Supplementary Material for Open-Vocabulary HOI Detection with Interaction-aware Prompt and Concept Calibration

Ting Lei[1] Shaofeng Yin[1] Qingchao Chen[2] Yuxin Peng[1] Yang Liu[1*]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]National Institute of Health Data Science, Peking University
{ting_lei, qingchao.chen, pengyuxin, yangliu}@pku.edu.cn yin_shaofeng@stu.pku.edu.cn

In this supplementary material, we provide a comprehensive evaluation of our INP-CC approach. In Sec. 1, we present a detailed description of the implementation details of the proposed method. Next, in Sec. 2, we compare the model efficiency of our approach with existing methods. In Section 3, we perform an in-depth analysis of the decomposition mechanism, the adaptive selection process for interaction-aware prompts, and the impact of different interaction classifier choices within our model. Sec. 4 showcases additional qualitative results to further illustrate the characteristics of our approach. In Sec. 5, we discuss the limitations of our method and suggest potential future research directions.

1. Implementation Details

In this section, we present a comprehensive description of the implementation details of our model. We employ the ViT-B/16 version as our visual encoder following [1, 3]. We set the size of the interaction-aware prompt to 128 and 8 on SWIG-HOI and HICO-DET datasets, respectively. We set k in the selection mechanism to 2. We set the cluster number J to 64 and select 10 hard negative samples during each iteration. The number of layers of the HOI decoder is set to 4. We set the cost weights λ_b, λ_{iou}, λ_{cls} and λ_d to 5, 2, 5, and 5 during training. We use focal loss [2] for interaction classification to counter the imbalance between positive and negative examples. We set γ to 2 during inference. We introduce 8 prefix tokens and 2 conjunctive tokens to connect the words of human actions and objects following [3] when constructing T_{hoi} introduced in Sec. 3.1. We set the batch size as 16 with a learning rate of 10^{-4} and use the Adam optimizer with decoupled weight decay regularization. We train our model for 80 epochs with a batch size of 128 on 2 NVIDIA 3090 GPUs.

Method	#Params	#FLOPs	mAP
THID [3]	239.08M	53.28G	13.26
CMD-SE [1]	227.79M	42.99G	15.75
INP-CC (Ours)	228.98M	28.45G	16.74

Table 1. Model statistics.

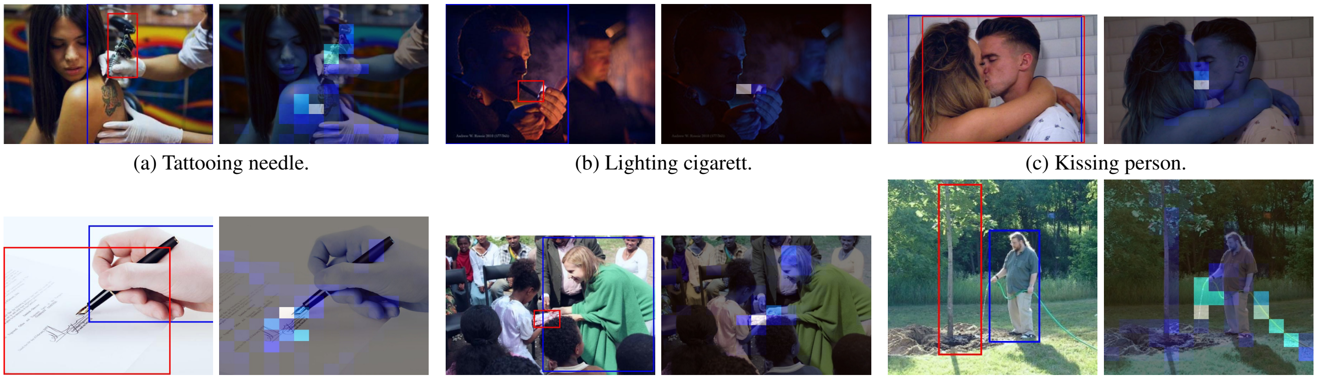
P_C	$\hat{P}_{IT}$	Non-rare	Rare	Unseen	Full
$h.r.$	-	21.41	15.02	10.12	15.30
$l.r.$	-	20.57	15.11	10.53	15.25
$h.r.$	$h.r.$	22.01	16.13	10.71	16.31
$h.r.$	$l.r.$	**22.84**	**16.74**	**11.02**	**16.74**
$l.r.$	$l.r.$	21.66	16.20	10.73	16.19

Table 2. Ablation of the low-rank decomposition of interaction prompts on the SWIG-HOI dataset. P_C: common prompts. $\hat{P}_{IT}$: interaction prompts. $h.r.$: naive high-rank implementation. $l.r.$: low-rank decomposition.

2. Model Efficiency

We further compare the computational cost of our model with previous open-vocabulary HOI detectors [1, 3] on the SWIG-HOI dataset in Tab. 1, emphasizing the advantages of our approach. Specifically, we use fvcore to compute the FLOPs of each model and count their parameter numbers. For consistency, a batch size of 1 is used for all models when calculating FLOPs. Our proposed INP-CC demonstrates superior efficiency, requiring fewer FLOPs than CMD-SE [1], despite having a similar number of parameters. While CMD-SE's multi-level decoding results in higher FLOPs, our model leverages interactive-aware prompts to achieve better performance with a more computationally efficient design.

[*] Corresponding author



(a) Tattooing needle.

(b) Lighting cigarett.

(c) Kissing person.

(d) Signing document.

(e) Distributing gift.

(f) Watering tree.

Figure 1. Qualitative Examples.

k	Non-rare	Rare	Unseen	Full
1	22.30	15.93	11.36	16.29
2	22.84	**16.74**	**11.02**	**16.74**
4	**22.89**	16.17	10.86	16.42

Table 3. Ablation of the selection mechanism. k: the number of selected interaction prompts.

Classifier	Non-rare	Rare	Unseen	Full
T_{hoi}	21.90	16.22	10.73	16.25
$T_{hoi} + T_{vis}$	**22.84**	**16.74**	**11.02**	**16.74**

Table 4. Ablation of the interaction classifier choice.

Knowledge Base	Non-rare	Rare	Unseen	Full
ConceptNet	21.02	14.47	10.63	14.83
WordNet	21.15	14.62	10.71	14.92
GPT (Ours)	**22.84**	**16.74**	**11.02**	**16.74**

Table 5. Performance comparison on different knowledge bases.

Strategy	Non-rare	Rare	Unseen	Full
Object-based [4]	22.24	15.81	10.36	15.96
Semantic-based	**22.84**	**16.74**	**11.02**	**16.74**

Table 6. Ablation of the hard negative strategy.

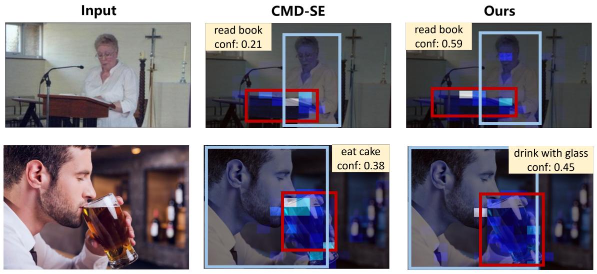


Figure 2. Qualitative examples to previous SOTA (CMD-SE).

3. Ablation Study

In this section, we empirically investigate the sensitivity of the proposed method to the low-rank decomposition and the selection mechanism, and the effect of different classifiers on the open-vocabulary SWIG-HOI dataset. Specifically, besides the four aspects of INP-CC we have discussed in Sec. 4.3, we further ablate on (1) the importance of the low-rank decomposition mechanism, (2) the topk selection mechanism, (3) the interaction classifier choice, (4) the external knowledge base, and (5) the hard negative strategy.

Low-rank Decomposition. We analyze the significance of the low-rank decomposition mechanism in Tab. 2. Our findings are as follows: (1) As seen in lines 1–2, when only common prompts are used, the high-rank ($h.r.$) method achieves slightly better performance than the low-rank ($l.r.$) method on non-rare interactions, but performs slightly worse on rare and unseen interactions. Overall, $h.r.$ achieves a full mAP of 15.30, marginally outperforming $l.r.$'s 15.25. This suggests that, in the absence of interaction-specific prompts, the high-rank method has a slight edge, likely due to the loss of expressive power in the low-rank method when not tuned for specific interactions. (2) Incorporating interaction-specific prompts significantly boosts performance across all categories, regardless of the decomposition method. For example, using $h.r.$+$h.r.$ improves full performance from 15.30 to 16.31. Notably, the combination of $h.r.$ for P_C and $l.r.$ for $\hat{P}_{IT}$ yields the best overall results, demonstrating that low-rank decomposition for interaction-specific prompts enhances the representation of nuanced interactions, especially for rare and unseen cases. (3) When both common and interaction-specific prompts are processed with low-rank decomposition ($l.r.$+$l.r.$), there is a slight drop in full performance, from 16.74 ($h.r.$+$l.r.$) to 16.19. This indicates that while low-rank decomposition is effective for interaction-specific prompts, high-rank decomposition for common prompts provides a stronger foundational representation, helping to preserve global context.

Topk Selection Mechanism. Tab. 3 illustrates the impact of the number of selected interaction prompts (k) on model performance. Our findings reveal the following trends: (1) Non-rare classes: Performance improves as k increases, peaking at $k = 4$ (22.89), indicating that additional prompts better capture common interaction patterns. (2) Rare and Unseen classes: The best performance is achieved at $k = 2$. A single prompt ($k = 1$) lacks sufficient diversity, while too many prompts ($k = 4$) introduce redundancy or noise, leading to suboptimal outcomes. (3) Overall performance (all classes): $k = 2$ yields the highest overall score (16.74), effectively balancing the trade-offs across Non-rare, Rare, and Unseen categories. In conclusion, selecting $k = 2$ offers the optimal balance between specificity and generalization, particularly benefiting Rare and Unseen scenarios while maintaining competitive performance for Non-rare classes.

Interaction Classifier Choice. In Tab. 4, we present the ablation results for the interaction classifier choice, comparing the performance of two configurations: (1) T_{hoi} only: This uses the CLIP text encoder to encode action and object tokens along with learnable tokens following [3]. (2) T_{hoi} + T_{vis}: This combines the T_{hoi} embeddings with the Instructor Embedding derived from the language model description. We find that the addition of the Instructor Embedding (T_{vis}) consistently improves performance across the different categories. The overall gains suggest that T_{vis} enhances the model's ability to handle a broader range of interactions, especially those it hasn't encountered during training.

External knowledge bases. We compare different external knowledge sources to assess their suitability for HOI contexts in Tab. 5. While ConceptNet and WordNet provide structured commonsense knowledge, they are not specifically designed for HOI tasks. By evaluating their coverage using SWIG-HOI labels, we observe that 11.75% of HOI concepts are missing in ConceptNet, and 11.54% are absent in WordNet. For the uncovered cases, we default to using the original HOI names. As shown in Tab. 5, our GPT-based generative approach outperforms both ConceptNet and WordNet, demonstrating superior contextual reasoning and generalization capabilities.

Hard negative strategy. We investigate the impact of the hard negative sampling strategy with [4]. Unlike the object-based negative sampling of [4], which focuses on object-level interactions, our method utilizes semantic clustering to identify similar interactions—across different object types (e.g., "hold paintbrush" vs. "hold pen"). These hard negatives encourage the model to focus on fine visual details, such as hand position or object shape, which helps it prioritize more essential cues for distinguishing between similar interactions. This approach is particularly beneficial when generalizing to unseen interactions that involve subtle differences between categories. As shown in Tab. 6, our method outperforms [4] by 0.78%, validating the advantage of using semantic-based negative sampling.

4. Qualitative Examples

In Fig. 1, we showcase more scenarios to demonstrate our model's robust performance in detecting diverse HOIs. The left images show accurate predictions with well-localized bounding boxes, while the right attention maps highlight regions critical for each interaction. For instance, the model successfully focuses on the tattoo needle and skin (Fig. 1a), the cigarette under low lighting (Fig. 1b), the pen-paper interaction during signing (Fig. 1d), and the hose used for watering a tree (Fig. 1f). These results highlight the model's ability to generalize across scenes and capture fine-grained human-object relationships, enabled by interaction-aware prompts and concept calibration.

We further present additional qualitative comparisons with the previous state-of-the-art CMD-SE model in Fig. 2. The figure demonstrates that our model exhibits superior attention to critical regions, such as the eyes (row 1), and more effectively distinguishes fine-grained actions, such as drinking versus eating (row 2).

5. Limitations

While negative sampling enhances the model's ability to distinguish between visually similar but semantically distinct actions, it may introduce challenges if the selected negative samples fail to effectively represent real-world scenarios. This could result in biases or overfitting to specific types of negative samples. In the future, we plan to explore semi-supervised or self-supervised approaches for enhanced inter-modal similarity modeling, which could improve performance. Future research could also focus on improving the model's ability to generalize across various domains (e.g., social media images, surveillance footage, etc.), which often present different visual characteristics and interaction types. We believe these improvements will lead to better overall performance, making our method more practical for real-world applications.

References

[1] 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] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1

[3] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detectors with natural language supervision. In *CVPR*, 2022. 1, 3

[4] Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13234–13243, 2021. 2, 3