HOID-R1: Reinforcement Learning for Open-World Human-Object Interaction Detection Reasoning with Multimodal Large Language Model

Zhenhao Zhang[1*], **Hanqing Wang**[2,3†*], **Xiangyu Zeng**[3,4*],
Ziyu Cheng[5], **Jiaxin Liu**[2], **Zhirui Liu**[2], **Kaiyang Ji**[2],
Tianyang Gui[2], **Ke Hu**[2], **Kangyi Chen**[2], **Yahao Fan**[2], **Mokai Pan**[2]

[1]ShanghaiTech University, [2]The Hong Kong University of Science and Technology (GZ), [3]Shanghai AI Lab,
[4]Nanjing University [5]University of Wisconsin, Madison
zhangzhh2024@shanghaitech.edu.cn,hwang201@connect.hkust-gz.edu.cn

Abstract

Understanding and recognizing human–object interaction (HOI) is a pivotal application in AR/VR and robotics. Recent open-vocabulary HOI detection approaches depend exclusively on large language models for richer textual prompts, neglecting their inherent 3D spatial understanding capabilities. To address this shortcoming, we introduce **HOID-R1**, the first HOI detection framework that integrates chain-of-thought (CoT) guided supervised fine-tuning (SFT) with group relative policy optimization (GRPO) within a reinforcement learning (RL) paradigm. Specifically, we initially apply SFT to imbue the model with essential reasoning capabilities, forcing the model to articulate its thought process in the output. Subsequently, we integrate GRPO to leverage multi-reward signals for policy optimization, thereby enhancing alignment across diverse modalities. To mitigate hallucinations in the CoT reasoning, we introduce an "MLLM-as-a-judge" mechanism that supervises the CoT outputs, further improving generalization. Extensive experiments show that **HOID-R1** achieves state-of-the-art performance on HOI detection benchmarks and outperforms existing methods in open-world generalization to novel scenarios.

1 Introduction

Human–object interaction (HOI) detection seeks not only to localize human and object instances in visual scenes, but also to characterize the semantic and functional relationships that define their interactions. As a foundational component of human-centric AI, accurate HOI detection underpins a diverse range of downstream applications—among them dexterous assistive and collaborative robotics, immersive augmented and virtual reality, surveillance and anomaly detection, advanced video understanding, and anticipatory activity forecasting. By modeling affordances, intentions, and social context, HOI detection endows autonomous agents with the perceptual and reasoning capabilities required for safe, effective operation in complex, human-populated environments.

Existing HOI detection methods are predominantly confined to small-scale closed-set benchmarks (e.g., (Gao, Zou, and Huang 2018), (Liao et al. 2020)). In these benchmarks, models are trained and evaluated on a fixed set of interaction categories. This constraint yields limited out-of-distribution generalization. When faced with novel verbs, unseen objects, or previously unobserved interaction combinations, performance degrades sharply.

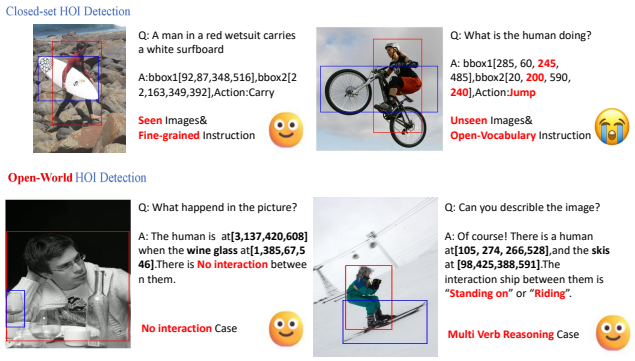

Figure 1: **Motivation.** Closed-set HOI detectors fail to generalize to novel verbs, objects, or interaction combinations in real scenes; adopting an open-world paradigm enables structured reasoning and semantic supervision for robust zero-shot inference under free-form instructions.

Recent work has begun to exploit large vision language models to generate richer interaction prompts and enable zero-shot inference (e.g., (Ning et al. 2023)). These prompt-based methods leverage only the linguistic priors of the underlying model and largely ignore its inherent reasoning capabilities. As a result, they remain sensitive to the precise phrasing of queries and struggle to disambiguate fine-grained or underspecified interactions. In contrast, our framework applies supervised fine-tuning followed by targeted post-training to fully harness the model's reasoning power. The resulting open world HOI detector generalizes robustly across novel verbs, unseen objects, and fuzzy natural language queries.

To address these challenges, we propose **HOID-R1**, an open-world HOI detection framework that integrates multi-stage reasoning, visual grounding, and policy learning un-

*These authors contributed equally.
†Project Leader.

der continual semantic supervision. Given an input image and a free-form language query, our reasoning module produces a structured chain of thought(Kojima et al. 2023),(Wei et al. 2023) comprising sequential hypotheses about potential human-object interactions, while a parallel segmentation module localizes regions relevant to the task in the visual scene. These symbolic cues and pixel-level signals are then combined by a policy model trained with Generalized Reward Policy Optimization to generate spatial coordinates and interaction labels. During training, multiple reward functions assess physical plausibility, spatial consistency, and task accuracy, and a collection of vision language models acting as multimodal judges provides iterative feedback on intermediate reasoning steps. This supervision mechanism leverages the reasoning capacity of large-scale vision language models to identify and correct hallucinated or unsupported chains of thought trajectories, ensuring that every inference is grounded in both visual evidence and linguistic context. As a result, HOID-R1 achieves robust open vocabulary generalization and maintains high accuracy on novel verbs, unseen objects, and underspecified queries in real-world HOI scenarios.

In summary, our contributions are as follows:

- **The first RL-based Chain-of-Thought framework for HOI detection.** We introduce the first reinforcement learning paradigm that integrates a chain-of-thought reasoning process directly into HOI detection, enabling the model to decompose complex interaction queries into a sequence of sub-reasoning steps and learn to optimize each step via policy gradient.
- **Open-World HOI detection reasoning and multi-level dataset.** Through supervised fine-tuning and GRPO-based post-training of the VLM, our model acquires Open-World human-object interaction reasoning capabilities and demonstrates strong generalization to open-vocabulary instructions and unseen images. To rigorously evaluate its open-world generalization, we hierarchically annotated a new HOI detection dataset
- **MLLM-as-a-Judge for chain-of-thought process.** We leverage a pre-trained 3D multimodal large language model as a soft "judge" to evaluate and guide each reasoning step, preventing the model from arriving at correct conclusions through flawed reasoning processes. We have enhanced its reasoning capabilities.
- **State-of-the-art performance.** Extensive experiments on HICO-DET and SWIG-HOI show that our approach outperforms existing baselines by a significant margin across all key metrics, achieving new state-of-the-art results in both seen- and unseen-object HOI detection.

2 Related Work

2.1 HOI Detection

Current HOI detection methods can be mainly divided into two categories: two-stage paradigm(Gao, Zou, and Huang 2018),(Cao et al. 2023) and one-stage paradigm(Kim et al. 2023),(Liao et al. 2020),(Chen et al. 2021),(Tamura, Ohashi, and Yoshinaga 2021). Two-stage strategy first detects all human and object instances using a pre-trained object detector, such as Faster R-CNN(Gao, Zou, and Huang 2018), followed by a second-stage module that enumerates possible human-object pairs and predicts their interactions. Although this design benefits from the maturity and robustness of standalone object detectors, it suffers from the inefficiency of exhaustive pairwise matching and limited ability to model contextual dependencies among entities. In contrast, one-stage approach detects (human-object-interaction) triplets directly through different perspectives, e.g., point-based detection(Liao et al. 2020) formulates HOI triplets as pairs of keypoints, such as the center points of human and object bounding boxes, and each interaction is modeled by predicting a pair of spatial points along with verb classification, anchor-based detection(Kim et al. 2023) extends the concept of anchor boxes from object detection to interaction modeling, human and object entities are predicted based on predefined anchor regions, and interactions are inferred using features extracted from the union area of the predicted boxes and set prediction methods(Chen et al. 2021),(Tamura, Ohashi, and Yoshinaga 2021) reformulate the HOI detection problem as a set-to-set matching problem, thus avoiding human-object pairing.

2.2 Large Reasoning Model

Unlike traditional LLMs, which can only take textual inputs, Multimodal Large Language Models (MLLMs) extend the capabilities of traditional large language models by integrating information from multiple modalities, such as text, images, audio, and video. By jointly modeling cross-modal interactions, MLLMs enable a wide range of tasks, including visual question answering, image captioning, and multimodal reasoning. The emergence of Large Reasoning Models (LRMs)(DeepSeek-AI et al. 2025),(Shao et al. 2024),(Li et al. 2025a),(Ouyang et al. 2025),(Shen et al. 2025), which are explicitly designed to enhance reasoning capabilities beyond language generation. One prominent example is Deepseek R1, a reasoning-centric model that integrates both pretraining on reasoning-oriented data and instruction tuning to excel at tasks requiring systematic thought. Compared to traditional LLMs, LRMs like R1 are better at decomposing complex problems, following long-term logical chains, and aligning intermediate steps with final outputs. LRMs have now being used in many topics: in motion generation tasks(Ouyang et al. 2025), LRMs can reasoning over temporal sequences and physical constraints, thus synthesize realistic and controllable human motion trajectories from abstract instructions or sparse keyframes, in computer vision(Shen et al. 2025), VLM-R1 demonstrates strong generalization across diverse tasks such as visual question answering, image captioning, and referring expression comprehension and also by modeling multi-entity interactions through structured reasoning, LRMs can better capture the semantic dependencies between humans, objects, and actions in HOI detection tasks(Li et al. 2025a).

3 Method

We propose a unified framework that integrates three components. An HOI detection network learns to localize inter-

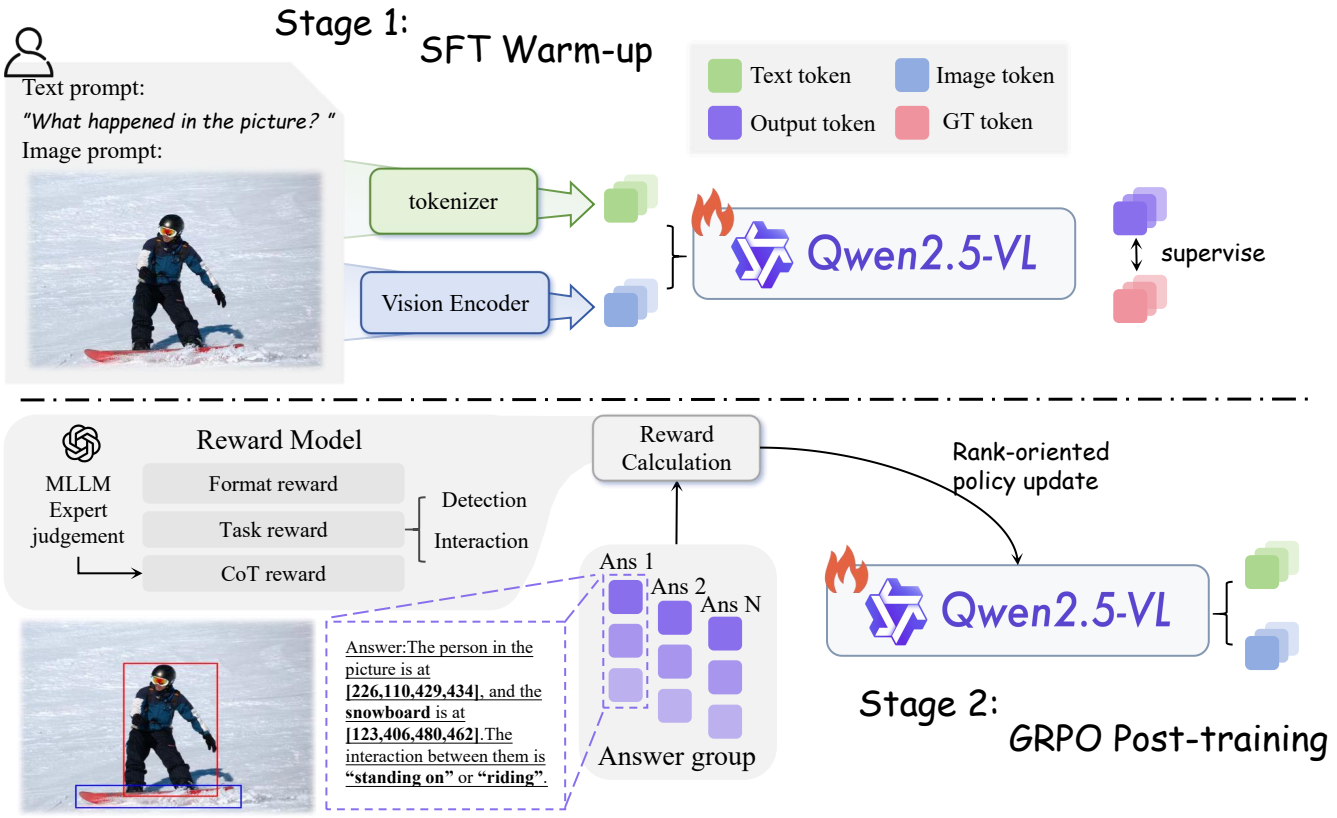

Figure 2: **Pipeline.** HOID-R1 first encodes the input image and free-form query into multimodal embeddings and warms up via SFT to generate chain-of-thought–annotated HOI triplets. During GRPO post-training, it samples candidate triplets, uses an MLLM judge to compute a composite reward on format compliance, detection accuracy, interaction classification, and CoT coherence, and updates the policy for precise localization and faithful reasoning.

actions and classify actions and objects using intersection-over-union and classification rewards. A chain of thought generator produces intermediate reasoning steps that are scored by a Process Reward Model and a Generalizable Reward Model to form a reasoning reward. A pretrained multimodal LLM judge uses these two models to provide mixed supervision at both individual steps and groups of steps. All rewards are combined into a single objective and optimized via reinforcement learning, yielding precise localization, accurate classification, and coherent reasoning.

3.1 Supervised Fine-Tuning

To bridge the gap between general reasoning capabilities and the structured requirements of HOI detection outputs, we employ Supervised Fine-Tuning (SFT) as an essential preparatory stage. Directly applying general-purpose reasoning models such as DeepSeek-R1 to HOI-related tasks often leads to inconsistently formatted outputs, especially when generating scene graph-style captions or structured interaction tuples. To address this, we design an SFT stage where the model is guided to adhere strictly to the desired output format (e.g., ⟨subject, verb, object⟩ triplets(Chao et al. 2018)), using curated demonstrations.

However, relying solely on rigid format supervision may inadvertently suppress the model's intrinsic reasoning ability, reducing it to pattern-matching instead of true inference. To counteract this, we introduce Cognitive Chain-of-Thought (CoT)(Li et al. 2025b) prompting within the SFT stage. Here, stepwise reasoning processes grounding them in the visual input, and inferring their interaction, are explicitly annotated using special ¡think¿ tags. This encourages the model to internalize a cognitively grounded reasoning procedure while still learning to output syntactically and semantically structured HOI predictions. Empirically, this hybrid supervision strategy enhances both interpretability and output consistency.

We impose task-specific constraints to reflect the domain characteristics of HOI detection, we specifies the domain of the thinking process: 1) human detection(the first entity of the CoT is human); 2) object detection(the second entity is restrict to objects); 3) relation existence(we tried to limit the relation within several verbs). This guides the model to focus its reasoning within a valid interaction space, reducing hallucination and improving output faithfulness. Furthermore, this constraint-aware design helps prevent overfitting to idiosyncratic CoT patterns, ensuring better general-

ization across diverse scenes.

Empirically, this hybrid strategy—combining structured format supervision with constrained yet expressive CoT reasoning—significantly improves both the interpretability and consistency of HOI predictions.

3.2 MLLM-as-a-Judge

Large language models often produce correct final answers while their intermediate reasoning contains semantic deviations or logical flaws. To address this problem, we introduce MLLM-as-a-Judge(Chen et al. 2024), which uses a pretrained multimodal large language model to provide mixed supervision over the chain of thought generation.

Process Reward Model (PRM): The PRM assigns a score or a binary judgment ("Is this step correct?") at each reasoning step in the generated CoT, and uses that as the training signal to fine-tune the model. In other words, the PRM provides correctness/incorrectness labels or scores for every intermediate step, teaching the model to be reliable at each stage of its reasoning.

Generalizable Reward Model (GRM): Building on standard preference learning, the GRM additionally preserves and regularizes the reward model's generative ability—applying a language-modeling loss to its hidden states—so that it can both evaluate the quality of reasoning and generate coherent intermediate steps like a language model. This generative supervision significantly boosts the reward model's generalization to unseen tasks or out-of-distribution samples.

During training, the student model's generated reasoning chain is fed into the MLLM judge, which generates step-level and generalizable feedback signals through PRM and GRM, and combines these signals into a composite reward within a reinforcement learning framework. By enforcing constraints on the intermediate reasoning process, the MLLM-as-a-Judge mechanism effectively suppresses tendencies to arrive at correct outcomes through flawed logic, thereby improving the reliability and interpretability of the model's reasoning.

3.3 Group Relative Policy Optimization

To optimize the R1 framework, the group relative policy optimization is adopted. GRPO is a reinforcement learning (RL) algorithm designed for optimizing policy models. As proposed in DeepSeek-R1(DeepSeek-AI et al. 2025), GRPO eliminates dependency on critic networks by leveraging direct response comparisons within stochastically sampled output groups. This approach reduces computational overhead while maintaining optimization stability. The objective function of GRPO leverages direct pairwise comparisons within the sampled group through implicit reward normalization. By using the group mean as a dynamic baseline, it reduces gradient variance while maintaining optimization stability. Crucially, it eliminates per-token value estimation, cutting computational overhead by $\mathcal{O}(n)$ for sequence length n compared to critic-dependent methods:

$$A = \frac{1}{G}\sum_{i=1}^{G} \min\left(\rho_i A_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)A_i\right) \quad (1)$$

$$B = \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \quad (2)$$

$$J_{\text{GRPO}}(\theta) = \mathbf{E}_{q\sim\mathcal{Q},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}}\left[A - B\right] \quad (3)$$

where $\rho_i = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}$ quantifies policy shift between current policy π_θ and behavioral policy $\pi_{\theta_{\text{old}}}$., ϵ controls clipping thresholds and β indicates the deviations via the KL divergence term. The advantage score A_i mitigates reward scale sensitivity and reduces gradient variance, which can standardize rewards to stabilize training:

$$A_i = \frac{r_i - \text{mean}(\{r_1, ..., r_c\})}{\text{std}(\{r_1, ..., r_c\})} \quad (4)$$

r_i represents the reward for response o_i. The KL Divergence is defined to control exploration without too much divergence from π_{ref} as follows:

$$D_{KL}(\pi_\theta||\pi_{ref}) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log(\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)}) - 1 \quad (5)$$

Our reward design generally consists of four parts: Format Reward, Detection Reward, Interaction Reward, and CoT Reward.

Format Reward This reward is designed to restrict the reasoning template format: Thus, the reward design is:

$$r_{format}(o_i) = \begin{cases} 1 & \text{if } o_i\text{format is right} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

This binary function thus enforces format requirements on syntax while granting flexibility in content.

Detection Reward Inspired by established object detection practices, we devise the sample-level IoU score R_{IoU} and the sample-level regression accuracy R_{reg} in the predicted anchor boxes for each sample(Carion et al. 2020),(Ren et al. 2016). The former measures the fraction of anchors whose Intersection over Union with their ground-truth boxes reaches or exceeds 0.5; the latter measures the fraction of anchors whose normalized L1 coordinate error falls below a threshold δ. Specifically, an anchor prediction is deemed correct if:

- **Overlap Accuracy** The predicted anchor box achieves an Intersection over Union of at least 0.5 with its ground-truth counterpart.
- **Coordinate Precision** The normalized L1 distance between predicted and ground-truth box coordinates is below the threshold δ.

The final detection reward is computed as a weighted combination of these two metrics:

$$r_{\text{det}}(o) = \beta \cdot R_{\text{IoU}} + (1-\beta) \cdot R_{\text{reg}}, \quad (7)$$

where β balances the trade-off between overlap accuracy and coordinate precision.

Model	Seen				Unseen			
	H-mIOU↑	**O-mIOU↑**	**A-ACC↑**	**mAP↑**	**H-mIOU↑**	**O-mIOU↑**	**A-ACC↑**	**mAP↑**
Fine-grained annotation								
HOI-Trans	$0.56^{\pm 0.018}$	$0.54^{\pm 0.015}$	$0.60^{\pm 0.028}$	$33.64^{\pm 1.27}$	$0.51^{\pm 0.021}$	$0.52^{\pm 0.019}$	$0.56^{\pm 0.025}$	$31.28^{\pm 1.34}$
DiffHOI	$0.63^{\pm 0.023}$	$0.61^{\pm 0.020}$	$0.67^{\pm 0.027}$	$44.06^{\pm 1.91}$	$0.61^{\pm 0.024}$	$0.60^{\pm 0.018}$	$0.63^{\pm 0.026}$	$42.25^{\pm 1.68}$
PAFR	$0.65^{\pm 0.025}$	$0.66^{\pm 0.022}$	$0.72^{\pm 0.029}$	$51.22^{\pm 2.39}$	$0.63^{\pm 0.026}$	$0.62^{\pm 0.020}$	$0.67^{\pm 0.031}$	$47.51^{\pm 1.89}$
HORP	$0.67^{\pm 0.021}$	$0.69^{\pm 0.024}$	$0.70^{\pm 0.032}$	$52.75^{\pm 1.84}$	$0.66^{\pm 0.027}$	$0.66^{\pm 0.023}$	$0.69^{\pm 0.030}$	$49.03^{\pm 2.05}$
Ours	$\mathbf{0.72}^{\pm 0.026}$	$\mathbf{0.73}^{\pm 0.028}$	$\mathbf{0.79}^{\pm 0.031}$	$\mathbf{55.07}^{\pm 2.11}$	$\mathbf{0.71}^{\pm 0.025}$	$\mathbf{0.70}^{\pm 0.027}$	$\mathbf{0.75}^{\pm 0.030}$	$\mathbf{52.98}^{\pm 2.18}$
Precise annotation								
HOI-Trans	$0.55^{\pm 0.019}$	$0.54^{\pm 0.021}$	$0.58^{\pm 0.028}$	$32.46^{\pm 1.12}$	$0.50^{\pm 0.022}$	$0.51^{\pm 0.017}$	$0.54^{\pm 0.024}$	$29.55^{\pm 1.26}$
DiffHOI	$0.62^{\pm 0.020}$	$0.61^{\pm 0.025}$	$0.65^{\pm 0.031}$	$42.17^{\pm 1.53}$	$0.60^{\pm 0.019}$	$0.59^{\pm 0.024}$	$0.61^{\pm 0.027}$	$40.05^{\pm 1.78}$
PAFR	$0.64^{\pm 0.022}$	$0.65^{\pm 0.026}$	$0.70^{\pm 0.034}$	$49.68^{\pm 2.27}$	$0.62^{\pm 0.021}$	$0.61^{\pm 0.022}$	$0.65^{\pm 0.030}$	$46.33^{\pm 2.05}$
HORP	$0.66^{\pm 0.018}$	$0.68^{\pm 0.020}$	$0.69^{\pm 0.028}$	$51.46^{\pm 1.76}$	$0.64^{\pm 0.023}$	$0.65^{\pm 0.021}$	$0.67^{\pm 0.031}$	$47.65^{\pm 1.99}$
Ours	$\mathbf{0.70}^{\pm 0.025}$	$\mathbf{0.72}^{\pm 0.027}$	$\mathbf{0.76}^{\pm 0.030}$	$\mathbf{53.81}^{\pm 2.13}$	$\mathbf{0.71}^{\pm 0.026}$	$\mathbf{0.69}^{\pm 0.025}$	$\mathbf{0.72}^{\pm 0.029}$	$\mathbf{51.19}^{\pm 2.21}$
Open-Vocabulary annotation								
HOI-Trans	$0.54^{\pm 0.020}$	$0.52^{\pm 0.018}$	$0.57^{\pm 0.025}$	$29.17^{\pm 1.13}$	$0.49^{\pm 0.022}$	$0.50^{\pm 0.019}$	$0.52^{\pm 0.023}$	$28.90^{\pm 1.07}$
DiffHOI	$0.61^{\pm 0.023}$	$0.60^{\pm 0.024}$	$0.64^{\pm 0.030}$	$40.95^{\pm 1.55}$	$0.59^{\pm 0.025}$	$0.58^{\pm 0.021}$	$0.60^{\pm 0.026}$	$39.44^{\pm 1.65}$
PAFR	$0.63^{\pm 0.025}$	$0.63^{\pm 0.022}$	$0.68^{\pm 0.031}$	$45.28^{\pm 2.06}$	$0.61^{\pm 0.021}$	$0.60^{\pm 0.020}$	$0.63^{\pm 0.030}$	$43.78^{\pm 1.74}$
HORP	$0.65^{\pm 0.020}$	$0.66^{\pm 0.025}$	$0.67^{\pm 0.029}$	$46.09^{\pm 2.12}$	$0.63^{\pm 0.024}$	$0.63^{\pm 0.021}$	$0.65^{\pm 0.028}$	$45.12^{\pm 2.04}$
Ours	$\mathbf{0.69}^{\pm 0.024}$	$\mathbf{0.71}^{\pm 0.027}$	$\mathbf{0.75}^{\pm 0.032}$	$\mathbf{51.68}^{\pm 2.36}$	$\mathbf{0.68}^{\pm 0.023}$	$\mathbf{0.68}^{\pm 0.022}$	$\mathbf{0.72}^{\pm 0.030}$	$\mathbf{50.03}^{\pm 2.14}$

Table 1: **Main Result on HICO-DET.**

Interaction Reward: Inspired by established classification practices, we define the sample-level action accuracy R_{act} and the sample-level object accuracy R_{obj} for each sample. The first metric measures the fraction of samples whose predicted action label matches the ground truth action. The second metric measures the fraction of samples whose predicted object label matches the ground truth object. Specifically, an interaction prediction is correct if:

- **Action Correctness**: The model's predicted action label (for example "pick up", "pour", or "rotate") must match the ground truth action label exactly. A sample is counted as an action correct only when the predicted category corresponds to the annotated category.
- **Object Correctness**: The model's predicted object label (for example "cup", "bottle", or "book") must match the ground truth object label exactly. A sample is considered object correct only when the predicted name aligns with the annotated object name unambiguously.

The final interaction reward is computed as a weighted combination of these two metrics:

$$r_{\text{int}}(o) = \gamma \cdot R_{\text{act}} + (1 - \gamma) \cdot R_{\text{obj}}, \tag{8}$$

where γ balances the importance of action correctness and object correctness.

CoT Reward To encourage both fine-grained relevance and high-level coherence in the model's intermediate reasoning(Wang et al. 2023), we design a *Chain-of-Thought* reward r_{CoT} that integrates two complementary signals:

- **Process Reward Model (PRM).** (Wang et al. 2025) Let N be the number of steps in the generated CoT, and let $s_i \in [0, 1]$ be the PRM score for step i. We define the step-level reward

$$R_{\text{prm}} = \frac{1}{N} \sum_{i=1}^{N} s_i. \tag{9}$$

- **Generalizable Reward Model (GRM).** Partition the chain into M groups of consecutive steps, and let $g_j \in [0, 1]$ be the GRM score for group j(Ouyang et al. 2022),(Rafailov et al. 2024). We define the group-level reward

$$R_{\text{grm}} = \frac{1}{M} \sum_{j=1}^{M} g_j. \tag{10}$$

These two signals are combined into a single scalar reward:

$$r_{\text{CoT}} = \lambda\, R_{\text{prm}} \;+\; (1 - \lambda)\, R_{\text{grm}}, \tag{11}$$

where $\lambda \in [0, 1]$ balances the emphasis between step-level accuracy and group-level coherence.

By optimizing this reward within a reinforcement learning framework, the model is encouraged to produce reasoning trajectories that are both semantically aligned with the prompt and logically coherent throughout.

4 Experiments

Our model is trained in two phases: (1) supervised fine-tuning on human–object interaction data; and (2) post-training using GRPO. We employ diverse datasets and evaluation metrics to assess our approach, demonstrating its capability to detect HOI under high-level instructions in both seen and unseen scenarios. Comparative experiments and

Model	Seen				Unseen			
	H-mIOU↑	O-mIOU↑	A-ACC↑	mAP↑	H-mIOU↑	O-mIOU↑	A-ACC↑	mAP↑
Fine-grained annotation								
HOI-Trans	$0.38^{\pm 0.012}$	$0.36^{\pm 0.014}$	$0.41^{\pm 0.019}$	$22.94^{\pm 1.08}$	$0.33^{\pm 0.016}$	$0.34^{\pm 0.015}$	$0.38^{\pm 0.016}$	$21.39^{\pm 1.06}$
DiffHOI	$0.41^{\pm 0.015}$	$0.39^{\pm 0.019}$	$0.43^{\pm 0.017}$	$29.93^{\pm 1.47}$	$0.40^{\pm 0.015}$	$0.39^{\pm 0.014}$	$0.41^{\pm 0.015}$	$28.47^{\pm 1.29}$
PAFR	$0.45^{\pm 0.016}$	$0.45^{\pm 0.014}$	$0.49^{\pm 0.021}$	$34.34^{\pm 1.50}$	$0.42^{\pm 0.015}$	$0.41^{\pm 0.019}$	$0.44^{\pm 0.018}$	$32.39^{\pm 1.42}$
HORP	$0.46^{\pm 0.018}$	$0.47^{\pm 0.015}$	$0.48^{\pm 0.015}$	$35.99^{\pm 1.71}$	$0.44^{\pm 0.017}$	$0.44^{\pm 0.014}$	$0.46^{\pm 0.018}$	$34.28^{\pm 1.58}$
Ours	$\mathbf{0.49}^{\pm 0.018}$	$\mathbf{0.49}^{\pm 0.019}$	$\mathbf{0.55}^{\pm 0.021}$	$\mathbf{37.74}^{\pm 1.88}$	$\mathbf{0.47}^{\pm 0.017}$	$\mathbf{0.46}^{\pm 0.017}$	$\mathbf{0.49}^{\pm 0.021}$	$\mathbf{35.51}^{\pm 1.73}$
Precise annotation								
HOI-Trans	$0.37^{\pm 0.012}$	$0.36^{\pm 0.013}$	$0.39^{\pm 0.018}$	$21.75^{\pm 1.07}$	$0.33^{\pm 0.015}$	$0.34^{\pm 0.015}$	$0.36^{\pm 0.017}$	$19.81^{\pm 0.96}$
DiffHOI	$0.40^{\pm 0.017}$	$0.39^{\pm 0.014}$	$0.42^{\pm 0.016}$	$28.02^{\pm 1.39}$	$0.39^{\pm 0.015}$	$0.38^{\pm 0.015}$	$0.40^{\pm 0.019}$	$26.23^{\pm 1.23}$
PAFR	$0.43^{\pm 0.016}$	$0.44^{\pm 0.016}$	$0.46^{\pm 0.020}$	$33.28^{\pm 1.66}$	$0.41^{\pm 0.017}$	$0.40^{\pm 0.015}$	$0.42^{\pm 0.021}$	$30.93^{\pm 1.44}$
HORP	$0.44^{\pm 0.019}$	$0.46^{\pm 0.017}$	$0.46^{\pm 0.019}$	$34.47^{\pm 1.59}$	$0.42^{\pm 0.017}$	$0.44^{\pm 0.017}$	$0.44^{\pm 0.020}$	$31.77^{\pm 1.43}$
Ours	$\mathbf{0.47}^{\pm 0.017}$	$\mathbf{0.48}^{\pm 0.018}$	$\mathbf{0.50}^{\pm 0.021}$	$\mathbf{36.05}^{\pm 1.81}$	$\mathbf{0.47}^{\pm 0.017}$	$\mathbf{0.45}^{\pm 0.016}$	$\mathbf{0.48}^{\pm 0.021}$	$\mathbf{34.19}^{\pm 1.54}$
Open-Vocabulary annotation								
HOI-Trans	$0.35^{\pm 0.015}$	$0.34^{\pm 0.014}$	$0.37^{\pm 0.017}$	$19.73^{\pm 0.99}$	$0.32^{\pm 0.015}$	$0.32^{\pm 0.013}$	$0.34^{\pm 0.017}$	$18.79^{\pm 0.92}$
DiffHOI	$0.40^{\pm 0.015}$	$0.39^{\pm 0.016}$	$0.42^{\pm 0.018}$	$27.02^{\pm 1.38}$	$0.38^{\pm 0.017}$	$0.37^{\pm 0.016}$	$0.39^{\pm 0.019}$	$25.63^{\pm 1.30}$
PAFR	$0.41^{\pm 0.017}$	$0.41^{\pm 0.017}$	$0.44^{\pm 0.019}$	$29.43^{\pm 1.51}$	$0.40^{\pm 0.016}$	$0.39^{\pm 0.016}$	$0.41^{\pm 0.017}$	$28.02^{\pm 1.38}$
HORP	$0.42^{\pm 0.017}$	$0.43^{\pm 0.016}$	$0.45^{\pm 0.019}$	$30.17^{\pm 1.61}$	$0.41^{\pm 0.017}$	$0.41^{\pm 0.017}$	$0.42^{\pm 0.018}$	$29.33^{\pm 1.47}$
Ours	$\mathbf{0.45}^{\pm 0.018}$	$\mathbf{0.47}^{\pm 0.017}$	$\mathbf{0.49}^{\pm 0.020}$	$\mathbf{33.64}^{\pm 1.72}$	$\mathbf{0.44}^{\pm 0.017}$	$\mathbf{0.44}^{\pm 0.017}$	$\mathbf{0.46}^{\pm 0.020}$	$\mathbf{32.52}^{\pm 1.59}$

Table 2: **Main Result on SWIG-HOI.**

ablation studies validate our design choices. All experiments are conducted on eight NVIDIA A800 GPUs.

4.1 Dataset

In our experiments, we selected two of the most commonly used benchmark datasets in the human-object interaction (HOI) detection field and re-annotated them.

HICO-DET(Chao et al. 2015) HICO-DET includes 47,776 images (38,118 for training, 9,658 for testing) annotated with 600 HOI categories formed by 117 verb classes and 80 object classes, totaling over 150,000 human–object pairs. Following standard zero-shot protocols, 120 of the rarest interaction triplets are withheld during training to assess a model's ability to recognize unseen images

SWIG-HOI(Pratt et al. 2020) Assembled from the SWIG and DOH datasets, SWIG-HOI comprises roughly 45,000 training images and 14,000 test images, covering 406 human actions and 1,000 object categories. Its test split contains about 5,500 human–object interaction instances, of which nearly 1,800 interactions are not seen during training, making it a challenging benchmark for open-vocabulary HOI detection

We annotate two datasets using three distinct schemes: fine-grained annotation, precise annotation, and open-vocabulary annotation. Each successive scheme places increasingly greater demands on the model's generalization capabilities.

- **Fine-grained annotation**: accurately describing the diverse attributes and actions of the people and objects depicted in the image. (e.g., "A man wearing black clothes is drinking a blue cup of water")

Figure 3: **Qualitative result.**More Qualitative results in the Appendix.

- **Precise annotation**: describing only the interactive actions and object depicted in the image. (e.g., "A man is drinking water")
- **Open-vocabulary annotation**: providing only an open-ended description of the person or object in the image (e.g., "What is the man doing?", "What action is the cup performing?"), or posing a broad query (e.g., "What is happening in the image?").

In addition, we partition the images into seen and unseen subsets and evaluate our method on each. The results show that our model achieves state-of-the-art performance across all six experimental settings defined by the three annotation schemes. Notably, it demonstrates strong generalization under open-vocabulary annotation on unseen images.

Model	Seen				Unseen			
	H-mIOU↑	O-mIOU↑	A-ACC↑	mAP↑	H-mIOU↑	O-mIOU↑	A-ACC↑	mAP↑
Open-Vocabulary annotation								
W/O PT	$0.54^{\pm 0.017}$	$0.53^{\pm 0.018}$	$0.57^{\pm 0.021}$	$39.34^{\pm 1.92}$	$0.54^{\pm 0.016}$	$0.53^{\pm 0.019}$	$0.57^{\pm 0.023}$	$37.74^{\pm 1.75}$
W/O FR	$0.59^{\pm 0.020}$	$0.62^{\pm 0.025}$	$0.67^{\pm 0.031}$	$43.94^{\pm 1.85}$	$0.61^{\pm 0.019}$	$0.60^{\pm 0.020}$	$0.62^{\pm 0.029}$	$42.91^{\pm 1.68}$
W/O DR	$0.60^{\pm 0.022}$	$0.60^{\pm 0.019}$	$0.65^{\pm 0.030}$	$45.23^{\pm 2.14}$	$0.58^{\pm 0.021}$	$0.58^{\pm 0.018}$	$0.64^{\pm 0.027}$	$43.89^{\pm 1.96}$
W/O IR	$0.62^{\pm 0.023}$	$0.62^{\pm 0.022}$	$0.64^{\pm 0.026}$	$44.18^{\pm 2.20}$	$0.61^{\pm 0.022}$	$0.60^{\pm 0.020}$	$0.64^{\pm 0.028}$	$44.35^{\pm 2.03}$
W/O CoTR	$0.60^{\pm 0.018}$	$0.64^{\pm 0.022}$	$0.65^{\pm 0.029}$	$45.35^{\pm 2.15}$	$0.61^{\pm 0.018}$	$0.60^{\pm 0.021}$	$0.64^{\pm 0.025}$	$43.97^{\pm 1.79}$
Ours	$\mathbf{0.69}^{\pm 0.024}$	$\mathbf{0.71}^{\pm 0.027}$	$\mathbf{0.75}^{\pm 0.032}$	$\mathbf{51.68}^{\pm 2.36}$	$\mathbf{0.68}^{\pm 0.023}$	$\mathbf{0.68}^{\pm 0.022}$	$\mathbf{0.72}^{\pm 0.030}$	$\mathbf{50.03}^{\pm 2.14}$

Table 3: **Ablation study on HICO-DET of Open-Vocabulary annotations.**

4.2 Evaluation metric

To accurately assess our model's performance, we adopt four evaluation metrics distilled from prior work.

- **H-mIOU** This metric is used to compute the mean Intersection over Union between the predicted person bounding boxes and the ground-truth.
- **O-mIOU** This metric calculates the mean Intersection over Union (mIoU) between the predicted object bounding boxes and their ground-truth.
- **A-ACC** This metric computes the probability of successfully predicting the interaction action.
- **mAP** A detection is deemed successful when both H-mIoU and O-mIoU exceed 0.5, and this metric measures the corresponding success rate.

Higher H-mIOU, O-mIOU, A-ACC and mAP mean powerful detection and reasoning ability of the method

4.3 Main Result

Compare to SOTA method We evaluate the four metrics on seen and unseen images from both datasets under all three annotation schemes, comparing our approach with recent state-of-the-art methods (e.g., HOI-Trans (Zou et al. 2021), DiffHOI(Yang et al. 2023), PAFR(Wu et al. 2024), HORP(Geng, Yang, and Zhang 2025)). The results are summarized in Tables 1 and 2.

Experimental results demonstrate that our model achieves state-of-the-art performance across all evaluation metrics. By leveraging the combined strengths of supervised fine-tuning and GRPO-based post-training, our approach exhibits robust open-world generalization, especially for combinations of open-vocabulary descriptions and unseen images.

Qualitative results We present visualized results that demonstrate our model's open-world generalization capabilities with Figure 3, showing strong performance under open-vocabulary descriptions on unseen images. More Qualitative results in the Appendix.

4.4 Ablation Study

We perform ablation studies on the open-vocabulary annotations of the HICO-DET dataset to examine the contribution of each component to open-world HOI detection. The detailed results are presented in Table 3.

W/O Post-training(W/O PT). We remove the GRPO-based post-training stage, which leads to a substantial drop in all four evaluation metrics. Without this reinforcement-learning fine-tuning, the model cannot leverage the multi-reward signals to refine its policy beyond the supervised fine-tuning stage, resulting in poorer localization (H-mIOU/O-mIOU) and degraded detection and action classification accuracy.

W/O Format Reward(W/O FR). We omit the format compliance reward, causing the model to generate HOI predictions with inconsistent or malformed ⟨subject, verb, object⟩ structures. Without this binary constraint, syntactic errors proliferate-missing tags, malformed tuples, and irregular delimiters-which in turn disrupt downstream parsing and degrade overall performance.

W/O Detection Reward(W/O DR). We disable the detection reward (RIoU and Rreg), removing the incentive for precise bounding-box overlap and coordinate accuracy. Consequently, both H-mIOU and O-mIOU suffer significant declines, leading to lower mAP as the model neglects fine-grained localization in favor of other objectives. This shows that the detection reward is indispensable for driving accurate spatial predictions.

W/O Interaction Reward(W/O IR). We turn off the interaction reward, so the model no longer receives feedback on correct action and object classification. As a result, A-ACC drops markedly, with the model frequently mislabeling verbs or objects despite reasonable bounding boxes. This indicates that the interaction reward is crucial for aligning the model's predictions with the semantic ground truth.

W/O CoT Reward(W/O CoTR). We remove the chain-of-thought reward, preventing optimization of the reasoning quality and coherence. The generated reasoning chains become semantically misaligned with the input prompt and logically disjointed across steps, leading to more hallucinations and less interpretable intermediate outputs. This validates that the CoT reward is vital for ensuring semantically meaningful and logically coherent reasoning trajectories.

5 Conclusion

This work presents HOID-R1, a unified framework for open-world human–object interaction detection that integrates chain-of-thought supervised fine-tuning with Group Relative Policy Optimization. By enforcing structured output for-

mats and employing an MLLM-based judge to supervise intermediate reasoning, the approach mitigates hallucinations and grounds predictions in meaningful affordance cues. Extensive evaluations on HICO-DET and SWIG-HOI demonstrate that HOID-R1 outperforms existing methods in both seen and unseen settings under open-vocabulary evaluation, while ablation studies confirm the contribution of each component. Future research will focus on enhancing computational efficiency, extending the framework to video-based HOI detection for improved temporal coherence.

References

Cao, Y.; Tang, Q.; Yang, F.; Su, X.; You, S.; Lu, X.; and Xu, C. 2023. Re-mine, Learn and Reason: Exploring the Cross-modal Semantic Correlations for Language-guided HOI detection. arXiv:2307.13529.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. arXiv:2005.12872.

Chao, Y.-W.; Liu, Y.; Liu, X.; Zeng, H.; and Deng, J. 2018. Learning to Detect Human-Object Interactions. arXiv:1702.05448.

Chao, Y.-W.; Wang, Z.; He, Y.; Wang, J.; and Deng, J. 2015. HICO: A Benchmark for Recognizing Human-Object Interactions in Images. In *Proceedings of the IEEE International Conference on Computer Vision*.

Chen, D.; Chen, R.; Zhang, S.; Liu, Y.; Wang, Y.; Zhou, H.; Zhang, Q.; Wan, Y.; Zhou, P.; and Sun, L. 2024. MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark. arXiv:2402.04788.

Chen, M.; Liao, Y.; Liu, S.; Chen, Z.; Wang, F.; and Qian, C. 2021. Reformulating HOI Detection as Adaptive Set Prediction. arXiv:2103.05983.

DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; Zhang, X.; Yu, X.; Wu, Y.; Wu, Z. F.; Gou, Z.; Shao, Z.; Li, Z.; Gao, Z.; Liu, A.; Xue, B.; Wang, B.; Wu, B.; Feng, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; Dai, D.; Chen, D.; Ji, D.; Li, E.; Lin, F.; Dai, F.; Luo, F.; Hao, G.; Chen, G.; Li, G.; Zhang, H.; Bao, H.; Xu, H.; Wang, H.; Ding, H.; Xin, H.; Gao, H.; Qu, H.; Li, H.; Guo, J.; Li, J.; Wang, J.; Chen, J.; Yuan, J.; Qiu, J.; Li, J.; Cai, J. L.; Ni, J.; Liang, J.; Chen, J.; Dong, K.; Hu, K.; Gao, K.; Guan, K.; Huang, K.; Yu, K.; Wang, L.; Zhang, L.; Zhao, L.; Wang, L.; Zhang, L.; Xu, L.; Xia, L.; Zhang, M.; Zhang, M.; Tang, M.; Li, M.; Wang, M.; Li, M.; Tian, N.; Huang, P.; Zhang, P.; Wang, Q.; Chen, Q.; Du, Q.; Ge, R.; Zhang, R.; Pan, R.; Wang, R.; Chen, R. J.; Jin, R. L.; Chen, R.; Lu, S.; Zhou, S.; Chen, S.; Ye, S.; Wang, S.; Yu, S.; Zhou, S.; Pan, S.; Li, S. S.; Zhou, S.; Wu, S.; Ye, S.; Yun, T.; Pei, T.; Sun, T.; Wang, T.; Zeng, W.; Zhao, W.; Liu, W.; Liang, W.; Gao, W.; Yu, W.; Zhang, W.; Xiao, W. L.; An, W.; Liu, X.; Wang, X.; Chen, X.; Nie, X.; Cheng, X.; Liu, X.; Xie, X.; Liu, X.; Yang, X.; Li, X.; Su, X.; Lin, X.; Li, X. Q.; Jin, X.; Shen, X.; Chen, X.; Sun, X.; Wang, X.; Song, X.; Zhou, X.; Wang, X.; Shan, X.; Li, Y. K.; Wang, Y. Q.; Wei, Y. X.; Zhang, Y.; Xu, Y.; Li, Y.; Zhao, Y.; Sun, Y.; Wang, Y.; Yu, Y.; Zhang, Y.; Shi, Y.; Xiong, Y.; He, Y.; Piao, Y.; Wang, Y.; Tan, Y.; Ma, Y.; Liu, Y.; Guo, Y.; Ou, Y.; Wang, Y.; Gong, Y.; Zou, Y.; He, Y.; Xiong, Y.; Luo, Y.; You, Y.; Liu, Y.; Zhou, Y.; Zhu, Y. X.; Xu, Y.; Huang, Y.; Li, Y.; Zheng, Y.; Zhu, Y.; Ma, Y.; Tang, Y.; Zha, Y.; Yan, Y.; Ren, Z. Z.; Ren, Z.; Sha, Z.; Fu, Z.; Xu, Z.; Xie, Z.; Zhang, Z.; Hao, Z.; Ma, Z.; Yan, Z.; Wu, Z.; Gu, Z.; Zhu, Z.; Liu, Z.; Li, Z.; Xie, Z.; Song, Z.; Pan, Z.; Huang, Z.; Xu, Z.; Zhang, Z.; and Zhang, Z. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Gao, C.; Zou, Y.; and Huang, J.-B. 2018. iCAN: Instance-Centric Attention Network for Human-Object Interaction Detection. arXiv:1808.10437.

Geng, P.; Yang, J.; and Zhang, S. 2025. HORP: Human-Object Relation Priors Guided HOI Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 25325–25335.

Kim, B.; Choi, T.; Kang, J.; and Kim, H. J. 2023. UnionDet: Union-Level Detector Towards Real-Time Human-Object Interaction Detection. arXiv:2312.12664.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2023. Large Language Models are Zero-Shot Reasoners. arXiv:2205.11916.

Li, L.; Chen, W.; Li, J.; Cheng, K.-T.; and Chen, L. 2025a. Relation-R1: Progressively Cognitive Chain-of-Thought Guided Reinforcement Learning for Unified Relation Comprehension. arXiv:2504.14642.

Li, L.; Chen, W.; Li, J.; Cheng, K.-T.; and Chen, L. 2025b. Relation-R1: Progressively Cognitive Chain-of-Thought Guided Reinforcement Learning for Unified Relation Comprehension. arXiv:2504.14642.

Liao, Y.; Liu, S.; Wang, F.; Chen, Y.; Qian, C.; and Feng, J. 2020. PPDM: Parallel Point Detection and Matching for Real-time Human-Object Interaction Detection. arXiv:1912.12898.

Ning, S.; Qiu, L.; Liu, Y.; and He, X. 2023. HOICLIP: Efficient Knowledge Transfer for HOI Detection with Vision-Language Models. arXiv:2303.15786.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Ouyang, R.; Li, H.; Zhang, Z.; Wang, X.; Zhu, Z.; Huang, G.; and Wang, X. 2025. Motion-R1: Chain-of-Thought Reasoning and Reinforcement Learning for Human Motion Generation. arXiv:2506.10353.

Pratt, S.; Yatskar, M.; Weihs, L.; Farhadi, A.; and Kembhavi, A. 2020. Grounded Situation Recognition. arXiv:2003.12058.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv:1506.01497.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300.

Shen, H.; Liu, P.; Li, J.; Fang, C.; Ma, Y.; Liao, J.; Shen, Q.; Zhang, Z.; Zhao, K.; Zhang, Q.; Xu, R.; and Zhao, T. 2025. VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model. arXiv:2504.07615.

Tamura, M.; Ohashi, H.; and Yoshinaga, T. 2021. QPIC: Query-Based Pairwise Human-Object Interaction Detection with Image-Wide Contextual Information. arXiv:2103.05399.

Wang, W.; Gao, Z.; Chen, L.; Chen, Z.; Zhu, J.; Zhao, X.; Liu, Y.; Cao, Y.; Ye, S.; Zhu, X.; Lu, L.; Duan, H.; Qiao, Y.; Dai, J.; and Wang, W. 2025. VisualPRM: An Effective Process Reward Model for Multimodal Reasoning. arXiv:2503.10291.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Wu, E. Z. Y.; Li, Y.; Wang, Y.; and Wang, S. 2024. Exploring Pose-Aware Human-Object Interaction via Hybrid Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17815–17825.

Yang, J.; Li, B.; Yang, F.; Zeng, A.; Zhang, L.; and Zhang, R. 2023. Boosting Human-Object Interaction Detection with Text-to-Image Diffusion Model. arXiv:2305.12252.

Zou, C.; Wang, B.; Hu, Y.; Liu, J.; Wu, Q.; Zhao, Y.; Li, B.; Zhang, C.; Zhang, C.; Wei, Y.; and Sun, J. 2021. End-to-End Human Object Interaction Detection with HOI Transformer. arXiv:2103.04503.