A. Proof of Lemma 3.1

Proof.

$$\mathbb{P}\left(f(x,\mathcal{Y})=y_i \mid x\in \text{OOD}\right)=\frac{\mathbb{P}\left(x\in \text{OOD} \mid f(x,\mathcal{Y})=y_i\right)\mathbb{P}(f(x,\mathcal{Y})=y_i)}{\mathbb{P}(x\in \text{OOD})}\propto \mathbb{P}(f(x,\mathcal{Y})=y_i).$$

□

B. Additional Evaluations

B.1. Hard OOD Tasks

We evaluate CLIPScope on hard OOD tasks. The results are shown in Table 3. Our approach shows performance comparable to NegLabel.

Table 3. Comparisons on hard OOD tasks. In each case, ID dataset is shown in the top, whereas OOD dataset is shown in the bottom. N/A represents that the corresponding results are not provided in the original paper.

	CLIPScope		NegLabel	
	AUROC	FPR95	AUROC	FPR95
ImageNet-10 ImageNet-20	98.41	7	98.86	5.1
ImageNet-10 ImageNet-100	98.89	2	99.51	1.68
ImageNet-20 ImageNet-10	98.42	6.8	98.81	4.6
ImageNet-20 ImageNet-100	97.43	11.98	N/A	N/A
ImageNet-100 ImageNet-10	92.11	25.8	90.19	40.2
ImageNet-100 ImageNet-20	89.83	27.6	N/A	N/A

B.2. FPR95 for Tests in Fig. 3

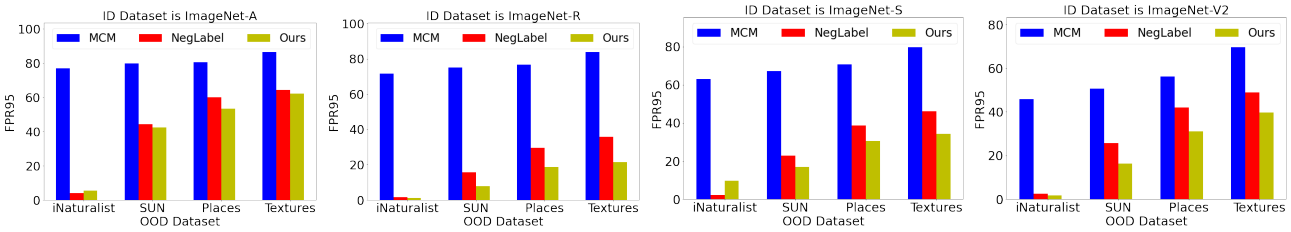

Figure 4. FPR95 (%) on domain-shifted ID datasets. A lower FPR95 implies a better performance.

B.3. Small ID Datasets

We conducted further experiments with smaller ID datasets. The results are shown in Fig. 5. Our approach consistently maintained its effectiveness across these smaller datasets. Specifically, Table 4 presents the FPR95 and AUROC values corresponding to Fig. 5. Our approach maintains consistently high AUROC across all small ID datasets. The lower FPR95 values observed in CUB-200, Oxford-Pet, and Food-101 can be attributed to their focus on fine-grained categories (e.g., specific bird species, pet breeds, or food types). These datasets contain highly specific and detailed visual features within each class, distinguishing them from OOD datasets. For ImageNet-10, ImageNet-20, and ImageNet-100 as ID datasets, the

average FPR95 ranges from 5% to 9%. While still relatively low, this slight performance drop can be attributed to their diverse classes with limited images per class. The results show the ability of our approach to deliver reliable OOD detection performance regardless of the ID dataset size.

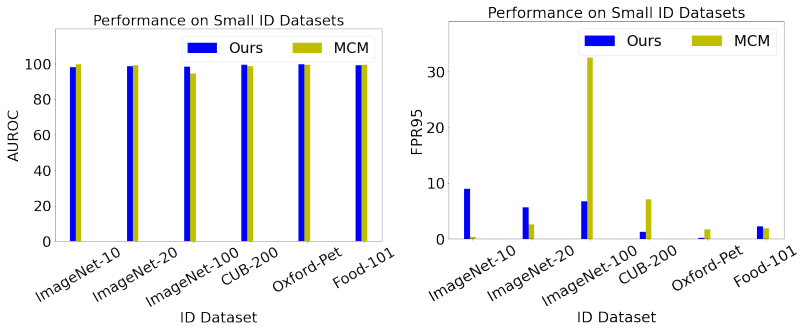

Figure 5. Performance (in %) of CLIPScope when applied to small ID datasets. The OOD datasets include iNaturalist, SUN, Places, and Textures. The reported numbers represent average results across these four OOD datasets.

Table 4. Performance of CLIPScope on each small ID dataset.

OOD Dataset	iNaturalist		SUN		Places		Textures		Average	
Small ID Datasets	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
ImageNet-10	99.85	0.64	97.84	10.63	96.33	18.80	98.62	5.81	98.159	8.965
ImageNet-20	99.90	0.42	98.74	7.24	97.84	10.96	98.69	4.07	98.789	5.671
ImageNet-100	99.67	1.28	98.66	5.65	97.44	10.64	97.55	9.37	98.329	6.734
CUB-200	99.78	0.65	99.68	0.87	99.23	2.56	99.67	1.11	99.589	1.294
Oxford-Pet	99.99	0.02	99.97	0.04	99.88	0.36	99.85	0.33	99.923	0.185
Food-101	99.97	0.11	99.83	0.61	99.63	1.40	97.26	6.91	99.169	2.255

B.4. Robustness Against Mining Parameters M and η

We assessed CLIPScope across various sizes M of OOD label space and percentile distances η. To mitigate the effects of randomness, we employed a reversing order. The findings are detailed in Table 5. M and η exert only a mild influence on our approach since only p_2 utilizes OOD labels.

B.5. Convergence

We conducted a series of experiments by varying the numbers of OOD samples to assess the impact on performance stability. The ratio of ID to OOD samples is maintained at 1:1. The results are shown in the first two subplots of Fig. 6 and Table 6. These numbers indicate that our approach reaches a performance plateau after processing approximately 1200 OOD samples, which is about 12% of the total OOD samples included in our test sets. Similarly, we conducted experiments to explore how varying proportions affect our approach's efficacy. We fix the number of OOD samples at 1600 and vary the number of ID samples. The results are shown in the last two subplots of Fig. 6 and Table 7. Our methodology shows performance plateaus at 2000 samples in this case.

B.6. Performance of CLIPScope with Various Backbones

We also evaluated the performance of CLIPScope using different backbones. Fig.7 shows the results. Compared to NegLabel, CLIPScope consistently exhibits comparable or superior performance across most of the tested models. This diverse set of model evaluations demonstrates that CLIPScope is adaptable to different architectural frameworks.

B.7. Performance on Unbalanced Datasets

Table 8 presents the performance of CLIPScope on subsets of ImageNet, using ImageNet-1K as the ID labels. This setup creates an unbalanced class distribution, with some in-distribution classes having no samples. Our approach shows a

Table 5. Performance (%) of CLIPScope with various M (the top table) and η (the bottom table). The ID dataset is ImageNet-1k.

OOD Dataset	iNaturalist		SUN		Places		Textures		Average	
Metric	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Different Sizes M of OOD Label Space $\mathcal{Y}^-$ (Nearest & Farthest)										
$M = 0$	97.98	8.23	95.79	18.38	91.81	30.68	92.36	31.52	94.488	22.204
$M = 50$	98.41	5.28	96.63	15.48	93.03	27.48	93.55	27.26	95.405	18.875
$M = 100$	98.68	4.87	96.7	15.23	93.19	27.23	93.69	26.93	95.565	18.565
$M = 500$	99.29	2.27	97.06	13.55	93.70	25.57	93.81	28.58	95.965	17.493
$M = 1000$	99.45	1.52	97.12	13.64	93.85	25.43	93.65	30.3	96.018	17.723
$M = 2000$	99.53	1.35	97.25	13.76	94.1	25.54	93.44	32.09	96.080	18.185
$M = 5000$	99.60	1.28	97.34	13.52	94.20	26.32	93.04	34.41	96.045	18.883
$M = 7000$	99.60	1.21	97.41	12.91	94.27	26.14	92.85	35.12	96.033	18.845
$M = 10000$	99.60	1.23	97.47	12.83	94.3	25.69	92.91	34.23	96.070	18.495
Different Percentile Distance η										
$\eta = 0.001$	99.49	1.68	97.10	12.25	95.12	21.12	92.90	31.39	96.153	16.610
$\eta = 0.05$	99.60	1.28	97.34	13.52	94.20	26.32	93.04	34.41	96.045	18.883
$\eta = 0.25$	99.52	1.56	96.92	13.68	94.75	23.06	92.55	32.56	95.935	17.715
$\eta = 0.5$	99.51	1.52	96.94	13.37	94.80	22.93	92.56	32.68	95.953	17.625
$\eta = 0.75$	99.58	1.3	97.31	13.51	94.19	26.41	93.24	33.70	96.080	18.730
$\eta = 0.95$	99.39	2.03	97.52	11.77	94.36	24.87	93.33	31.64	96.150	17.578
$\eta = 0.999$	99.27	2.91	97.57	11.02	93.97	25.56	92.95	32.5	95.940	17.998

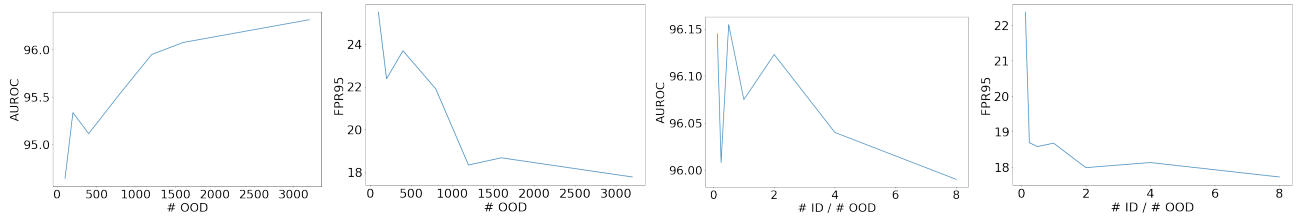

Figure 6. Performance (%) of CLIPScope across different quantities of OOD samples (top), and varying ratios of ID to OOD samples (bottom). The ID dataset is ImageNet-1k. The OOD datasets include iNaturalist, SUN, Places, and Textures. The figures presented are the average results from these four cases.

Table 6. Performance of CLIPScope on each small ID dataset. Each case contains 50% ID samples and 50% OOD samples.

OOD Dataset	iNaturalist		SUN		Places		Textures		Average	
# Samples	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
200	99.02	4	93.91	30	90.3	39	95.33	29	94.640	25.500
400	99.35	3.5	94.9	26.5	90.9	38.5	96.19	21	95.335	22.375
800	99.16	3.75	95.06	22	94.16	24.75	92.07	44.25	95.113	23.688
1600	99.33	2.5	96.63	16.37	94.56	25.25	91.63	43.5	95.538	21.905
2400	99.42	1.66	96.96	15.08	94.58	23	92.83	33.58	95.948	18.330
3200	99.44	1.5	96.98	13.75	94.98	24.75	92.90	34.68	96.075	18.670
6400	99.52	1.34	97.30	13	95.12	23.53	93.32	33.21	96.315	17.770
12800	99.55	1.23	97.43	12.68	94.34	26.73	93.15	35	96.118	18.910

slight decrease in AUROC and an increase in FPR95 in most cases due to this imbalance. As discussed in the limitations section, potential misleading information, such as providing redundant ID labels, could negatively impact detection accuracy. However, as shown in Table 4, our approach is effective if the ID labels are correctly provided.

Table 7. Performance of CLIPScope on various ID/OOD ratios. The number of OOD samples is fixed at 1600.

OOD Dataset	iNaturalist		SUN		Places		Textures		Average	
# ID / # OOD	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
1/8	99.19	23.12	96.95	11.5	95.06	22.37	93.38	32.5	96.145	22.373
1/4	99.20	1.93	96.77	13.37	94.83	24.37	93.23	35.06	96.008	18.683
1/2	99.43	1.87	96.97	14	95.15	22.75	93.07	35.68	96.155	18.575
1/1	99.44	1.5	96.98	13.75	94.98	24.75	92.90	34.68	96.075	18.670
2/1	99.46	1.5	97.06	14.62	95.06	22.87	92.91	32.93	96.123	17.980
4/1	99.49	1.5	97.01	14	94.95	23.56	92.71	33.43	96.040	18.123
8/1	99.53	1.5	97.01	13.43	94.85	22.75	92.57	33.18	95.990	17.715

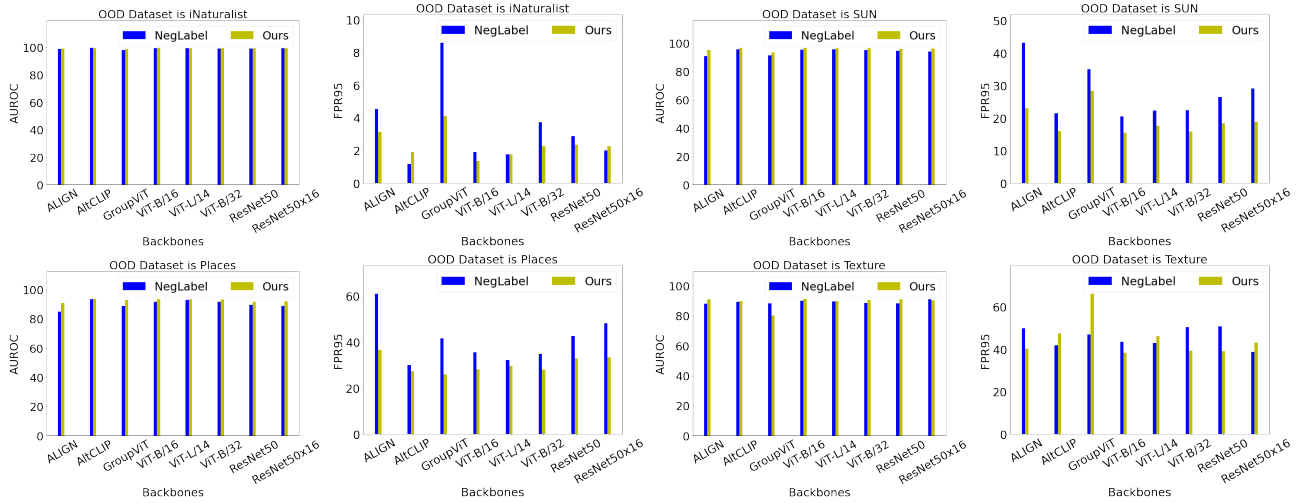

Figure 7. Performance (%) of CLIPScope with various backbones. The ID dataset is ImageNet-1k.

Table 8. Performance of CLIPScope on unbalanced ID datasets.

OOD Dataset	iNaturalist		SUN		Places		Textures		Average	
ID Datasets	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
ImageNet-1K	99.60	1.28	97.34	13.52	94.20	26.32	93.04	34.41	96.045	18.883
ImageNet-10	98.98	5.11	97.41	9.66	92.46	23.39	90.82	32.34	94.918	17.625
ImageNet-20	99.61	1.50	96.43	15.01	90.90	33.41	88.67	38.63	93.903	22.138
ImageNet-100	99.35	2.14	95.75	17.15	90.70	31.34	87.46	44.02	93.315	23.663

C. Further Discussions

C.1. Computation Complexity

The computational complexity of CLIPScope is $\mathcal{O}(2MD)$ per image, where M is the number of negative labels and D is the dimension of the embedding feature. This complexity is the same as NegLabel's. Both methods use around 10,000 OOD labels and CLIP as the feature extractor, resulting in an efficient OOD detection time of about 1ms per sample. The mining algorithm, which processes large corpora like WordNet, takes only a few minutes on a single GPU machine and is performed before the inference phase, not affecting the inference speed. Importantly, CLIPScope calculates the confidence score for each input instance only once, eliminating the need for repeated scoring and improving computational efficiency.

C.2. Overlap Between Mined and Actual OOD Labels

Our OOD label mining strategy does not assume access to OOD test data, ensuring an unbiased selection of OOD labels without prior knowledge of the test data's OOD classes. This approach is similar to NegLabel's. Any overlap between the

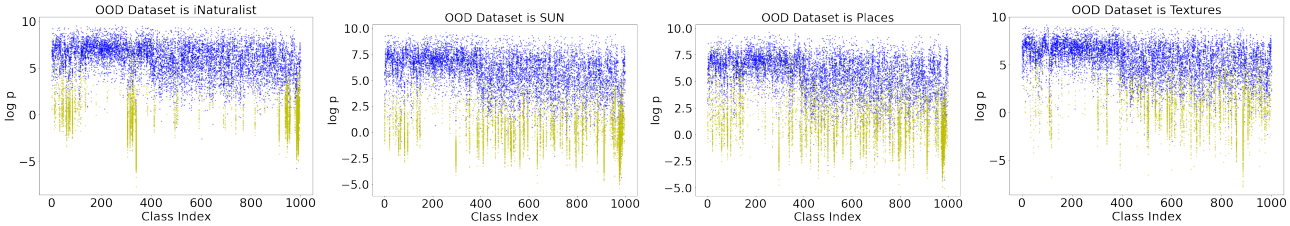

Figure 8. The logarithm confidence scores log p of ID (blue) and OOD (yellow) samples. The ID dataset is ImageNet-1k.

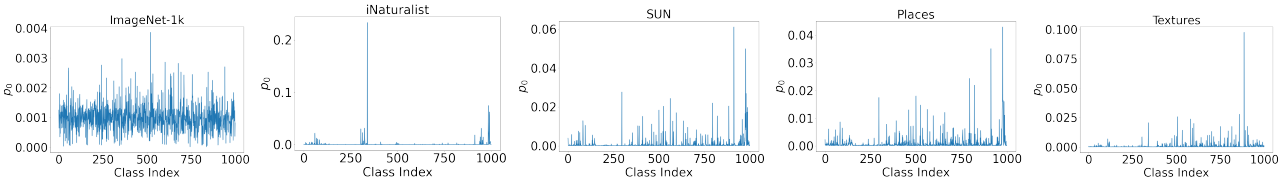

Figure 9. The classification behavior of CLIP on ID dataset is different from the classification behavior on OOD datasets.

mined OOD labels and the actual OOD classes in the test data highlights the effectiveness of our mining strategy rather than being a drawback. NegLabel has previously justified using a wide range of concepts, potentially including the semantic labels of OOD samples, as a reasonable approach. This justification holds, especially when the corpus is large, similar to how vision-language models (VLMs) are considered suitable for evaluation in zero-shot tasks despite potential exposure to task-relevant data. When developers have specific insights into likely OOD labels, these can be intentionally included in the negative label space to further improve OOD detection effectiveness. Furthermore, Table 5 demonstrates that our approach remains effective even with a small number of M (e.g., 0, 50, or 100). For smaller values of M, the mined OOD labels are less likely to overlap with actual OOD labels.

C.3. ID Instances Classified into High Likelihood Classes

ID instances that are classified into classes with high likelihood are influenced by the elevated class likelihood values. This effect is reflected in their confidence scores. Despite this influence, the confidence scores of these ID instances are still likely to surpass the threshold because the numerator of their confidence scores is usually high. Indeed, Fig. 8 shows the logarithm confidence scores log p of ID and OOD samples for different datasets. Most ID instances have higher scores than OOD instances even in the high likelihood classes. Fig. 9 shows which classes have the highest likelihood.

The p_0 in Fig. 9 for each dataset D is calculated as follows:

$$p_0(y_i) = \frac{1}{|D|} \sum_{x \in D} \mathbb{1}(f(x, \mathcal{Y}) = y_i) \quad \forall y_i \in \mathcal{Y} \tag{7}$$

where $\mathcal{Y}$ is the ImageNet-1K labels. Given the ID dataset D_I and the OOD dataset D_O, we have

$$\mathbb{P}(x \in \text{OOD} \mid f(x, \mathcal{Y}) = y_i) = \frac{\sum_{x \in D_O} \mathbb{1}(f(x, \mathcal{Y}) = y_i)}{\sum_{x \in D_O} \mathbb{1}(f(x, \mathcal{Y}) = y_i) + \sum_{x \in D_I} \mathbb{1}(f(x, \mathcal{Y}) = y_i)}. \tag{8}$$

Based on Fig. 9, $\mathbb{P}(x \in \text{OOD} \mid f(x, \mathcal{Y}) = y_i)$ varies significantly between classes. However, CLIPScope provides very good performance in this general case, as evidenced by the results shown in Table 1.

C.4. Training-Based Methods

Training-based or tuning-based methods may improve their performance by using historical test samples. However, compared to training-based methods, our approach does not rely on ground-truth labels from historical test data and offers substantial advantages in terms of efficiency. It requires minimal memory, as it uses only histogram information based on the empirical output of CLIP rather than ground-truth labels, and it has faster computation compared to fine-tuning models with numerous parameters. These benefits make our method more practical for applications requiring frequent updates.

C.5. Broader Impact

This paper presents work whose goal is to advance the field of machine learning. It demonstrates a positive impact in the realm of zero-shot OOD detection by leveraging posterior information from historical instances. This approach has shown a considerable improvement in detection accuracy, setting a precedent for other OOD detection methods. The integration of posterior information into confidence score calculations could potentially enhance the performance of various OOD detection models, not limited to zero-shot approaches. However, the potential misleading information within the historical data could adversely affect detection accuracy, compromising the reliability of open-world deployed machine learning systems.

C.6. Future Works

While CLIPScope currently utilizes only class likelihood as its form of posterior information, future explorations could delve into other types of posterior data. This expansion could uncover new dimensions of accuracy and efficiency in OOD detection. Furthermore, the development of new OOD detection scores remains a valuable and promising avenue of research. It would be interesting to investigate how existing OOD detectors could benefit from the incorporation of CLIPScope's approach to using posterior information.