Parameter-Efficient Semantic Augmentation for Enhancing Open-Vocabulary Object Detection

Supplementary Material

In this supplementary, we provide more detailed results for the experiments discussed in the paper. We also provide results for additional ablation studies.

1. More Details of OV-COCO$^+$

To further evaluate the generalization ability of OVOD models, we construct an extended benchmark, OV-COCO$^+$, which integrates COCO with downstream dataset and enlarges the unified category space to assess overall cross-domain generalization capability. Specifically, we merge three downstream datasets with substantial domain shift (i.e., ArTaxOr [1], DIOR [4], and UODD [3]) into the COCO dataset. During the integration, we combine all images and annotations into a single dataset and append the categories from each downstream dataset to the original COCO categories. If a category name already exists in COCO, we reuse the category ID in COCO to maintain consistency.

For example, the original COCO category set can be written as:

$$\{0 : \text{person},\ 1 : \text{bicycle},\ \ldots,\ 79 : \text{toothbrush}\},$$

and the ArTaxOr dataset contains 7 insect-related categories:

$$\{0 : \text{Araneae},\ 1 : \text{Coleoptera},\ \ldots,\ 6 : \text{Odonata}\}.$$

After integration, the unified category space of OV-COCO$^+$ simply appends these categories:

$$\{0{:}\text{person},\ 1{:}\text{bicycle},\ \ldots,\ 79{:}\text{toothbrush},\ 80{:}\text{Araneae},\ \ldots,\ 86{:}\text{Odonata}\}.$$

Similarly, the DIOR categories are appended in the same manner (where *airplane* follows the COCO naming convention):

$$\{0{:}\text{person},\ 1{:}\text{bicycle},\ \ldots,\ 79{:}\text{toothbrush},\ 80{:}\text{Expressway-Service-area},\ \ldots,\ 98{:}\text{windmill}\}.$$

For UODD, the unified category list extends to:

$$\{0{:}\text{person},\ 1{:}\text{bicycle},\ \ldots,\ 79{:}\text{toothbrush},\ 80{:}\text{seacucumber},\ 81{:}\text{seaurchin},\ 82{:}\text{scallop}\}.$$

During evaluation, we use the unified category labels as text input and perform open-vocabulary detection over all categories. The evaluation metric is mAP, computed as the mean AP50–95 across all categories.

Method	OV-COCO$^+$		
	w ArTaxOr	w DIOR	w UODD
Predefined	22.9	12.1	45.4
CoOp	6.2	17.3	24.0
AttriCLIP	6.5	15.0	2.0
MSPB	1.8	14.4	1.2
Predefined + SAR	49.5	45.4	50.3
CoOp + SAR	47.3	46.4	46.4
AttriCLIP + SAR	47.6	46.7	46.5
MSPB + SAR (ours)	**52.3**	**50.1**	**50.5**

Table 1. More results of the ablation on different prompt methods on OV-COCO$^+$.

Method	OV-COCO$^+$		
	w ArTaxOr	w DIOR	w UODD
MSPB + DDAS	40.7	40.8	49.0
MSPB + SAR (ours)	**52.3**	**50.1**	**50.5**

Table 2. More results of the ablation on different routing mechanisms on OV-COCO$^+$.

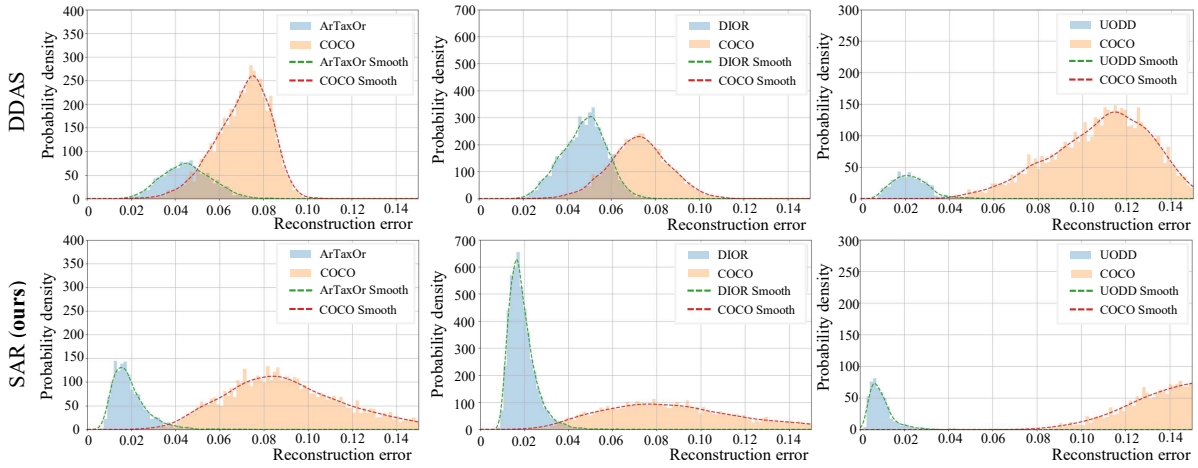
2. More Results of OV-COCO$^+$

Comparison of different textual semantic augmentation. We further compare MSPB with three textual semantic augmentation strategies on OV-COCO$^+$: (1) Predefined prompts ("a photo of a [CLS]") [7], (2) Learnable prompt as CoOp [10], and (3) Learnable prompt as AttriCLIP [8]. For a fair comparison, we replace only the textual augmentation strategy while keeping V-LoRA (*i.e.*, applying LoRA to the image encoder) unchanged. Here, "w" denotes the additional dataset included for joint evaluation (e.g., "w ArTaxOr" indicates evaluating OV-COCO jointly with ArTaxOr). As shown in Table 1, the performance remains relatively low under this setting regardless of the textual semantic augmentation strategy used. However, once SAR is incorporated, all methods exhibit clear performance improvements. Notably, the combination of our proposed MSPB with SAR achieves the highest mAP across all datasets.

Comparison of different routing mechanisms. Similarly, we replace the proposed SAR with its baseline routing mechanism, DDAS [9], on OV-COCO$^+$. As shown in Table 2, our SAR consistently achieves the highest performance across all datasets.

Designs of MSPB	$\text{mAP}_{\text{ArTaxOr}}$	mAP_{DIOR}	mAP_{UODD}	mAP_{avg}
Using image feature maps from all scales (*i.e.*, $S=4$)	78.5	57.9	48.0	61.5
Reversing the order of prompts prepended to the category labels	78.2	56.1	48.0	60.8
Non-repetitive selection of (key, prompt) pairs	77.6	**58.8**	48.1	61.5
3 (key, prompt) pairs per scale	78.5	54.8	48.1	60.5
4 (key, prompt) pairs per scale	78.4	55.5	48.6	60.8
5 (key, prompt) pairs per scale	76.4	57.3	48.3	60.7
6 (key, prompt) pairs per scale	77.5	57.5	47.9	61.0
Ours	**79.1**	57.7	**48.9**	**61.9**

Table 3. Ablation on more designs of the multi-scale prompt bank.



(a) Trained on ArTaxOr (b) Trained on DIOR (c) Trained on UODD

Figure 1. The reconstruction error distributions of autoencoders trained on different downstream datasets (*i.e.*, ArTaxOr, DIOR and UODD) for samples from downstream datasets or COCO.

3. More Designs of the Multi-Scale Prompt Bank

Beyond the design of the multi-scale prompt bank described in the paper, we further explore additional variations, focusing on the following four aspects:

1) Whether the multi-scale prompt bank utilizes feature maps from all scales of the image encoder (*i.e.*, $S=4$). In the paper, we only use feature maps from the first 3 scales of the image encoder to select the most relevant keys and prompts (*i.e.*, $S=3$). This leads to the length of the learnable vectors prepended to the category label embeddings as actually $S\times M=3\times 12=36$. For fair comparison, we keep this length unchanged, so M is set to 9 in this experimental setting (*i.e.*, $S\times M=4\times 9=36$).

2) The ordering of the scale-aware prompts prepended to the embeddings of the downstream category labels. In the paper, prompts selected for larger-scale image feature maps are concatenated in earlier positions, *i.e.*, $\mathbf{t}_k(\tilde{\mathcal{P}}) = \text{concat}(\mathbf{P}_{i_1};\ldots;\mathbf{P}_{i_S};[\text{CLS}]_k)$. Here, we invert this order, *i.e.*, prompts selected from smaller-scale image feature maps are placed in earlier positions, forming $\mathbf{t}_k(\tilde{\mathcal{P}}) = \text{concat}(\mathbf{P}_{i_S};\ldots;\mathbf{P}_{i_1};[\text{CLS}]_k)$.

3) Whether prompts can be repeatedly selected by image feature maps of different scales. In the main text, $\tilde{\mathbf{z}}^s$ of different scales can select the same (key, prompt) pairs. Here, we test an alternative approach where a (key, prompt) pair can be selected by $\tilde{\mathbf{z}}^s$ only one time. Specifically, $\tilde{\mathbf{z}}^1$ first selects the most relevant key and prompt. Subsequent selections by $\{\tilde{\mathbf{z}}^s\}_{s=2}^{S}$ then exclude the key and prompt chosen by $\tilde{\mathbf{z}}^1$, and so on. This ensures that $\tilde{\mathbf{z}}^s$ of different scales do not select the same (key, prompt) pair.

4) Whether to assign independent prompt banks to each scale. In the paper, all (key, prompt) pairs are selectable by $\tilde{\mathbf{z}}^s$ of different scales. Here, we group (key, prompt) pairs, with each group corresponding to $\tilde{\mathbf{z}}^s$ of a specific scale. For the number of (key, prompt) pairs in each group, we test 3, 4, 5, and 6 respectively.

The results of the above experiments are presented in Table 3. We analyze these results as follows:

1) For using image feature maps from all scales to select (key, prompt) pairs, the results indicate that introducing the additional final scale does not further enhance the semantic representation capability of the bank. Instead, it leads to a

τ	ArTaxOr			DIOR			UODD			H_{mean}
	mAP_{tgt}	mAP_{coco}	H	mAP_{tgt}	mAP_{coco}	H	mAP_{tgt}	mAP_{coco}	H	
0.01	5.3	50.6	9.6	3.7	50.6	6.90	20.7	50.6	29.4	15.3
0.02	51.7	50.6	51.1	40.7	50.6	45.1	45.4	50.6	47.9	48.0
0.03	71.8	50.4	59.2	55.2	50.2	52.6	48.0	50.6	49.3	53.7
0.035	75.5	50.2	60.3	56.8	49.9	53.1	48.6	50.6	49.6	54.3
0.036	75.6	50.1	60.3	57.0	49.8	53.2	48.6	50.6	49.6	54.3
0.037	76.0	50.1	60.4	57.1	49.6	53.1	48.6	50.6	49.6	54.4
0.038	76.5	50.0	60.5	57.2	49.4	53.0	48.6	50.6	49.6	54.4
0.039	76.8	49.9	60.5	57.3	49.3	53.0	48.6	50.6	49.6	54.4
0.04	76.8	49.8	60.4	57.4	49.2	53.0	48.6	50.6	49.6	54.3
0.05	78.5	47.6	59.3	57.3	46.8	51.6	48.9	50.6	49.7	53.5
0.06	78.8	43.8	56.3	57.7	43.2	49.4	48.9	50.6	49.7	51.8
0.07	78.9	38.9	52.1	57.7	38.3	46.0	48.9	50.6	49.7	49.3
0.08	79.0	32.3	45.9	57.7	33.1	42.1	48.9	50.5	49.7	45.9
0.09	79.0	24.7	37.6	57.7	28.3	38.0	48.9	50.4	49.6	41.7
0.10	79.1	18.3	29.7	57.7	23.9	33.8	48.9	49.9	49.4	37.6
0.11	79.1	13.3	22.8	57.7	20.3	30.0	48.9	49.1	49.0	33.9
0.12	79.1	9.4	16.8	57.7	17.3	26.6	48.9	47.3	48.1	30.5
0.13	79.1	6.5	12.0	57.7	15.2	24.1	48.9	44.4	46.5	27.5
0.14	79.1	4.7	8.9	57.7	13.7	22.1	48.9	40.8	44.5	25.2
0.15	79.1	2.9	5.6	57.7	11.9	19.7	48.9	35.9	41.4	22.2

Table 4. Ablation on the routing decision threshold τ. H_{mean} denotes the mean value of H across all downstream datasets (*i.e.*, ArTaxOr, DIOR, UODD).

Components			ArTaxOr			DIOR			UODD		
V-LoRA	MSBP	SAR	mAP_{tgt}	mAP_{coco}	H	mAP_{tgt}	mAP_{coco}	H	mAP_{tgt}	mAP_{coco}	H
			1.4	50.6	2.7	3.0	50.6	5.7	2.9	50.6	5.5
✓			61.6	22.7	33.2	47.8	5.5	9.9	45.5	46.2	45.8
	✓		22.5	0.2	0.4	17.4	20.2	18.7	21.5	0.7	1.4
✓	✓		79.1	0.5	1.0	57.7	9.2	15.9	48.9	1.5	2.9
✓		✓	59.5	50.4	54.6	45.2	49.5	47.3	45.1	50.6	47.7
	✓	✓	22	50.3	30.6	17.3	49.9	25.7	21.4	50.6	30.1
✓	✓	✓	76.8	49.9	**60.5**	57.3	49.3	**53.0**	48.6	50.6	**49.6**

Table 5. More detailed results of the ablation on different components of our framework.

marginal degradation in the model's adaptability to downstream tasks.

2) For reversing the order of prompts prepended to the category labels, we observe that it leads to a decline in the model's adaptability. This is likely because large-scale features capture global semantics and provide stable initial guidance for the model, while small-scale features focus on local details. Starting the text sequence without comprehensive global information tends to introduce noise, adversely affecting the overall performance.

3) Compared to the non-repetitive selection of (key, prompt) pairs, we observe that allowing $\tilde{\mathbf{z}}^s$ of different scales to repeatedly select the same (key, prompt) pair yields more balanced and robust results.

4) Compared to assigning a specific group of prompt banks independently to $\tilde{\mathbf{z}}^s$ of each scale, we observe that the ungrouped prompt bank achieves better performance. This likely stems from the ungrouped prompt bank's ability to share information across scales, allowing each (key, prompt) pair to learn semantics from multiple feature levels. This enhances the learning of domain-specific knowledge while maintaining semantic diversity and reusability. In contrast, using independent prompt banks tends to fragment semantic information, ultimately limiting the model's adaptability to downstream domain knowledge.

Method	ArTaxOr			DIOR			UODD		
	mAP_{tgt}	mAP_{coco}	H	mAP_{tgt}	mAP_{coco}	H	mAP_{tgt}	mAP_{coco}	H
Predefined	61.6	22.7	33.2	47.8	5.5	9.9	45.5	46.2	45.8
CoOp	74.2	0.9	1.8	55.5	10.3	17.4	47.4	23.6	31.5
AttriCLIP	78.8	1.0	2.0	56.9	7.1	12.6	48.5	0.6	1.2
MSPB	79.1	0.5	1.0	57.7	9.2	15.9	48.9	1.5	2.9
Predefined + SAR	59.5	50.4	54.6	45.2	49.5	47.3	45.1	50.6	47.7
CoOp + SAR	71.9	47.4	57.1	55.1	47.7	51.1	47.2	48.7	48.0
AttriCLIP + SAR	76.5	47.8	58.8	56.5	47.4	51.6	48.3	48.7	48.5
MSPB + SAR **(ours)**	76.8	49.9	**60.5**	57.3	49.3	**53.0**	48.6	50.6	**49.6**

Table 6. More detailed results of the ablation on different prompt methods.

Method	ArTaxOr			DIOR			UODD		
	mAP_{tgt}	mAP_{coco}	H	mAP_{tgt}	mAP_{coco}	H	mAP_{tgt}	mAP_{coco}	H
MSPB + DDAS	75.7	36.2	49.0	54.9	40.2	46.4	48.9	38.5	43.1
MSPB + SAR **(ours)**	76.8	49.9	**60.5**	57.3	49.3	**53.0**	48.6	50.6	**49.6**

Table 7. More detailed results of the ablation on different routing mechanisms.

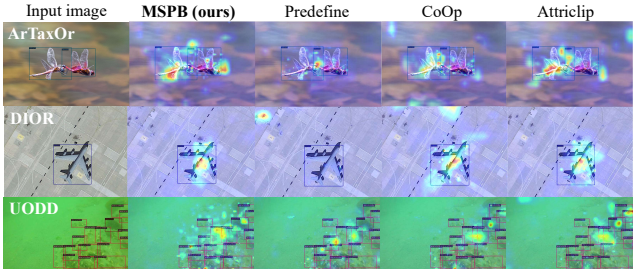


Figure 2. Visualization of the selected prompts at different textual semantic augmentation using Grad-CAM[6].

4. More Ablations on the Semantic-Aware Router

The reconstruction error distribution. We visualize the reconstruction error distributions of autoencoders trained on different downstream datasets (*i.e.*, ArTaxOr [1], DIOR [4], and UODD [3]) for samples from both downstream datasets and the general domain (*i.e.*, COCO [5]). Fig. 1 compares our proposed SAR with the baseline method, DDAS [9]. The results show that SAR consistently reduces the overlap in reconstruction errors between the target datasets and COCO, making the two data distributions clearly separated and more distinguishable. These findings further validate our conclusion: explicit modeling of content and style information significantly enhances the router's ability to distinguishing between different data distributions, thereby improving routing accuracy and effectively balancing the model's adaptability and open-vocabulary generalization.

Ablation on τ with more detailed numbers. As shown in Table 4, we comprehensively evaluate the model's detection performance across different SAR routing decision thresholds τ. The results indicate that the model's performance initially improves with increasing τ, reaching a peak at $\tau = 0.039$, after which performance begins to decline. Therefore, we adopt $\tau = 0.039$ as the optimal threshold for all other experiments.

5. More Detailed Results for Other Experiments

Effectiveness of different components. In addition to the ArTaxOr dataset, we also evaluate the effectiveness of different components (*i.e.*, V-LoRA [2], MSPB and SAR) in our framework on the DIOR and UODD datasets. The results are reported in Table 5. The results on these datasets all align with the conclusions analyzed in the paper, demonstrating the effectiveness of these core designs in our framework.

Different prompt methods. Table 6 presents detailed results for various prompt methods (*i.e.*, T-LoRA [2], CoOp [10], AttriCLIP [8], and our proposed MSPB) across all three datasets. The observed trends are consistent with previous discussions in the paper, emphasizing that our proposed MSPB consistently achieves the best balance between domain-specific adaptation and open-vocabulary generalization, resulting in the highest H across all datasets. We also provide results without the proposed SAR module in Table 6, clearly demonstrating that incorporating SAR improves the harmonic mean for all prompt strategies. Notably, MSPB combined with SAR delivers the best overall performance. These results further highlight the effectiveness of our method.

Different routing mechanisms. We provide more detailed numbers in Table 7, which compares the proposed SAR with the baseline DDAS. SAR consistently outper-

forms DDAS across all datasets.

Visualization of prompts. We provide additional visualizations across more datasets and methods in Fig. 2. The observations are consistent with those reported in the paper.

References

[1] Geir Drange. Arthropod taxonomy orders object detection dataset, 2019. 1, 4

[2] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 1(2): 3, 2022. 4

[3] Lihao Jiang, Yi Wang, Qi Jia, Shengwei Xu, Yu Liu, Xin Fan, Haojie Li, Risheng Liu, Xinwei Xue, and Ruili Wang. Underwater species detection using channel sharpening attention. In *Proceedings of the ACM International Conference on Multimedia*, pages 4259–4267, 2021. 1, 4

[4] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020. 1, 4

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[6] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2017. 4

[7] Hao Wang, Pengzhen Ren, Zequn Jie, Xiao Dong, Chengjian Feng, Yinlong Qian, Lin Ma, Dongmei Jiang, Yaowei Wang, Xiangyuan Lan, et al. Ov-dino: Unified open-vocabulary detection with language-aware selective fusion. *arXiv preprint*, 2024. 1

[8] Runqi Wang, Xiaoyue Duan, Guoliang Kang, Jianzhuang Liu, Shaohui Lin, Songcen Xu, Jinhu Lü, and Baochang Zhang. Attriclip: A non-incremental learner for incremental knowledge learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3654–3663, 2023. 1, 4

[9] Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23219–23230, 2024. 1, 4

[10] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 4