Supplementary Materials to: LAPT: Label-driven Automated Prompt Tuning for OOD Detection with Vision-Language Models

Yabin Zhang[1,2], Wenjie Zhu[1], Chenhang He[1], and Lei Zhang[1,2⋆]

[1] The Hong Kong Polytechnic University
[2] OPPO Research Institute
{csybzhang,cslzhang}@comp.polyu.edu.hk

The following materials are provided in this supplementary file:

- More results on small-scaled ID datasets (*cf*. Section 4.1 in the main paper).
- More detailed results on OpenOOD benchmark (*cf*. Section 4.2 in the main paper).
- More analyses and discussions (*cf*. Section 4.3 in the main paper).

A Results on Small-scaled ID Datasets

In addition to experiments with the large-scale ImageNet-1K ID dataset, we also report results on smaller-scaled ID datasets following [5]. As shown in Tab. A1, performance on these smaller datasets is nearing saturation, with our method and NegLabel both achieving nearly perfect ID/OOD discrimination. On more challenging tasks (*e.g.*, using ImageNet-1K as the ID dataset and in the near-OOD detection scenario), our method shows a more pronounced improvement, as shown in the main paper.

B More Detailed Results on OpenOOD Benchmark

The detailed OOD detection and full-spectrum OOD detection results on the OpenOOD benchmark are presented in Tab. A2 and Tab. A3, respectively.

C More Analyses and Discussions

Stage Ablation. We have ablated the roles of different stages, including sample collection, data mixing and prompt tuning in Fig. 5, Tab. 4 and Fig. 4 of the main paper, respectively. To highlight their individual contributions, we reorganize the NearOOD results from the OpenOOD benchmark into Tab. A4.

Impact of Text Prompts in Data Collection. Our generation or retrieval-based image collection process exhibits high robustness to text prompts, as evidenced in Tab. A6. The robustness stems from two facts. First, images retrieved

⋆ Corresponding author.

Table A1: OOD detection results with small-scaled ID datasets, where results are based on a VITB/16 CLIP encoder.

Methods	OOD datasets									
	INaturalist		Sun		Places		Textures		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
ID dataset: Stanford-Cars [6]										
NegLabel [5]	99.99	0.01	99.99	0.01	99.99	0.03	99.99	0.01	99.99	0.01
LAPT (Ours)	99.99	0.01	99.99	0.01	99.99	0.02	99.99	0.01	99.99	0.01
ID dataset: CUB-200 [11]										
NegLabel [5]	99.96	0.18	99.99	0.02	99.90	0.33	99.99	0.01	99.96	0.13
LAPT (Ours)	99.98	0.15	99.99	0.01	99.95	0.16	99.99	0.01	99.98	0.08
ID dataset: Oxford-Pet [8]										
NegLabel [5]	99.99	0.01	99.99	0.02	99.96	0.17	99.97	0.11	99.98	0.07
LAPT (Ours)	99.99	0.01	99.99	0.01	99.97	0.11	99.97	0.12	99.98	0.06
ID dataset: Food-101 [2]										
NegLabel [5]	99.99	0.01	99.99	0.01	99.99	0.01	99.60	1.61	99.90	0.40
LAPT (Ours)	99.99	0.01	99.99	0.01	99.99	0.01	99.71	1.25	99.92	0.32
ID dataset: ImageNet-10										
NegLabel [5]	99.83	0.02	99.88	0.20	99.75	0.71	99.94	0.02	99.85	0.24
LAPT (Ours)	99.85	0.01	99.91	0.14	99.82	0.63	99.95	0.02	99.88	0.20
ID dataset: ImageNet-20										
NegLabel [5]	99.95	0.15	99.51	1.93	98.97	4.40	99.11	2.41	99.39	2.22
LAPT (Ours)	99.96	0.13	99.58	1.78	99.02	3.97	99.23	2.21	99.45	2.02
ID dataset: ImageNet-100										
NegLabel [5]	99.87	0.57	97.89	11.26	96.25	19.15	96.00	20.37	97.50	12.84
LAPT (Ours)	99.91	0.49	98.12	9.21	97.01	15.37	96.89	18.13	97.98	10.80

or generated with different prompts show high consistency. For example, the top 10 images retrieved with different prompts overlap by more than a half, and images generated from various prompts, while not identical, maintain a high cosine similarity (around 0.9) in CLIP space. Second, our training strategy is inherently robust, yielding consistent results when the training data are highly similar. Practically, we employed the simplest prompt '<label>' in the data collection process.

Data Mixing Strategies. As illustrated in Tab. A6, the application of cross-modal data mixing results in a significant performance improvement, indicating that mitigating image noise is crucial when using automatically collected images. Cross-distribution mixing also yields a certain degree of performance enhancement, which is consistent with findings reported in [14].

Hyper-parameter for data mixing. Results with different α in Eq. (10) and β in Eq. (12) are illustrated in Fig. A1a and Fig. A1b, respectively. By default, $\alpha = 1.0$ and $\beta = 0.3$ are used in our experiments.

Different VLMs. Results with various VLMs architectures are illustrated in Tab. A7. A stronger backbone typically leads to better performance.

Visualization of Learned Prompts. Following [15], we interpret the learned continuous prompts by searching their closest words within the vocabulary, where the Euclidean distance is adopted. It's important to note that CLIP employs BEP for its tokenization process, which means that its vocabulary consists

Table A2: Detailed OOD detection results of our LAPT on the OpenOOD benchmark, where ImageNet-1k is adopted as ID dataset.

Near-/Far-OOD	Datasets	FPR95 ↓	AUROC ↑
Near-OOD	SSB-hard [10]	65.47	80.09
	NINCO [1]	52.40	85.16
	Mean	58.94	82.63
Far-OOD	iNaturalist [9]	1.17	99.63
	Textures [3]	38.40	89.72
	OpenImage-O [12]	35.00	93.45
	Mean	24.86	94.26

Table A3: Detailed full-spectrum OOD detection results of our LAPT on the OpenOOD benchmark, where ImageNet-1k is adopted as ID dataset.

Near-/Far-OOD	Datasets	FPR95 ↓	AUROC ↑
Near-OOD	SSB-hard [10]	76.44	71.73
	NINCO [1]	65.92	77.80
	Mean	71.18	74.77
Far-OOD	iNaturalist [9]	1.25	99.56
	Textures [3]	53.09	86.23
	OpenImage-O [12]	44.85	90.63
	Mean	33.07	92.14

of subwords that are commonly found within larger words. For example, the subword 'bi' could be included in the vocabulary due to its frequent occurrence in English words such as 'billion' and 'bionic'. As shown in Tab. A8, the prompts learned for ID differ from those for OOD. Interestingly, the OOD prompts we learned share similarities with the best text prompts found by NegLabel. For instance, they both include the term 'nice'.

Time complexity. We analyze the time complexity of our method and competitors in Tab. A9. Compared to methods that tune a heavy module, such as CLIPN [13], which learns an additional text encoder, our approach only learns lightweight prompts, resulting in faster training speed. Regarding test speed, most methods perform similarly because, regardless of how the text branch is designed, the text encoding typically executes only once. Therefore, the primary time consumption lies in encoding the test image with the image encoder. ZOC [4] is an exception, as it passes the image feature through an additional caption generator module, leading to slower test speed.

Table A4: where 'NegL', 'SamC', 'DisAwareP' and 'DMix' represent the negative label mining, automated sample collection, distribution-aware prompt tuning, and data mixing, respectively.

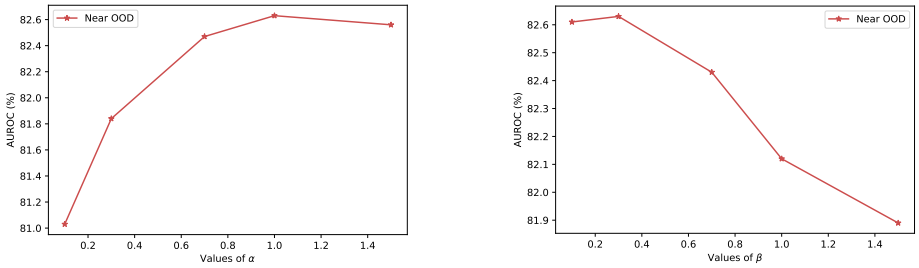
NegL	SamC	DisAwareP	DMix	AUROC ↑
✓	✗	✗	✗	76.92
✓	✓	✗	✗	77.85
✓	✓	✗	✓	79.01
✓	✓	✓	✗	79.20
✓	✓	✓	✓	82.63

Table A5: Results with different text prompts in data collection.

Prompts for data collection	FPR95 ↓	AUROC ↑
'<label>'	58.94	82.63
'The nice <label>'	58.89	82.57
'A photo of a <label>'	58.91	82.65

Table A6: AUROC results with different data mixing strategies.

Vanilla	Cross-modal Mixing	Cross-distribution Mixing	Near-OOD	Far-OOD
✓	✗	✗	79.20	93.15
✓	✓	✗	82.07	93.68
✓	✓	✓	82.63	94.26



(a) Values of α **(b)** Values of β

Fig. A1: Results with different values of α and β.

Table A7: OOD detection results of our LAPT with different VLMs architectures, where ImageNet-1K is used as the ID dataset.

Backbone	OOD datasets									
	INaturalist		Sun		Places		Textures		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
ResNet50	99.31	2.59	95.17	25.32	90.64	38.49	89.57	47.89	93.67	28.57
VITB/16	99.63	1.16	96.01	19.12	92.01	33.01	91.06	40.32	94.68	23.40
VITL/14	99.69	1.07	96.36	19.87	93.83	31.25	90.89	40.18	95.19	23.09

Table A8: The closest words to each of the four context vectors learned by LAPT, along with their distances in parentheses.

#	ID	OOD
1	given (1.5288)	nice (0.9968)
2	cot (1.3776)	eats (1.3191)
3	{ (1.9539)	bi (1.2240)
4	wooden (1.1061)	etsyshop (1.7584)

Table A9: Analyses on the time complexity of our LAPT and competitors. 'Training' measures the training time, and 'Param.' presents the number of learnable parameters. 'Test' reports the inference speed, measured with a batch size of 32. Results are achieved with a NVIDIA L40 GPU.

Methods	**Training**	**Test**	**Param.**	**FPR95** ↓
MCM [7]	–	10.2ms	–	43.93
NegLabel [5]	–	10.5ms	–	25.40
ZOC [4]	>24h	50.6ms	336M	85.19
CLIPN [13]	>24h	10.3ms	37.8M	31.10
LAPT (Ours)	1h	10.5ms	4K	23.40

References

1. Bitterwolf, J., Mueller, M., Hein, M.: In or out? fixing imagenet out-of-distribution detection evaluation. In: ICML (2023), https://proceedings.mlr.press/v202/bitterwolf23a.html
2. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13. pp. 446–461. Springer (2014)
3. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3606–3613 (2014)
4. Esmaeilpour, S., Liu, B., Robertson, E., Shu, L.: Zero-shot out-of-distribution detection based on the pre-trained model clip. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 6568–6576 (2022)
5. Jiang, X., Liu, F., Fang, Z., Chen, H., Liu, T., Zheng, F., Han, B.: Negative label guided OOD detection with pretrained vision-language models. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=xUO1HXz4an
6. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013)
7. Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., Li, Y.: Delving into out-of-distribution detection with vision-language representations. Advances in Neural Information Processing Systems **35**, 35087–35102 (2022)
8. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012)
9. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018)
10. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Open-set recognition: A good closed-set classifier is all you need? arXiv preprint arXiv:2110.06207 (2021)
11. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
12. Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4921–4930 (2022)
13. Wang, H., Li, Y., Yao, H., Li, X.: Clipn for zero-shot ood detection: Teaching clip to say no. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1802–1812 (2023)
14. Zhang, J., Inkawhich, N., Linderman, R., Chen, Y., Li, H.: Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5531–5540 (January 2023)
15. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)