Harnessing Large Language and Vision-Language Models for Robust Out-of-Distribution Detection
Appendix

Pei-Kang Lee R11922116@CSIE.NTU.EDU.TW
Jun-Cheng Chen PULLPULL@CITI.SINICA.EDU.TW
Ja-Ling Wu WJL@CMLAB.CSIE.NTU.EDU.TW

Editors: Hung-yi Lee and Tongliang Liu

Appendix A. More Related Work

Besides CLIP-based OOD detection, we also include a brief review of relevant research works of traditional OOD detection for a comprehensive survey as follows. **Tradtional OOD Detection** Traditional OOD detection primarily utilizes CNN-based or Vision Transformer (ViT)-based approaches Dosovitskiy et al. (2021) for detection by leveraging visual features. Some methods attempt to enhance the distinction between ID and OOD samples by designing proper scoring functions. MSP Hendrycks and Gimpel (2017) uses the maximum softmax score as a criterion, positing that the model exhibits higher confidence levels for ID samples. ODIN Liang et al. (2018) further enhances the separation between ID and OOD scores through temperature scaling and input preprocessing.

However, the study Nguyen et al. (2015) indicates that discriminative models can exhibit overconfidence. Instead, the energy score Liu et al. (2020), theoretically aligns with the probability density function, seeking to mitigate the issue of over-confidence. The Mahalanobis score Lee et al. (2018) approaches the problem from the feature space, assuming that the class-conditional distribution follows a multivariate Gaussian distribution and uses the Mahalanobis distance as the scoring function. RMD Ren et al. (2021) further improves the Mahalanobis distance by removing the influence of background statistics, enhancing the detection performance of Near-OOD samples. ViM Wang et al. (2022) argues that considering the probability space, logits space, or feature space individually has limitations and proposes a virtual logit to integrate perspectives from all three. Beyond designing different scoring functions, some studies have sought to improve OOD detection by examining the model's behavior. ReAct Sun et al. (2021) identifies differences of the activation patterns between OOD data and ID data and further improves performance by implementing clipping methods. Meanwhile, LogitNorm Wei et al. (2022) addresses the problem of over-confidence by enforcing a constant vector norm on the logits in training.

Appendix B. More Implementation Details

B.1. Training Details

B.1.1. Negative Mining.

We use the NLTK library to extract nouns and adjectives from the WordNet 3.1 Fellbaum (1998) corpus dataset. To mitigate semantic overlap, we select only the first word for each lexname. Following

Table 1: Examples of LLM inputs, prompts, and outputs for superclass labels and background descriptions generation. The example LLM model is Claude 3.5 Sonnet with temperature setting to zero.

Target Output	**LLM Input**	**LLM Prompt**	**Example Output**

Superclass labels	Class labels	Output immediate superclass of {class label} in its classification hierarchy. Provide only the lower case superclass name, no explanation.	`{`
			`"tench": "cyprinidae",`
			`"goldfish": "cyprinidae",`
			`"absorbent paper": "paper towel",`
			`"bed": "cradle",`
			`}`

Background descriptions	Superclass labels	Provide description of a {superclass label}: 10 precise, scenic-setting details, 3-word phrases. Each phrase should contain environment and related object. Avoid using adjectives, provide concrete descriptions. Separate with commas. Single line output.	`{`
			`"absorbent paper":`
			`"Paper soaks spill,`
			`napkin catches drip,`
			`towel dries hands,`
			`blotter absorbs ink,`
			`tissue wipes nose,`
			`filter traps sediment,`
			`coaster protects table,`
			`sponge cleans counter,`
			`pad collects grease,`
			`liner prevents leaks",`
			
			`"bed":`
			`"Sheets whisper breeze,`
			`pillows catch moonlight,`
			`headboard frames window,`
			`nightstand holds lamp,`
			`rug absorbs footsteps,`
			`curtains filter sunlight,`
			`blanket drapes floor,`
			`clock ticks softly,`
			`books stack nearby,`
			`mirror reflects bed",`
			`}`

NegLabel Jiang et al. (2024), we designate the lowest 15% similarity scores (0.95-quantile) in nouns and adjectives as negative labels. We follow the same prompt template setting as in NegLabel.

B.1.2. PROMPT TUNING.

Prompt tuning Zhou et al. (2022) with learnable vector length of 16 is performed for 200 epochs, with few-shot training image batch size 256 and OOD image batch size 512, using a learning rate of 0.025 (SGD optimizer).

B.1.3. VISUAL PROMPT TUNING (VPT).

VPT Jia et al. (2022) with learnable vector length of 12 is trained for 5 epochs, with few-shot training image batch size 32 and OOD image batch size 64, using a learning rate of 0.2 (SGD optimizer).

B.2. Prompts For Superclass Labels and Background Descriptions

Table 1 presents our prompt design. For each class label, we employ LLMs to identify corresponding superclass labels. Subsequently, after obtaining the set of superclass labels, we utilize LLMs to generate relevant background descriptions. Based on the current price-per-token for Claude 3.5 Sonnet, the generation of superclass labels and background descriptions for ImageNet-1K would incur an approximate cost of 1.2 US dollars.

Table 2: Comparison of the proposed zero-shot OOD detection method and NegLabel Jiang et al. (2024) on robustness against covariate shift. The best result in each column is in bold. The values for Superclass-BG are derived from the average of three independent generations using identical prompts. All methods are based on CLIP-B/16, which employs a ViT-B/16 as the image encoder and a masked self-attention Transformer as the text encoder. All values are measured in percentages. The shaded part represents our method.

ID Dataset	Method	iNaturalist		SUN		Places		Texture		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
ImageNet-Sketch	NegLabel	1.89	99.41	22.13	94.98	38.35	90.87	49.49	**88.68**	27.96	93.49
	Superclass-BG	**1.34**	**99.58**	**15.51**	**96.04**	**30.12**	**92.68**	**49.08**	87.86	**24.01**	**94.04**
ImageNet-A	NegLabel	**3.23**	**98.97**	36.85	**90.69**	**54.21**	**83.96**	**66.28**	**79.88**	**40.14**	**88.38**
	Superclass-BG	3.32	98.89	**36.55**	90.08	55.46	82.55	71.81	72.52	41.78	86.03
ImageNet-R	NegLabel	1.44	99.63	14.84	96.10	28.20	92.09	37.91	**90.01**	20.60	94.46
	Superclass-BG	**0.82**	**99.79**	**10.40**	**97.09**	**21.60**	**94.06**	**37.11**	89.48	**17.48**	**95.11**
ImageNet-V2	NegLabel	2.05	99.48	23.64	94.77	40.34	90.41	51.93	**87.99**	29.49	93.17
	Superclass-BG	**1.50**	**99.65**	**17.16**	**95.98**	**32.76**	**92.41**	**51.74**	87.36	**25.79**	**93.85**

Appendix C. Further Experiments and Discussions

C.1. Datasets and Benchmarks.

We utilize the ImageNet-1K OOD benchmark Huang et al. (2021) to compare our method with existing zero-shot and few-shot training-based OOD detection approaches. The ImageNet-1K OOD benchmark employs ImageNet-1K Deng et al. (2009) as the ID dataset and uses iNaturalist Van Horn et al. (2018), SUN Xiao et al. (2010), Places Zhou et al. (2018), and Texture Cimpoi et al. (2014) as OOD datasets, which have no class overlap with the ID dataset. Following the settings of the NegLabel approach, we utilize ImageNet-A Hendrycks et al. (2021b), ImageNet-Sketch Wang et al. (2019), ImageNet-R Hendrycks et al. (2021a), and ImageNet-V2 Recht et al. (2019) to evaluate

Table 3: Comparison of the proposed zero-shot OOD detection method and NegLabel Jiang et al. (2024) on the fine-grained dataset.

ID Dataset	Method	iNaturalist		SUN		Places		Texture		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CUB-200	NegLabel	**0.12**	**99.98**	**0.02**	**100.00**	**0.27**	**99.92**	**0.00**	**100.00**	**0.10**	**99.97**
	Superclass-BG	0.14	99.97	0.02	99.99	0.28	99.91	0.00	99.99	0.11	99.96
Oxford-Pet	NegLabel	**0.00**	**100.00**	0.02	**100.00**	0.16	**99.96**	**0.12**	**99.97**	0.08	**99.98**
	Superclass-BG	**0.00**	**100.00**	**0.00**	**100.00**	**0.15**	**99.96**	0.14	99.96	**0.07**	**99.98**
Stanford-Cars	NegLabel	**0.00**	**100.00**	**0.01**	**100.00**	**0.02**	**99.99**	**0.00**	**100.00**	**0.01**	**100.00**
	Superclass-BG	**0.00**	**100.00**	**0.01**	**100.00**	0.03	**99.99**	**0.00**	**100.00**	**0.01**	**100.00**
Food-101	NegLabel	**0.00**	**100.00**	**0.00**	**100.00**	0.01	**100.00**	1.67	99.61	0.42	99.90
	Superclass-BG	**0.00**	**100.00**	**0.00**	**100.00**	**0.00**	**100.00**	**1.38**	**99.75**	**0.35**	**99.94**

Table 4: Comparison of the proposed zero-shot and few-shot OOD detection methods with MCM Ming et al. (2022) and NegLabel Jiang et al. (2024) on the ImageNet-100 dataset.

Method	iNaturalist		SUN		Places		Texture		Average	
	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
zero-shot										
MCM	14.93	97.27	31.64	95.29	30.58	95.10	38.32	93.15	28.86	95.20
NegLabel	0.49	99.87	8.43	98.20	18.54	96.39	23.28	95.63	12.69	97.52
gray!20 Superclass-BG	0.19	99.92	7.22	98.31	14.52	97.00	23.55	95.38	11.37	97.65
16-shot										
gray!20 Ours (Train)	**0.14**	**99.94**	**5.40**	**98.64**	**9.52**	**97.88**	**4.72**	**98.85**	**4.94**	**98.82**

Table 5: Performance of few-shot training for our proposed method across various datasets under different numbers of training samples. The ID dataset is ImageNet-1K.

# of Samples	iNaturalist		SUN		Places		Texture		SSB-hard		NINCO	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
1	0.89	99.77	15.16	96.43	30.81	92.74	50.87	88.91	78.61	76.61	63.10	83.84
4	0.65	99.83	12.07	96.91	23.21	94.47	28.10	93.80	58.82	84.94	54.01	87.16
8	**0.50**	**99.86**	**11.64**	**97.18**	21.82	94.88	18.67	95.70	47.19	88.92	49.83	88.39
16	0.68	99.81	13.15	96.94	**16.03**	**96.55**	**11.74**	**97.38**	**27.97**	**94.24**	**45.97**	**89.73**

the robustness of our approach against covariate shift. In addition, we also evaluate our proposed zero-shot OOD detection method on fine-grained datasets, including CUB-200 Wah et al. (2011), Food-101 Bossard et al. (2014), Stanford-Cars Krause et al. (2013) and Oxford-Pets Parkhi et al. (2012).

Beyond the aforementioned datasets, we further evaluate the proposed method on OpenOOD V1.5 ImageNet-1K benchmark Zhang et al. (2023); Vaze et al. (2022). These benchmarks introduce more challenging Near-OOD datasets, such as SSB-hard Vaze et al. (2022) and NINCO Bitterwolf et al. (2023), and robustness evaluations against covariate shifts.

C.2. Robustness Against Covariate Shift

In Table 2, superclass-BG method demonstrates significant improvements in both FPR95 and AUROC metrics across all ID datasets, with the exception of ImageNet-A. It is crucial to note that the

Table 6: The impact of varying α on zero-shot OOD detection performance using the Superclass-BG$_\alpha$. The ID dataset is ImageNet-1K.

α	iNaturalist		SUN		Places		Texture		SSB-hard		NINCO	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
0	1.27	99.68	20.70	94.86	35.74	91.60	47.66	88.27	74.82	78.30	57.98	83.94
0.1	1.25	99.69	20.54	94.90	35.55	91.65	47.62	88.35	**74.58**	**78.39**	57.94	83.93
0.2	1.27	99.69	20.23	95.01	35.23	91.73	47.09	88.42	74.70	78.27	57.59	84.08
0.3	1.25	99.70	19.82	95.10	34.68	91.80	46.56	88.57	74.97	78.22	57.60	84.15
0.4	1.23	99.70	19.32	95.26	34.03	92.00	46.42	88.56	74.92	78.31	57.45	84.21
0.5	1.21	99.70	18.89	95.42	33.37	92.15	46.17	88.59	75.19	78.25	57.28	84.32
0.6	1.19	**99.71**	17.98	95.60	32.36	92.43	46.37	88.58	75.64	78.21	57.59	84.29
0.7	1.15	**99.71**	17.57	95.73	31.55	92.59	45.32	88.81	75.90	78.15	57.47	84.43
0.8	**1.11**	**99.71**	16.50	96.02	30.14	92.91	44.65	89.12	76.42	78.04	57.47	84.43
0.9	1.12	**99.71**	14.74	96.34	28.04	93.29	44.18	89.30	76.52	77.92	**57.23**	**84.59**
1.0	1.15	**99.71**	12.91	96.72	26.26	93.76	44.08	89.50	76.88	77.79	57.37	84.52
1.1	1.23	99.69	11.32	97.10	24.45	94.26	44.13	89.73	78.72	77.08	58.44	83.80
1.2	1.51	99.63	**10.66**	**97.25**	**23.38**	**94.38**	**43.69**	**90.04**	81.52	74.60	59.88	82.98

Table 7: The impact of varying β on the performance of enhanced positive embeddings, which are applied after the phase 1 prompt tuning. The ID dataset is ImageNet-1K.

β	Near-OOD						Far-OOD							
	SSB-hard		NINCO		Average		iNaturalist		Texture		OpenImage-O		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
OpenOOD V1.5 ImageNet-1K														
0	53.10	89.59	58.72	86.51	55.91	88.05	0.46	99.79	15.36	96.89	20.06	96.66	11.96	97.78
0.1	52.00	**89.60**	57.44	86.98	54.72	88.29	0.39	99.82	**15.01**	96.97	19.19	96.73	11.53	97.84
0.2	51.34	89.50	55.98	87.36	53.66	88.43	0.32	99.84	15.22	**96.99**	**18.77**	**96.76**	**11.44**	**97.87**
0.3	50.64	89.25	54.52	87.63	52.58	**88.44**	0.28	99.86	15.39	96.93	18.85	96.73	11.51	97.84
0.4	**49.93**	88.81	52.28	**87.78**	51.11	88.30	**0.25**	**99.87**	16.02	96.78	19.56	96.63	11.95	97.76
0.5	49.95	88.12	51.83	87.77	50.89	87.94	**0.25**	**99.87**	18.28	96.48	20.80	96.44	13.11	97.60
0.6	50.93	87.09	**50.81**	87.57	**50.87**	87.33	0.26	99.86	21.08	96.01	22.51	96.12	14.62	97.33
0.7	52.32	85.64	50.94	87.17	51.63	86.40	0.31	99.85	25.24	95.27	25.02	95.63	16.86	96.92
0.8	55.58	83.66	51.54	86.52	53.56	85.09	0.38	99.82	30.32	94.17	28.75	94.93	19.82	96.31
0.9	59.70	81.06	53.44	85.63	56.57	83.35	0.52	99.78	37.02	92.55	33.46	93.92	23.67	95.42
OpenOOD V1.5 ImageNet-1K Full-Spectrum														
0	67.43	80.43	71.92	76.09	69.67	78.26	5.63	98.47	31.12	90.93	36.60	92.12	24.45	93.84
0.1	66.39	80.60	70.70	76.96	68.54	78.78	4.79	98.74	30.29	91.28	35.12	92.42	23.40	94.15
0.2	65.62	80.72	69.37	77.82	67.50	79.27	3.96	98.98	29.96	91.60	34.11	92.69	22.68	94.42
0.3	64.63	**80.75**	67.87	78.64	66.25	79.69	3.26	99.18	29.45	91.85	33.52	92.92	22.08	94.65
0.4	63.61	80.65	65.55	79.43	64.58	80.04	2.57	99.35	**29.34**	92.03	**33.44**	93.09	**21.79**	94.82
0.5	62.98	80.39	64.54	80.14	63.76	80.27	2.01	99.48	30.88	**92.10**	33.78	**93.18**	22.22	**94.92**
0.6	**62.89**	79.90	62.78	80.77	62.83	**80.33**	1.54	99.57	32.82	92.01	34.37	93.16	22.91	**94.92**
0.7	63.14	79.11	61.83	81.27	**62.48**	80.19	1.19	99.64	35.96	91.70	35.74	93.01	24.30	94.78
0.8	64.78	77.95	**61.18**	81.63	62.98	79.79	0.95	99.67	39.92	91.06	38.23	92.67	26.37	94.47
0.9	67.15	76.33	61.41	**81.81**	64.28	79.07	**0.80**	**99.68**	44.96	89.98	41.46	92.07	29.07	93.91

superclass-BG method considers candidate labels without filtering. Under comparable conditions, the NegLabel method (i.e., with filtering) achieves an average FPR95 of 47.45% and an AUROC of 83.88% on the ImageNet OOD benchmark. These results indicate that the superclass-BG approach not only help maintain robustness against covariate shift but also exhibits enhanced effectiveness in this regard.

Table 8: Performance comparison of the proposed Superlcass-BG zero-shot OOD detection method with different VLM architectures. The ID dataset is ImageNet-1K.

Backbone	iNaturalist		SUN		Places		Texture		SSB-hard		NINCO	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CLIP												
RN50	1.84	99.53	18.95	95.47	34.30	91.99	46.99	88.54	78.66	77.47	70.76	79.86
RN101	1.86	99.55	20.85	95.07	39.77	90.42	48.17	87.70	80.69	75.38	61.69	83.64
ViT-B/32	1.99	99.55	15.53	96.44	28.44	93.59	48.87	88.51	79.85	75.69	59.88	82.74
ViT-B/16	1.15	99.71	12.91	96.72	26.26	93.76	44.08	**89.50**	76.88	77.79	57.37	84.52
ViT-L/14	**1.09**	**99.72**	**13.80**	**96.88**	**22.48**	**94.93**	**42.59**	89.12	**72.46**	**79.82**	**54.46**	**86.21**
ALIGN												
EfficientNet-B7	2.90	99.35	16.33	95.90	29.61	92.64	44.31	88.42	79.55	77.37	69.68	78.57

C.3. Performance Comparison on Fine-Grained Dataset

In Table 3, we present a comparative analysis of four fine-grained datasets. It is noteworthy that Oxford-Pet, Food-101, and CUB-200 encompass class labels pertaining to animal and food categories. Under unfiltered conditions, our experimental results demonstrate that the proposed Superclass-BG method exhibits competitive performance over NegLabel. Moreover, when utilizing Food-101 as the ID dataset, we observe an improvement in performance for our approach with respect to NegLabel.

C.4. Performance Comparison on ImageNet-100

We also conduct experiments on the smaller ImageNet-100 dataset. For fair comparisons, we use the same training subset of ImageNet-1K classes as the ID dataset used in MCM while removing 16 images randomly per class for both zero-shot and few-shot OOD evaluations. In addition, the 16 removed images are served as the few-shot training dataset. As observed in Table 4, the Superclass-BG method demonstrates notable improvements across all datasets except for the Texture dataset, where it performs slightly inferior to NegLabel. Post-training results exhibit superior performance across all datasets.

C.5. Comparison of Different Numbers of Training Samples in Few-Shot learning

In Table 5, we present a comparative analysis of the impact of varying quantities of training samples. The results reveal that for the iNaturalist and SUN datasets, an increase in training samples does not significantly affect overall performance. Conversely, for the Places, Texture, SSB-hard, and NINCO datasets, we observe a consistent improvement in performance correlating with an increase in training data. Notably, the 4-shot training results surpass the current state-of-the-art performance of both CNN-based and CLIP-based models on the ImageNet OOD benchmark. This outcome underscores the efficacy of our proposed methodology.

C.6. The Influence of Removing the Background Feature

Let SC denote the superclass labels generated by LLMs, and $BG(SC)$ represent the background descriptions corresponding to each superclass label, also generated by LLMs. Our proposed Superclass-BG method isolates the core semantic features of the ID label space by removing background features from the superclass features. In this section, we examine the impact of varying the proportion

of background feature removal on zero-shot OOD detection performance. Let $\mathcal{T}$ be the text encoder of CLIP, we define the refined representation as:

$$\text{Superclass-BG}_\alpha = \mathcal{T}(SC) - \alpha \cdot \mathcal{T}(BG(SC)). \tag{1}$$

Analysis of the results presented in Table 6 reveals a notable trend: as the value of α increases, corresponding to a greater proportion of background feature removal, performance on the ImageNet OOD benchmark improves significantly. Conversely, a slight decrease in performance is observed on the SSB-hard dataset, while the NINCO dataset shows a marginal improvement. In the context of the relatively more distinguishable ImageNet OOD benchmark, the removal of background features allows for a more focused representation characterizing the core semantics within the ID label space. This refinement enhances the model's ability to differentiate OOD samples that are inherently distant in semantics from the ID label space. However, this process may inadvertently eliminate certain features crucial for identifying more challenging Near-OOD samples, explaining the slight performance degradation observed in the SSB-hard dataset. The NINCO dataset, specifically designed to ensure no overlap in background or object features with any class in ImageNet-1K, necessitates a more comprehensive understanding of the ID label space. We posit that the capacity of the Superclass-BG method to produce a more refined representation of the ID label space contributes to the observed improvement in performance on this particular OOD dataset. Based on the empirical evidence, setting $\alpha = 1$ yields a balanced performance across both the ImageNet OOD and Near-OOD datasets. Consequently, we have opted to standardize the value of α at 1 for our experiments.

C.7. Pilot Study of Candidate Label Filtering

Figures 1 and 2 illustrate the implications of deliberately excluding labels belonging to the *animal* and *food* categories from the candidate labels to avoid conflicts with the ImageNet-1K major categories. This exclusion reveals an inherent entanglement between two objectives: ensuring greater discriminative power between major categories and negative labels, and accurately identifying Near-OOD samples (primary instances from the animal and the food categories that are visually similar to ID samples). In the absence of semantically proximate negative labels, Near-OOD samples are more likely to exhibit affinity towards visually similar ID labels, potentially leading to misclassification. Conversely, the inclusion of the animal and food category negative labels may adversely affect ID sample classification, as evidenced in Figure 3. However, the fundamental principle of negative-mining involves identifying negative labels that are semantically distant from the ID label space to enhance ID-OOD discrimination. Consequently, the selected negative labels are inherently those with greater semantic distance from ID labels within the candidate label set. We posit that the impact on ID samples is relatively minor compared to the substantial improvement in Near-OOD performance. Experimental results support this hypothesis. In the challenging ImageNet OOD benchmark, while unfiltered results of our approach show a marginal decrease in performance, there is a marked improvement in Near-OOD performance. Moreover, our proposed Superclass-BG method demonstrates the capacity to enhance performance across both scenarios simultaneously by refining the ID label space.

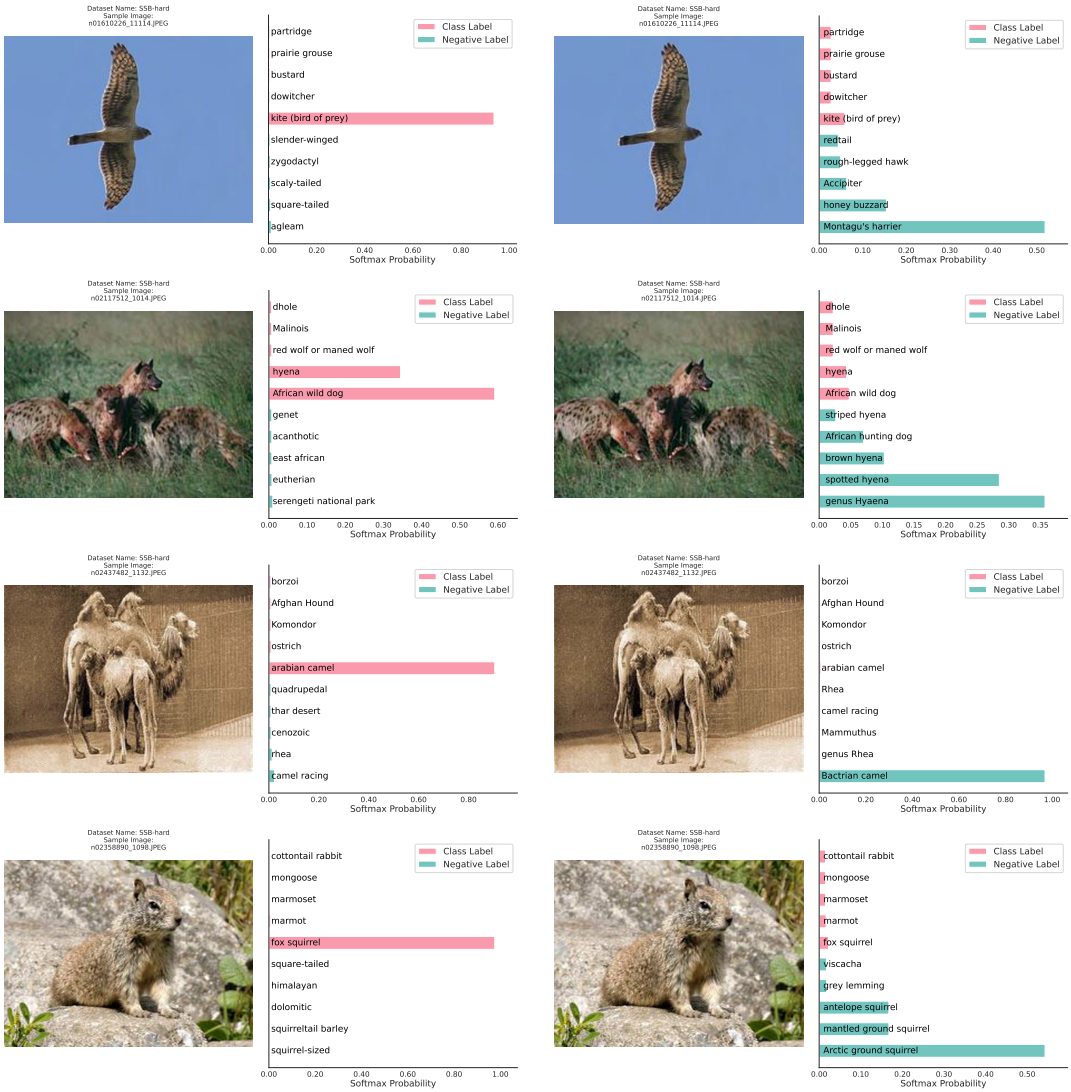

Figure 1: Comparative analysis of the impact of candidate label filtering. Each row represents an OOD image erroneously classified as ID under filtering conditions. The left column displays results with filtering applied (removal of animal and food categories), while the right column shows results without filtering. We examine the top 5 class labels with the highest cosine similarity to the image embeddings, as well as the top 5 negative labels with the highest cosine similarity to the image embeddings. These negative labels are obtained through negative mining from candidate labels. Our observations indicate that for the SSB-hard dataset, a challenging Near-OOD dataset more closely aligned with ImageNet-1K, the inclusion of animal and food candidate labels proves beneficial for these difficult OOD samples. The selected negative labels demonstrate increased affinity between the OOD images and these negative labels.

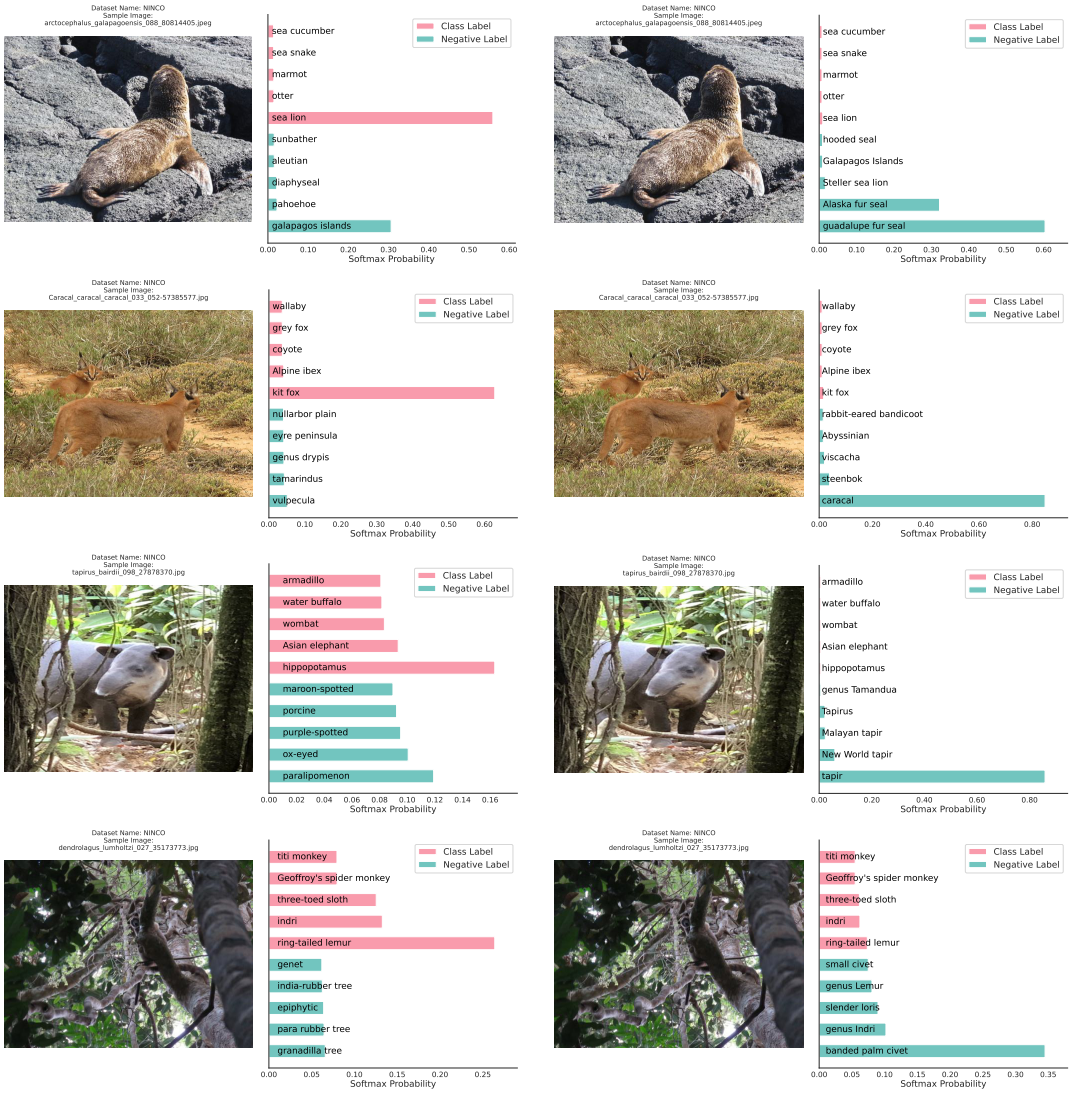

Figure 2: Comparative analysis of the impact of candidate label filtering. Each row represents an OOD image erroneously classified as ID under filtering conditions. The left column displays results with filtering applied (removal of animal and food categories), while the right column shows results without filtering. We examine the top 5 class labels with the highest cosine similarity to the image embeddings, as well as the top 5 negative labels with the highest cosine similarity to the image embeddings. These negative labels are obtained through negative mining from candidate labels. Our observations indicate that for the NINCO dataset, a challenging Near-OOD dataset more closely aligned with ImageNet-1K, the inclusion of animal and food candidate labels proves beneficial for these difficult OOD samples. The selected negative labels demonstrate increased affinity between the OOD images and these negative labels.

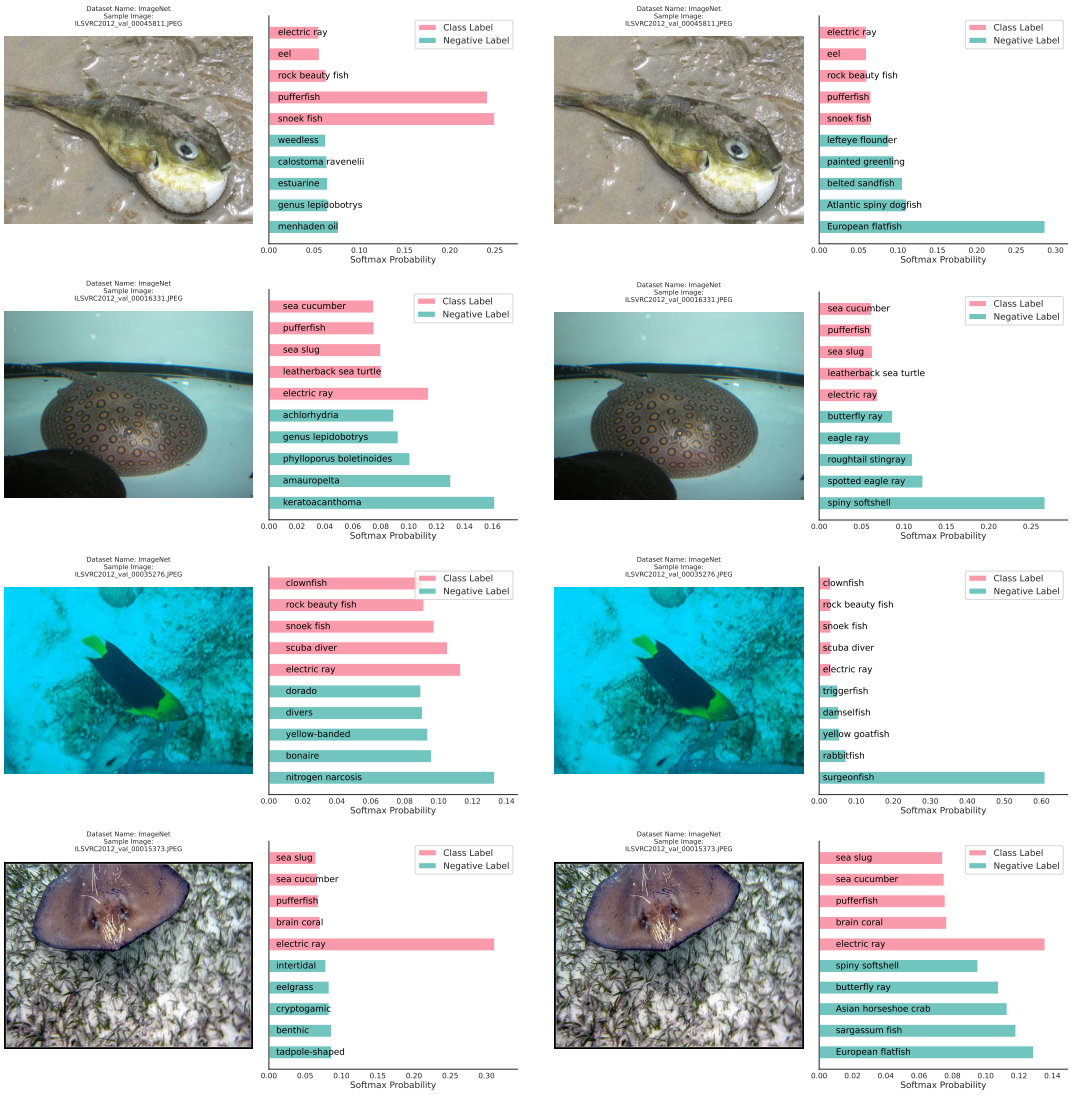

Figure 3: Comparative analysis of the impact of candidate label filtering. Each row represents an ID image erroneously classified as OOD under conditions without label filtering. The left column displays results with filtering applied (removal of animal and food categories), while the right column shows results without filtering. We examine the top 5 class labels with the highest cosine similarity to the image embeddings, as well as the top 5 negative labels with the highest cosine similarity to the image embeddings. These negative labels are obtained through negative mining from candidate labels. The experimental results reveal that, although the essence of negative-mining lies in selecting labels with lower similarity to the ID label space, these candidate labels, which belong to the majority category of ID labels, may exert an adverse influence on ID images. Consequently, ID images pertaining to the animals category might exhibit an increased affinity towards negative labels within the animal category.

Table 9: Full experiment results of our proposed zero-shot and few-shot OOD detection method on the OpenOOD V1.5 benchmark.

Method	Near-OOD				Far-OOD					
	SSB-hard		NINCO		iNaturalist		Texture		OpenImage-O	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
ImageNet-1K Benchmark										
Ours (Superclass-BG)	64.93	77.74	56.27	84.48	0.75	99.71	44.64	90.25	39.55	92.52
Ours (Train)	26.59	94.21	42.22	89.69	0.51	99.81	9.41	98.04	9.68	98.13
ImageNet-1K Full Spectrum Benchmark										
Ours (Superclass-BG)	70.52	74.17	62.51	81.80	0.85	99.65	50.98	88.32	45.77	91.14
Ours (Train)	43.80	86.22	58.89	79.85	4.71	98.88	23.48	93.46	23.88	94.85

Table 10: Performance comparison of proposed Superclass-BG zero-shot OOD detection method on different LLMs across multiple OOD datasets. The ID dataset is ImageNet-1K.

LLM	iNaturalist		SUN		Places		Texture		SSB-hard		NINCO	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Run 1												
Claude 3.5 Sonnet	1.16	99.71	13.73	96.53	26.85	93.60	44.54	89.50	76.61	77.81	57.50	84.30
gpt-4o	1.23	99.70	11.40	97.04	24.48	94.24	51.60	87.39	76.14	78.42	58.42	83.82
gpt-4-turbo	1.16	99.71	13.66	96.59	26.70	93.74	44.89	89.42	75.98	78.09	56.79	84.44
gpt-4o-mini	1.21	99.70	11.24	97.03	24.48	94.16	44.36	89.44	78.50	77.26	58.91	83.54
Run 2												
Claude 3.5 Sonnet	1.15	99.71	12.91	96.72	26.26	93.76	44.08	89.50	76.88	77.79	57.37	84.52
gpt-4o	1.20	99.70	11.39	97.05	24.46	94.24	50.73	87.50	76.78	78.23	59.02	83.69
gpt-4-turbo	1.21	99.71	13.45	96.59	26.61	93.70	43.67	89.84	76.61	77.60	57.72	84.51
gpt-4o-mini	1.21	99.70	11.42	96.99	24.52	94.11	43.23	89.93	78.72	77.05	59.27	83.25
Run 3												
Claude 3.5 Sonnet	1.14	99.71	13.44	96.67	26.76	93.75	44.17	89.72	77.37	77.63	57.64	84.21
gpt-4o	1.20	99.70	11.46	97.01	24.82	94.20	50.92	87.70	76.07	78.57	58.25	83.61
gpt-4-turbo	1.13	99.71	13.63	96.54	26.77	93.67	43.78	89.70	75.65	78.09	56.86	84.23
gpt-4o-mini	1.20	99.70	11.13	97.09	24.14	94.22	44.15	89.54	78.40	77.19	58.63	83.32

C.8. Enhancement of Utilizing Learned Positive Labels with Class Labels

Let β denote the weight of the class label embedding when constructing the enhanced positive embedding. Formally, let C represent the class labels and P denote the positive labels learned through phase 1 prompt tuning. The enhanced positive embedding P' is then defined as:

$$P' = (1 - \beta) \cdot \mathcal{T}(P) + \beta \cdot \mathcal{T}(C) \tag{2}$$

To investigate the impact of different β values on the robustness against covariate shift, the results presented in Table 7 demonstrate the performance of the enhanced positive embedding after phase 1 prompt tuning but without undergoing phase 2 visual prompt tuning. From these findings, we can observe that while smaller β values yield superior performance on the OpenOOD V1.5 ImageNet-1K benchmark, which does not account for covariate shift. This indicates that the positive labels obtained through phase 1 prompt tuning have indeed adapted to the target distribution. However, the opposite trend is observed in the full-spectrum benchmark that considers covariate shift. Upon increasing the weight of the class label embedding, we note a reduction in the model susceptibility to covariate shift, suggesting that the class labels inherently retain CLIP powerful generalization

Table 11: Performance comparisons of various configurations in the proposed methods. S-BG denotes Superclass-BG. The ID dataset is ImageNet-1K.

S-BG	Filter	PT	VPT	iNaturalist		SUN		Places		Texture		SSB-hard		NINCO	
				FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
zero-shot															
×	×			1.32	99.66	21.73	94.51	36.99	91.22	48.46	88.19	75.78	78.00	59.36	83.73
✓	×			1.15	99.71	13.36	96.64	26.62	93.70	44.26	89.57	76.95	77.74	57.50	84.34
×	✓			1.65	99.56	18.84	95.68	34.15	91.93	44.91	89.94	84.00	70.59	69.05	77.25
✓	✓			1.46	99.59	10.94	97.33	23.86	93.97	39.04	91.18	84.94	69.05	69.93	75.99
few-shot															
×	×	✓	×	0.58	99.85	18.79	95.30	34.71	91.91	22.20	94.92	48.31	88.01	50.73	86.93
×	×	×	✓	1.46	99.67	23.47	93.76	34.26	91.35	42.09	89.89	71.42	79.96	57.72	84.87
✓	×	✓	×	0.47	99.87	11.95	97.03	26.87	93.98	19.08	95.64	48.55	88.14	48.88	87.79
×	✓	✓	×	0.71	99.81	14.61	96.62	30.40	93.20	18.58	95.93	58.81	83.80	61.77	82.68
✓	✓	✓	×	0.75	99.79	8.04	98.02	20.00	95.24	16.99	96.30	67.60	79.01	62.35	81.68
×	×	✓	✓	0.71	99.80	21.38	95.12	23.47	94.84	15.14	96.60	31.04	93.43	47.69	89.07
✓	×	✓	✓	0.68	99.81	13.15	96.94	16.03	96.55	11.74	97.38	27.97	94.24	45.97	89.73
×	✓	✓	✓	0.92	99.77	16.42	96.48	20.87	95.53	12.61	97.24	43.29	90.15	58.59	84.71
✓	✓	✓	✓	1.01	99.77	8.12	98.06	16.03	96.20	14.04	97.06	60.82	82.48	60.87	82.51

capabilities. Through the weighted sum of these two components, we aim to maintain robust generalization while simultaneously improving performance on the target distribution. Based on our experimental findings, we have chosen to set β to 0.5, as this value provides the best balanced results, effectively preserving generalization capabilities while enhancing performance on the target distribution.

C.9. Various Vision Language Model Architectures

For a more comprehensive comparison, we also provide the zero-shot OOD detection performance of Superclass-BG with different VLMs, CLIP and ALIGN Jia et al. (2021) models. The experimental results are presented in Table 8.

Appendix D. Full Experimental Results

The greater details of the experimental results of Table 2, Table 4, Figure 3 and Table 5 in the main paper can be found in Table 9, Table 10, Table 11 and Table 12, respectively. Table 9 presents the full experimental results on the OpenOOD V1.5 ImageNet-1K benchmark and full-spectrum benchmark. Table 10 demonstrates the experimental outcomes of Superclass-BG generated by LLMs. Table 11 provides an ablation study on different components of the experimental design. Finally, Table 12 presents the experimental results for various combinations of background description numbers and description lengths.

Table 12: Superclass-BG zero-shot OOD detection performance for various configurations. The ID dataset is ImageNet-1K. *Length of description* denotes the length of a single background description generated by LLMs, while *number of descriptions* indicates the total number of generated descriptions. For all results, we use the same superclass labels generated by the LLM prompt as described in Table 1, with the temperature setting at zero.

Number of	**iNaturalist**		**SUN**		**Places**		**Texture**		**SSB-hard**		**NINCO**	
Description	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Length of Description: 1												
5	1.33	99.68	10.99	97.26	23.11	94.57	51.52	87.83	77.06	77.81	59.19	82.48
6	1.31	99.67	10.13	97.45	21.79	94.90	51.67	87.60	76.59	77.94	60.36	82.15
7	1.30	99.67	10.35	97.41	22.28	94.72	48.88	88.84	77.87	76.19	60.82	81.84
8	1.22	99.70	10.18	97.43	22.28	94.66	49.49	88.65	78.16	76.47	61.18	81.93
9	1.28	99.68	10.59	97.38	22.87	94.54	48.07	89.30	78.71	75.54	61.45	81.72
10	1.39	99.66	10.26	97.46	22.40	94.65	47.29	89.37	78.85	75.71	62.27	81.35
11	1.40	99.66	10.15	97.52	21.94	94.75	47.98	89.28	78.21	75.98	62.69	81.02
12	1.32	99.67	9.67	97.56	21.38	94.92	50.34	88.54	77.41	76.85	61.31	81.15
13	1.40	99.65	10.10	97.54	22.24	94.76	49.04	89.19	78.09	75.47	61.79	80.85
14	1.34	99.67	9.86	97.54	21.57	94.88	49.15	88.88	77.43	76.38	61.57	81.03
15	1.32	99.67	11.18	97.27	23.54	94.45	48.19	88.99	77.62	76.51	61.14	82.10
Length of Description: 2												
5	1.24	99.70	11.14	97.16	23.56	94.40	47.48	89.06	77.93	77.30	58.64	83.57
6	1.15	99.71	11.02	97.21	23.47	94.50	47.57	89.10	78.06	77.07	58.52	82.69
7	1.14	99.71	11.05	97.23	23.67	94.47	48.62	88.78	77.49	77.58	58.85	82.87
8	1.17	99.71	10.88	97.33	23.28	94.56	46.91	89.40	77.86	77.17	59.19	82.28
9	1.18	99.71	10.89	97.28	23.48	94.52	47.94	89.10	78.39	76.96	58.78	83.16
10	1.18	99.71	11.92	97.03	24.86	94.20	46.61	89.33	77.21	77.38	58.32	83.00
11	1.14	99.71	10.55	97.31	22.54	94.69	48.24	88.90	76.84	77.85	58.78	82.90
12	1.17	99.71	11.43	97.21	24.10	94.40	45.80	89.62	77.50	77.35	58.40	82.70
13	1.15	99.72	11.38	97.22	23.68	94.46	48.81	88.65	77.46	77.63	58.29	83.22
14	1.13	99.72	11.07	97.24	23.56	94.50	48.63	88.75	77.25	77.59	58.13	82.94
15	1.13	99.72	11.90	97.03	24.54	94.28	46.56	89.40	77.18	77.72	58.34	83.24
Length of Description: 3												
5	1.22	99.70	13.35	96.54	26.46	93.62	45.90	89.08	76.94	77.73	57.93	84.22
6	1.19	99.71	12.74	96.82	25.71	93.98	47.16	88.49	76.61	78.22	57.62	83.98
7	1.16	99.71	12.63	96.77	25.53	93.91	47.39	88.47	76.90	77.87	57.67	84.31
8	1.13	99.71	13.36	96.69	26.42	93.79	43.89	89.67	77.40	77.58	57.37	84.20
9	1.15	99.71	13.51	96.57	27.05	93.59	45.57	89.12	77.15	77.64	57.08	84.42
10	1.15	99.71	12.91	96.72	26.26	93.76	44.08	89.50	76.88	77.79	57.37	84.52
11	1.15	99.70	12.94	96.72	26.08	93.79	47.02	88.71	76.82	78.01	57.71	84.55
12	1.15	99.71	12.88	96.82	25.52	94.03	46.88	88.69	76.46	78.19	57.23	83.91
13	1.09	99.71	13.33	96.57	26.35	93.66	45.14	88.93	76.66	77.89	57.66	84.31
14	1.10	99.71	13.61	96.62	26.67	93.71	45.43	89.14	76.65	77.91	57.15	84.27
15	1.12	99.71	14.05	96.49	27.38	93.57	44.80	89.40	76.46	77.84	57.13	84.03
Length of Description: 4												
5	1.17	99.71	14.78	96.25	28.72	93.21	45.69	88.82	76.84	77.92	57.77	84.45
6	1.13	99.71	14.36	96.38	28.26	93.38	47.23	88.53	76.46	77.99	57.40	84.46
7	1.16	99.71	14.67	96.35	28.42	93.32	47.71	88.48	76.19	78.05	57.49	84.68
8	1.13	99.71	15.07	96.15	29.08	93.11	46.08	88.70	76.27	77.95	57.55	84.36
9	1.11	99.72	14.49	96.41	28.07	93.39	47.29	88.33	75.92	78.25	57.08	84.25
10	1.13	99.71	15.05	96.20	29.28	93.07	44.82	89.27	76.42	77.85	57.64	84.64
11	1.12	99.71	14.89	96.26	29.02	93.22	44.36	89.24	76.82	77.88	57.88	84.41
12	1.16	99.71	14.11	96.37	27.71	93.39	47.09	88.42	76.09	78.13	57.43	84.50
13	1.12	99.71	14.24	96.43	28.19	93.43	48.05	88.40	76.42	78.22	57.62	84.34
14	1.10	99.72	15.55	96.11	30.10	92.98	46.72	88.45	75.81	78.15	57.28	84.40
15	1.10	99.71	15.49	96.07	29.89	92.91	46.01	88.70	75.56	78.04	57.52	84.76

References

Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? Fixing ImageNet Out-of-Distribution Detection Evaluation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 105–140, 2023.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101-mining discriminative components with random forests. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing Textures in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3606–3613, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021b.

Rui Huang, Andrew Geng, and Yixuan Li. On the Importance of Gradients for Detecting Distributional Shifts in the Wild. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 677–689, 2021.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual Prompt Tuning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 709–727, 2022.

Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative Label Guided OOD Detection with Pretrained Vision-Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR, 2024.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2013.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7167–7177, 2018.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based Out-of-distribution Detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:21464–21475, 2020.

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 (NeurIPS)*, 35:35087–35102, 2022.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015. doi: 10.1109/CVPR.2015.7298640.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 97, pages 5389–5400, 2019.

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

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

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8769–8778, 2018.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-Set Recognition: A Good Closed-Set Classifier is All You Need. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10506–10518, 2019.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-Of-Distribution with Virtual-logit Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *Proceedings of the 39th International Conference on Machine Learning*, pages 23631–23644, 2022.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN Database: Large-Scale Scene Recognition from Abbey to Zoo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492, 2010.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection. *arXiv preprint arXiv:2306.09301*, 2023.

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, 2018.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision (IJCV)*, 130(9):2337–2348, 2022.