[Re] Negative Label Guided OOD Detection with Pretrained Vision-Language Models

Jeonghyeon Kim
Dept. of Data Science
Seoul National University of Science and Technology
mawjdgus@ds.seoultech.ac.kr

Seulbi Lee
Dept. of Data Science
Seoul National University of Science and Technology
seulbi@ds.seoultech.ac.kr

Uichan Lee
Dept. of Data Science
Seoul National University of Science and Technology
uichan@ds.seoultech.ac.kr

Reproducibility Summary

Unlike traditional visual out-of-distribution (OoD) detection methods, Vision-Language Models (VLMs) have the capability to leverage multimodal information. However, the potential of such multimodal information in VLM-based OoD detection remains largely untapped. In this report, we aim to reproduce the results reported by [7], which propose a method to more effectively leverage the knowledge embedded in VLMs for OoD detection. Our objective is to evaluate whether this methodology, which leverages VLMs knowledge, can exhibit generalizability beyond the benchmark datasets utilized in their study, extending to more challenging benchmarks. To achieve this, we employ the MOS benchmark [6], one of the most widely used in recent studies, as well as the more challenging OpenOOD v1.5 benchmark [13] for our reproduction efforts. We observed the generalizability of [7] in leveraging the capabilities of VLMs knowledge through reproducing experiments and conducting extended benchmark dataset experiments.

1 Introduction

1.1 VLMs for OoD detection

Out-of-Distribution (OoD) detection is a critical challenge in the field of machine learning, especially for models deployed in dynamic and unpredictable environments. Traditional OoD detection techniques primarily use only visual inputs and do not leverage textual information. Recently, with the development of Vision Language Models (VLMs) such as CLIP [9], the capability of OoD detection has been significantly enhanced due to the rich textual and visual knowledge embedded within these models.

There has been considerable research interest in leveraging the textual information provided by VLMs for zero-shot OoD detection. For instance, ZOC [4] introduces the novel task of zero-shot OoD detection and employs a trainable captioner to produce candidate OoD labels that correspond to OoD images. Nonetheless, when applied to extensive datasets with numerous in-distribution (ID) classes, such as ImageNet-1k, the captioner may fail to generate effective candidate OoD labels, leading to subpar performance. On the other hand, MCM [8] identifies OoD images using the maximum logit of scaled softmax. However, approaches like MCM rely solely on the information from the ID label space and do not fully exploit the textual information of VLMs. As a result, there is significant untapped potential to improve OoD detection by more effectively leveraging the rich textual knowledge contained within VLMs.

1.2 Proposed Methods : How to fully leverage VLMs for OoD detection

The paper we aim to reproduce proposes a method to enhance the utilization of knowledge embedded in VLMs for OoD detection. The authors introduce a substantial number of negative labels to enable the model to differentiate OoD samples with greater nuance and detail. Their method, termed NegLabel, identifies OoD samples by analyzing the affinities between ID labels and negative labels. The paper also presents the NegMining algorithm, designed to select high-quality negative labels from extensive corpus databases. This algorithm measures the distance between a

Preprint. Under review.

negative label and ID labels to evaluate their suitability based on semantic divergence. By choosing negative labels with significant semantic differences from ID labels, the algorithm improves the separability between ID and OoD samples. Additionally, the paper introduces a novel scoring scheme for OoD detection. This score is positively correlated with the affinities between images and ID labels and negatively correlated with the affinities between images and negative labels. By integrating knowledge from both the ID and negative label spaces, this approach better leverages the VLMs' text comprehension capabilities. The authors also provide a theoretical analysis to enhance understanding of the role and mechanism of negative labels.

We aim to replicate the findings reported in the paper, which outlines a method for enhancing OoD detection by leveraging VLMs through the application of negative labels. The contributions of the original paper can be summarized as follows:

- **Introduction of Negative Labels**. The authors propose the use of negative labels to significantly enhance the ability of VLMs to distinguish between ID and OoD samples.
- **NegMining Algorithm**. They introduce the NegMining algorithm, which effectively selects high-quality negative labels that increase the semantic divergence between ID and OoD samples, thereby improving detection accuracy.
- **Novel Scoring Scheme**. A new scoring scheme is developed that combines the affinities between images and both ID and negative labels. This approach better leverages the text comprehension capabilities of VLMs, resulting in superior OoD detection performance.

Our goal is to determine if the proposed methodology, which effectively leverages the knowledge of VLMs, can generalize beyond the benchmark datasets originally used in their study, extending to more demanding benchmarks. To this end, we utilize the MOS benchmark [6], a prominent benchmark in recent research, as well as the more challenging OpenOOD v1.5 benchmark [13] for our reproduction experiments. Our findings indicate impressive zero-shot performance even on the OpenOOD v1.5 benchmark.

2 OoD detection with negative labels

To utilize negative labels for OoD detection, we follow the approach as described in [7]. We start by selecting a high-quality set of negative labels using the NegMining algorithm. These negative labels, which exhibit significant semantic divergence from the ID labels, are then merged with the ID labels to create an extended label space. This extended label space is used to compute text embeddings via the CLIP text encoder. The OoD score for each image is calculated by evaluating the cosine similarities between the image embeddings and both the ID and negative text embeddings. Finally, the scores are aggregated to determine the overall confidence of whether an image belongs to the ID set, effectively distinguishing between ID and OoD samples.

NegMining. In the original paper, NegMining utilizes a vast collection of words from a lexical database, such as WordNet [5], to generate a candidate label space $\mathcal{T}^c = \{\tilde{t}_1, \tilde{t}_2, \dots, \tilde{t}_C\}$, where C represents the total number of candidate texts. For each candidate label $\tilde{t}_i$ (where $i = 1, 2, ..., C$), the authors compute negative cosine similarities as the distance measure between the candidate label and the entire ID label space:

$$D_i = \text{percentile}_\eta \left(\{D_{ik}\}_{k=1}^K\right) = \text{percentile}_\eta \left(\left\{-\cos(\tilde{T}_i, T_k)\right\}_{k=1}^K\right) \tag{1}$$

Here, $\tilde{T}_i = \mathbf{f}^{\text{text}}(\text{prompt}(\tilde{t}_i))$ denotes the text embedding of the candidate word $\tilde{t}_i$, with $\mathbf{f}^{\text{text}}$ being the text encoder. The variable K represents the number of ID classes, and percentile_η, where $\eta \in [0, 1]$, represents the 100η-th percentile of the data. The negative label selection criteria is then defined as $\mathcal{T}^- = \text{topk}(\{D_1, D_2, \dots, D_C\}, \mathcal{T}^C, M)$. The operator $\text{topk}(A, B, M)$ identifies the indices of the top-M elements in set A and selects the corresponding elements from set B based on these indices. This approach systematically narrows down the candidate labels by selecting those that exhibit the highest negative cosine similarities, thereby enhancing the identification of OoD samples through a refined negative label space.

NegLabel Score. Given the selected negative label set $\mathcal{T}^-$, it is combined with the ID labels $\mathcal{T}$ to form an extended label space $\mathcal{T}^{\text{ext}} = \mathcal{T} \cup \mathcal{T}^-$. These extended labels are then fed into the CLIP text encoder to generate their respective text embeddings. The cosine similarities between these text embeddings and the image embeddings are subsequently calculated. The proposed OoD score is formulated as:

$$S(\mathcal{I}) = S^*(\mathrm{sim}(\mathcal{I}, \mathcal{T}), \mathrm{sim}(\mathcal{I}, \mathcal{T}^-)), \tag{2}$$

where $S^*(\cdot, \cdot)$ is a fusion function that integrates the similarity of the sample with the ID labels $\mathrm{sim}(\mathcal{I}, \mathcal{T})$ and the similarity with the negative labels $\mathrm{sim}(\mathcal{I}, \mathcal{T}^-)$.

In this extended label space $\mathcal{T}^{\mathrm{ext}}$, the OoD detection task can be interpreted as the model's confidence in the image belonging to $\mathcal{T}$. A natural approach is to apply a normalization function to evaluate the proportion of the similarity $\mathrm{sim}(\mathcal{I}, \mathcal{T})$ within the extended label space $\mathcal{T}^{\mathrm{ext}}$. The paper proposes a NegLabel score in a sum-softmax form as follows:

$$S(\mathcal{I}) = \frac{\sum_{i=1}^{K} \exp(\cos(I \cdot T_i)/\tau)}{\sum_{i=1}^{K} \exp(\cos(I \cdot T_i)/\tau) + \sum_{j=1}^{M} \exp(\cos(I \cdot \tilde{T}_j)/\tau)} \tag{3}$$

where K is the number of ID labels, τ is the temperature coefficient of the softmax function, and M is the number of negative labels. This method effectively quantifies the OoD score by leveraging the relationships between image embeddings and both ID and negative text embeddings, ensuring a robust and accurate detection mechanism.

Grouping Strategy. The authors begin by evenly dividing the selected M negative labels into n_g groups, each containing approximately $[M/n_g]$ negative labels, with any remaining labels being discarded. They then compute a separate NegLabel score for each group. The final NegLabel score is obtained by summing the scores of all groups and calculating the average.

3 Implementation Details

To replicate the method proposed by [7], we endeavored to faithfully follow the methodologies presented in the paper. The specific details of implementation are provided in Table 1. To accurately reproduce the original paper results, we employed the MOS benchmark, which was used as the main benchmark in [7]. Furthermore, to assess the paper's generalizability in more challenging scenarios, we utilized the OpenOOD v1.5 dataset. These experiments aim to validate the generalizability and performance of the proposed method under diverse and demanding conditions.

Datasets and Benchmarks We assess NegLabel using the ImageNet-1k OoD detection benchmarks, as well as MOS [6] and OpenOOD v1.5 [13]. In line with MOS, we employ ImageNet-1k [3] as the ID dataset, and include a range of semantically diverse OoD datasets such as iNaturalist [10], SUN [12], Places [14], and Textures [2]. Additionally, we use the OpenOOD v1.5 benchmarks to evaluate performance in Near-OoD and Far-OoD scenarios, also utilizing ImageNet-1k as the ID dataset. For Near-OoD, we incorporate datasets like SSB-hard [11] and NINCO [1], which are semantically more challenging to distinguish from the ID dataset. In the Far-OoD category, we utilize datasets including iNaturalist, Textures, and OpenImage-O.

OoD Detection Metrics False Positive Rate at 95% True Positive Rate (FPR95): This metric measures the false positive rate of OoD images when the true positive rate of ID images is set at 95%. It provides insight into the model's ability to minimize false alarms while maintaining high sensitivity to true positive ID samples. Area Under the Receiver Operating Characteristic Curve (AUROC): This metric calculates the area under the ROC curve, which plots the true positive rate against the false positive rate at various threshold settings. AUROC provides a comprehensive measure of the model's discriminative performance across all classification thresholds.

4 Replication Experiments

4.1 Results in original paper : MOS benchmark datasets

We evaluated the MOS benchmark datasets to replicate the results presented in [7]. As shown in Table 2, our findings confirm that the NegLabel methodology effectively leverages the richer textual information of VLMs compared to methodologies such as MCM. Additionally, our reproduced performance surpasses that reported in the original paper. We attribute this improvement to the difference in the number of nouns and adjectives used. While the original paper mentioned the use of 10,000 negative labels, they extracted 10,499 nouns and adjectives from the negative labels, resulting in diminished performance.

Table 1: Fine-grain details for replicating the experiments

Property	Values	Remark
Framework	Pytorch	Version 1.12.0
Models	CLIP-ViT-B/16	OpenAI 400M
Datasets	OoD : iNaturalist, SUN, Places, Textures, SSB-hard, NINCO, iNaturalist, Textures, OpenImage-o	ID : Imagenet-1k MOS, OpenOOD v1.5
Corpus	WordNet	Only semantic information, using only nouns and adjectives.
GPU resources	NVIDIA RTX A6000 x1	
Prompt template	The nice <label>	default template in [7]
Percentile	0.05	default distance percentile in [7]
Group numbers, n_g	50, 450	MOS, OpenOOD v1.5, 100 in [7]
Neglabel numbers	Noun : 8500, Adj : 1500	Noun : 8578, Adj : 1921 in [7]

Table 2: This table presents the results of experiments conducted on the MOS benchmark dataset. The utilized ID data is from ImageNet-1k, with all results expressed as percentages. ↑ denotes that higher values indicate improved performance, while ↓ shows that lower values are more desirable.

Methods	iNaturalist		SUN		Places		Textures		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MCM	32.28	94.40	39.33	92.28	44.94	89.83	57.98	85.99	43.63	90.63
NegLabel	1.48	99.59	17.24	95.90	32.76	92.02	44.95	89.72	24.11	94.31

As shown in Table 3, similar to the experiment in [7], we conducted NegLabel inference using multiple prompts to evaluate whether NegLabel was dependent on prompt engineering. Consistent with the performance reported in [7], our results indicate that NegLabel's performance is significantly influenced by the choice of prompts. Following the original paper's methodology, we used the prompt "The nice <label>", which yielded the best performance among the tested prompts.

Table 4 shows the performance variations according to the number of groups in the grouping strategy. In the original paper, the best performance was achieved with 100 groups, which was adopted as the standard. However, in our replication, the optimal performance was obtained with 50 groups. Table 5 illustrates how performance changes depending on the selected percentile when measuring the distance between WordNet and ID text embeddings. Our experiments demonstrate that OoD detection performance improves as the selected percentile decreases. This finding suggests that selecting the closest distances between ID texts and WordNet results in more effective text choices for OoD detection in the MOS benchmark.

4.2 Validation of Generalizability and Performance : OpenOOD v1.5 benchmark datasets

In this subsection, we aim to validate the generalizability and performance of the original paper under more challenging conditions. As shown in Table 6, contrary to prior understanding, the prompt "The nice <label>" does not exhibit the best performance in Near OoD and Far OoD scenarios. In contrast, as seen in Table 6, the prompt "a cropped photo of a

Table 3: Prompt engineering : MOS Benchmark.

Prompt	iNaturalist		SUN		Places		Textures		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
a dark photo of a <label>	1.99	99.36	28.67	94.37	48.34	89.16	73.05	79.05	38.01	90.48
a blurry photo of a <label>	2.16	99.32	25.42	93.79	45.30	88.17	69.79	80.74	35.67	90.51
a low resolution photo of a <label>	1.62	99.55	23.86	94.61	43.08	89.82	70.46	81.09	34.76	91.27
a cropped photo of a <label>	1.16	99.66	24.79	94.97	43.70	90.35	67.07	82.29	34.18	91.82
a good photo of a <label>	1.37	99.62	22.11	95.19	42.46	90.35	61.03	85.58	31.74	92.68
<label>	1.48	99.60	17.26	96.04	30.84	92.70	58.78	87.12	27.09	93.87
The nice <label>	1.48	99.59	17.24	95.90	32.76	92.02	44.95	89.72	24.11	94.31

Table 4: Impact of the group numbers in the grouping strategy on OoD detection: MOS benchmark.

Group Number	iNaturalist		SUN		Places		Textures		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
1	2.30	99.37	23.23	95.14	39.85	90.98	46.49	89.64	27.97	93.78
50	1.48	99.59	17.24	95.90	32.76	92.02	44.95	89.72	24.11	94.31
100	1.55	99.58	17.95	95.82	33.53	91.97	44.34	89.86	24.34	94.31
150	1.56	99.58	18.25	95.76	33.68	91.93	43.87	89.92	24.34	94.30
200	1.56	99.58	18.52	95.74	33.87	91.92	43.76	89.97	24.43	94.30
250	1.59	99.57	18.77	95.72	34.12	91.90	43.58	90.00	24.52	94.30
300	1.59	99.57	18.77	95.72	34.12	91.90	43.58	90.00	24.52	94.30
350	1.60	99.57	19.08	95.69	34.26	91.88	43.55	90.00	24.62	94.29
400	1.60	99.57	19.07	95.68	34.36	91.88	43.37	90.04	24.60	94.29
450	1.61	99.57	19.06	95.68	34.26	91.88	43.30	90.04	24.56	94.29
500	1.61	99.57	19.14	95.67	34.37	91.87	43.39	90.06	24.63	94.29
1000	1.64	99.56	19.49	95.64	34.64	91.85	43.28	90.08	24.76	94.28

Table 5: The choice of distances for the selection of negative labels: MOS benchmark.

Percentile	iNaturalist		SUN		Places		Textures		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
0.95	1.86	99.48	17.88	95.72	32.70	91.83	46.56	89.30	24.75	94.08
0.90	1.70	99.50	18.06	95.77	32.61	91.80	45.66	89.51	24.51	94.15
0.85	1.67	99.52	17.87	95.80	33.12	91.82	45.35	89.70	24.50	94.21
0.80	1.61	99.54	17.99	95.80	33.11	91.86	45.73	89.45	24.61	94.16
0.75	1.58	99.55	17.88	95.77	33.40	91.78	45.76	89.46	24.66	94.14
0.70	1.56	99.57	17.36	95.84	32.97	91.88	45.44	89.60	24.33	94.22
0.65	1.52	99.58	17.49	95.84	32.95	91.88	45.21	89.61	24.29	94.23
0.60	1.50	99.58	17.47	95.84	33.14	91.89	45.76	89.49	24.47	94.20
0.55	1.48	99.59	17.26	95.87	32.93	91.95	45.48	89.58	24.29	94.25
0.50	1.49	99.58	17.23	95.87	32.93	91.97	46.10	89.45	24.44	94.22
0.45	1.47	99.59	17.40	95.86	33.09	91.94	45.66	89.55	24.40	94.23
0.40	1.48	99.59	17.43	95.86	33.02	91.96	45.89	89.50	24.45	94.22
0.35	1.48	99.59	17.27	95.88	33.02	91.97	45.66	89.54	24.36	94.24
0.30	1.48	99.59	17.42	95.88	33.01	91.99	45.71	89.57	24.40	94.26
0.25	1.49	99.59	17.43	95.87	32.90	92.00	45.73	89.54	24.39	94.25
0.20	1.47	99.59	17.35	95.95	32.79	92.03	45.82	89.50	24.36	94.27
0.15	1.48	99.59	17.20	95.90	32.76	92.04	45.67	89.57	24.28	94.27
0.10	1.48	99.59	17.09	95.93	32.68	92.07	45.71	89.58	24.24	94.29
0.05	1.48	99.59	17.24	95.90	32.76	92.02	44.95	89.72	24.11	94.31
0.00	1.47	99.59	17.15	95.91	33.08	91.91	44.70	89.76	24.10	94.29

Table 6: Prompt engineering : OpenOOD v1.5 Benchmark.

	SSB-hard		NINCO		Near-OOD		iNaturalist		Textures		Openimage-O		Far-OOD	
Prompt	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
a dark photo of a <label>	79.77	72.58	75.05	73.12	77.41	72.85	4.42	99.02	77.46	77.31	37.36	91.03	39.75	89.12
a blurry photo of a <label>	82.97	69.85	77.43	68.98	80.20	69.41	4.25	99.03	68.18	80.79	35.31	91.29	35.91	90.37
a low resolution photo of a <label>	81.66	73.38	74.57	74.17	78.11	73.78	1.89	99.52	72.50	81.47	33.63	92.44	36.01	91.14
a cropped photo of a <label>	73.76	76.95	70.95	76.24	72.35	76.60	1.31	99.64	66.34	83.11	28.69	93.46	32.11	92.07
a good photo of a <label>	76.75	75.80	70.78	78.55	73.77	77.18	1.53	99.60	57.71	86.92	30.33	93.53	29.86	93.35
<label>	82.31	71.94	68.23	78.28	75.27	75.11	1.60	99.58	55.99	88.09	32.69	92.68	30.09	93.45
The Nice <label>	82.78	71.18	68.75	77.21	75.77	74.20	1.61	99.57	41.30	90.95	29.04	93.59	23.98	94.70

<label>" achieves the best performance based on FPR95 in Near-OoD scenarios. This demonstrates that the choice of prompt can significantly affect the results, highlighting the importance of prompt selection in the NegLabel process.

Nevertheless, Table 7 illustrates that NegLabel more effectively utilizes VLM's textual information compared to MCM. These findings suggest that although the NegLabel methodology effectively leverages textual information, further adjustments are required depending on the specific problem conditions being addressed.

As shown in Table 8, we observed that the performance of Near- and Far-OoD scenarios improve as Group Number increases. Unlike MOS, the best performance is shown at a large group number of 450. In Table 9, distance percentile also shows a tendency for performance to improve as it gets closer to 0. Through these extended experiments, we show that when selecting these hyper parameters, which result in different OoD performance depending on the dataset, specific criteria are needed to select them. However, the original paper does not provide any major criteria for selecting the hyperparameters.

5 Discussion

Through this reproduce challenge, we aimed to determine whether NegLabel fully leverages the pre-trained knowledge of Vision-Language Models (VLMs) and to investigate if this method maintains its generalizability across more challenging benchmark datasets, such as OpenOOD v1.5. Our findings indicate that not only does NegLabel possess significant generalizability, but it also highlights the performance differences across datasets based on prompts or hyperparemters. Traditional zero-shot methodologies often overlook the impact of prompts; however, recent advancements in Prompt Learning, which focus on parameter-efficient fine-tuning by solely using prompts, are gaining traction. Ultimately, while using negative text shows potential for fully leveraging VLMs' knowledge, our results suggest that there is still room for improvement.

Table 7: This table presents the results of experiments conducted on the OpenOOD v1.5 benchmark dataset. The terms Near-OoD and Far-OoD refer to the average OoD detection performance across benchmark datasets within each respective scenario.

	SSB-hard		NINCO		Near-OOD		iNaturalist		Textures		Openimage-O		Far-OOD	
Methods	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MCM	89.45	64.11	82.70	69.82	86.08	66.97	61.94	87.62	54.26	87.71	53.80	88.60	56.67	87.98
NegLabel	82.78	71.18	68.75	77.21	75.77	74.20	1.61	99.57	41.30	90.95	29.04	93.59	23.98	94.70

Table 8: Impact of the group numbers in the grouping strategy on OoD detection : OpenOOD v1.5 benchmark.

Group Number	SSB-hard		NINCO		Near-OOD		iNaturalist		Textures		Openimage-O		Far-OOD	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
1	81.87	71.32	69.82	77.09	75.85	74.21	2.32	99.36	44.98	90.56	31.10	93.10	26.14	94.34
50	82.67	71.16	69.09	77.04	75.88	74.10	1.49	99.58	42.87	90.69	29.99	93.47	24.78	94.58
100	82.83	71.17	68.97	77.09	75.90	74.13	1.55	99.58	42.23	90.80	29.66	93.52	24.48	94.63
150	82.83	71.18	68.92	77.12	75.88	74.15	1.56	99.57	41.88	90.85	29.37	93.55	24.27	94.66
200	82.82	71.17	68.89	77.13	75.85	74.15	1.57	99.57	41.80	90.89	29.25	93.56	24.21	94.68
250	82.86	71.18	68.89	77.13	75.87	74.16	1.59	99.57	41.69	90.92	29.16	93.57	24.15	94.69
300	82.91	71.17	68.84	77.20	75.87	74.19	1.61	99.57	41.67	90.91	29.18	93.58	24.15	94.69
350	82.84	71.18	68.82	77.22	75.83	74.20	1.60	99.57	41.53	90.92	29.09	93.58	24.08	94.69
400	82.84	71.18	68.84	77.22	75.84	74.20	1.61	99.57	41.38	90.96	29.06	93.58	24.01	94.70
450	82.78	71.18	68.75	77.21	75.77	74.20	1.61	99.57	41.30	90.95	29.04	93.59	23.98	94.70
500	82.83	71.17	68.74	77.21	75.78	74.19	1.61	99.56	41.40	90.96	29.06	93.59	24.02	94.71
1000	82.95	71.16	68.70	77.27	75.83	74.21	1.64	99.56	41.28	90.99	29.08	93.59	24.00	94.71

Table 9: The choice of distances for the selection of negative labels : OpenOOD v1.5 benchmark.

Percentile	SSB-hard		NINCO		Near-OOD		iNaturalist		Textures		Openimage-O		Far-OOD	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
0.95	84.07	68.56	69.40	75.96	76.74	72.26	2.15	99.45	42.83	90.64	29.55	93.64	24.84	94.58
0.90	84.18	68.32	69.48	75.54	76.83	71.93	1.90	99.48	41.47	90.92	29.45	93.66	24.27	94.69
0.85	84.10	68.98	69.47	75.85	76.79	72.42	1.78	99.50	41.12	91.08	29.61	93.62	24.17	94.73
0.80	83.68	69.55	69.33	75.92	76.51	72.74	1.75	99.52	41.84	90.81	29.37	93.65	24.32	94.66
0.75	83.40	70.07	69.16	76.37	76.28	73.22	1.70	99.53	41.49	90.83	29.41	93.65	24.20	94.67
0.70	83.30	70.23	69.09	76.62	76.20	73.43	1.68	99.55	41.63	90.88	29.21	93.64	24.17	94.69
0.65	83.20	70.37	69.14	76.85	76.17	73.61	1.63	99.55	41.67	90.86	29.20	93.63	24.17	94.68
0.60	83.01	70.67	69.03	76.89	76.02	73.78	1.63	99.56	41.82	90.77	29.04	93.63	24.17	94.65
0.55	83.00	70.79	68.92	76.96	75.96	73.87	1.62	99.56	41.53	90.84	29.06	93.63	24.07	94.68
0.50	83.03	70.81	68.86	77.11	75.94	73.96	1.63	99.56	41.76	90.77	29.14	93.63	24.18	94.65
0.45	82.97	70.89	68.91	77.13	75.94	74.01	1.62	99.56	41.74	90.81	29.01	93.63	24.13	94.67
0.40	82.92	71.00	68.92	77.18	75.92	74.09	1.63	99.56	42.09	90.79	29.12	93.62	24.28	94.66
0.35	82.96	70.99	68.84	77.13	75.90	74.06	1.61	99.56	42.02	90.79	29.02	93.62	24.21	94.65
0.30	82.90	71.04	68.86	77.16	75.88	74.10	1.62	99.56	41.84	90.82	29.06	93.62	24.17	94.67
0.25	82.83	71.11	68.80	77.20	75.82	74.16	1.63	99.56	42.02	90.78	29.06	93.61	24.24	94.65
0.20	82.87	71.07	68.91	77.19	75.89	74.13	1.63	99.56	41.94	90.76	29.12	93.59	24.23	94.64
0.15	82.90	71.15	69.04	77.20	75.97	74.18	1.63	99.56	41.86	90.82	29.11	93.58	24.20	94.65
0.10	82.88	71.15	68.87	77.18	75.87	74.16	1.63	99.56	41.80	90.85	29.15	93.57	24.19	94.66
0.05	82.78	71.18	68.75	77.21	75.77	74.20	1.61	99.57	41.30	90.95	29.04	93.59	23.98	94.70
0.00	83.00	70.82	68.82	77.25	75.91	74.04	1.61	99.57	40.97	91.03	29.01	93.67	23.86	94.75

6 Conclusion

In this report, we examined the generalizability of [7] in leveraging the capabilities of Vision-Language Models (VLMs) knowledge through replicating experiments and conducting extended benchmark dataset experiments. Through these experiments, we observed the performance variations of OoD detection based on changes in prompt engineering, distance percentile, and group strategy for different datasets within NegLabel. While in most cases, we confirmed that NegLabel can fully leverage the capabilities of VLMs, we also noted performance variations due to changes in prompts, group numbers, and percentiles in the the extended experiment OpenOOD v1.5 benchmark datasets.

These findings suggest that using Negative text for OoD detection generally yields good performance; however, this can be suboptimal. Future research should focus on these detailed aspects to further advance VLMs OoD detection. Additionally, by making our code publicly available on our GitHub repository, we demonstrate the reproducibility and validity of our experiments.

References

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

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

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

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

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

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

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

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

[10] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[11] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations*, 2021.

[12] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[13] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, et al. Openood v1. 5: Enhanced benchmark for out-of-distribution detection. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*, 2023.

[14] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.