Learning to Diversify and Focus: A Reinforcement Framework for Open-Vocabulary HOI Detection

Yongchao Xu[1] Jiawei Liu[1*] Junfeng Wang[1] Sen Tao[2] Na Jiang[3] Zheng-Jun Zha[1]
[1]University of Science and Technology of China, Hefei, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3]Capital Normal University, Beijing, China
{yongchaoxu, jfwang2002}@mail.ustc.edu.cn taosen23@mail.ucas.ac.cn
jiangna@cnu.edu.cn {jwliu6, zhazj}@ustc.edu.cn

In the supplementary material, we provide a comprehensive evaluation of our SD-IF framework. Specifically, Section 1 introduces the experimental setting, including datasets and evaluation metric. Section 2 presents additional experimental results, including hyper-parameter analysis and the transferability of our proposed modules when integrated into existing OV-HOI approaches.

1. Experimental Setting

Datasets. We evaluate the proposed SD-IF on two public benchmarks, HICO-DET [1] and SWIG-HOI [8]. The HICO-DET labels 80 object classes, 117 interaction classes, and 600 HOI categories. Following previous works[3–5, 9], we hold out 120 rare HOIs to simulate the open vocabulary setting. The SWIG-HOI has unseen HOI classes naturally in the test set. At inference time, SWIG-HOI requires a comprehensive evaluation of performance on 1391 unseen, 2841 rare, and 1307 non-rare HOI classes, which puts a high demand on the generalization of the model.

Evaluation Metric. We use the mean Average Precision (mAP) as the evaluation metric on these datasets. Following prior works [2, 3, 6, 7, 9–11], a predicted HOI triplet is regarded as a true-positive HOI sample when the Intersection over Union (IoU) between the detected human and object bounding boxes and the corresponding ground-truth boxes is larger than 0.5, and the interaction class prediction is correct simultaneously.

2. More Experimental Results

Hyperparameters of the SD module. In the Semantic Diversification (SD) module, the hyperparameter κ controls the magnitude of semantic perturbation applied to each query, γ balances semantic deviation against visual-feature consistency, and β regulates the trade-off between exploration and stability during policy learning. As reported in Table 1, we conduct a systematic ablation to assess the influence of these three hyperparameters. The results show that setting $\kappa = 0.2$, $\gamma = 0.5$, and $\beta = 0.5$ yields the best overall performance. This configuration introduces sufficiently diverse semantic variations while maintaining visual coherence, enabling SD-IF to achieve the most effective semantic expansion without destabilizing training.

Hyperparameters of the IF module. In the Interaction Focusing (IF) module, the coefficient η controls the balance between spatial focusing and semantic consistency within the hybrid reward. As shown in Table 2, the model achieves the best performance when $\eta = 0.2$, indicating that spatial focusing plays a more dominant role in enhancing the model's generalization ability. This highlights the importance of accurately attending to interaction-critical regions in open-vocabulary HOI detection.

Module transferability analysis. To further validate the effectiveness and generality of the proposed modules, we integrate SD and IF into the CMD-SE [3] framework and evaluate their contributions. As shown in Table 3, adding either SD or IF individually yields consistent performance gains, while incorporating both leads to an additional improvement, demonstrating that our modules can be seamlessly integrated into existing OV-HOI methods. It is worth noting that CMD-SE equipped with both modules still performs slightly below our SD-IF framework. We attribute this to the fine-grained semantic information introduced by CMD-SE via LLM, which may not align well with the semantic exploration strategy of SD, thereby limiting the overall generalization benefit. In contrast, SD-IF does not rely on any external LLM-based priors and achieves stronger generalization in open-world scenarios with a more lightweight and self-contained design.

Performance under varying numbers of Unseen

*Corresponding author

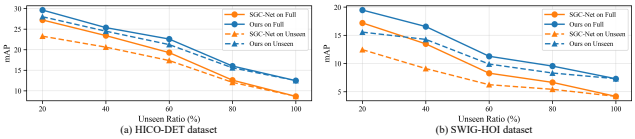

Figure 1. Performance comparisons under different Unseen ratios.

κ	γ	β	Non-rare	Rare	Unseen	Full
0.0	0.0	0.0	22.42	16.25	13.03	16.90
0.2	0.0	0.0	22.77	17.01	13.51	17.49
0.5	0.0	0.0	22.59	16.92	13.21	17.32
1.0	0.0	0.0	22.51	16.85	13.18	17.26
0.2	0.5	0.0	24.78	17.01	13.79	18.03
0.2	1.0	0.0	23.50	16.95	13.82	17.71
0.2	2.0	0.0	23.43	16.97	13.57	17.64
0.2	0.5	0.5	**25.01**	**18.86**	**15.57**	**19.48**
0.2	0.5	1.0	24.97	18.80	15.44	19.41

Table 1. Ablation studies of the hyperparameters of the SD module on the SWIG-HOI [8] dataset.

η	Non-rare	Rare	Unseen	Full
0.0	24.83	18.70	15.39	19.32
0.2	**25.01**	**18.86**	**15.57**	**19.48**
0.3	24.89	18.80	15.49	19.41
0.4	24.92	18.77	15.43	19.38
0.5	24.88	18.79	15.40	19.38
1.0	24.85	18.82	15.45	19.40
2.0	24.73	18.71	15.41	19.30

Table 2. Ablation studies of the hyperparameters of the IE module on the SWIG-HOI dataset.

classes. We evaluate varying seen/Unseen proportions by increasing the Unseen HOI ratio from 20% to 100% under the same training protocol. As shown in Figure 1, performance drops as the Unseen ratio increases, yet our method consistently outperforms the SOTA SGC-Net on both Unseen and Full across all settings. The gain is most pronounced at moderate ratios (20%–60%) and remains at 100% Unseen, indicating robust generalization across different openness levels.

Method	Non-rare	Rare	Unseen	Full
CMD-SE [3]	21.46	14.64	10.70	15.26
+ *SD*	22.39	16.58	14.29	17.38
+ *IF*	24.37	16.50	13.18	17.52
+ *SD+IF*	24.61	18.42	15.04	19.03
SD-IF (Ours)	**25.01**	**18.86**	**15.57**	**19.48**

Table 3. Module transferability analysis of the proposed SD-IF.

References

[1] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 1

[2] Bumsoo Kim, Jonghwan Mun, Kyoung-Woon On, Minchul Shin, Junhyun Lee, and Eun-Sol Kim. Mstr: Multi-scale transformer for end-to-end human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19578–19587, 2022. 1

[3] Ting Lei, Shaofeng Yin, and Yang Liu. Exploring the potential of large foundation models for open-vocabulary hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16657–16667, 2024. 1, 2

[4] Ting Lei, Shaofeng Yin, Qingchao Chen, Yuxin Peng, and Yang Liu. Open-vocabulary hoi detection with interaction-aware prompt and concept calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23945–23957, 2025.

[5] Xin Lin, Chong Shi, Zuopeng Yang, Haojin Tang, and Zhili Zhou. Sgc-net: Stratified granular comparison network for open-vocabulary hoi detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4539–4549, 2025. 1

[6] Jiawei Liu, Yongchao Xu, Sen Tao, Yuexuan Qi, and Zheng-Jun Zha. Mamba-driven comprehensive context learning for zero-shot hoi detection. *International Journal of Computer Vision*, 134(1):10, 2026. 1

[7] Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23507–23517, 2023. 1

[8] Suchen Wang, Kim-Hui Yap, Henghui Ding, Jiyan Wu, Junsong Yuan, and Yap-Peng Tan. Discovering human interactions with large-vocabulary objects via query and multi-scale detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13475–13484, 2021. 1, 2

[9] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2022. 1

[10] Yongchao Xu, Jiawei Liu, Sen Tao, Qiang Zhang, and Zheng-Jun Zha. Hoimamba: Efficient mamba-based disentangled progressive learning for hoi detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8987–8995, 2025.

[11] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20104–20112, 2022. 1