HOI-R1: Exploring the Potential of Multimodal Large Language Models for Human-Object Interaction Detection

Junwen Chen[0000−0003−1355−4128], Peilin Xiong[0009−0007−0112−3634], and Keiji Yanai[0000−0002−0431−183X]

Department of Informatics, The University of Electro-Communications, Tokyo, Japan
{chen-j, xiong-p, yanai}@mm.inf.uec.ac.jp

Abstract. Recent human-object interaction detection (HOID) methods highly require prior knowledge from vision-language models (VLMs) to enhance the interaction recognition capabilities. The training strategies and model architectures for connecting the knowledge from VLMs to the HOI instance representations from the object detector are challenging, and the whole framework is complex for further development or application. On the other hand, the inherent reasoning abilities of multimodal large language models (MLLMs) on human-object interaction detection are under-explored. Inspired by the recent success of training MLLMs with reinforcement learning (RL) methods, we propose HOI-R1 and first explore the potential of the language model on the HOID task without any additional detection modules. We introduce an HOI reasoning process and HOID reward functions to solve the HOID task by pure text. Experiments on HICO-DET across multiple open-source MLLMs, including the Qwen-VL family (Qwen2.5-VL and Qwen3-VL) and Rex-Omni, show consistent improvements. Especially, HOI-R1 boosts Qwen2.5-VL-3B 2× accuracy with great generalization ability.

Keywords: Human-object Interaction Detection · Multimodal Large Language Model · Reinforcement Learning.

1 Introduction

Human–Object Interaction Detection (HOID) aims to detect interactions between humans and surrounding objects in an image. Given an image, the goal is to predict a set of HOI instances represented as $\{B_h, B_o, \textit{Object Class}, \textit{Interaction Class}\}$. Beyond recognizing the relationship category, HOID requires precise localization and pairing of human–object regions, making it a challenging form of fine-grained relationship understanding with grounding.

Most existing HOID approaches [20,23,11] follow a detector-centric paradigm: an object detector provides human/object proposals, and a dedicated interaction head (often transformer-based) predicts HOI labels. Due to long-tailed HOI annotations, recent state-of-the-art methods [16,4,6] increasingly rely on VLM priors (e.g., CLIP [17]/BLIP [14]-style representations) to enhance interaction

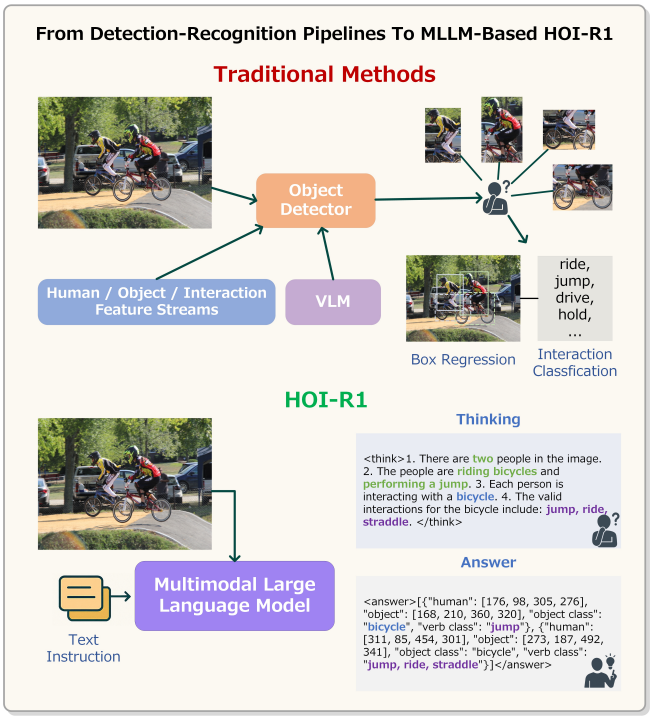


Fig. 1: Comparison of the pipeline of traditional HOID methods and our proposed HOI-R1. Traditional HOID methods rely on object detectors to extract HOI embeddings, while HOI-R1 directly interprets interactions through natural language reasoning using MLLMs.

recognition. However, transferring such priors typically introduces additional modules, training objectives, and architectural coupling between detectors and interaction decoders, leading to complex pipelines that are hard to extend or reuse.

In contrast, modern multimodal large language models (MLLMs) [13,3,22] have shown strong capabilities in fine-grained visual perception, spatial grounding, and multimodal reasoning. In particular, the Qwen-VL family (Qwen2.5-VL [3] and the more recent Qwen3-VL [2]) provides general-purpose grounded generation, while Rex-Omni [9] reformulates visual perception tasks as next-point prediction and exhibits strong localization priors. DeepSeek-R1 [7] exemplifies this trend, demonstrating that RL can induce powerful reasoning behaviors even without supervised fine-tuning (SFT) as a preliminary step. Recent studies [8,21,19] also validate RL's effectiveness for aligning MLLMs with visual reasoning tasks. Despite these advances, the capability of MLLMs for structured HOID—simultaneously reasoning about interactions and localizing HOI pairs in a standardized evaluation protocol—remains underexplored.

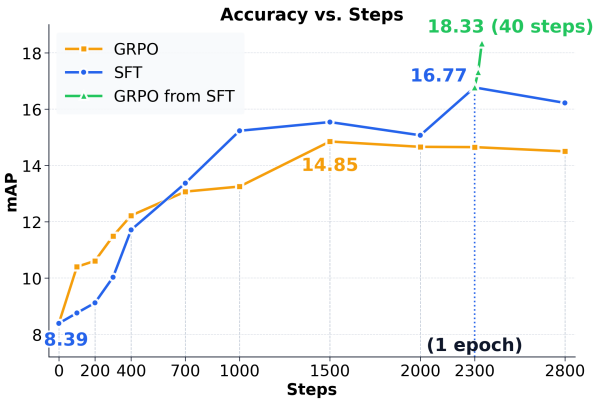


Fig. 2: Training convergence of HOI-R1 with Qwen2.5-VL-3B on HICO-DET. The mAP of *Full* category on *Default Setting* is shown. HOI-R1 achieves more than 2x performance boost with only 1 epoch SFT and 40 steps RL training.

To this end, we first explore their tremendous potential in HOID tasks, as shown in Fig. 1. We propose **HOI-R1**, a radical shift: replacing detectors with natural language reasoning, using MLLMs to directly interpret interactions through holistic scene understanding in both visual and textual modalities. Solving HOID purely through natural language reasoning requires simultaneous prediction of multiple bounding boxes, precise pairing of objects with their interactions, and accurate relationship recognition—all within a complex, structured reasoning pipeline. In practice, we design a systematic prompt structure to guide the reasoning, which injects HOI knowledge through SFT with thinking distillation. As shown in Fig. 2, with HOI knowledge distillation, MLLM shows a significant performance boost. Then, we introduce RL for further alignment with four reward functions, including format rewards for output structure, object/interaction label accuracy, and a one-to-one matching HOI IoU reward, and the performance can be improved with only 100 training steps.

- We introduce HOI-R1, the first MLLM framework that solves HOID end-to-end via natural language, eliminating object detectors.
- We introduce an SFT with thinking distillation to extend the HOI knowledge and a reinforcement learning (RL) paradigm to align the MLLM on HOID with our HOI reward functions to further enhance the performance.
- We demonstrate consistent gains across multiple open-source MLLMs on a standard HOID benchmark, establishing new baselines for MLLM-based HOID.

2 Related Work

MLLMs for vision tasks. Recent MLLMs have made rapid progress in fine-grained visual understanding and grounded generation. The Qwen-VL fam-

ily provides strong multimodal perception and spatial grounding for both images and videos [3,2]. Qwen2.5-VL [3] combines dynamic-resolution visual encoding and absolute time-aware video modeling to achieve fine-grained perception, robust structured understanding of complex images and long-horizon videos. Qwen3-VL [2] supports up to 256K-token interleaved multimodal contexts and, through architectural upgrades such as enhanced interleaved-MRoPE, DeepStack-based multi-level ViT fusion, and text-aligned temporal grounding, delivers superior pure-text understanding, long-context multimodal comprehension, and advanced visual–math reasoning. Rex-Omni [10] reformulates object detection and related perception tasks as discrete coordinate token prediction, leveraging quantized 0–999 coordinate tokens, achieve state-of-the-art zero-shot object perception while enabling versatile language-aware visual grounding capabilities such as referring, pointing, GUI grounding, and OCR.

RL enhanced MLLMs on Visual Reasoning Tasks. Vision-R1 [8] uses an MLLM that combines cold-start initialization with RL to enhance reasoning capabilities on math benchmarks while generating human-like reasoning processes. Reason-RFT [21] introduces a novel two-phase reinforcement fine-tuning framework that combines SFT with Chain-of-Thought (CoT) reasoning activation and Group Relative Policy Optimization (GRPO) [18] to enhance generalization in visual reasoning tasks. VLM-R1 [19] focuses on tasks like Referring Expression Comprehension (REC) and Open-Vocabulary Object Detection (OVD), leveraging deterministic ground-truth annotations for stable reward computation. The study also highlights key insights such as reward hacking in object detection and the emergence of the "OD aha moment," where models first reason about object presence before localization. Based on these successful experiences, we extend the capabilities of MLLM to the more complex HOI task, design input instructions, training strategies, and reward functions for HOID, and provide new possible directions for further development in this field.

3 Metholodology

3.1 Language-based HOID Prediction

As shown in Fig. 1, unlike conventional HOI detection methods that rely on bounding box regression and interaction classification, we propose a novel language-based paradigm that directly outputs all HOI instances in natural language format. Without modifying the model architecture or compromising its original capabilities, we design the input question template to effectively elicit the model's HOI detection potential. We illustrate the detailed design of the question template in Fig. 4. The question template consists of three key components to guide the model:

Task Instruction: We first establish the model's role ("You are an HOI detection model") and provide the complete vocabulary for both objects and interactions in the HOID dataset. The exhaustive list of `<VALID OBJECT CLASSES>` and `<VALID INTERACTIONS>` serves as a constrained output space, ensuring the

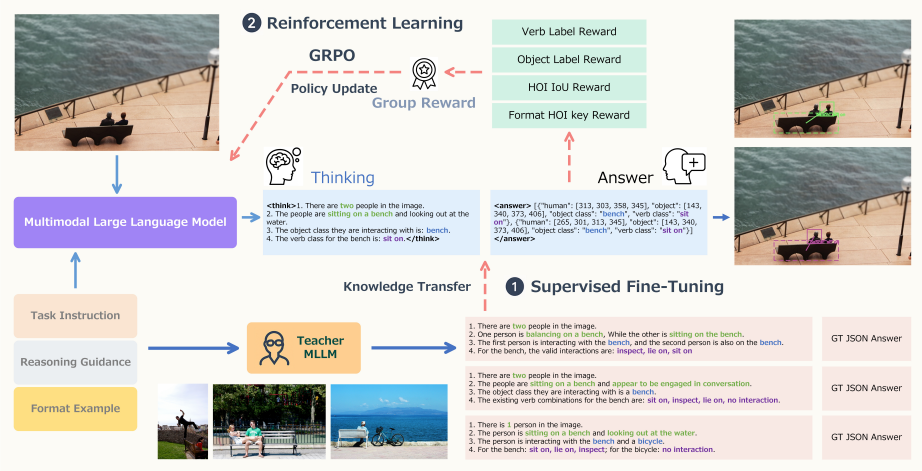


Fig. 3: Overview of our HOI-R1 framework. The input consists of two modalities: image and text. The question text consists of three parts: the task instruction includes basic information about the task, the reasoning guidance provides hints for the reasoning process, and the format example regularizes the output. First, a Teacher MLLM model is used to generate reasoning steps for Supervised Fine-tuning (SFT). Then, in the Reinforcement Learning (RL) stage, the student MLLM model, as the policy model, is trained with four reward signals.

model's predictions align with standard HOI benchmarks while preventing hallucination of irrelevant categories.

Reasoning Guidance: The "Thinking Process" breaks down the complex HOI detection task into sequential reasoning steps, mirroring human cognitive processes. First, the MLLM is required to identify humans in the scene, then analyze their actions, and finally determine their interactions with surrounding objects. This step-by-step decomposition enables the model to handle the compositional nature of HOI relationships systematically.

Format Example: The output template demonstrates the expected JSON structure containing both the reasoning chain (`<think>` tag) and final HOI predictions (`<answer>` tag). As recent MLLMs are trained to represent bounding boxes, we directly incorporate the spatial coordinates into the language output.

The template design offers several advantages: (1) The explicit verb-object compatibility list (`<VALID INTERACTIONS>`) avoids hallucination of unlikely verb-object combinations; (2) The thinking process prompts enable the model to leverage its inherent reasoning capabilities for complex scene understanding; (3) The structured output format bridges the gap between free-form language generation and standardized HOI detection requirements.

```
You are an HOI detection model. Follow this exact structure:

<VALID OBJECT CLASSES>:
<object class>:person, bicycle, car, motorcycle, airplane ...

<VALID INTERACTIONS> (object: allowed_verbs):
<object class>: <verb class>, <verb class>, ···
airplane: board, direct, exit, fly, inspect, load, ride, sit on, wash, no interaction
bicycle: carry, hold, inspect, jump, hop on, park, push, repair, ride, sit on,
straddle, walk, wash, no interaction
...

Thinking Process:
1. How many people are in the image?
2. What are those people doing?
3. What <object class> in <VALID OBJECT CLASSES> are each people
interacting with?
4. Find existing <verb class> combinations from <VALID INTERACTIONS> for
each <object class>.

EXAMPLE OUTPUT FORMAT:
<think>
reasoning process here
</think>
<answer>
[
   {
      "human":  [x1, y1, x2, y2],
      "object": [x1, y1, x2, y2],
      "object class": "<object class>",
      "verb class": "<verb class>, <verb class>, ... "
   },
   ...
]
</answer>
```

Fig. 4: The input question template for HOI-R1. The template consists of three key components: Task Instruction, Reasoning Guidance, and Format Example.

3.2 Thinking Distillation via SFT

To transfer task-specific knowledge to the student MLLM, we employ a supervised fine-tuning (SFT) stage with a teacher-student knowledge transfer paradigm. This process, termed *thinking distillation*, leverages a powerful teacher model to generate reasoning traces that guide the student model's learning process.

Teacher Reasoning Generation: We utilize GPT4o-mini [1] as the teacher model to generate step-by-step reasoning traces. For each image in the training set of HICO-DET, we input the image along with the structured prompt's *Reasoning Guidance* (as defined in Section 3.1). The teacher model produces natural language reasoning sequences enclosed within `<think>` tags. These distilled reasoning traces capture the implicit logical process of HOI detection that traditional supervised learning fails to explicitly teach.

Thinking Distillation and Answer Supervision: With the ground-truth annotations from the dataset, we supervise the student model to learn both the reasoning traces and the final HOI predictions. The student model is trained to predict two components: (1) predicting the teacher-generated `<think>` sequences, internalizing the step-by-step HOI reasoning logic. (2) `<answer>` component is directly supervised using ground-truth HOI triplets from HICO-DET annotations. This ensures precise alignment with the target task objectives while maintaining output fidelity. Given image x, question template q, teacher-generated reasoning r, and ground-truth answer a, the training objective is the autoregressive

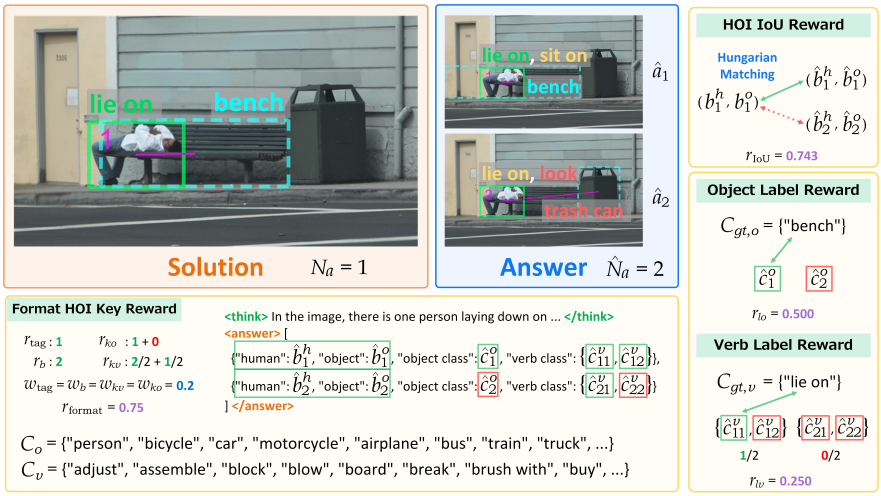


Fig. 5: The reward functions of HOI-R1. We design key format reward, label reward, and label reward to ensure the structural, semantic, and geometric alignment of the model outputs with the ground truth.

negative log-likelihood over the supervised tokens:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x,q,r,a)\sim\mathcal{D}}\Bigg[\underbrace{\sum_{t=1}^{T_r}\log\pi_\theta\big(r_t \mid x,q,r_{<t}\big)}_{\text{reasoning supervision}} + \underbrace{\sum_{t=1}^{T_a}\log\pi_\theta\big(a_t \mid x,q,r,a_{<t}\big)}_{\text{answer supervision}}\Bigg] \tag{1}$$

where π_θ is the student MLLM parameterized by θ, and $\mathcal{D}$ is the training dataset. As shown in Fig. 2, the SFT stage establishes a strong foundation for the subsequent reinforcement learning alignment, which further refines the model's outputs using HOID-specific rewards.

3.3 HOID Reinforcement Learning

After establishing foundational capabilities via SFT, we further align the student MLLM through Reinforcement Learning (RL) to enforce structural, semantic, and geometric alignment with ground truth. Following recent successes, we employ the Group Relative Policy Optimization (GRPO) [18] algorithm, which is efficient for post-training LLMs and MLLMs. For each input image x with question template q, GRPO samples G outputs $\{o_i\}_{i=1}^G$ from the old policy $\pi_{\theta_{old}}$ and optimizes the policy with the following objective:

$$\begin{aligned}\mathcal{J}_{\text{GRPO}} = &-\mathbb{E}_{(x,q)\sim D,\{o_i\}_{i=1}^G\sim\pi_{\theta_{old}}(O|x,q)} \\ &\frac{1}{G}\sum_{i=1}^{G}\Big\{\min[s_1\cdot\hat{A}_i, s_2\cdot\hat{A}_i] - \beta\mathbb{D}_{\text{KL}}[\pi_\theta||\pi_{\text{ref}}]\Big\}\end{aligned} \tag{2}$$

$$s_1 = \frac{\pi_\theta(o_i|x,q)}{\pi_{\theta_{old}}(o_i|x,q)} \tag{3}$$

$$s_2 = \text{clip}(\frac{\pi_\theta(o_i|x,q)}{\pi_{\theta_{old}}(o_i|x,q)}, 1-\epsilon, 1+\epsilon) \tag{4}$$

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})} \tag{5}$$

where $\hat{A}_i$ is the advantage for output o_i, and $\mathbb{D}_{\text{KL}}$ is the KL divergence. The reward function is crucial for guiding the model towards predictions. As an HOI instance involves multiple elements (e.g., human, object, object class, and verb class), to ensure fine-grained alignment, we design element-specific rewards that guide the learning process more comprehensively. In Fig. 5, we illustrate the detailed design of our reward functions, which consist of three components: (1) HOI key format reward, (2) object and verb label reward, and (3) HOI IoU reward. Each component is described in detail below.

HOI Key Format Reward: Since the predictions are generated in plain text, it is necessary to ensure the correctness of the output format. Therefore, we design a format reward for each key-value pair in the dict of every HOI instance within the prediction list. First, the model output must contain the `<answer>` tag; otherwise, all rewards are set to zero. We divide the rewards into five components: the reward for the `<think>` tag r_{tag}, the reward for human and object boxes r_b, the reward for object label r_{ko}, and the reward for verb label r_{kv}. Specifically, for the entire output text, the thinking tag reward is defined as:

$$r_{\text{tag}} = \mathbf{1}[\texttt{"<think>"} \in \hat{y}] \tag{6}$$

where $\mathbf{1}[\cdot]$ is the indicator function, and $\hat{y}$ is the model output text. Then, considering the i-th predicted HOI instance dict $\hat{a}_i$ in the answer list, the box format reward r_{b_i} is defined as:

$$\begin{aligned} r_{b_i} = \mathbf{1}\Big[&\{\texttt{"human"}, \texttt{"object"}\} \subseteq \text{keys}(\hat{a}_i) \wedge \\ &\forall b \in \tilde{B}_i : \text{IoU}(\hat{b}_i^o, b^o) \leq 0.5 \ \wedge \ \text{IoU}(\hat{b}_i^h, b^h) \leq 0.5\Big] \end{aligned} \tag{7}$$

$$B_i = B_{i-1} \cup \{(\hat{b}_i^h, \hat{b}_i^o)\}, \quad B_0 = \emptyset \tag{8}$$

where $\hat{b}_i^h$ and $\hat{b}_i^o$ are the predicted human and object bounding boxes, respectively, and $\tilde{B}$ is the set of all previously predicted boxes such that both the human and object box IoU are less than 0.5 with each other. Both the keys "human" and "object" must exist in the dict $\hat{a}_i$, and we defined $\tilde{B}$ that records all unique boxes to avoid reward hacking by duplicated boxes. Moreover, if an instance's boxes are duplicated, no further rewards are computed for that instance. For the object-label reward r_{ko_i}, the key "object class" must exist, and its value $\hat{c}_i^o$ must belong to the predefined object-class set C_a.

$$r_{ko_i} = \mathbf{1}[\texttt{"object class"} \in \text{keys}(\hat{a}_i) \ \wedge \ \hat{c}_i^o \in C_o] \tag{9}$$

Different from the object class, since a single HOI instance may involve multiple interactions, we compute the ratio between the number of distinct labels belonging to the verb-class set C_v and the total number of predicted verb labels as the reward.

$$r_{kv_i} = \frac{|\{\tilde{c}_i^v\}|}{|\{\hat{c}_i^v\}|} \cdot \mathbf{1}[\texttt{"verb class"} \in \text{keys}(\hat{a}_i)] \tag{10}$$

$$\tilde{c}_i^v = \text{Unique}(\{\hat{c}_{ij}^v \mid \hat{c}_{ij}^v \in C_v, j = 1, 2, \ldots, |\{\hat{c}_i^v\}|\}) \tag{11}$$

where $\hat{c}_{ij}^v$ is the j-th predicted verb label in the i-th HOI instance, and Unique$(\cdot)$ returns the set of unique elements. In addition, we introduce a key penalty to avoid duplicate keys in the dict of each HOI instance as follows:

$$\alpha_i = \frac{N_k}{N_k + |\hat{N}_{k_i} - N_k|} \tag{12}$$

where $N_k = 4$ is the standard number of keys, and $\hat{N}_{k_i}$ is the number of keys in the i-th predicted HOI instance. Finally, all component rewards are combined with weights to form the overall format reward:

$$r_{\text{format}} = w_{\text{tag}} r_{\text{tag}} + \frac{\sum_{i=1}^{\hat{N}_a} \alpha_i \left[w_b r_{b_i} + w_{ko} r_{ko_i} + w_{kv} r_{kv_i} \right]}{\max(N_a, \hat{N}_a)} \tag{13}$$

where $\hat{N}_a$ and N_a are the numbers of predicted and ground-truth HOI instances, respectively, and w_{tag}, w_b, w_{ko}, and w_{kv} are the weights for each component.

Object and Verb Label Reward: Different from the format reward, which is irrelevant to specific labels, the purpose of label reward is to encourage the model to make more accurate predictions of the HOIs that appear in the ground truth. In practice, we compare the predicted labels in each HOI instance a_i with the ground-truth label set C_{gt} one by one in sequential order with a drop-on-match strategy. The object label rewards are defined as:

$$r_{lo} = \frac{\sum_{i=1}^{\hat{N}_a} \alpha_i \mathbf{1}[\hat{c}_i^o \in C_{gt,o}^{(i-1)}]}{\max(N_a, \hat{N}_a)} \tag{14}$$

where

$$C_{gt,o}^{(i)} = \begin{cases} C_{gt,o}^{(i-1)} \setminus \{\hat{c}_i^o\}, & \text{if } \hat{c}_i^o \in C_{gt,o}^{(i-1)}, \\ C_{gt,o}^{(i-1)}, & \text{otherwise.} \end{cases} \tag{15}$$

and the verb label reward is defined as:

$$r_{lv} = \frac{\sum_{i=1}^{\hat{N}_a} \frac{\alpha_i}{|\{\hat{c}_i^v\}|} \sum_{j=1}^{|\{\hat{c}_i^v\}|} \mathbf{1}[\hat{c}_{ij}^v \in C_{gt,v}^{(i-1)}]}{\max(N_a, \hat{N}_a)} \tag{16}$$

where

$$C_{gt,v}^{(i)} = C_{gt,v}^{(i-1)} \setminus \{\bar{c}_i^v\} \tag{17}$$

$$\{\bar{c}_i^v\} = \{\hat{c}_{ij}^v \mid \hat{c}_{ij}^v \in C_{gt,v}^{(i-1)}, j = 1, 2, \ldots, |\{\hat{c}_i^v\}|\} \tag{18}$$

HOI IoU Reward: Inspired by recent transformer-based HOID methods [20], we leverage the Hungarian algorithm [12] to match the predicted HOI boxes and ground-truth boxes for accurate spatial alignment. The cost matrix, considering the Intersection over Union (IoU) of HOI pairs, is defined as:

$$C_{ij} = 1 - s_{ij} \tag{19}$$

$$s_{ij} = \frac{1}{2}\left[\text{IoU}(\hat{b}_i^h, b_j^h) + \text{IoU}(\hat{b}_i^o, b_j^o)\right] \tag{20}$$

where $\hat{b}_i^h$ and $\hat{b}_i^o$ are the predicted human and object bounding boxes, respectively. The one-to-one matching $\mathcal{M}^*$ is obtained by solving the linear assignment problem:

$$\mathcal{M}^* = \text{argmin}_{\mathcal{M}} \sum_{(i,j)\in\mathcal{M}} C_{ij} \tag{21}$$

The final reward is defined as:

$$r_{\text{IoU}} = \frac{1}{N_a} \sum_{(i,j)\in\mathcal{M}^*} s_{ij} \tag{22}$$

Note that the one-to-one matched predicted HOI instances can not be more than the ground-truth HOI instances, i.e., $|\mathcal{M}^*| \leq N_a$, we use N_a is used to normalize the reward.

Finally, the overall reward considering all components is defined as:

$$r = r_{\text{format}} + r_{lo} + r_{lv} + r_{\text{IoU}} \tag{23}$$

With our HOID-specific rewards, the model is effectively guided to produce accurate and well-structured HOI predictions.

4 Experiments

In this section, we present the experimental setup, datasets, and evaluation metrics used to assess the performance of our proposed method. We also provide a detailed analysis of the results obtained from various experiments conducted to validate our approach.

4.1 Experimental Settings

Dataset and Metric. We conduct experiments on the HICO-DET [5] dataset, a widely used benchmark for Human-Object Interaction (HOI) detection. The dataset consists of 38,118 training images and 9,658 test images, encompassing 600 HOI categories formed by 117 verbs and 80 objects. The HOI categories are further divided into three subsets based on the number of instances: *Full*, *Rare*, and *Non-Rare*. In addition, the evaluation is split into two settings: *Default* and *Known Object*, where the latter do not include unknown objects. The mean Average Precision (mAP) is employed as the primary evaluation metric, calculated using an IoU threshold of 0.5 for both human and object bounding boxes, and the object and verb label must be correctly predicted.

Method	Training Sessions	Default			Known Object		
		Full	Rare	Non-Rare	Full	Rare	Non-Rare
— Traditional HOID Methods —							
HO-RCNN [5]	150k	7.81	5.37	8.54	10.41	8.94	10.85
HOTR [11]	100 epoch	25.10	17.34	27.42	-	-	-
HOI-Trans [23]	150 epoch	26.61	19.15	28.84	29.13	20.98	31.57
QPIC [20]	150 epoch	29.07	21.85	31.23	31.68	24.14	33.93
— MLLMs —							
Qwen2.5-VL-3B (baseline)	-	8.39	9.60	8.03	8.96	9.83	8.70
Qwen2.5-VL-7B	-	10.46	14.30	9.31	11.01	14.63	9.93
Qwen2.5-VL-32B-AWQ	-	18.12	24.56	16.20	19.90	25.77	18.15
Qwen2.5-VL-72B-AWQ	-	20.71	29.62	18.05	22.93	31.87	20.26
Qwen3-VL-4B (baseline)	-	16.16	22.31	14.32	17.55	22.87	15.96
Qwen3-VL-8B	-	17.41	21.96	16.05	19.17	23.39	17.91
— Supervised Fine-Tuning (SFT) or Reinforcement Learning (RL) —							
Qwen2.5-VL-3B-SFT	400 steps	11.71	10.52	12.07	12.60	10.70	13.17
Qwen2.5-VL-3B-SFT	1000 steps	15.23	12.26	16.11	16.41	12.60	17.54
Qwen2.5-VL-3B-SFT	1 epoch	16.77	14.20	17.53	18.05	14.45	19.13
Qwen2.5-VL-3B-GRPO	400 steps	12.22	14.56	11.52	13.12	14.84	12.60
Qwen2.5-VL-3B-GRPO	1000 steps	13.25	15.22	12.66	14.18	15.60	13.75
Qwen2.5-VL-3B-GRPO	1 epoch	14.65	15.90	14.28	15.48	16.14	15.28
Qwen3-VL-4B-SFT	1 epoch	22.46	21.53	22.73	24.04	21.97	24.66
Rex-Omni-3B [10] (HOI-SFT)	1 epoch	**25.30**	21.24	**26.51**	26.97	21.79	**28.51**
— HOI-R1 (SFT+GRPO) —							
HOI-R1-Qwen2.5-VL-3B	1 epoch + 40 steps	18.33	16.03	19.02	19.83	16.25	20.90
HOI-R1-Qwen3-VL-4B	1 epoch + 40 steps	23.35	**23.58**	23.28	25.12	**24.11**	25.42
HOI-R1-Rex-Omni-3B	1 epoch + 40 steps	25.24	22.57	26.04	**27.00**	23.21	28.14

Table 1: Comparison on the HICO-DET dataset.

Settings		Default			Known Object		
Thinking	Task Description	Full	Rare	Non-Rare	Full	Rare	Non-Rare
✓	✓	**8.39**	**9.60**	**8.03**	**8.96**	**9.83**	**8.70**
	✓	7.91	8.30	7.79	8.60	8.60	8.60
✓		3.06	2.82	3.13	3.11	2.82	3.20
		2.75	1.86	3.13	2.88	1.86	3.19

Table 2: Ablation studies of prompt design. The original Qwen2.5-VL-3B-Instruct model is used.

Methods	Default			Known Object		
Full	Rare	Non-Rare	Full	Rare	Non-Rare	
HOI-R1	**13.25**	**15.22**	**12.66**	**14.18**	**15.60**	**13.75**
w/o label reward	12.49	14.78	11.80	13.54	15.26	13.02
w/o IoU reward	9.63	11.12	9.19	10.29	11.38	9.96

Table 3: Ablation studies of reward functions. All of the Qwen2.5-VL-3B-Instruct models are trained for 1,000 steps by GRPO.

4.2 Implementation Details

Our main experiments are conducted on the Qwen2.5-VL-3B-Instruct [3] model, and we also validate the effectiveness of our method on the Qwen3-VL-4B-Instruct [2] and Rex-Omni-3B [9] models. In the SFT stage, the GPT-4o-mini [1] is used to generate the thinking process for each training image in the HICO-DET dataset. The ground-truth annotations are converted into our desired output format. For an HOI pair with multiple interactions, we merge them into a

single dict entry with a list of verbs. The full model is trained for 1 epoch with a batch size of 8. The AdamW optimizer [15] is used with a learning rate of 1e-6, and a cosine learning rate scheduler is applied. Next, in the RL stage, the model is trained for 40 steps with a batch size of 16. The group size G for GRPO is set to 4. The AdamW optimizer is used with a learning rate of 1e-6, and a linear learning rate scheduler is applied. In the reward functions, the weights for the HOI key format reward are all set to 0.2.

4.3 Comparison to Baselines and HOID Methods

In Table 1, we compare our proposed HOI-R1 method with traditional HOID methods and MLLMs with different scales. First, we evaluate the original Qwen-VL models with our designed HOID prompt in a training-free manner. As a result, our baseline models, Qwen2.5-VL-3B and Qwen3-VL-4B, achieve 8.39 and 16.16 mAP on the *Full* category under the *Default* setting, which is higher than the traditional HOID method HO-RCNN [5]. We also find that the performance on the *Rare* category is better than that on the *Full* and *Non-Rare* categories, which is different from traditional HOID methods. We consider the reason is that the image contains rare HOI categories, which usually have fewer HOI instances, making it easier for the MLLM to focus on the relevant interactions and utilize its strong reasoning ability to identify them. Especially, the 32B and 72B models outperform the transformer-based HOID method, QPIC [20], demonstrating the strong prior knowledge of large MLLMs.

Next, we conduct SFT or RL training on the baseline model. From the results of Qwen2.5-V-3B 400-step training, GRPO outperforms SFT by 0.51 mAP, and also from the training curve in Fig. 2, RL training increases the performance more effectively than SFT in the early training stage. As our SFT training introduces HOI-specific knowledge from a teacher model, after 1 epoch training, with additional prior knowledge, the SFT model achieves 16.77 mAP, which is 2x higher than the baseline model. For the Rex-Omni model, which does not have reasoning ability, we do SFT training without the thinking process part in the prompt. The Rex-Omni-HOI-SFT model achieves 25.30 mAP on the *Full* category, which is higher than HOTR [11]. Notably, our HOI-R1 model only trains for 1 epoch, which converges much faster than traditional HOID methods that require training for hundreds of epochs. The SFT training increases task-specific knowledge, and the RL training enhances the generalization capabilities. Our HOI-R1 combines the benefits of the two training stages. The Rex-Omni-3B model achieves 27.00 mAP on the *Full* categories under the *Known Object* setting, which is 17.7% higher than Qwen2.5-VL-72B-AWQ, and Qwen3-VL-4B achieves 23.58 mAP on the *Rare* categories, which is 7.9% higher than QPIC [20].

4.4 Ablation Study

Reward Functions. To isolate and evaluate the effectiveness of our prompt design, in Table 2, we conduct the ablation study on the original Qwen2.5-VL-3B baseline model. Comparing line 1 and line 2 and line 3 and line 4, we find



Fig. 6: An example of the reward advantage of GRPO. The predicted HOI instances are visualized on the left with the rewards on the right.

that removing the chain-of-thought part mainly decreases the performance on the *Rare* categories, indicating that the thinking process helps the model to reason about less common interactions. Moreover, removing the task description part significantly degrades the performance on all categories, demonstrating that providing clear instructions is crucial for the model to understand and perform the HOID task effectively.

Reward Functions. Our reward functions are designed specifically for the HOID task. To evaluate their effectiveness, we conduct drop-one experiments on our full model in Table 3. From the result, removing the label reward results in a decrease of 0.76 mAP on the *Full* category under the *Default* setting, indicating that the label reward, which encourages correct verb and object predictions, is crucial for improving the model's performance. Furthermore, removing the IoU reward leads to a more significant drop of 3.62 mAP, highlighting the importance of accurate localization in HOID tasks. These results demonstrate that each component of our reward functions contributes to the overall performance, and their combination is essential for achieving optimal results.

4.5 Qualitative Results

In Fig. 6, we present a visualization of the results of a group of reward cases. From the result, answer 1 has a higher reward than answer 2, as it correctly identifies more human-object interactions with accurate bounding boxes, while answer 3 misidentifies the interaction as "wait" and has less precise bounding boxes. The reward reflects these differences, showing that our designed reward functions effectively guide the model to generate more accurate HOI predictions. This example highlights the advantage of our GRPO training in enhancing the model's ability to accurately detect human-object interactions.

5 Conclusion

In this paper, we present HOI-R1, the first pure MLLM framework for HOID tasks, which eliminates the need for object detectors. We introduce a novel two-stage training paradigm that combines supervised fine-tuning (SFT) and reinforcement learning (RL) to effectively adapt MLLMs for HOID tasks. The SFT stage focuses on enhancing the model's ability to recognize human-object interactions through carefully designed instruction templates and data augmentation techniques. The RL stage further refines the model's performance by optimizing it for specific HOID metrics, ensuring that the model not only understands the interactions but also excels in practical evaluation scenarios. With our proposed SFT and RL paradigm, HOI-R1 achieves a significant performance boost on the HICO-DET dataset. Our results demonstrate the potential of MLLMs in structured tasks like HOID, paving the way for future research in this direction.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 6, 11
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025) 2, 4, 11
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025) 2, 4, 11
4. Cao, Y., Tang, Q., Su, X., Chen, S., You, S., Lu, X., Xu, C.: Detecting any human-object interaction relationship: Universal hoi detector with spatial prompt learning on foundation models (2024) 1
5. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: WACV (2018) 10, 11, 12
6. Chen, J., Wang, Y., Yanai, K.: Focusing on what to decode and what to train: Sov decoding with specific target guided denoising and vision language advisor. In: WACV (2025) 1

7. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025) 2
8. Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749 (2025) 2, 4
9. Jiang, Q., Huo, J., Chen, X., Xiong, Y., Zeng, Z., Chen, Y., Ren, T., Yu, J., Zhang, L.: Detect anything via next point prediction. arXiv preprint arXiv:2510.12798 (2025) 2, 11
10. Jiang, Q., Huo, J., Chen, X., Xiong, Y., Zeng, Z., Chen, Y., Ren, T., Yu, J., Zhang, L.: Detect anything via next point prediction. arXiv preprint arXiv:2510.12798 (2025) 4, 11
11. Kim, B., Lee, J., Kang, J., Kim, E.S., Kim, H.J.: HOTR: End-to-end human-object interaction detection with transformers. In: CVPR (2021) 1, 11, 12
12. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Res. Logist. Quart pp. 83–97 (1955) 10
13. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024) 2
14. Li, J., Li, D., Savarese, S., Hoi, S.: 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) 1
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2018) 12
16. Ning, S., Qiu, L., Liu, Y., He, X.: Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In: CVPR (2023) 1
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 1
18. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024) 4, 7
19. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., et al.: VLM-R1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615 (2025) 2, 4
20. Tamura, M., Ohashi, H., Yoshinaga, T.: QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information. In: CVPR (2021) 1, 10, 11, 12
21. Tan, H., Ji, Y., Hao, X., Chen, X., Wang, P., Wang, Z., Zhang, S.: Reason-RFT: Reinforcement fine-tuning for visual reasoning of vision language models. In: NeurIPS (2025) 2, 4
22. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025) 2
23. Zou, C., Wang, B., Hu, Y., Liu, J., Wu, Q., Zhao, Y., Li, B., Zhang, C., Zhang, C., Wei, Y., et al.: End-to-end human object interaction detection with hoi transformer. In: CVPR (2021) 1, 11