
InstructHOI: Context-Aware Instruction for Multi-Modal Reasoning in Human-Object Interaction Detection

Supplementary Materials

The supplementary materials are organized as follows. In Section 1, we further investigate the effects of different Large Foundation Models (LFMs), HOI-domain fine-tuning strategies, advanced object detector and the number of representative tokens, respectively. In Section 2, we further compare the zero-shot HOI detection capabilities of fine-tuned and pretrained LFM. In Section 3, we present the qualitative results of Interest Token Selector (ITS), two interaction prediction branches of InstructHOI, failure case analysis and in-the-wild HOI detection. In Section 4, we provide additional details about the HOI-domain fine-tuning process. In Section 5, we further offer discussion on the limitation and social impact of InstructHOI.

1 Ablation Study

For a fair comparison, all HOI detectors in the experiments use the same “Resnet50” as backbone.

Effect of Different LFMs To investigate the impact of different Large Foundation Models (LFMs) on the performance of our InstructHOI, we develop variants of InstructHOI that utilize different LFMs, including LLaVA-OV [1] and InternVL2 [2]. Additionally, for each LFM variant, we further evaluate the effect of foundation model scales. As summarized in Table 1, the results on the V-COCO dataset indicate that different LFMs can influence HOI detection performance. Overall, by leveraging the reasoning capabilities of LFMs, our InstructHOI achieves state-of-the-art performance with both LLaVA-OV and InternVL2. Furthermore, as the scale of LFMs increases, InstructHOI demonstrates performance improvement, likely due to enhanced reasoning capabilities at larger model scales.

Effect of HOI-Domain Fine-tuning Strategies In Table 2, we evaluate the impact of fine-tuning strategies on the performance of InstructHOI. The “pretrained” indicates pretrained LFM without fine-tuning; “fine-tune[†]” indicates fine-tuning on the original training dataset (i.e., HICO-DET or V-COCO); “fine-tune[‡]” indicates fine-tuning on the generated dataset (see subsection 3.2). Specifically, the “fine-tune[†]” strategy outperforms the “pretrained” strategy by 1.68 mAP in the full settings of HICO-DET and by 1.0 mAP in $AP_{role}^{\#1}$ of V-COCO, respectively. The results indicate that the LoRA fine-tuning can effectively bridge the knowledge gap between general and HOI domains, thus enhancing the interaction reasoning capabilities of LFMs. Furthermore, the “fine-tune[‡]” strategy outperforms the “fine-tune[†]” strategy by 0.59 mAP and 3.31 mAP in the full and rare settings on the HICO-DET dataset, respectively. This highlights that fine-tuning on the generated large-scale dataset can improve the generalization capabilities of InstructHOI, enabling it to better recognize rare HOI categories.

Effect of Representative Token Number in ITS To evaluate the effect of representative token number for instructions in ITS, we also conduct experiments on the two benchmarks, as shown in Table 3. As observed, compared to *All* (which represents selecting all the local image tokens as representative tokens), the best performance is achieved when the token number is set to 20. This indicates that the representative token selection process could effectively filters out irrelevant local image tokens from the instructions, providing effective context guidance for interaction reasoning.

Effect of Advanced Object Detector To ensure a fair comparison, InstructHOI utilizes the most commonly adopted object detector (i.e., DETR [6]) in all the experiments of the manuscript. To investigate the impact of a more advanced detector, we replace DETR with DINO [7]. As shown in Table 4, incorporating a more advanced detector significantly improves InstructHOI’s performance,

Table 1: Results on V-COCO dataset for InstructHOI using different LFM. “**TP**” denotes the trainable parameters of InstructHOI modules. “*” indicates the trainable parameters reported in paper [3]. “FT511B” represents the language model FlanT5XXL_{11B}.

Methods - <i>LFMs</i>	Foundation Models	TP	$AP_{role}^{\#1}$	$AP_{role}^{\#2}$
UniHOI [4] - <i>BLIP-2</i> [5]	ViT-L OPT _{2.7B}	52.3M*	65.6	68.3
UniHOI [4] - <i>BLIP-2</i> [5]	ViT-G FT5 _{11B}	-	65.8	68.6
InstructHOI - <i>LLaVA-OV</i> [1]	ViT-L Qwen2 _{0.5B}	34.6M	70.3	73.4
InstructHOI - <i>LLaVA-OV</i> [1]	ViT-L Qwen2 _{7B}	46.3M	<u>71.4</u>	<u>74.9</u>
InstructHOI - <i>InternVL2</i> [2]	ViT-L Qwen2 _{0.5B}	36.2M	70.8	74.2
InstructHOI - <i>InternVL2</i> [2]	ViT-L InternLM2 _{7B}	48.5M	72.1	75.4

Table 2: Results of different HOI-domain fine-tuning strategies on both HICO-DET and V-COCO datasets. The “pretrained” indicates pretrained LFM without fine-tuning; “fine-tune[†]” indicates fine-tuning on the original training dataset (i.e., HICO-DET or V-COCO); “fine-tune[‡]” indicates fine-tuning on the generated dataset (see subsection 3.2).

Strategy	HICO-DET (Default)			V-COCO	
	Full	Rare	Non-Rare	$AP_{role}^{\#1}$	$AP_{role}^{\#2}$
pretrained	43.68	41.75	44.26	69.4	71.7
fine-tune [†]	45.36	43.20	46.01	70.4	73.5
fine-tune [‡]	45.95	46.51	45.78	70.8	74.2

Table 3: Effect of representative token number in the Interest Token Selector. *All* represents that all the local image tokens are selected as representative ones.

Token Number	HICO-DET (Default)			V-COCO	
	Full	Rare	Non-Rare	$AP_{role}^{\#1}$	$AP_{role}^{\#2}$
10	45.54	46.04	45.39	70.4	73.4
15	45.79	46.32	45.63	70.7	73.9
20	45.95	46.51	45.78	70.8	74.2
<i>All</i>	45.86	46.58	45.64	70.8	74.0

Table 4: Effect of Advanced Object Detector. Both DETR and DINO use “Resnet50” as backbone.

Object Detector	HICO-DET (Default)		
	Full	Rare	Non-Rare
DETR [6]	45.95	46.51	45.78
DINO [7]	48.26	48.76	48.11

yielding a notable gain of **2.31** mAP (from 45.95 to 48.26). Furthermore, better detectors can also enhance context-aware guidance learning (see Sec. 3.3).

Efficiency Comparison Following prior work, we add the most commonly used metrics ("Large Foundation Model" (LFM), "Training Data" (TD), "Full Parameters" (FP), "Trainable Parameters" (TP), "Inference Time" (IT), and "Training Time" (TT)) in Table 6, to provide a clearer efficiency comparison, as shown below. All measurements are obtained using Tesla A800 GPUs (or GPUs with comparable performance). Compared to existing methods, InstructHOI demonstrates notable performance improvements while maintaining comparable model parameters (1.13B full parameters and 42.3M trainable parameters) and computational costs (267ms inference time and a total training

Table 5: Performance comparisons of fine-tuning and zero-shot LFM, under RF-UC and NF-UC zero-shot settings. InternVL2_{1b} is taken as LFM.

Methods	Type	Full	Seen	Unseen
Pre-trained	RF-UC	40.19	41.60	34.54
Fine-tuned	RF-UC	43.25	44.86	36.82
Pre-trained	NF-UC	34.97	35.40	33.24
Fine-tuned	NF-UC	38.34	38.82	36.42

Table 6: Efficiency Comparison. Large Foundation Model (LFM), Training Data (TD), Full Parameters (FP), Trainable Parameters (TP), Inference Time (IT), and Training Time (TT))

Method	LFM	TD(K)	FP(B)	TP(M)	IT(ms)	TT(h)
DiffusionHOI [8]	Stable Diffusion & CLIP	48	1.02	27.6	105	17.2
UniHOI [4]	BLIP-2	48	1.14	52.3	82	18.7
MP-HOI [9]	Stable Diffusion & CLIP	286	1.05	41.5	196	23.8
SICHOI [10]	ChatGPT & BLIP	48	-	-	-	-
InstructHOI	InternVL2 & CLIP	140	1.13	42.3	267	(7.2+11.1)

time of 18.3 hours). Specifically, the 7.2 hour training time refers to the HOI-SFT time for the MLLMs, while the 11.1 hours represents the second-stage training time for the full model’s component.

2 Additional Zero-shot Experiments

To bridge HOI-domain knowledge gap and align interaction descriptions, pre-trained LFMs are fine-tuned on the HOI reasoning dataset (see Sec. 3.2). In this section, we evaluate the zero-shot HOI detection performance of InstructHOI when equipped with either fine-tuned or pre-trained LFMs. As shown in Table 5, fine-tuned LFMs significantly outperform their pre-trained counterparts, achieving notable improvements of **2.28** mAP in the RF-UC unseen setting and **3.18** mAP in the NF-UC unseen setting. These improvements can be attributed to LoRA finetuning, which efficiently updates LFMs in a lightweight manner, preserving their generalization ability while enhancing interaction reasoning capabilities.

3 Qualitative Results

Interest Token Selector In InstructHOI, we develop a Interest Token Selector (ITS) to select the informative image tokens for each human-object pair, thereby aligning the reasoning process with interaction regions. As illustrated in Fig. 1, the input image is dynamically split into multiple local images, and ITS predicts interaction relevance of local images for human-object each pair. Finally, ITS selects the informative image patch local images for each pair using the interaction relevance. We illustrate two human-object pairs as examples, where the important areas for each pair are selected and highlighted.

Interaction Prediction in Two Branches InstructHOI involves two interaction prediction branches: the Visual Interaction Decoder (VID) branch and the Multi-Modal Reasoning (MMR) branch. We visualize the prediction results of the two branches in Fig. 2, under both supervised and zero-shot settings. By introducing the multi-modal interaction reasoning capabilities of LFMs, InstructHOI can uncover low confidence HOI detections that are discarded by traditional VID, especially for rare and unseen HOI categories.

Failure Case Analysis To provide a more comprehensive analysis of InstructHOI’s performance, Fig. 3 illustrates several representative failure cases. The reasoning mechanism of LFM integrates multi-modal contextual information to infer human-object interactions (HOIs) within complex scenes, akin to human speculative reasoning. However, the advanced reasoning capabilities of LFM may

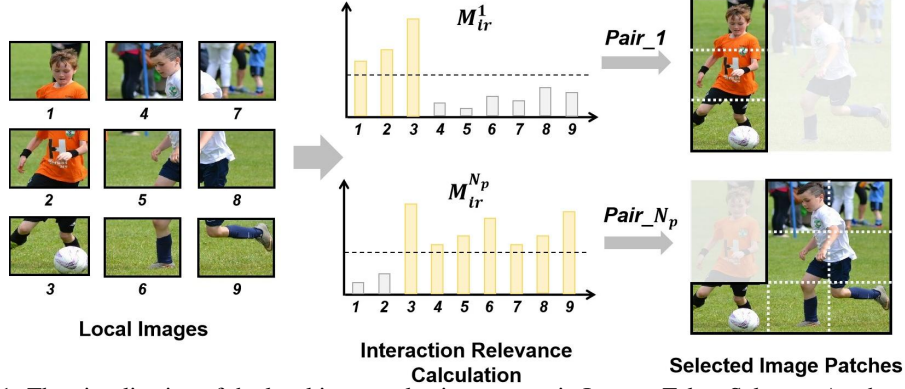


Figure 1: The visualization of the local image selection process in Interest Token Selector. As observed, the important areas for each pair are selected and highlighted.

result in InstructHOI predicting potential interactions, such as “talk”, “chase” and “catch,” which are not represented in the ground truth labels.

In-the-wild HOI Detection In addition to conducting zero-shot experiments, we perform in-the-wild human-object interaction (HOI) detection using arbitrary images sourced from the Internet, as shown in Fig. 4. By leveraging HOI-domain Large Foundation Models (LFMs) for interaction reasoning, InstructHOI demonstrates the capability to accurately detect complex HOIs in real-world scenarios. For instance, it can identify interactions such as “lasso” and “shear” in grassland scenarios, as well as “fill” and “pour” actions in meal contexts. Moreover, InstructHOI is capable of detecting actions from a first-person perspective, such as “stir.” The visualizations highlight the open-world interaction recognition capabilities of our InstructHOI.

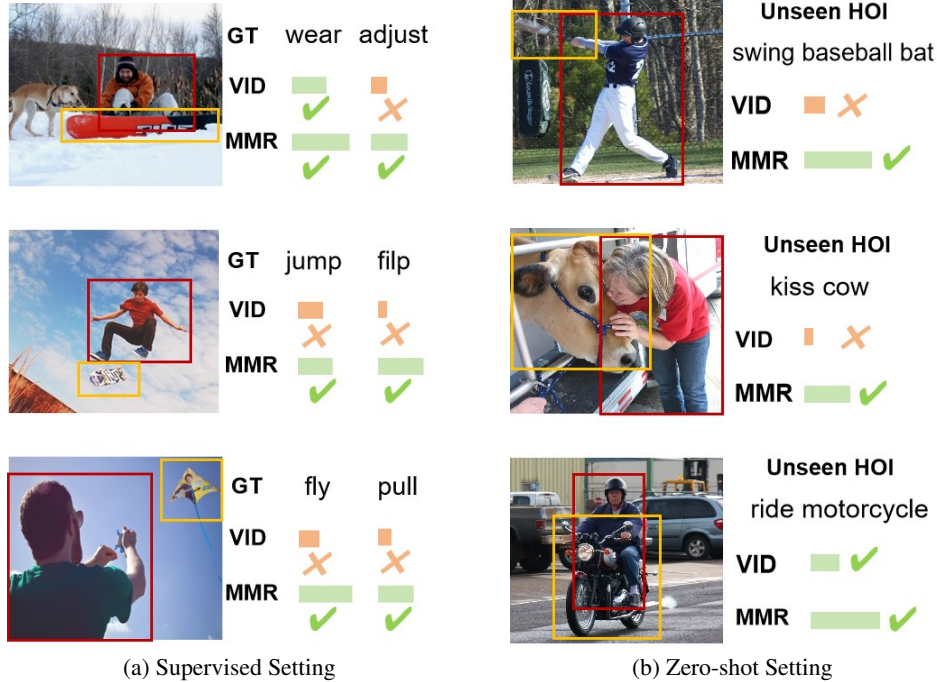


Figure 2: Visual results of two interaction prediction branches in InstructHOI: the Visual Interaction Decoder (VID) branch and the Multl-Modal Reasoning (MMR) branch, under both supervised and zero-shot settings. “GT” indicates the groundtruth, and the length of the color block represents the predicted probability. By introducing the interaction reasoning capabilities of LFMs, InstructHOI can find missed or ambiguous HOI detections.



Figure 3: The visualization of several failure cases. By introducing the reasoning capabilities of LFM, Instruc-tHOI may predict potential interactions, akin to human speculative reasoning.

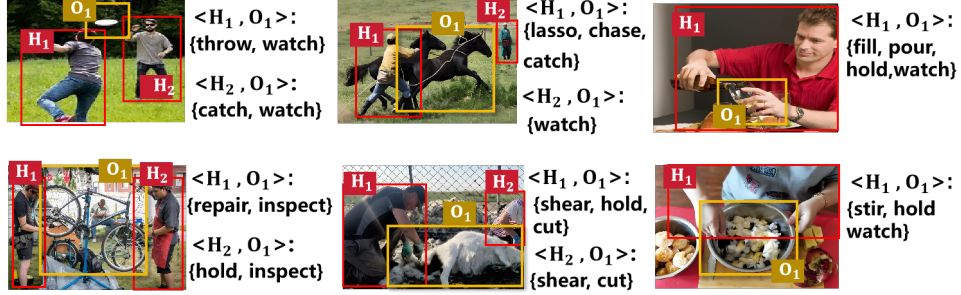


Figure 4: The visualization of in-the-wild HOI detection. Humans and objects are represented by red and orange bounding boxes respectively, and interaction predictions are shown alongside corresponding images.

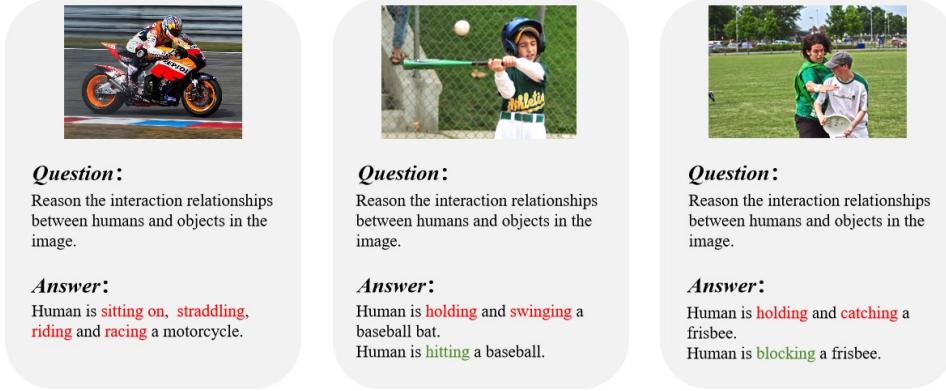


Figure 5: The visualization of generated interaction-reasoning dataset with image-text pairs. The interactions of different human-object pairs are in distinct colors.

4 HOI-Domain Fine-tuning

Fine-tuning Dataset To address the limited availability of HOI reasoning data, we aggregate five existing image-only datasets to create an interaction-reasoning dataset consisting of 140K image-text pairs. Specifically, to generate interaction reasoning texts, we transform the original one-hot labels into Question-and-Answer conversations, as illustrated in Fig. 5.

Implementation Details Here, we provide detailed implementation information for **LoRA** fine-tuning. Taking InternVL2_{1b} as an example, we freeze the entire pretrained model (approximately 1B parameters) while training around 8M parameters (0.8%) for the language model of InternVL2. The fine-tuning process is conducted on a single Tesla A800 GPU for 1 epoch, with a batch size of 8 and a learning rate of 10^{-5} . The image size is set to 448×448 , and the dynamic patch number of local images is configured within the range of (4, 12).

5 Discussion

5.1 Limitation

To adapt the pre-trained LFM to the HOI domain, the paper construct a large-scale dataset consisting of 140K image-text pairs aggregated from existing datasets. However, the diversity and representativeness of this dataset may still be limited, particularly in terms of rare or previously unseen interaction categories. Such limitations could constrain the generalization capability of the LFM in real-world scenarios. As a result, the potential to explore and apply interaction detection in more complex and diverse environments may be hindered.

5.2 Broader Impact

This paper presents a feasible method for comprehending the intricate interactions between humans and objects, demonstrating significant potential for application across diverse domains, including human-computer interaction, healthcare, and autonomous driving. Nevertheless, there is a risk that InstructHOI would be used inappropriately, as it may be susceptible to misuse in contexts such as surveillance or the tracking of individuals, thereby raising legitimate concerns regarding the infringement of user privacy. Consequently, it is essential to rigorously consider ethical imperatives and ensure legal compliance when addressing issues pertaining to personal privacy. Furthermore, to mitigate potential adverse societal impacts, it is necessary to establish robust safety protocols and systems. Ethical review processes should be instituted prior to the development and deployment of such technologies to guarantee adherence to prevailing societal and moral standards.

References

- [1] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, *et al.*, “Llava-onevision: Easy visual task transfer,” *arXiv preprint arXiv:2408.03326*, 2024.
- [2] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma, *et al.*, “How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites,” *arXiv preprint arXiv:2404.16821*, 2024.
- [3] Q. Lei, B. Wang, and T. Robby T., “Ez-hoi: Vlm adaptation via guided prompt learning for zero-shot hoi detection,” *Advances in Neural Information Processing Systems*, 2024.
- [4] Y. Cao, Q. Tang, X. Su, S. Chen, S. You, X. Lu, and C. Xu, “Detecting any human-object interaction relationship: Universal hoi detector with spatial prompt learning on foundation models,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 739–751, 2023.
- [5] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *International Conference on Machine Learning*, pp. 19730–19742, PMLR, 2023.
- [6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European Conference on Computer Vision*, pp. 213–229, Springer, 2020.
- [7] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, and H.-Y. Shum, “Dino: Detr with improved denoising anchor boxes for end-to-end object detection,” in *International Conference on Learning Representations*, 2023.
- [8] L. Li, W. Wang, and Y. Yang, “Human-object interaction detection collaborated with large relation-driven diffusion models,” *Advances in Neural Information Processing Systems*, 2024.
- [9] J. Yang, B. Li, A. Zeng, L. Zhang, and R. Zhang, “Open-world human-object interaction detection via multi-modal prompts,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16954–16964, 2024.
- [10] J. Luo, W. Ren, W. Jiang, X. Chen, Q. Wang, Z. Han, and H. Liu, “Discovering syntactic interaction clues for human-object interaction detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 28212–28222, 2024.