Activation Matters: Test-time Activated Negative Labels for OOD Detection with Vision-Language Models

Supplementary Material

6. Appendix

A6.1. Dataset

We perform extensive experiments using the large-scale ImageNet-1k dataset [8] as the ID dataset. In line with previous works [26, 29, 48], we evaluate the method on four OOD datasets, including iNaturalist [65], SUN [73], Places [85], and Textures [7]. Additionally, we validate our approach on the OpenOOD benchmark [75, 81], where OOD datasets are categorized into near-OOD (e.g., SSB-hard [67], NINCO [3]) and far-OOD (e.g., iNaturalist [65], Textures [7], OpenImage-O [68]) based on their similarity to the ImageNet dataset. Besides the traditional OOD detection with semantic-shift, we also study a more practical full-spectrum OOD detection setting [76], where the model is additionally challenged by the robustness to covariate shifts. Following [81], we adopt ImageNet-C [19], ImageNet-R [24], and ImageNet-V2 [57] as the covariate-shifted test data in the full-spectrum setting.

A6.2. Detailed Results on ImageNet Dataset

We compare against traditional methods [12, 20, 27, 43, 44, 62, 64, 68] by fine-tuning CLIP-encoders with labeled training samples, as described in [29], and report results of [2, 6, 29, 30, 40, 40, 51, 54, 78, 83] based on their original papers.

The detailed OOD detection results with OOD datasets of INaturalist/Sun/Places/Textures are illustrated in Tab. A6.

The detailed OOD detection results and full-spectrum OOD detection results under the OpenOOD setting are presented in Tab. A7 and Tab. A8, respectively.

A6.3. Results on CIFAR10/100

Besides ImageNet, we also assess our method on the smaller CIFAR10/100 datasets [32] under the OpenOOD framework. Specifically, with CIFAR10/100 as the ID datasets, we utilize near-OOD datasets such as CIFAR100/10 and TIN [35], and far-OOD datasets including MNIST [9], SVHN [52], Texture [7], and Places365 [85]. As illustrated in Tab. A9, our advantage still holds.

A6.4. Results with Medical Images

We also validate our TANL method with medical images following the OpenOOD setup [81]. Specifically, we use BIMCV as the ID dataset, where the task is to distinguish COVID-19 patients from healthy individuals using chest X-ray images [66]. The OOD dataset is constructed using the CT-SCAN and X-Ray-Bone datasets. The CT-SCAN dataset includes computed tomography (CT) images of COVID-19 patients and healthy individuals, while the X-Ray-Bone dataset contains X-ray images of hands. As shown in Tab. A11, our method significantly outperforms existing competitors.

A6.5. More Analyses

Queue length. We formulate $\mathcal{X}_{pos/neg}$ as fixed-length FIFO queues, where the queue length L determines the number of cached samples in dynamic environments. As shown in Fig. A5a, appropriately increasing the queue length can reduce error, and performance saturates when the queue length exceeds 300. Therefore, we set the queue length to 300 by default.

γ values. As shown in Fig. A5b, the optimal threshold γ differs between near-OOD and far-OOD scenarios because the distance between OOD samples and ID samples varies significantly. Instead of manually searching for the best γ, we adopt an automated approach to set it dynamically. This is achieved by modeling the score $\mathcal{S}_{aa}$ of all historical samples as a bimodal distribution of two clusters (*e.g.*, ID Vs. OOD) and identifying the threshold that minimizes the intra-cluster variations of the two clusters [41]:

$$\min_{\gamma} \quad \text{var}(\mathcal{O}^{+}_{\text{his}}) + \text{var}(\mathcal{O}^{-}_{\text{his}}) \tag{A6.17}$$

$$\text{where} \quad \mathcal{O}^{+}_{\text{his}} = \{\mathcal{S}_{aa}(\boldsymbol{v}) \mid \mathcal{S}_{aa}(\boldsymbol{v}) \geq \gamma,\ \boldsymbol{v} \in \mathcal{X}_{\text{his}}\}\,, \tag{A6.18}$$

$$\mathcal{O}^{-}_{\text{his}} = \{\mathcal{S}_{aa}(\boldsymbol{v}) \mid \mathcal{S}_{aa}(\boldsymbol{v}) < \gamma,\ \boldsymbol{v} \in \mathcal{X}_{\text{his}}\}\,, \tag{A6.19}$$

where $\text{var}(A)$ measures the variation of the score set A, and $\mathcal{X}_{\text{his}}$ is also a FIFO queue that stores the most recent $20,000$ test samples, initialized with positive and negative samples as defined in Eq. 10.

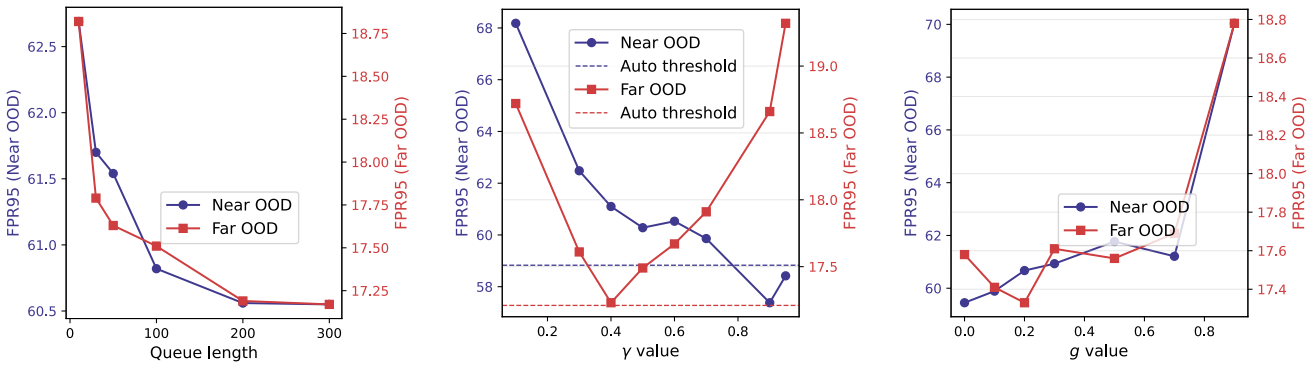
This dynamic thresholding typically achieves results comparable to those with manually searched γ and is adopted as the default setting.

g values. As shown in Fig. A5c, an appropriately small gap g helps reduce the detection error in the Far OOD setting, whereas the Near OOD setting benefits from a smaller value of g. By default, we set $g = 0.2$ for all experiments.

VLM Backbones. As shown in Tab. A12, our TANL achieves excellent results across various VLM backbones, where a stronger backbone generally leads to better performance. Interestingly, unlike NegLabel, which heavily relies on a strong backbone to achieve high results, our method

Table A6. Complete OOD detection results with ImageNet-1k, where a VITB/16 CLIP encoder is adopted.

Methods	OOD datasets									
	INaturalist		Sun		Places		Textures		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Methods requiring training (or fine-tuning)										
MSP [20]	87.44	58.36	79.73	73.72	79.67	74.41	79.69	71.93	81.63	69.61
ODIN [43]	94.65	30.22	87.17	54.04	85.54	55.06	87.85	51.67	88.80	47.75
Energy [44]	95.33	26.12	92.66	35.97	91.41	39.87	86.76	57.61	91.54	39.89
GradNorm [27]	72.56	81.50	72.86	82.00	73.70	80.41	70.26	79.36	72.35	80.82
ViM [68]	93.16	32.19	87.19	54.01	83.75	60.67	87.18	53.94	87.82	50.20
KNN [62]	94.52	29.17	92.67	35.62	91.02	39.61	85.67	64.35	90.97	42.19
VOS [12]	94.62	28.99	92.57	36.88	91.23	38.39	86.33	61.02	91.19	41.32
NPOS [64]	96.19	16.58	90.44	43.77	89.44	45.27	88.80	46.12	91.22	37.93
ZOC [14]	86.09	87.30	81.20	81.51	83.39	73.06	76.46	98.90	81.79	85.19
LSN [54]	95.83	21.56	94.35	26.32	91.25	34.48	90.42	38.54	92.96	30.22
CLIPN [69]	95.27	23.94	93.93	26.17	92.28	33.45	90.93	40.83	93.10	31.10
LoCoOp [51]	96.86	16.05	95.07	23.44	91.98	32.87	90.19	42.28	93.52	28.66
TagOOD [38]	98.97	5.00	92.22	27.70	87.81	40.40	90.60	36.31	92.40	27.85
FA_{GL} [47]	96.48	14.49	93.46	27.65	92.44	31.09	92.93	29.50	93.82	25.68
LAPT [83]	99.63	1.16	96.01	19.12	92.01	33.01	91.06	40.32	94.68	23.40
NegPrompt [40]	98.73	6.32	95.55	22.89	93.34	27.60	91.60	35.21	94.81	23.01
CMA [30]	99.62	1.65	96.36	16.84	93.11	27.65	91.64	33.58	95.13	19.93
AdaND [5]	99.49	1.91	95.49	20.53	94.55	20.95	93.01	21.76	95.58	16.00
SynOOD [39]	99.57	1.57	95.82	20.46	93.37	12.12	95.29	22.94	97.01	14.27
Training-free & non-adaptive methods										
MCM [48]	94.59	32.20	92.25	38.80	90.31	46.20	86.12	58.50	90.82	43.93
CoVer [79]	95.98	22.55	93.42	32.85	90.27	40.71	90.14	43.39	92.45	34.88
CMA [37]	96.89	23.84	93.69	30.11	93.17	29.86	88.47	47.35	93.05	32.79
EOE [4]	97.52	12.29	95.73	20.04	92.95	30.16	85.64	57.53	92.96	30.09
NegLabel [29]	99.49	1.91	95.49	20.53	91.64	35.59	90.22	43.56	94.21	25.40
Peng *et al.* [55]	99.64	1.04	96.32	18.45	95.81	31.15	92.15	38.79	96.00	22.36
DualCnst [34]	99.70	0.98	95.66	18.13	91.81	31.77	91.73	34.91	94.72	21.45
CSP [6]	99.60	1.54	96.66	13.66	92.90	29.32	93.86	25.52	95.76	17.51
Training-free & test-time adaptation methods										
CLIPScope [16]	99.61	1.29	96.77	15.56	93.54	28.45	91.41	38.37	95.30	20.88
AdaNeg [82]	99.71	0.59	97.44	9.50	94.55	34.34	94.93	31.27	96.66	18.92
OODD [78]	99.79	0.85	97.17	12.94	92.51	30.68	94.51	30.67	96.00	18.79
TANL (Ours)	**99.84**	**0.42**	**99.07**	**3.53**	**95.87**	**21.90**	**97.11**	**13.38**	**97.97**	**9.81**



(a) Queue length (b) γ value (c) g values

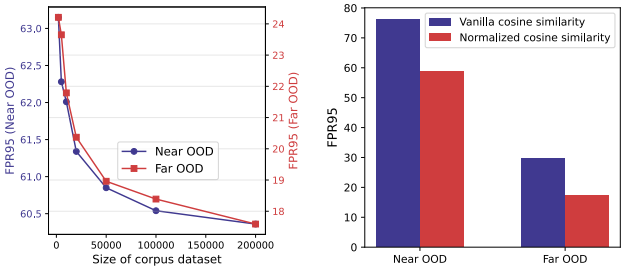
Figure A5. Analyses on (a) the queue length L, (b) threshold γ, and (c) gap value g in Eq. 9 under the OpenOOD setup.

Table A7. Detailed OOD detection results on the OpenOOD benchmark, where ImageNet-1k is adopted as ID dataset.

Settings	OOD Datasets	FPR95 ↓	AUROC ↑
Near-OOD	SSB-hard [67]	62.51	83.96
	NINCO [3]	57.61	85.10
	Mean	60.06	84.53
Far-OOD	iNaturalist [65]	0.47	99.83
	Textures [7]	11.22	97.53
	OpenImage-O [68]	39.95	91.92
	Mean	17.21	96.43

Table A8. Detailed full-spectrum OOD detection results on the OpenOOD benchmark, where ImageNet-1k, ImageNet-C, ImageNet-R, ImageNet-V2 are used as ID datasets.

Settings	OOD Datasets	FPR95 ↓	AUROC ↑
Near-OOD	SSB-hard [67]	70.50	79.29
	NINCO [3]	66.92	78.51
	Mean	68.71	78.90
Far-OOD	iNaturalist [65]	1.70	99.58
	Textures [7]	19.68	94.81
	OpenImage-O [68]	46.07	88.66
	Mean	22.48	94.35



(a) Size of corpus dataset (b) Activation Metric

Figure A6. Analyses on (a) size of corpus dataset and (b) activation metric under the OpenOOD setup.

works well even with a smaller ResNet-50 backbone, demonstrating its effectiveness and practicality.

Corpus Datasets. The corpus dataset provides candidates for activated labels, and its diversity is a prerequisite for selecting effective activated labels. We construct the corpus dataset by selecting nouns and adjectives from WordNet [1] and Part-of-Speech-Tags [2], removing duplicate words within the ID set. This results in corpus datasets containing 140.5K and 270.2K words, respectively. For a fair comparison, we also adopt the WordNet subset filtered by NegLabel as the corpus dataset, which contains 70K adjectives and nouns.

As illustrated in Tab. A13, different corpus datasets lead to similarly good results, demonstrating the robustness of our method to the choice of corpus dataset. Beyond comparing different corpus datasets, we also analyze the sensitivity of our method to the size of the corpus dataset by randomly sampling smaller subsets from the Part-of-Speech-Tags dataset. As shown in Fig. A6a, using an appropriately large-sized corpus dataset generally results in better performance, with results saturating when the corpus dataset exceeds 100K. For fair comparison, we adopt the WordNet subset filtered by NegLabel as the corpus dataset by default.

Activation Metric. The activation metric we use is the cosine similarity normalized by a softmax function, as shown in Eq. 5. One may wonder whether directly using vanilla cosine similarity, *i.e.*, modifying Eq. 5 to $Act(\mathcal{X}, \widehat{y}_i) = \frac{1}{|\mathcal{X}|}\sum_{\boldsymbol{x}\in\mathcal{X}} f_{img}(\boldsymbol{x}) f_{txt}(\rho(\widehat{y}_i))$, could also work. As shown in Fig. A6b, using the normalized similarity leads to a significant advantage. This is likely because unnormalized similarity has a lower discriminative capacity. For example, same-class similarities often cluster around 0.3, while different-class similarities typically cluster around 0.1 [48], making it difficult to distinguish activated labels effectively.

Statistical Significance. We control the input sequence using a random seed and report OOD detection results on the ImageNet-1k dataset with three random seeds as follows: AUROC: 97.97 ± 0.01 and FPR95: 9.81 ± 0.01. These results validate the robustness of our method to variations in the input sequence.

Mutual Enhancement. In our method, improving the accuracy of the activation score in Eq. 8 helps select more effective negative labels in Eq. 6, thereby enhancing the OOD detection capability of the activation-aware score in Eq. 15. Similarly, the improved OOD detection capability of the activation-aware score contributes to selecting more accurate negative and positive images in Eq. 9, which in turn enhances the estimation of the activation score in Eq. 8. Overall, there exists a mutual enhancement relationship between the estimation of the activation score and the activation-aware OOD score function. Such a relationship is evidenced in Tab. A14, where we remove this mutual enhancement by fixing $\mathcal{S}_{aa}$ in Eq. 9 using activated labels estimated via the initialized $\mathcal{X}_{pos/neg}$ in Eq. 10. This leads to a significant performance drop.

Temporal Shift. Since our method dynamically explores activated negative labels in a test-time adaptation manner, it is crucial to investigate its stability under temporal shifts, where OOD environments evolve over time. Specifically, we use ImageNet as the ID dataset and assume that OOD datasets change sequentially over time (*e.g.*, I-S-P-T represents OOD datasets transitioning from INaturalist to Sun to Places to Textures). In implementation, we do not re-initialize the queue $\mathcal{X}_{pos/neg}$ for each OOD dataset. As

[1] https://wordnet.princeton.edu/

[2] https://www.kaggle.com/datasets/ruchi798/part-of-speech-tagging

Table A9. OOD detection results with CIFAR10/100 on the OpenOOD benchmark. Full results are provided in Tables A10.

Methods	FPR95 ↓		AUROC ↑	
	Near-OOD	Far-OOD	Near-OOD	Far-OOD
Methods requiring training (or fine-tuning)				
PixMix [25] + KNN [62]	–	–	93.10	95.94
OE [21] + MSP [20]	–	–	94.82	96.00
PixMix [25] + RotPred [22]	–	–	94.86	98.18
Training-free & non-adaptive methods				
MCM [48]	30.86	17.99	91.92	95.54
NegLabel [29]	28.75	6.60	94.58	98.39
Training-free & test-time adaptation methods				
AdaNeg [82]	20.40	2.79	94.78	99.26
TANL (Ours)	**19.31**	**2.43**	**94.91**	**99.26**

(a) Results with ID dataset of CIFAR10

Methods	FPR95 ↓		AUROC ↑	
	Near-OOD	Far-OOD	Near-OOD	Far-OOD
Methods requiring training (or fine-tuning)				
GEN [45]	–	–	81.31	79.68
VOS [13] + EBO [44]	–	–	80.93	81.32
SCALE [74]	–	–	80.99	81.42
OE [21] + MSP [20]	–	–	88.30	81.41
Training-free & non-adaptive methods				
MCM [48]	75.20	59.32	71.00	76.00
NegLabel [29]	71.44	40.92	70.58	89.68
Training-free & test-time adaptation methods				
AdaNeg [82]	59.07	29.35	84.62	95.25
TANL (Ours)	**57.74**	**24.55**	**85.06**	**95.41**

(b) Results with ID dataset of CIFAR100.

Table A10. Detailed OOD detection results on the OpenOOD benchmark.

Near/Far-OOD	Datasets	FPR95 ↓	AUROC ↑
Near-OOD	CIFAR100 [32]	33.63	91.17
	TIN [35]	4.99	98.65
	Mean	19.31	94.91
Far-OOD	MNIST [9]	0.13	99.96
	SVHN [52]	0.04	99.97
	Texture [7]	0.04	99.86
	Places365 [85]	9.51	97.25
	Mean	2.43	99.26

(a) Detailed results with ID dataset of CIFAR10.

Near/Far-OOD	Datasets	FPR95 ↓	AUROC ↑
Near-OOD	CIFAR10 [32]	56.32	80.47
	TIN [35]	59.16	89.65
	Mean	57.74	85.06
Far-OOD	MNIST [9]	0.54	99.81
	SVHN [52]	5.81	98.53
	Texture [7]	31.36	95.36
	Places365 [85]	60.49	90.94
	Mean	24.55	95.41

(b) Detailed results with ID dataset of CIFAR100.

shown in Tab. A15, our method consistently maintains a large advantage over the NegLabel baseline, validating its robustness to temporal shifts.

Sample Order. In our experiments, ID and OOD samples are randomly shuffled during testing, corresponding to the "Random Shuffled" scenario. To analyze the impact of sample order, we have conducted additional experiments. Specifically, we consider two extreme cases: "ID First" (all ID samples are tested before any OOD samples) and "OOD First" (all OOD samples are tested before any ID samples). As shown in Tab. A16, while performance does drop under these extreme settings compared to the "Random Shuf-

Table A11. OOD detection results with medical images following the OpenOOD setting.

Methods	OOD datasets CT-SCAN		X-Ray-Bone		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
NegLabel [29]	63.53	100.0	99.68	0.56	81.61	50.28
AdaNeg [82]	93.48	100.0	99.99	0.11	96.74	50.06
TANL (Ours)	**93.94**	**39.06**	**100.0**	**0.0**	**96.97**	**19.53**

Table A12. OOD detection results of TANL with different VLMs backbones, where ImageNet-1K is used as the ID dataset.

Methods	OOD datasets INaturalist		Sun		Places		Textures		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
ResNet50										
NegLabel [29]	99.24	2.88	94.54	26.51	89.72	42.60	88.40	50.80	92.97	30.70
TANL (Ours)	**99.73**	**0.99**	**99.04**	**4.00**	**95.97**	**23.76**	**97.01**	**12.71**	**97.88**	**10.37**
VITB/32										
NegLabel [29]	99.11	3.73	95.27	22.48	91.72	34.94	88.57	50.51	93.67	27.92
TANL (Ours)	**99.76**	**0.87**	**99.24**	**3.09**	**96.07**	**20.97**	**96.50**	**16.28**	**97.89**	**10.30**
VITB/16										
NegLabel [29]	99.49	1.91	95.49	20.53	91.64	35.59	90.22	43.56	94.21	25.40
TANL (Ours)	**99.84**	**0.42**	**99.07**	**3.53**	**95.87**	**21.90**	**97.11**	**13.38**	**97.97**	**9.81**
VITL/14										
NegLabel [29]	99.53	1.77	95.63	22.33	93.01	32.22	89.71	42.92	94.47	24.81
TANL (Ours)	**99.88**	**0.29**	**99.15**	**3.42**	**96.11**	**20.79**	**97.12**	**13.57**	**98.07**	**9.52**

fled" scenario, our method still significantly outperforms the NegLabel baseline. This demonstrates the robustness and effectiveness of our method, even when the order of test samples is highly imbalanced.

Early Errors. We dynamically estimate the activation scores using the cached images in the memory queues during test time. A natural question arises: in extreme situations—such as early testing stages where samples may be more challenging or when the initial OOD detector is relatively weak—the memory queues may contain a high fraction of misclassified images. In a test-time adaptation setting, this raises the concern of whether such errors may accumulate and eventually compromise the accuracy of activation estimation.

To investigate this issue, we manually control the OOD

Table A13. OOD detection results with different corpus datasets on the OpenOOD benchmark, where ImageNet-1k is adopted as ID dataset.

Corpus Datasets	FPR95 ↓		AUROC ↑	
	Near-OOD	Far-OOD	Near-OOD	Far-OOD
Vanilla WordNet	60.52	17.34	83.24	95.92
Part-of-speech-tagging	60.36	17.24	83.74	96.15
WordNet-subset Filtered by NegLabel	60.06	17.21	84.53	96.43

Table A14. Analyses on the mutual enhancement between the estimation of activation information and the activation-aware OOD score function.

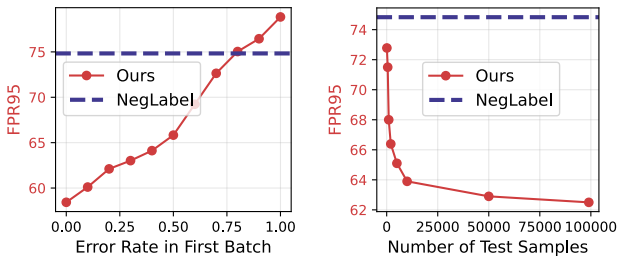
Methods	FPR95 ↓	
	Near-OOD	Far-OOD
W/o mutual enhancement	63.00	19.04
With mutual enhancement (Ours)	60.06	17.21

Table A15. OOD detection results under the temporal shifts, where ImageNet-1k ID dataset and a VITB/16 CLIP encoder are adopted.

Methods	OOD datasets									
	INaturalist		Sun		Places		Textures		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
NegLabel [29]	99.49	1.91	95.49	20.53	91.64	35.59	90.22	43.56	94.21	25.40
TANL under Temporal Shifts										
I-S-P-T	99.84	0.45	98.77	4.75	95.93	22.16	96.67	14.91	97.80	10.56
S-P-T-I	99.83	0.46	99.04	3.65	95.92	21.37	96.70	15.09	97.87	10.14
P-T-I-S	99.84	0.45	98.73	5.28	95.87	21.67	96.66	15.33	97.87	11.68
T-I-S-P	99.83	0.47	98.77	4.69	95.94	21.30	97.04	13.51	97.89	9.99

Table A16. FPR95 (↓) with different orders of ID and OOD samples.

Order of Test Samples	INaturalist	SUN	Places	Textures	Average
NegLabel (Baseline)	1.91	20.53	35.59	43.56	25.40
ID First	2.97	8.35	29.63	24.91	16.47
OOD First	2.74	9.06	36.00	19.14	16.73
Random Shuffled	0.42	3.53	21.90	13.38	9.81



(a) Early Error Rate (b) Number of Test Samples

Figure A7. Analyses on (a) error rates in the first batch and (b) number of test samples with ImageNet (ID) and SSB-hard (OOD).

misclassification rate of images inserted into the memory queues in the first batch, where batch size 256 is adopted. Specifically, a rate of 0.1 indicates that 10% of the images cached into the ID/OOD queues are intentionally misclassified. As shown in Fig. A7a, performance indeed degrades as the misclassification rate increases. When the misclassification rate exceeds 80%, test-time adaptation begins to have a negative impact compared to the NegLabel baseline. This demonstrates that our method is reasonably robust to early-stage errors. Specifically, it indicates that only minimal initial classifier quality is needed—our approach continues to behave positively even when the initial model is quite weak (∼20% accuracy).

Number of Test Samples. As a test-time method, our approach is expected to improve as more test samples become available, eventually converging to a stable performance level. Here, we empirically verify this behavior by varying the number of test samples and examining how many samples are typically required for the method to stabilize and reach peak performance.

Specifically, we use ImageNet (50K test samples) as ID and SSB-hard (49K samples) as OOD, as both datasets provide a large and approximately balanced number of test images. For each target number of test samples, we randomly draw that many samples from the mixed dataset, repeat the sampling process three times, and report the average performance. As shown in Fig. A7b, the FPR95 consistently decreases as more test samples are included and roughly converges around 10K samples. In addition, our method yields a substantial improvement over NegLabel in the low-sample regime (fewer than 1K test samples), validating its applicability in practical scenarios where only limited test data are available.

Activation-aware Score Variant with Explicit Weighting. Our activation-aware score implicitly re-weights negative labels through rank repetition. Here, we analyze an alternative that applies explicit activation-based weighting. Specifically, we introduce a weighting parameter into the score function in Eq. 4, yielding the following two variants:

$$S_{aa}^{ew1}(\boldsymbol{v}) = \sum_{i=1}^{C} \frac{\exp(\boldsymbol{v}\boldsymbol{t}_i)}{\sum_{j=1}^{C}\exp(\boldsymbol{v}\boldsymbol{t}_j) + \sum_{j=1}^{M}\exp(w_j \boldsymbol{v}\widetilde{\boldsymbol{t}}_j)}, \tag{A6.20}$$

$$S_{aa}^{ew2}(\boldsymbol{v}) = \sum_{i=1}^{C} \frac{\exp(\boldsymbol{v}\boldsymbol{t}_i)}{\sum_{j=1}^{C}\exp(\boldsymbol{v}\boldsymbol{t}_j) + \sum_{j=1}^{M} w_j \exp(\boldsymbol{v}\widetilde{\boldsymbol{t}}_j)}, \tag{A6.21}$$

where each weight w_j is obtained by normalizing the activa-

tion score $\widehat{Act_b}(\widehat{y}_j)$:

$$w_j = M \frac{\widehat{Act_b}(\widehat{y}_j)}{\sum_i^M \widehat{Act_b}(\widehat{y}_i)}. \tag{A6.22}$$

This normalization ensures that the average weight of negative labels remains 1, while assigning relatively larger weights to those with higher activation scores.

As shown in Tab. A17, $S_{aa}^{ew1}(\boldsymbol{v})$ performs poorly because the exponential operation amplifies the effect of w_j, causing negative labels with large activation to dominate the denominator. Although $S_{aa}^{ew2}(\boldsymbol{v})$ surpasses the vanilla $S_{nl}(\boldsymbol{v})$, demonstrating the effectiveness of explicit weighting, our implicit weighting strategy still provides a clear advantage. We believe that more sophisticated explicit weighting formulations beyond Eq. A6.22 could potentially achieve comparable or even superior performance. Nonetheless, all these variants consistently confirm our central insight: **incorporating activation information into the OOD score is crucial for improving OOD detection performance.**

Limitations. The limitation of our method lies in the assumption that the corpus dataset covers words related to the OOD distribution and that the pre-trained text encoder understands these words. This assumption may not hold in certain domains. For example, WordNet primarily contains everyday vocabulary and lacks sufficient medical terms, while the vanilla CLIP model has limited understanding of medical images. As a result, improvements in medical OOD detection can be restricted. This limitation suggests that domain-specific tasks may require a tailored corpus dataset and pre-trained models, which is left for future investigation.

Visualization of Activated Labels. We visualize the activated labels under different OOD datasets and activation criteria in Fig. A8. We observe that the selected activated labels indeed exhibit a high degree of similarity with the OOD dataset. For example, when the OOD dataset is *Textures*, the labels with high activation scores are 'grating' and 'polka dot', which align well with the visual characteristics of the *Textures* dataset, as shown in Fig. A8a. Furthermore, we find that the highly activated scores selected using $Act(\mathcal{X}_{neg})$ and $\widehat{Act_d}$ are quite similar, which explains the effectiveness of $Act(\mathcal{X}_{neg})$ as a criterion, as demonstrated in Fig. 3b. In contrast, the activated labels selected using only positive images do not show a clear relationship with the target OOD dataset, which corresponds to its lower results in Fig. 3b.

Table A17. FPR95 (↓) with activation-aware score variant of explicit weighting.

OOD Score	INaturalist	SUN	Places	Textures	Average
NegLabel (Baseline)	1.91	20.53	35.59	43.56	25.40
$S_{nl}(\boldsymbol{v})$	1.30	6.67	25.20	18.37	12.91
$S_{aa}^{ew1}(\boldsymbol{v})$	99.85	100	100	100	99.96
$S_{aa}^{ew2}(\boldsymbol{v})$	0.62	4.08	24.06	15.04	10.95
$S_{aa}(\boldsymbol{v})$	0.42	3.53	21.90	13.38	9.81

OOD: INaturalist

ID: ImageNet

Ranked Corpus Dataset

$\widehat{Act}_d$	Labels
0.0056	western poison oak
0.0049	blue-eyed grass
0.0034	poison oak
	...
-0.00041	scaly-tailed
-0.00042	siberian crab
-0.00046	iditarod

OOD: Textures

ID: ImageNet

Ranked Corpus Dataset

$\widehat{Act}_d$	Labels
0.0022	grating
0.0018	polka dot
0.0015	pothole
	...
-0.00061	dwarf birch
-0.00066	capuchin
-0.00072	iditarod

(a) Ranked corpus dataset in far-ood setting

OOD: SSB-hard

ID: ImageNet

Ranked Corpus Dataset

$\widehat{Act}_d$	Labels
0.0051	yellow-banded
0.0025	square-tailed
0.0023	nyala
...	...
-0.00068	capuchin
-0.00086	chelonian
-0.00092	nonvenomous

OOD: Ninco

ID: ImageNet

Ranked Corpus Dataset

$\widehat{Act}_d$	Labels
0.0041	cephalopod
0.0033	fire extinguisher
0.0025	caecilian
...	...
-0.00065	dwarf birch
-0.00067	capuchin
-0.00077	iditarod

(b) Ranked corpus dataset in near-ood setting

OOD: Textures

ID: ImageNet

Ranked Corpus Dataset

$\widehat{Act}_d$	Labels
0.0022	grating
0.0018	polka dot
0.0015	pothole
...	...
-0.00061	dwarf birch
-0.00066	capuchin
-0.00072	iditarod

Ranked Corpus Dataset

$Act(\chi_{neg})$	Labels
0.0022	grating
0.0018	polka dot
0.0015	pothole
...	...
7.6×10^{-8}	churchill downs
7.5×10^{-8}	locomotive
7.0×10^{-8}	dwarf-white trillium

Ranked Corpus Dataset

$-Act(\chi_{pos})$	Labels
-1.3×10^{-8}	saxifrage
-1.6×10^{-8}	anthony dollar
-1.8×10^{-8}	habenaria
...	...
-0.00062	square-tailed
-0.00066	capuchin
-0.00072	iditarod

(c) Ranked corpus dataset with different ranking criteria

Figure A8. Visualization of the ranked corpus dataset with (a) far-ood setting, (b) near-ood setting, and (c) different ranking criteria.