x16)	23.03	95.45	20.00	95.92	31.94	92.52	52.54	87.65	**31.88** (+2.52)	**92.89** (+0.73)
Energy (RN101)	97.82	71.11	87.81	81.10	85.43	77.92	95.96	62.32	91.75 (-0.00)	73.11(-0.16)
MaxLogit (RN101)	92.65	77.38	74.77	84.67	75.96	81.30	90.90	68.66	83.57 (-1.72)	78.00 (-0.61)
MCM (RN101)	60.90	88.14	39.37	91.96	48.62	88.08	59.49	85.34	52.09 (-1.44)	88.38 (-0.11)
EOE (Ours) (RN101)	33.46	93.86	23.25	95.19	33.48	91.62	58.53	85.50	**37.18** (+1.10)	**91.54** (+0.36)

Table 20: Additional empirical results with different CLIP vision encoders on ImageNet-1K benchmark. The **black bold** indicates the best performance. The gray indicates that the comparative methods require training or an additional massive auxiliary dataset. Energy (FT) requires fine-tuning, while Energy is post-hoc.

Method	OOD Dataset								Average	
	iNaturalist		SUN		Places		Texture			
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Fort et al. (ViT-B/16)	15.07	96.64	54.12	86.37	57.99	85.24	53.32	84.77	45.12	88.25
Energy (FT) (ViT-B/16)	21.59	95.99	34.28	93.15	36.64	91.82	51.18	88.09	35.92	92.26
MSP (ViT-B/16)	40.89	88.63	65.81	81.24	67.90	80.14	64.96	78.16	59.89	82.04
CLIPN (ViT-B/16)	19.13	96.20	25.69	94.18	32.14	92.26	44.60	88.93	30.39	92.89
Fort et al. (ViT-L/14)	15.74	96.51	52.34	87.32	55.14	86.48	51.38	85.54	43.65	88.96
Energy (FT) (ViT-L/14)	10.62	97.52	30.46	93.83	32.25	93.01	44.35	89.64	29.42	93.50
MSP (ViT-L/14)	34.54	92.62	61.18	83.68	59.86	84.10	59.27	82.31	53.71	85.68
CLIPN (ViT-L/14)	25.09	94.59	24.76	94.93	30.89	93.14	48.97	87.01	32.43	92.42
Energy (ViT-B/16)	81.08	85.09	79.02	84.24	75.08	83.38	93.65	65.56	82.21	79.57
MaxLogit (ViT-B/16)	61.66	89.31	64.39	87.43	63.67	85.95	86.61	71.68	69.08	83.59
MCM (ViT-B/16)	30.92	94.61	37.59	92.57	44.71	89.77	57.85	86.11	42.77	90.77
EOE (Ours) (ViT-B/16)	12.29	97.52	20.40	95.73	30.16	92.95	57.53	85.64	**30.09**	**92.96**
Energy (ViT-B/32)	80.16	83.75	77.21	83.85	61.28	87.77	92.96	65.19	77.90	80.14
MaxLogit (ViT-B/32)	64.13	87.86	65.16	86.64	51.92	89.10	86.63	70.57	66.96	83.54
MCM (ViT-B/32)	33.91	93.61	41.82	91.42	45.66	89.56	61.67	84.67	45.77	89.82
EOE (Ours) (ViT-B/32)	14.97	97.03	21.88	95.40	29.91	92.97	60.10	84.61	**31.72**	**92.50**
Energy (ViT-L/14)	78.84	85.87	78.87	83.51	70.30	86.44	93.90	63.74	80.48	79.89
MaxLogit (ViT-L/14)	58.96	90.13	63.97	87.75	57.00	89.05	88.90	69.37	67.21	84.08
MCM (ViT-L/14)	28.35	94.95	28.93	94.14	35.34	92.00	59.79	84.88	38.10	91.49
EOE (Ours) (ViT-L/14)	11.79	97.63	17.14	96.46	24.93	94.23	59.55	84.69	**28.35**	**93.25**
Energy (RN50)	91.53	80.26	83.70	82.78	84.64	81.30	95.51	62.76	88.84	76.78
MaxLogit (RN50)	76.17	85.78	73.09	85.46	76.43	83.26	91.03	69.37	79.18	80.97
MCM (RN50)	32.06	93.86	46.14	90.74	60.60	85.67	60.09	85.71	49.72	88.99
EOE (Ours) (RN50)	14.44	97.05	24.85	94.82	38.89	90.64	59.73	85.44	**34.47**	**91.98**
Energy (RN50x4)	83.73	83.14	79.73	84.70	71.95	85.21	96.10	62.29	82.88	78.84
MaxLogit (RN50x4)	69.80	87.02	63.44	88.27	58.87	87.66	91.37	69.89	70.87	83.21
MCM (RN50x4)	44.87	91.42	35.22	92.85	44.15	89.49	57.11	86.21	45.34	89.99
EOE (Ours) (RN50x4)	26.48	95.01	21.82	95.47	32.71	92.21	57.22	86.32	**34.56**	**92.25**
Energy (RN50x16)	68.41	88.91	63.01	89.49	71.03	85.61	92.29	69.43	73.69	83.36
MaxLogit (RN50x16)	50.80	91.32	49.84	91.13	61.08	87.23	84.89	75.08	61.65	86.19
MCM (RN50x16)	35.98	93.05	38.20	92.30	51.33	87.89	55.20	86.80	45.18	90.01
EOE (Ours) (RN50x16)	21.19	95.81	23.42	95.19	35.86	91.59	57.13	86.05	**34.40**	**92.16**
Energy (RN101)	95.66	74.97	87.35	81.14	85.97	78.65	98.03	58.33	91.75	73.27
MaxLogit (RN101)	85.97	80.62	72.23	85.35	75.19	82.37	94.02	66.09	81.85	78.61
MCM (RN101)	51.90	89.78	40.09	91.68	50.69	87.50	59.93	85.00	50.65	88.49
EOE (Ours) (RN101)	32.57	93.97	24.60	94.73	36.43	90.98	59.53	85.07	**38.29**	**91.19**

Table 21: Additional empirical results with different vision-language models with ImageNet-1K as the ID dataset. The **black bold** indicates the best performance.

Method	OOD Dataset								Average	
	iNaturalist		SUN		Places		Texture			
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Energy (GroupViT)	85.33	79.29	83.49	79.78	80.05	77.79	93.72	59.34	85.65	74.05
MaxLogit (GroupViT)	81.80	79.60	83.01	79.03	80.42	76.91	93.49	59.78	84.68	73.83
MCM (GroupViT)	53.20	88.33	60.30	84.77	66.48	80.93	61.52	81.50	60.38	83.88
EOE (Ours) (GroupViT)	20.48	95.75	28.52	93.13	38.26	89.27	61.37	81.44	**37.16**	**89.90**
Energy (AltCLIP)	58.47	91.57	50.29	92.29	51.29	90.51	94.73	64.36	63.69	84.68
MaxLogit (AltCLIP)	49.81	92.35	41.29	93.11	45.73	91.04	89.50	68.98	56.58	86.37
MCM (AltCLIP)	43.55	92.88	28.39	94.56	38.22	91.57	53.58	87.00	40.94	91.50
EOE (Ours) (AltCLIP)	18.35	96.75	15.50	97.05	25.38	94.34	54.68	86.36	**28.48**	**93.63**
Energy (ALIGN)	92.18	83.62	74.14	86.28	70.48	84.64	68.65	81.76	76.36	84.08
MaxLogit (ALIGN)	83.47	84.95	68.46	86.63	67.08	84.77	65.32	81.89	71.08	84.56
MCM (ALIGN)	60.63	89.39	53.20	89.17	61.05	85.58	59.63	83.94	58.63	87.02
EOE (Ours) (ALIGN)	25.32	95.45	26.67	94.50	38.57	90.95	57.32	84.41	**36.97**	**91.33**

Table 22: Additional ablation studies on score functions. The **bold** indicates the best performance on each dataset.

Score Funtion	ID OOD	ImageNet-1K Texture		ImageNet-1K NINCO		Food-50 Food-50		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
S_{MAX}		60.58	85.48	100.00	73.18	100.00	85.39	86.86	81.35
S_{Energy}		78.82	78.43	78.98	75.64	45.89	88.28	67.90	80.78
S_{MaxLogit}		75.65	77.72	74.21	75.97	44.90	87.89	64.92	80.53
S_{MSP}		**55.88**	**86.56**	78.05	74.87	41.04	91.61	58.32	84.35
S_{EOE}		57.55	85.64	**73.63**	**76.93**	**37.95**	**91.96**	**56.37**	**84.85**

The results for other vision-language models are shown in Table 21. We perform experiments with GroupViT[8] (Xu et al., 2022), AltCLIP[9] (Chen et al., 2022) and ALIGN[10] (Jia et al., 2021). When using GroupViT as the backbone, our EOE's performance (37.16% in terms of FPR95) is significantly better than MCM's (60.38% in terms of FPR95). These results indicate that our EOE is more generalizable to different VLMs.

D.2. Score Functions

Here, we present the specific form of the score function designed in the ablation study. They are S_{MAX}, S_{MSP}, S_{Energy} and S_{MaxLogit}. First, we review the definition of label-wise matching score $s_i(x)$:

$$s_i(x) = \frac{\mathcal{I}(x) \cdot \mathcal{T}(t_i)}{\|\mathcal{I}(x)\| \cdot \|\mathcal{T}(t_i)\|}; \quad i \in [1, K+L],\ t_i \in \mathcal{Y}_{\text{id}} \cup \mathcal{Y}_{\text{outlier}}. \tag{7}$$

The specific form of S_{MAX} is as follows:

$$S_{\text{MAX}}(x; \mathcal{Y}_{\text{id}}, \mathcal{Y}_{\text{outlier}}, \mathcal{T}, \mathcal{I}) = \begin{cases} \frac{1}{K} & \max\limits_{i\in[1,K]} s_i < \max\limits_{j\in[K+1,L]} s_j \\ \max\limits_{i\in[1,K]} \frac{e^{s_i(x)}}{\sum_{j=1}^{K} e^{s_j(x)}} & \max\limits_{i\in[1,K]} s_i \geq \max\limits_{j\in[K+1,L]} s_j \end{cases}, \tag{8}$$

S_{MAX} indicates that if the s_j ($j \in [K+1, L]$) of an input sample is larger than the s_k ($j \in [1, K]$), this sample is recognized to be an OOD sample. This means that the highest degree of similarity observed between the input sample and the outlier label exceeds that between the input sample and any in-distribution (ID) class label. Otherwise, the input sample is calculated according to maximum softmax probabilities (MSP).

S_{MSP} is an adaptation of MSP, as defined in Eq. 5. Similarly, S_{Energy} and S_{MaxLogit} are modifications of the Energy and MaxLogit metrics, respectively, incorporating outlier classes into their secondary components.

$$S_{\text{Energy}}(x; \mathcal{Y}_{\text{id}}, \mathcal{Y}_{\text{outlier}}, \mathcal{T}, \mathcal{I}) = -T \left(\log \sum_{i=1}^{K} e^{f_i(x)/T} - \log \sum_{j=K+1}^{L} e^{f_j(x)/T} \right), \tag{9}$$

$$S_{\text{MaxLogit}}(x; \mathcal{Y}_{\text{id}}, \mathcal{Y}_{\text{outlier}}, \mathcal{T}, \mathcal{I}) = \max_{i\in[1,K]} s_i(x) - \max_{j\in[K+1,K+L]} s_j(x). \tag{10}$$

We conducted additional ablation study experiments for score functions under three more challenging scenarios, including two large-scale benchmark settings and one fine-grained OOD detection setting in Table 22. Specifically, for large-scale benchmarks, we evaluate far OOD detection (ImageNet-1K *vs.* Texture) and near OOD detection (ImageNet-1K *vs.* NINCO, organized by OpenOOD). For fine-grained OOD detection, we evaluate the setting of Food-50 *vs.* Food-51. Our S_{EOE} consistently outperform the commonly used scores of S_{MAX}, S_{Energy}, and S_{MaxLogit}. Compared between two variants of our scores, *i.e.*, S_{EOE} and S_{MSP}, our weighted version (S_{EOE}, β=0.25) achieves the best results in most cases and on average.

D.3. LLM Prompts

We performed additional ablation study experiments for LLM prompts on the same three datasets as in Section D.2. The results are shown in Table 23, and our resemble prompt consistently outperforms other prompts in these challenging scenarios.

[8] https://huggingface.co/nvidia/groupvit-gcc-yfcc
[9] https://huggingface.co/BAAI/AltCLIP
[10] https://huggingface.co/kakaobrain/align-base

Table 23: Additional ablation studies on LLM prompts. The **bold** indicates the best performance on each dataset.

LLM Prompt	ID OOD	ImageNet-1K Texture		ImageNet-1K NINCO		Food-50 Food-50		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Irrelevant		60.69	85.14	79.36	74.30	41.89	90.74	60.64	83.39
Dissimilar		58.86	85.29	75.10	76.78	40.00	91.67	57.98	84.58
Resemble		**57.55**	**85.64**	**73.63**	**76.93**	**37.95**	**91.96**	**56.37**	**84.85**

We provide specific examples of *'dissimilar'* and *'irrelevant'* LLM prompts in Figure 13 and Figure 14, respectively. The LLM chosen is GPT-3.5-turbo-16k.

Q: Given the image category [water jug], please suggest visually dissimilar categories that are not directly related or belong to the same primary group as [water jug]. Provide suggestions that do not share visual characteristics but are from broader and different domains than [water jug].
A: There are three classes dissimilar to [water jug], and they are from broader and different domains than [water jug]:
- trumpets
- helmets
- rucksacks

Q: Given the image category [horse], please suggest visually dissimilar categories that are not directly related or belong to the same primary group as [horse]. Provide suggestions that do not share visual characteristics but are from broader and different domains than [horse].
A: There are three classes dissimilar to [horse], and they are from broader and different domains than [horse]:

ID class label: horse

- pineapple
- laptop
- mountain

Figure 13: Instance of *'dissimilar'* LLM prompt for OOD detection, ID class label: *horse*. Note that, the gray is the LLM prompt for near OOD detection, and the green is the LLM actually returns.

Q: Given the image category [water jug], please suggest categories that are not directly related or belong to the same primary group as [water jug].
A: There are three classes from broader and different domains than [water jug]:
- trumpets
- helmets
- rucksacks

Q: Given the image category [horse], please suggest categories that are not directly related or belong to the same primary group as [horse].
A: There are three classes from broader and different domains than [horse]:

ID class label: horse

- pineapple
- laptop
- sunglasses

Figure 14: Instance of *'irrelevant'* LLM prompt for OOD detection, ID class label: *horse*. Note that gray is the LLM prompt for near OOD detection, and green is the LLM that actually returns.

D.4. Various LLMs

We conducted additional ablation studies on various LLMs using the same three datasets mentioned in Section D.2. The results, presented in Table 24, indicate that our strategy is effective across different LLMs. Note that due to safety precautions, Claude2 declines to process large-scale datasets such as ImageNet-1K. And LLaMA2-7B's responses do not adhere strictly to our predefined JSON format. Consequently, we employed alternative LLMs, including Mixtral-8x7B-Chat (Jiang et al.,

Table 24: Additional ablation studies on various LLMs. The **bold** indicates the best performance on each dataset.

LLM	ID OOD	ImageNet-1K Texture		ImageNet-1K NINCO		Food-50 Food-50		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MCM		57.85	86.11	79.40	74.27	43.48	91.75	60.21	84.04
EOE(Ours, Mixtral-8x7B-Chat)		56.95	86.02	73.70	75.50	33.48	93.28	54.71	84.93
EOE(Ours, Claude-3-Haiku)		54.4	86.58	73.80	75.99	37.08	92.44	55.09	85.00
EOE(Ours, Gemini-Pro)		56.56	86.85	72.88	75.67	33.73	93.02	54.39	85.19
EOE(Ours, GPT-3.5-turbo-16k)		57.55	85.64	73.63	76.93	37.95	91.96	56.37	84.85
EOE(Ours, GPT-4)		53.65	87.05	73.41	74.96	35.66	93.10	54.24	85.04

2024), Gemini-Pro (Reid et al., 2024), Claude-3-Haiku (Anthropic, 2024), and GPT-4 (OpenAI, 2023).

E. Limitations

The results are shown in Table 25, without prior knowledge of the OOD task. In the majority of scenarios, the performance of our EOE remains superior to that of the MCM.

Table 25: Using **far** OOD detection prompt to envision outlier class labels on Zero-shot **near** and **fine-grained** OOD detection tasks. The green indicates Our EOE outperforms the strong baseline MCM.

Method	ID OOD	near OOD detection				fine-grained OOD detection							
		ImageNet-10 ImageNet-20		ImageNet-20 ImageNet-10		CUB-100 CUB-100		Stanford-Cars-98 Stanford-Cars-98		Food-50 Food-51		Oxford-Pet-18 Oxford-Pet-19	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MCM		5.00	98.71	17.40	97.87	83.58	67.51	83.99	68.71	43.38	91.75	63.92	84.88
EOE (Ours)		4.20	99.09	13.93	98.10	74.74	73.41	76.83	71.60	37.95	91.96	52.55	90.33
EOE (Ours) (far prompt)		5.47	98.74	11.40	98.21	77.84	72.29	76.24	73.03	39.98	91.62	67.27	88.21

F. Specific example of LLM prompts

To obtain the outlier class labels automatically generated by the LLM, we provide a question and answer template for LLM and append '-' to the end of the answer to produce a bullet-point list output. It is worth noting that the provided template does not contain any ID class content. Below, we present specific examples for three LLM prompts. The LLM prompt for far OOD detection is depicted in Figure 15. The LLM prompts for near OOD detection and fine-grained OOD detection are showcased in Figure 16 and Figure 17, respectively.

To better understand the outlier class labels generated by LLMs, we show in Figure 18 the outlier class labels generated by GPT-3.5-turbo-16k for ImageNet-1K (ID), which were instructed under the far OOD prompt.

Q: I have gathered images of 4 distinct categories: ['Husky dog', 'Garfield cat', 'churches', 'truck']. Summarize what broad categories these categories might fall into based on visual features. Now, I am looking to identify 5 categories that visually resemble these broad categories but have no direct relation to these broad categories. Please list these 5 items for me.
A: These 5 items are:
- black stone
- mountain
- Ginkgo Tree
- river
- Rapeseed

Q: I have gathered images of 100 distinct categories: ['Apple pie', 'Baby back ribs', 'Baklava', 'Beef carpaccio', 'Beef tartare', 'Beet salad', 'Beignets', 'Bibimbap', 'Bread pudding', 'Breakfast burrito', 'Bruschetta', 'Caesar salad', 'Cannoli', 'Caprese salad', 'Carrot cake', 'Ceviche', 'Cheesecake', 'Cheese plate', 'Chicken curry', 'Chicken quesadilla', 'Chicken wings', 'Chocolate cake', 'Chocolate mousse', 'Churros', 'Clam chowder', 'Club sandwich', 'Crab cakes', 'Creme brulee', 'Croque madame', 'Cup cakes', 'Deviled eggs', 'Donuts', 'Dumplings', 'Edamame', 'Eggs benedict', 'Escargots', 'Falafel', 'Filet mignon', 'Fish and chips', 'Foie gras', 'French fries', 'French onion soup', 'French toast', 'Fried calamari', 'Fried rice', 'Frozen yogurt', 'Garlic bread', 'Gnocchi', 'Greek salad', 'Grilled cheese sandwich', 'Grilled salmon', 'Guacamole', 'Gyoza', 'Hamburger', 'Hot and sour soup', 'Hot dog', 'Huevos rancheros', 'Hummus', 'Ice cream', 'Lasagna', 'Lobster bisque', 'Lobster roll sandwich', 'Macaroni and cheese', 'Macarons', 'Miso soup', 'Mussels', 'Nachos', 'Omelette', 'Onion rings', 'Oysters', 'Pad thai', 'Paella', 'Pancakes', 'Panna cotta', 'Peking duck', 'Pho', 'Pizza', 'Pork chop', 'Poutine', 'Prime rib', 'Pulled pork sandwich', 'Ramen', 'Ravioli', 'Red velvet cake', 'Risotto', 'Samosa', 'Sashimi', 'Scallops', 'Seaweed salad', 'Shrimp and grits', 'Spaghetti bolognese', 'Spaghetti carbonara', 'Spring rolls', 'Steak', 'Strawberry shortcake', 'Sushi', 'Tacos', 'Takoyaki', 'Tiramisu', 'Tuna tartare', 'Waffles']. Summarize what broad categories these categories might fall into based on visual features. Now, I am looking to identify 50 classes that visually resemble these broad categories but have no direct relation to these broad categories. Please list these 50 items for me.
A: These 50 items are:

ID dataset: Food-101

- Orchid
- Watermelon
- Sunflower
- ...
- Neon graffiti

Figure 15: Instance of LLM prompt for far OOD detection, ID dataset: Food-101. Note that, the gray is the LLM prompt for far OOD detection, and the green is the LLM actually returns.

Q: Given the image category [water jug], please suggest visually similar categories that are not directly related or belong to the same primary group as [water jug]. Provide suggestions that share visual characteristics but are from broader and different domains than [water jug].
A: There are three classes similar to [water jug], and they are from broader and different domains than [water jug]:
- trumpets
- helmets
- rucksacks

Q: Given the image category [horse], please suggest visually similar categories that are not directly related or belong to the same primary group as [horse]. Provide suggestions that share visual characteristics but are from broader and different domains than [horse].
A: There are three classes similar to [horse], and they are from broader and different domains than [horse]:

ID class label: horse

- zebra
- giraffe
- deer

Figure 16: Instance of LLM prompt for near OOD detection, ID class label: *horse*. Note that, the gray is the LLM prompt for near OOD detection, and the green is the LLM actually returns.

Q: I have a dataset containing 10 unique species of dogs. I need a list of 10 distinct dog species that are NOT present in my dataset, and ensure there are no repetitions in the list you provide. For context, the species in my dataset are: ['husky dog', 'alaskan Malamute', 'cossack sled dog', 'golden retriever', 'German Shepherd', 'Beagle', 'Bulldog', 'Poodle', 'Dachshund', 'Doberman Pinscher']
A: The other 10 dog species not in the dataset are:
- Labrador Retriever
- Rottweiler
- Boxer
- Border Collie
- Shih Tzu
- Akita
- Saint Bernard
- Australian Shepherd
- Great Dane
- Boston Terrier

Q: I have a dataset containing 50 different species of food. I need a list of 50 distinct food species that are NOT present in my dataset, and ensure there are no repetitions in the list you provide. For context, the species in my dataset are: ['Prime Rib', 'Dumplings', 'Strawberry Shortcake', 'Frozen Yogurt', 'Seaweed Salad', 'Tiramisu', 'Red Velvet Cake', 'Omelette', 'Beef Carpaccio', 'Lasagna', 'Donuts', 'Sushi', 'Beignets', 'Chicken Wings', 'Carrot Cake', 'Gnocchi', 'Lobster Bisque', 'Spaghetti Bolognese', 'Greek Salad', 'Oysters', 'Caprese Salad', 'Panna Cotta', 'Shrimp And Grits', 'Baby Back Ribs', 'Creme Brulee', 'Gyoza', 'Escargots', 'Churros', 'Grilled Cheese Sandwich', 'Scallops', 'Breakfast Burrito', 'Cheesecake', 'Huevos Rancheros', 'Cheese Plate', 'Steak', 'Apple Pie', 'Mussels', 'Crab Cakes', 'Pancakes', 'Pulled Pork Sandwich', 'Bruschetta', 'Hot Dog', 'Risotto', 'Chicken Curry', 'Paella', 'Cannoli', 'Eggs Benedict', 'Fried Calamari', 'French Fries', 'Lobster Roll Sandwich']
A: The other 50 food species not in the dataset are:

ID dataset: Food50

- Lychee
- chicken tacos
- Durian
- ...
- affogato
- pizza

Figure 17: Instance of LLM prompt for fine-grained OOD detection, ID dataset: Food-50. Note that, the gray is the LLM prompt for fine-grained OOD detection, and the green is the LLM actually returns.

Outlier label for ImageNet-1K:"Sunflower", "Waterfall", "Desert", "Sunset", "Rainbow", "Snowflake", "Aurora borealis", "Lightning", "Galaxy", "Moon", "Star", "Cloud", "Forest", "Mountain range", "Beach sunset", "Autumn leaves", "Spring flowers", "Winter landscape", "Summer beach", "City skyline", "Countryside", "Ocean waves", "Water droplet", "Fireworks", "Hot air balloon", "Iceberg", "Sand dunes", "Tropical island", "Rainforest", "Canyon", "Glacier", "Tornado", "Volcanic eruption", "Underwater coral reef", "Safari wildlife", "Desert oasis", "Northern lights", "Full moon", "Shooting star", "Thunderstorm", "Milky Way", "Rolling hills", "Waterfall", "Sunset over the ocean", "Snow-capped mountains", "Lush green meadow", "Colorful autumn foliage", "Blooming cherry blossoms", "Frozen lake", "Sandy beach with palm trees- Desert cactus", "Tropical rainforest", "Snowy mountain peak", "Vibrant sunset sky", "Misty waterfall", "Sandy beachscape", "Autumn forest scene", "Lush green valley", "Colorful flower field", "Majestic ocean view", "Serene lake scene", "Rolling countryside hills", "Urban cityscape skyline", "Starry night sky", "Thunderstorm lightning", "Fiery sunset horizon", "Tranquil river scene", "Snowy winter landscape", "Blossoming cherry trees", "Vibrant coral reef", "Rocky canyon walls", "Frozen Arctic tundra", "Exotic tropical island", "Dense jungle foliage", "Majestic volcano eruption", "Serene moonlit night", "Milky Way galaxy", "Vibrant rainbow arc", "Sandy desert dunes", "Crystal clear waterfall", "Autumn foliage reflection", "Snow-covered pine forest", "Sunflower field", "Vibrant coral reef", "Rolling green hills", "Sandy beach with palm trees", "Colorful hot air balloons", "Frozen lake with mountains", "Lush green meadow", "Blooming cherry blossoms", "Rocky mountain range", "Serene ocean waves", "Thunderstorm with lightning", "Milky Way galaxy", "Vibrant sunset over water", "Snowy alpine landscape", "Peaceful countryside scene", "Colorful spring flowers", "Tropical island paradise", "Desert oasis with palm trees- Cherry blossom tree", "Snowy mountain range", "Golden sunset beach", "Misty forest scene", "Colorful tulip field", "Serene lake reflection", "Rolling green meadows", "Vibrant butterfly garden", "Majestic waterfall cascade", "Tranquil riverbank scene", "Snowy winter wonderland", "Blooming sunflower field", "Urban city skyline", "Starry night landscape", "Thunderstorm cloudscape", "Fiery sunset over mountains", "Crystal clear river", "Autumn foliage carpet", "Tropical coral reef", "Rocky desert landscape", "Frozen Arctic wilderness", "Exotic palm-fringed island", "Dense tropical rainforest", "Majestic volcanic crater", "Serene moonlit beach", "Milky Way starry sky", "Vibrant double rainbow", "Sandy desert oasis", "Cascading waterfall pool", "Autumn forest path", "Snow-covered alpine peaks", "Sunflower garden", "Vibrant underwater coral", "Rolling countryside scenery", "Sandy beach paradise", "Colorful hot air balloons", "Frozen lake panorama", "Lush green pasture", "Blooming cherry orchard", "Rocky mountain summit", "Serene ocean sunset", "Thunderstorm lightning bolts", "Milky Way galaxy cluster", "Vibrant sunset over mountains", "Snowy pine forest", "Peaceful countryside road", "Colorful spring garden", "Tropical island getaway", "Desert oasis with camels- Lavender field", "Snow-capped peaks", "Golden sunset sky", "Misty woodland scene", "Colorful wildflowers meadow", "Serene lake landscape", "Rolling green hills", "Vibrant butterfly garden", "Majestic waterfall view", "Tranquil riverbank scenery", "Snowy winter forest", "Blooming poppy field", "Urban cityscape view", "Starry night panorama", "Thunderstorm clouds", "Fiery sunset over ocean", "Crystal clear stream", "Autumn foliage trail", "Tropical coral reef", "Rocky desert canyon", "Frozen Arctic landscape", "Exotic palm-fringed beach", "Dense rainforest canopy", "Majestic volcanic eruption", "Serene moonlit lake", "Milky Way galaxy view", "Vibrant rainbow colors", "Sandy desert dunes", "Cascading waterfall pool", "Autumn forest path", "Snow-covered alpine scenery", "Sunflower bouquet", "Vibrant underwater life", "Rolling countryside fields", "Sandy beach paradise", "Colorful hot air balloons", "Frozen lake reflection", "Lush green meadow", "Blooming cherry blossom", "Rocky mountain range", "Serene ocean waves", "Thunderstorm lightning", "Milky Way galaxy cluster", "Vibrant sunset over water", "Snowy pine forest", "Peaceful countryside road", "Colorful spring garden", "Tropical island getaway", "Desert oasis with camels- Lavender field", "Snow-capped peaks", "Golden sunset sky", "Misty woodland scene", "Colorful wildflowers meadow", "Serene lake landscape", "Rolling green hills", "Vibrant butterfly garden", "Majestic waterfall view", "Tranquil riverbank scenery", "Snowy winter forest", "Blooming poppy field", "Urban cityscape view", "Starry night panorama", "Thunderstorm clouds", "Fiery sunset over ocean", "Crystal clear stream", "Autumn foliage trail", "Tropical coral reef", "Rocky desert canyon", "Frozen Arctic landscape", "Exotic palm-fringed beach", "Dense rainforest canopy", "Majestic volcanic eruption", "Serene moonlit lake", "Milky Way galaxy view", "Vibrant rainbow colors", "Sandy desert dunes", "Cascading waterfall pool", "Autumn forest path", "Snow-covered alpine scenery", "Sunflower bouquet", "Vibrant underwater life", "Rolling countryside fields", "Sandy beach paradise", "Colorful hot air balloons", "Frozen lake reflection", "Lush green meadow", "Blooming cherry blossom", "Rocky mountain range", "Serene ocean waves", "Thunderstorm lightning", "Milky Way galaxy cluster", "Vibrant sunset over water", "Snowy pine forest", "Peaceful countryside road", "Colorful spring garden", "Tropical island getaway", "Desert oasis with camels- Lavender field", "Snow-capped peaks", "Golden sunset sky", "Misty woodland scene", "Colorful wildflowers meadow", "Serene lake landscape", "Rolling green hills", "Vibrant butterfly garden", "Majestic waterfall view", "Tranquil riverbank scenery", "Snowy winter forest", "Blooming poppy field", "Urban cityscape view", "Starry night panorama", "Thunderstorm clouds", "Fiery sunset over ocean", "Crystal clear stream", "Autumn foliage trail", "Tropical coral reef", "Rocky desert canyon", "Frozen Arctic landscape", "Exotic palm-fringed beach", "Dense rainforest canopy", "Majestic volcanic eruption", "Serene moonlit lake", "Milky Way galaxy view", "Vibrant rainbow colors", "Sandy desert dunes", "Cascading waterfall pool", "Autumn forest path", "Snow-covered alpine scenery", "Sunflower bouquet", "Vibrant underwater life", "Rolling countryside fields", "Sandy beach paradise", "Colorful hot air balloons", "Frozen lake reflection", "Lush green meadow", "Blooming cherry blossom", "Rocky mountain range", "Serene ocean waves", "Thunderstorm lightning", "Milky Way galaxy cluster", "Vibrant sunset over water", "Snowy pine forest", "Peaceful countryside road", "Colorful spring garden", "Tropical island getaway", "Desert oasis with camels"

Figure 18: Outlier class labels generated by GPT-3.5-turbo-16k for ImageNet-1K. Due to space constraints, we show results for $L = 300$ here.