RRL-HOI: Reflective Reinforcement Learning for Open-Vocabulary HOI Detection

Yongchao Xu[1,*] **Junfeng Wang**[1,*] **Jiawei Liu**[1,†] **Yufei Zheng**[1] **Sen Tao**[2]
Tao Jiang[3] **Jiangbo Ai**[3] **Jin Zhang**[3] **Zheng-Jun Zha**[1]
[1]MoE Key Lab of BIPC, University of Science and Technology of China
[2]University of Chinese Academy of Sciences
[3]Alibaba Group
{yongchaoxu, jfwang2002, zyf2001}@mail.ustc.edu.cn
{jwliu6, zhazj}@ustc.edu.cn
taosen23@mail.ucas.ac.cn
{cigarette.jt, aijiangbo.ajb, zj146372}@alibaba-inc.com
[*]Equal contribution. [†]Corresponding author.

Abstract

Open-Vocabulary Human-Object Interaction (HOI) detection aims to localize interacting human-object pairs while generalizing to novel interaction categories beyond the training set. Although Multimodal Large Language Models (MLLMs) exhibit strong open-vocabulary understanding, directly applying MLLMs to HOI detection remains challenging, as language-prior hallucinations and coupled localization-semantic errors can destabilize structured HOI predictions. To address these issues, we propose RRL-HOI, a novel Reflective Reinforcement Learning framework that trains MLLMs to follow a verify-and-revise self-correction policy. Specifically, RRL-HOI first adapts MLLMs to produce box-grounded HOI triplets, and then introduces an executable reflection mechanism that targets the two failure modes through structured edits over localization, human-object pairing, and interaction semantics. In this way, localization and pairing edits mitigate localization-semantic coupling, while interaction-semantic edits suppress visually unsupported language-prior predictions. To further improve this reflection mechanism, RRL-HOI formulates reflection as a reinforcement learning objective trained with GRPO, where the reward jointly considers localization quality, pair-association correctness, interaction correctness, and edit consistency. Through reward-driven policy learning, RRL-HOI turns open-vocabulary recognition from a single-step generative prediction into an explicit verify-and-revise decision process, thereby bridging the gap between general multimodal understanding and precise interaction detection. Extensive experiments on two standard open-vocabulary HOI detection benchmarks demonstrate the effectiveness of our method. Code is publicly available.[1]

1 Introduction

Human-Object Interaction (HOI) detection aims to localize interacting human-object pairs and recognize their interactions, typically represented as a set of ⟨*human, action, object*⟩ triplets. However, real-world interactions are long-tailed, compositional, and constantly evolving, making it difficult for closed-vocabulary detectors [23, 34, 47, 48] to cover unseen actions, objects, and their novel combinations [20, 38]. This motivates Open-Vocabulary HOI (OV-HOI) detection, which requires recognizing novel interaction categories while preserving accurate human-object localization [19, 25].

[1]https://github.com/xuyongchao-bot/RRL-HOI

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

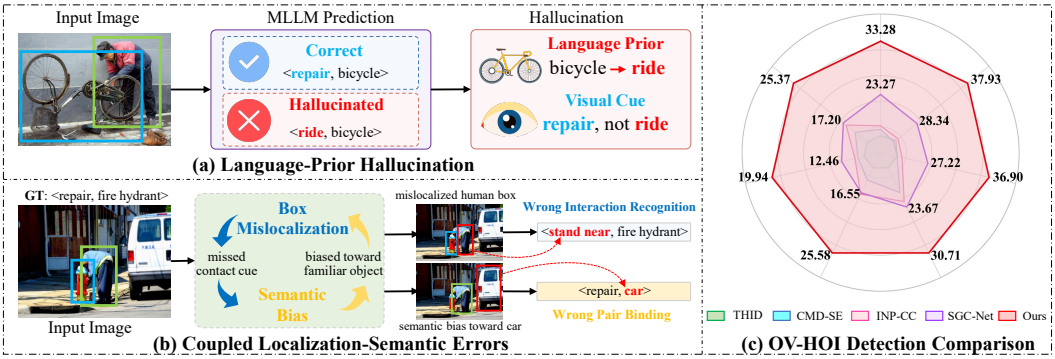


Figure 1: **Motivation and performance comparison of RRL-HOI.** MLLM-based OV-HOI detection suffers from two key issues: **(a)** *language-prior hallucination*, where visually unsupported interactions are predicted from common priors, and **(b)** *coupled localization-semantic errors*, where inaccurate boxes or pairings lead to wrong interaction recognition. **(c)** RRL-HOI addresses these issues through reflective verify-and-revise prediction, achieving stronger OV-HOI performance across benchmarks.

Recently, Multimodal Large Language Models (MLLMs) [1, 2, 7, 26, 36, 53] provide a promising way to address this open-vocabulary requirement. With rich linguistic knowledge, cross-modal alignment, and instruction-following abilities, MLLMs can interpret unseen interaction concepts through natural language rather than relying on fixed classifiers or predefined label spaces [18, 45, 50]. This makes them a natural foundation for flexible OV-HOI reasoning.

However, directly adapting MLLMs to OV-HOI detection still suffers from two key challenges, as observed in a directly adapted MLLM baseline: **(1) Language-Prior Hallucination**: Although MLLMs possess rich open-vocabulary knowledge, they tend to over-rely on linguistic priors when visual evidence is weak or ambiguous [6, 52]. As shown in Figure 1 (a), even though the image supports the interaction "repair bicycle", the model predicts the more common but visually unsupported interaction "ride bicycle". Such plausible hallucinations are especially harmful for rare or unseen interactions, undermining the reliability of structured HOI outputs. **(2) Coupled Localization-Semantic Errors**: HOI detection requires the model to localize the human and object, bind them as an interaction pair, and recognize the interaction semantics. As shown in Figure 1 (b), the ground-truth interaction is "repair fire hydrant", which relies on the contact between the human and the fire hydrant. In this case, the inaccurate human box misses the hand-contact region, causing the model to lose the key repair cue and misclassify the interaction as a weaker spatial relation, *e.g.*, "stand near fire hydrant". Meanwhile, driven by the semantic prior of "repair", the model is biased toward a more typical repairable object and incorrectly binds the action to the nearby car, producing "repair car". This example shows that localization, pair assignment, and interaction recognition are tightly coupled: an error in one component can trigger errors in the others [16, 39], making single-step MLLM generation unstable for box-grounded HOI prediction.

To address these issues, we propose RRL-HOI, a novel Reflective Reinforcement Learning framework that redefines MLLM-based OV-HOI detection as an explicit verify-and-revise decision process. Rather than treating OV-HOI detection as a single-step generative prediction, RRL-HOI first adapts the MLLM to produce box-grounded HOI triplets as an inspectable intermediate state. It then introduces an executable reflection mechanism that converts self-correction into structured edits over localization, human-object pairing, and interaction semantics, directly targeting *language-prior hallucinations* and *coupled localization-semantic errors*. To make this revision process learnable and outcome-driven, RRL-HOI further optimizes the reflection policy with GRPO [31], where verifiable rewards favor edit trajectories that improve localization quality, pair-association correctness, interaction correctness, and edit consistency. Through reward-driven reflection optimization, RRL-HOI bridges general multimodal understanding and precise interaction detection by transforming MLLM predictions into verifiable and revisable structured HOI decisions. This design is empirically substantiated by Figure 1 (c), where RRL-HOI consistently outperforms prior methods across various open-vocabulary metrics.

Our main contributions can be summarized as follows:

- We propose RRL-HOI, a reflective reinforcement learning paradigm for MLLM-based OV-HOI detection, which shifts OV-HOI recognition from direct triplet generation to a structured verify-and-revise decision process over box-grounded HOI predictions.

- We introduce an executable reflection mechanism that unifies box adjustment, pair rebinding, and interaction refinement to explicitly correct language-prior hallucinations and coupled localization-semantic errors, and we further train it with GRPO and verifiable rewards to favor corrections that genuinely improve final HOI predictions.
- Our method outperforms previous state-of-the-art methods on two public benchmarks HICO-DET [5] and SWIG-HOI [37], demonstrating its adaptability to various novel interactions.

2 Related Work

Open-Vocabulary HOI Detection. Open-Vocabulary Human-Object Interaction (OV-HOI) detection aims to recognize interaction categories absent from the training set [38, 40]. Traditional HOI methods typically rely on predefined category spaces [32, 47], focusing on human-object pair localization and interaction classification. To improve recognition of unseen interactions, THID [38] first learns transferable HOI representations with natural language supervision. Later OV-HOI methods exploit the open-vocabulary knowledge of Vision-Language Models [22, 30] and Large Language Models (LLMs) [3, 4, 8, 13] for interaction semantic modeling. For example, CMD-SE [20] uses LLMs to generate fine-grained human body-part descriptions. INP-CC [21] reduces semantic confusion with interaction-aware prompts and concept calibration. SGC-Net [25] improves open-vocabulary recognition via multi-granularity visual alignment and hierarchical semantic comparison. Despite these advances, existing methods still largely rely on CLIP-style task-specific frameworks, while the rich open-vocabulary reasoning potential of MLLMs for OV-HOI detection remains underexplored.

MLLMs for Visual Understanding. Multimodal Large Language Models (MLLMs) integrate visual encoders with LLMs, demonstrating strong capabilities in general visual understanding, visual question answering, and instruction following. LLaVA [26] builds a general-purpose vision-language dialogue model through visual instruction tuning, while MiniGPT-4 [53] and InstructBLIP [7] further enhance multimodal understanding via visual feature alignment and instruction tuning. The Qwen-VL family [1, 2, 36] shows strong multimodal perception and spatial localization capabilities for image and video understanding, supporting fine-grained visual comprehension and object grounding. Qwen2.5-VL [2] improves multimodal understanding with dynamic resolution processing, absolute time encoding, and Window Attention, while Qwen3-VL [1] extends the series to long-context text-image-video reasoning with enhanced spatial-temporal modeling.

Reinforcement Learning for MLLM Reasoning. Reinforcement learning has recently emerged as a key approach for eliciting reasoning abilities in large models [35, 41, 43, 44]. OpenAI-o1 [14] trains models with large-scale reinforcement learning to reason using chain-of-thought, significantly advancing the development of reasoning models. DeepSeek-R1 [9] further shows that reinforcement learning alone can induce complex behaviors such as self-reflection, verification, and multi-step reasoning, achieving strong performance on math, coding, and general reasoning tasks. Building on this trend, Kimi k1.5 [33] extends reinforcement learning to multimodal large model training, improving cross-modal reasoning through long-context scaling and enhanced policy optimization. In addition, SCoRe [17] improves the self-correction ability of language models through multi-turn online reinforcement learning, while Visual-RFT [28] introduces R1-style reinforcement learning into visual perception tasks with verifiable rewards for multimodal models. However, existing methods struggle with language-prior hallucinations [52] and limited cross-modal consistency [49], which hinder reliable structured prediction in OV-HOI detection.

3 Method

3.1 Overview

As illustrated in Figure 2, the training pipeline of RRL-HOI consists of three sequential stages: ***(1)*** *Initial Triplet Generation via Supervised Fine-Tuning* (Sec. 3.3) learns to generate initial triplets directly, ***(2)*** *Structured Triplet Reflection via Edit Instruction Tuning* (Sec. 3.4) learns structured edit-based reflection and revision, and ***(3)*** *Reflective Policy Optimization via GRPO* (Sec. 3.5) further optimizes the reflection policy by rewarding edit trajectories that improve final HOI prediction quality.

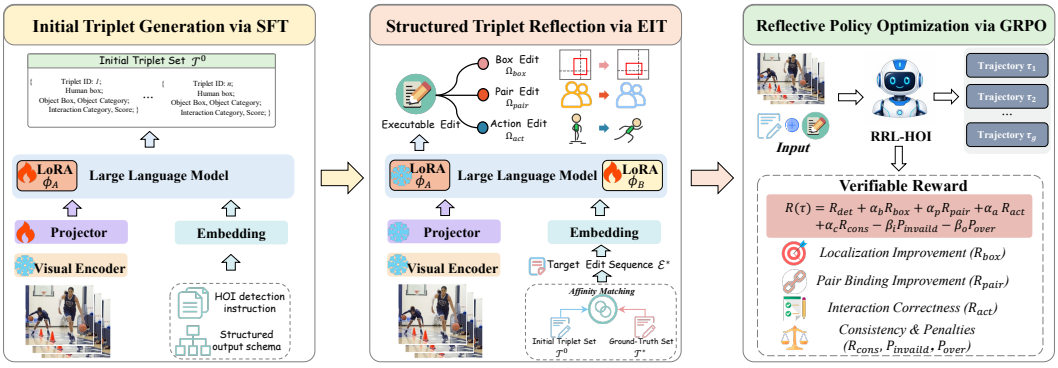


Figure 2: **Overview of our RRL-HOI framework.** RRL-HOI consists of three stages: **Stage-1** SFT generates initial box-grounded triplets, **Stage-2** EIT learns executable triplet edits, and **Stage-3** GRPO optimizes the reflection policy with verifiable rewards. This pipeline turns OV-HOI detection from direct triplet generation into a structured verify-and-revise process.

3.2 Problem Formulation

Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, open-vocabulary HOI detection aims to predict a structured set of human-object interaction triplets:

$$\mathcal{T} = \{T_i\}_{i=1}^{N}, \quad T_i = \langle B_i^h, B_i^o, \mathcal{A}_i, \mathbf{s}_i \rangle. \tag{1}$$

Here, B_i^h and B_i^o denote the human and object bounding boxes of the i-th interaction instance, $\mathcal{A}_i \subseteq \mathcal{V}$ is the set of interaction verbs assigned to the human-object pair, $\mathcal{V}$ is the open-vocabulary interaction space, and $\mathbf{s}_i$ denotes the corresponding confidence scores. Since RRL-HOI does not rely on an external detector, each box is represented by normalized coordinate tokens:

$$B = [x_1, y_1, x_2, y_2], \quad x_1, y_1, x_2, y_2 \in \{0, \dots, K\}, \tag{2}$$

where K is the coordinate quantization resolution.

RRL-HOI formulates OV-HOI prediction as an initial structured generation process followed by reflective edit-based revision. Specifically, the initial policy π_{init} first produces an initial triplet set $\mathcal{T}^0$ from the input image I, and the reflection policy π_θ then predicts an executable edit sequence $\mathcal{E}$ conditioned on both I and $\mathcal{T}^0$. The predicted edits are finally applied to obtain the revised triplet set $\mathcal{T}^1$:

$$\mathcal{T}^0 = \pi_{\text{init}}(I), \quad \mathcal{E} \sim \pi_\theta(\mathcal{E} \mid I, \mathcal{T}^0), \quad \mathcal{T}^1 = \text{Apply}(\mathcal{T}^0, \mathcal{E}). \tag{3}$$

Here, $\text{Apply}(\cdot)$ denotes the deterministic execution of the predicted edits on the initial triplets.

The objective of reflection is to improve the triplet-level quality with respect to the ground-truth triplet set $\mathcal{T}^*$:

$$Q(\mathcal{T}^1, \mathcal{T}^*) > Q(\mathcal{T}^0, \mathcal{T}^*), \tag{4}$$

where $Q(\cdot, \cdot)$ denotes a triplet-level quality function. This formulation casts open-vocabulary HOI detection as a "verify-and-revise" process, enabling localization, pair binding, and interaction semantics to be optimized within a unified box-level triplet space.

3.3 Initial Triplet Generation via Supervised Fine-Tuning

In the first stage, we train the MLLM to directly generate the initial triplet set $\mathcal{T}^0$. Given the frozen MLLM backbone Θ_0 and the trainable initial LoRA adapter ϕ_A, the initial prediction policy is:

$$\pi_{\text{init}}(I) = f_{\Theta_0, \phi_A}(I). \tag{5}$$

The first stage prompt contains the image I, the HOI detection instruction, and the required structured output schema. No proposals or candidate boxes are provided. The detailed prompt is given in Appendix A.1. The first stage objective $\mathcal{L}_{\text{init}}$ is defined as:

$$\mathcal{L}_{\text{init}} = \mathcal{L}_{\text{tri}} + \lambda_b \mathcal{L}_{\text{box}} + \lambda_a \mathcal{L}_{\text{act}}. \tag{6}$$

Here, $\mathcal{L}_{\text{tri}}$ supervises the structured triplet output format, $\mathcal{L}_{\text{box}}$ supervises normalized box-coordinate tokens, and $\mathcal{L}_{\text{act}}$ supervises interaction verbs. The coefficients λ_b and λ_a balance localization and interaction supervision.

This stage provides a cold-start triplet predictor. Its predictions do not need to be perfect, but they must be structured enough to support subsequent reflection. By grounding the initial prediction in normalized box-level triplets, this stage establishes an explicit localization-interaction space, enabling later reflection to operate on structured visual units rather than unconstrained textual responses.

3.4 Structured Triplet Reflection via Edit Instruction Tuning

Reflection Adapter. Stage 2 trains the reflection adapter ϕ_B to revise the initial triplet set $\mathcal{T}^0$. As illustrated in Figure 2, this stage constructs target edits by matching the initial triplets with ground-truth triplets, and then trains ϕ_B to predict executable triplet-level revisions. Given the image I and the serialized $\mathcal{T}^0$, the reflection policy π_θ implemented by ϕ_B is supervised to generate a structured edit sequence rather than free-form explanations.

Affinity Matching. To construct supervision for edit instruction tuning, we match each initial prediction $T_i^0 \in \mathcal{T}^0$ with a ground-truth triplet $T_j^* \in \mathcal{T}^*$. The affinity score combines human localization, object localization, and interaction-semantic similarity:

$$S_{ij} = \lambda_h \,\text{IoU}(B_i^{h,0}, B_j^{h,*}) + \lambda_o \,\text{IoU}(B_i^{o,0}, B_j^{o,*}) + \lambda_i \,\text{F1}(\mathcal{A}_i^0, \mathcal{A}_j^*). \tag{7}$$

Here, $B_i^{h,0}$ and $B_i^{o,0}$ are the predicted human and object boxes in T_i^0, while $B_j^{h,*}$ and $B_j^{o,*}$ are the corresponding ground-truth boxes in T_j^*. $\mathcal{A}_j^*$ denotes the ground-truth interaction verb set. $\text{F1}(\mathcal{A}_i^0, \mathcal{A}_j^*)$ measures the set-level overlap between the predicted and ground-truth verb sets, rewarding correct verbs while penalizing missing and false verbs. λ_h, λ_o, and λ_i balance the three matching terms.

The optimal matching is then obtained by selecting the prediction-ground-truth assignment with the maximum total affinity:

$$\mathcal{M}^0 = \arg\max_{\mathcal{M}} \sum_{(i,j)\in\mathcal{M}} S_{ij}. \tag{8}$$

Here, $\mathcal{M}$ denotes a candidate matching between initial predictions and ground-truth triplets, and each pair (i, j) matches the initial triplet T_i^0 to the ground-truth triplet T_j^*. The optimal matching $\mathcal{M}^0$ is then used to identify how each initial prediction should be revised.

Target Edit Sequence. Based on $\mathcal{M}^0$, localization mismatches generate box edits, wrong human-object associations generate pair edits, and missing or false verbs generate action edits. These automatically constructed edits form the target edit sequence:

$$\mathcal{E}^* = \{e_t^*\}_{t=1}^T, \quad e_t^* \in \Omega = \Omega_{\text{box}} \cup \Omega_{\text{pair}} \cup \Omega_{\text{act}}. \tag{9}$$

Here, each target edit e_t^* is an executable operation selected from the unified edit space Ω. The three edit subsets specify what part of a triplet is revised: Ω_{box} updates human/object boxes by keeping, adjusting, or replacing them; Ω_{pair} corrects human-object binding by rebinding, swapping, or removing pairs; and Ω_{act} refines interaction verbs by keeping, adding, removing, or replacing them. The detailed reflection prompt and output schema are provided in Appendix A.2.

For each triplet T_i^0, we update its interaction verbs through action edits. Let $\mathcal{A}_i^0$ denote the initial verb set selected from the open-vocabulary interaction space $\mathcal{V}$, where each verb a is kept if its confidence score $s_{i,a}^0$ is higher than the threshold γ. The revised verb set $\mathcal{A}_i^1$ is obtained by adding visually supported verbs Δ_i^+ and removing unsupported verbs Δ_i^-:

$$\mathcal{A}_i^1 = (\mathcal{A}_i^0 \cup \Delta_i^+) \setminus \Delta_i^-, \quad \mathcal{A}_i^0 = \{a \in \mathcal{V} \mid s_{i,a}^0 > \gamma\}. \tag{10}$$

In this way, box editing, pair rebinding, and interaction editing allow the model to correct localization errors, wrong pair assignments, and language-prior hallucinations.

Edit Instruction Tuning Objective. The reflection adapter ϕ_B is trained to imitate the target edit sequence $\mathcal{E}^*$ in an autoregressive manner:

$$\mathcal{L}_{\text{ref}} = -\mathbb{E}_{(I,\mathcal{T}^0,\mathcal{E}^*)} \left[\sum_{t=1}^{T} \log \pi_\theta(e_t^* \mid I, \mathcal{T}^0, e_{<t}^*) \right]. \tag{11}$$

Here, e_t^* is the t-th target edit, and $e_{<t}^*$ denotes the previous target edits. By predicting the next edit conditioned on the image, the initial triplets, and previous edits, the reflection adapter learns how to transform triplet-level prediction errors into executable revisions.

By converting triplet-level prediction discrepancies into executable edit supervision, RRL-HOI effectively builds a visually grounded revision space that explicitly connects localization correction, pair reassignment, and interaction verification.

3.5 Reflective Policy Optimization via GRPO

Stage 3 further optimizes the reflection policy with GRPO [31]. Different from Stage 2, which imitates automatically constructed edit labels, GRPO compares multiple reflection trajectories for the same initial prediction and reinforces those that lead to verifiable HOI improvements.

Given an input $x = (I, \mathcal{T}^0)$, where I is the image and $\mathcal{T}^0$ is the initial triplet set, the old reflection policy $\pi_{\theta_{\text{old}}}$ samples a group of G trajectories:

$$\{\tau_g\}_{g=1}^{G} \sim \pi_{\theta_{\text{old}}}(\tau \mid x), \quad \tau_g = (\mathcal{E}_g, \mathcal{T}_g^1), \quad \mathcal{T}_g^1 = \text{Apply}(\mathcal{T}^0, \mathcal{E}_g). \tag{12}$$

Here, each trajectory τ_g consists of a sampled edit sequence $\mathcal{E}_g$ and the revised triplet set $\mathcal{T}_g^1$ obtained after applying the edits. Each trajectory is assigned a verifiable reward $R(\tau_g)$ according to the quality of its revised HOI prediction.

Following GRPO, we compute the normalized advantage of each trajectory within the sampled group:

$$\hat{A}_g = \frac{R(\tau_g) - \text{mean}(\{R(\tau_1), \ldots, R(\tau_G)\})}{\text{std}(\{R(\tau_1), \ldots, R(\tau_G)\}) + \epsilon}. \tag{13}$$

The advantage $\hat{A}_g$ measures whether trajectory τ_g performs better or worse than other sampled trajectories for the same initial prediction. This group-relative normalization reduces reward variance and encourages the model to prefer more effective reflection strategies.

The reflection policy is optimized with the clipped GRPO objective:

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}\left[\frac{1}{G}\sum_{g=1}^{G} \min\left(\rho_g(\theta)\hat{A}_g, \text{clip}(\rho_g(\theta), 1-\epsilon_c, 1+\epsilon_c)\hat{A}_g\right) - \beta_{\text{KL}} D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\right]. \tag{14}$$

Here, $\rho_g(\theta) = \pi_\theta(\tau_g \mid x)/\pi_{\theta_{\text{old}}}(\tau_g \mid x)$ is the policy ratio between the current reflection policy π_θ and the sampling policy $\pi_{\theta_{\text{old}}}$. The clipping threshold ϵ_c stabilizes policy updates, while the KL term constrains π_θ to stay close to the frozen Stage-2 reflection model π_{ref}. This prevents the policy from drifting away from the supervised executable edit format.

Verifiable Reward. The reward function evaluates whether a sampled reflection trajectory truly improves the revised triplet set over the initial prediction:

$$R(\tau) = R_{\text{det}} + \alpha_b R_{\text{box}} + \alpha_p R_{\text{pair}} + \alpha_a R_{\text{act}} + \alpha_c R_{\text{cons}} - \beta_i P_{\text{invalid}} - \beta_o P_{\text{over}}. \tag{15}$$

Here, R_{det} measures the final HOI detection quality of the revised triplets. The three improvement rewards, R_{box}, R_{pair}, and R_{act}, compare the revised prediction $\mathcal{T}^1$ with the initial prediction $\mathcal{T}^0$ in terms of box localization, human-object pair binding, and interaction-verb correctness, respectively. In this way, the reflection policy is encouraged to make edits only when they lead to verifiable improvements over the initial triplets. In addition, R_{cons} encourages the predicted edits to be faithfully reflected in the revised triplet set, while P_{invalid} and P_{over} penalize invalid outputs and unnecessary over-editing. Detailed definitions of these reward components are provided in Appendix B.

Through group-relative optimization over multiple revision trajectories, the reflection policy progressively favors edits that improve localization, pair binding, and interaction correctness, rather than merely imitating heuristic edit labels.

3.6 Training and Inference

Training. Stage 1 trains the initial adapter ϕ_A, Stage 2 freezes ϕ_A and trains the reflection adapter ϕ_B, and Stage 3 further optimizes only ϕ_B with GRPO while keeping the base MLLM and ϕ_A frozen. This decoupling stabilizes RL training by fixing the initial prediction distribution.

Inference. At inference time, RRL-HOI performs prediction through a generate-and-revise process. It first generates an initial triplet set $\mathcal{T}^0$ from the input image, then predicts an executable edit sequence $\mathcal{E}$ to revise boxes, pairings, and interaction verbs, and finally applies the edits to obtain the final prediction:

$$\mathcal{T}^0 = \pi_{\text{init}}(I), \quad \mathcal{E} = \pi_\theta(I, \mathcal{T}^0), \quad \mathcal{T}^1 = \text{Apply}(\mathcal{T}^0, \mathcal{E}). \tag{16}$$

For standard HOI evaluation, multi-verb triplets are expanded into single-verb detections by sharing the same human-object boxes and assigning one verb and its confidence score to each entry.

4 Experiments

4.1 Experimental Setting

Datasets and Evaluation Metrics. We evaluate the proposed RRL-HOI on two public OV-HOI detection datasets, HICO-DET [5] and SWIG-HOI [37], and use the Mean Average Precision (mAP) metric on both datasets. Due to space limitations, a detailed description of these benchmarks and evaluation metrics can be found in the Appendix C.

Implementation Details. We adopt Qwen3-VL-8B [1] as the default base MLLM. To evaluate the generality of RRL-HOI across different MLLM backbones, we also report results with Qwen2.5-VL-7B [2] under the same training and inference recipe. Unless otherwise specified, all experiments and ablation studies are conducted with Qwen3-VL-8B. The visual encoder and the base language model are frozen throughout training, and LoRA [12] modules are inserted into the language decoder. We set the coordinate quantization resolution K to 1000. RRL-HOI is trained in three stages. In Stage 1, we train the initial triplet generation adapter for 2 epochs with a batch size of 8 and a learning rate of 1×10^{-5}. In Stage 2, we freeze the Stage-1 adapter and train the reflection adapter for 1 epoch on automatically constructed edit-instruction data, using batch size 8, learning rate 2×10^{-5}, and gradient clipping with a maximum norm of 1.0. In Stage 3, we optimize only the reflection adapter with GRPO for 1 epoch, sampling $G = 4$ reflection trajectories per input. The learning rate is 1×10^{-6}, the clipping coefficient ϵ_c is 0.2, and the KL coefficient β_{KL} is 0.01. During inference, we use greedy decoding with a maximum output length of 1024 tokens. All experiments are implemented on 4 NVIDIA A100 GPUs. More implementation details are provided in Appendix D.

4.2 Comparisons with State-of-the-Art Methods

Among detector-free OV-HOI detection methods, ImagineAgent [46] is the closest MLLM-based baseline. It improves OV-HOI reasoning by constructing CoT-style reasoning data, invoking external tools such as retrieval, cropping, outpainting, and viewpoint transformation, and fine-tuning the MLLM with full-parameter training. In contrast, our RRL-HOI is more efficient: it freezes the base MLLM, trains only lightweight LoRA adapters, and automatically constructs edit-instruction supervision from initial triplet predictions and ground-truth annotations. Thus, RRL-HOI avoids both full-parameter fine-tuning and costly CoT data construction, while directly optimizing structured box-level HOI revision.

HICO-DET. As shown in Table 1, we compare RRL-HOI with representative zero-shot HOI methods and recent open-vocabulary HOI methods on HICO-DET. Compared with previous detector-free OV-HOI approaches, RRL-HOI achieves clear improvements across unseen, seen, and full categories. Under the same Qwen2.5-VL-7B backbone as ImagineAgent, RRL-HOI obtains 32.05%, 35.59%, and 34.81% mAP on unseen, seen, and full categories, respectively, outperforming ImagineAgent by 2.34%, 6.82%, and 5.85%. With Qwen3-VL-8B, RRL-HOI further improves to 33.28%, 37.93%, and 36.90%, surpassing the strongest previous detector-free method by 3.57%, 6.94%, and 7.27%. These results show that our reflective revision mechanism achieves stronger box-grounded HOI prediction without relying on external detector proposals or heavy full-parameter training.

SWIG-HOI. We further evaluate RRL-HOI on the more diverse SWIG-HOI benchmark. As shown in Table 2, RRL-HOI consistently outperforms previous SOTA methods on non-rare, rare, unseen, and full categories. With Qwen2.5-VL-7B as the backbone, RRL-HOI achieves 28.73%, 22.64%, 17.79%, and 22.86% mAP, improving over ImagineAgent by 5.84%, 4.62%, 5.26%, and 5.11%, respectively. With Qwen3-VL-8B, RRL-HOI further reaches 30.71%, 25.58%, 19.94%, and 25.37% mAP. The consistent gains, especially on rare and unseen categories, indicate that reward-driven

Table 1: Comparison of our RRL-HOI with state-of-the-art methods on the HICO-DET dataset.

Method	Pretrained Detector	Unseen	Seen	Full
FCL [10] (CVPR'21)	✓	13.16	24.23	22.01
SCL [11] (ECCV'2022)	✓	19.07	30.39	28.08
GEN-VLKT [24] (CVPR'2022)	✓	21.36	32.91	30.56
OpenCat [51] (CVPR'2023)	✓	21.46	33.86	31.38
HOICLIP [29] (CVPR'2023)	✓	23.48	34.47	32.26
THID [38] (CVPR'2022)	✗	15.53	24.32	22.38
CMD-SE [20] (CVPR'2024)	✗	16.70	23.95	22.35
INP-CC [21] (ICCV'2025)	✗	17.38	24.74	23.12
SGC-Net [25] (CVPR'2025)	✗	23.27	28.34	27.22
SD-IF [40] (CVPR'2026)	✗	28.01	30.09	29.63
ImagineAgent (Qwen2.5-VL-7B) [46]	✗	29.71	28.77	28.96
RRL-HOI (Qwen2.5-VL-7B)	✗	32.05	35.59	34.81
RRL-HOI (Qwen3-VL-8B)	✗	**33.28**	**37.93**	**36.90**

Table 2: Comparison of our RRL-HOI with state-of-the-art methods on the SWIG-HOI dataset.

Method	Non-rare	Rare	Unseen	Full
CHOID [37] (ICCV'2021)	10.93	6.63	2.64	6.64
QPIC [32] (CVPR'2021)	16.95	10.84	6.21	11.12
GEN-VLKT [24] (CVPR'2022)	20.91	10.41	–	10.87
MP-HOI [42] (CVPR'2024)	20.28	14.78	–	12.61
THID [38] (CVPR'2022)	17.67	12.82	10.04	13.26
CMD-SE [20] (CVPR'2024)	21.46	14.64	10.70	15.26
INP-CC [21] (ICCV'2025)	22.84	16.74	11.02	16.74
SGC-Net [25] (CVPR'2025)	23.67	16.55	12.46	17.20
SD-IF [40] (CVPR'2026)	25.01	18.86	15.57	19.48
ImagineAgent (Qwen2.5-VL-7B) [46]	22.89	18.02	12.53	17.75
RRL-HOI (Qwen2.5-VL-7B)	28.73	22.64	17.79	22.86
RRL-HOI (Qwen3-VL-8B)	**30.71**	**25.58**	**19.94**	**25.37**

Table 3: Ablation studies of RRL-HOI on the SWIG-HOI dataset. SFT denotes Initial Triplet Generation via Supervised Fine-Tuning, EIT denotes Structured Triplet Reflection via Edit Instruction Tuning, and GRPO denotes Reflective Policy Optimization via GRPO.

(a) Overall Model Framework Design

Method	Non-rare	Rare	Unseen	Full
Base	4.79	10.61	5.26	7.89
Base+ SFT	22.08	13.17	5.69	13.39
Base+ SFT+ EIT	27.49	21.85	15.73	21.64
Base+ SFT+ EIT+GRPO	**30.71**	**25.58**	**19.94**	**25.37**

(b) Comparison of Different Reflection Formats

Method	Non-rare	Rare	Unseen	Full
Initial Prediction	22.08	13.17	5.69	13.39
Free-form Reflection	23.91	12.85	7.49	14.11
Structured Triplet Regeneration	25.88	17.60	12.45	18.26
Executable Edit Reflection	**27.49**	**21.85**	**15.73**	**21.64**

(c) Comparison of Different Edit Types

Box	Pair	Action	Non-rare	Rare	Unseen	Full
✗	✗	✗	22.08	13.17	5.69	13.39
✓	✗	✗	23.61	14.52	7.34	14.86
✗	✓	✗	23.28	14.11	6.98	14.48
✗	✗	✓	24.35	16.97	10.42	17.06
✓	✓	✗	24.76	15.83	8.65	16.13
✗	✓	✓	25.92	19.34	13.87	19.51
✓	✓	✓	**27.49**	**21.85**	**15.73**	**21.64**

(d) Comparison of Different Reward Designs

Method	Non-rare	Rare	Unseen	Full
w/o Localization R_{box}	28.84	23.71	18.42	23.59
w/o Pair R_{pair}	28.55	23.08	17.91	23.07
w/o Interaction R_{act}	29.26	22.64	16.80	22.73
w/o Consistency R_{cons}	29.78	24.39	18.73	24.24
w/o Invalid P_{invalid}	29.35	24.12	18.54	23.95
w/o Over-editing P_{over}	29.61	24.80	18.91	24.46
RRL-HOI (Ours)	**30.71**	**25.58**	**19.94**	**25.37**

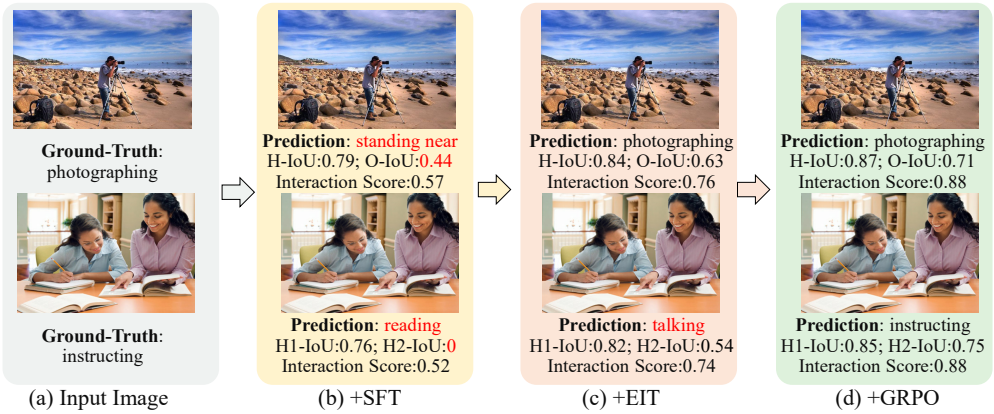
executable reflection provides an efficient and effective way to convert MLLM knowledge into reliable open-vocabulary HOI predictions.

Overall, RRL-HOI achieves stronger performance with a more efficient training paradigm. By replacing complex CoT/tool-augmented data construction with automatically generated executable edits, RRL-HOI can correct coupled localization-semantic errors and suppress language-prior hallucinations, leading to more accurate and generalizable OV-HOI detection.

4.3 Ablation Studies

To analyze the effectiveness of different designs in RRL-HOI, we conduct comprehensive ablation studies on the SWIG-HOI dataset. As shown in Table 3, we evaluate the proposed framework from four aspects: the overall training pipeline, the reflection format, the executable edit space, and the verifiable reward design. Additional hyperparameter ablations are provided in Appendix E. Beyond these ablations, Appendices H–J provide further controlled analyses: an equal-budget continued-training control and a prediction-transition study that isolate the contribution of GRPO, a same-family backbone-scaling study and a parameter-matched shared-versus-separate adapter ablation, and a ground-truth-blind visual-support evaluation that separates improved visual grounding from mere annotation alignment. We also clarify that P_{invalid} is an executability constraint rather than a semantic compatibility whitelist.

Overall Model Framework Design. As shown in Table 3 (a), directly applying the base MLLM to OV-HOI detection yields limited performance, with only 7.89% mAP on the full set. After introducing SFT, the full mAP improves from 7.89% to 13.39%, indicating that supervised fine-tuning is essential for adapting the MLLM to generate parseable box-grounded HOI triplets. When further adding EIT, the model achieves 27.49%, 21.85%, 15.73%, and 21.64% mAP on non-rare, rare, unseen, and



(a) Input Image (b) +SFT (c) +EIT (d) +GRPO

Figure 3: **Qualitative prediction evolution of RRL-HOI.** We visualize two representative examples from the initial SFT prediction to the EIT-refined prediction and the final GRPO-optimized prediction. For the **photographing** example, the SFT model produces a weak spatial relation with inaccurate object grounding, while EIT corrects the interaction verb and improves the object localization. For the **instructing** example, the initial prediction focuses on a local reading-related cue and misses the instructed person, and EIT partially corrects the interaction through structured reflection. GRPO further improves both examples by favoring edits with better localization, pair binding, and interaction correctness.

full categories, respectively. This demonstrates that structured edit supervision effectively enhances the model's ability to revise localization, human-object binding, and interaction semantics. Finally, GRPO further improves the full mAP from 21.64% to 25.37%, with consistent gains across all splits. These results validate the effectiveness of the proposed three-stage verify-and-revise training pipeline.

Comparison of Different Reflection Formats. Table 3 (b) compares different reflection formats under the same initial SFT predictions. Free-form reflection brings only limited improvement, increasing the full mAP from 13.39% to 14.11%. This suggests that natural-language self-reflection is insufficient for reliable structured HOI revision, since the model may produce plausible explanations without making valid box-level corrections. In contrast, structured triplet regeneration further improves the full mAP to 18.26%, showing that enforcing a structured output format benefits HOI prediction. However, it still underperforms executable edit reflection, which achieves 21.64% full mAP. This indicates that revising initial predictions through explicit edit operations is more stable than regenerating complete triplets from scratch. Therefore, the proposed executable edit reflection provides a more suitable self-correction format for structured OV-HOI detection.

Comparison of Different Edit Types. Table 3 (c) analyzes the contribution of each executable edit type. Without edits, the model reduces to the SFT-only baseline with 13.39% full mAP. Box and pair edits bring moderate gains, while action edits provide a larger improvement, especially increasing unseen mAP from 5.69% to 10.42%, indicating their importance for open-vocabulary generalization. Combining pair and action edits further improves full mAP to 19.51%, showing their complementarity. The best result, 21.64% full mAP, is achieved when box, pair, and action edits are all enabled, confirming that the three edit types jointly form an effective executable revision space.

Comparison of Different Reward Designs. Table 3 (d) evaluates the reward components in GRPO. Removing the localization, pair, or interaction reward consistently degrades performance, and the largest unseen mAP drop, from 19.94% to 16.80%, occurs when R_{act} is removed, highlighting the role of interaction-level verification in suppressing unsupported semantic predictions. Removing the consistency reward or penalty terms also reduces performance, indicating that edit consistency, output validity, and over-editing control are important for stable reflection. The full reward design achieves the best results across all metrics, validating the effectiveness of the verifiable reward formulation.

4.4 Qualitative Results

Figure 3 shows the progressive prediction evolution of RRL-HOI on two representative cases. In the photographing example, the SFT model predicts the weak relation "standing near" with inaccurate object grounding, while EIT corrects the action to "photographing" and improves the object IoU from

0.44 to 0.63; GRPO further raises the human/object IoUs to 0.87/0.71 and increases the interaction score to 0.88. In the instructing example, the SFT model is biased toward the local reading cue and fails to bind the second human target, resulting in an H2-IoU of 0. EIT partially recovers the two-person interaction structure but still predicts "talking", whereas GRPO finally revises the interaction to "instructing" and improves H2-IoU to 0.75. These results demonstrate that executable reflection progressively corrects localization, pair binding, and interaction semantics, while GRPO further favors edits that lead to higher-quality HOI predictions.

5 Conclusion

In this paper, we introduced RRL-HOI, a reflective reinforcement learning framework for open-vocabulary HOI detection. RRL-HOI reformulates MLLM-based HOI prediction from single-step triplet generation into a structured verify-and-revise process. It first generates box-grounded initial triplets, then revises them with executable edits over localization, pair binding, and interaction semantics, and finally optimizes the reflection policy with GRPO under verifiable rewards. This design enables the model to correct coupled localization-semantic errors and reduce visually unsupported language-prior hallucinations. Experiments on HICO-DET and SWIG-HOI show that RRL-HOI consistently outperforms prior methods, while ablation studies validate the effectiveness of the three-stage pipeline, executable edit space, and reward design. These results demonstrate that reward-driven reflection is an effective mechanism for bridging general multimodal understanding and precise structured interaction detection. Controlled analyses in the appendix further isolate the contribution of GRPO under an equal-budget control, verify that the gains are not a model-scale artifact, and distinguish reduced visually unsupported predictions from mere annotation alignment.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62476260, 62225207 and 62436008, the Fundamental Research Funds for the Central Universities under Grant WK2100000057, and the CCF-1688 Yuanbao Cooperation Fund under Grant No. 2025006.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, and Jun Tang. Qwen2.5-vl technical report. 2025.

[3] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023.

[4] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

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

[6] Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Jianing Yang, David F Fouhey, Joyce Chai, and Shengyi Qian. Multi-object hallucination in vision language models. *Advances in Neural Information Processing Systems*, 37:44393–44418, 2024.

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

[8] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[9] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[10] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14646–14655, 2021.

[11] Zhi Hou, Baosheng Yu, and Dacheng Tao. Discovering human-object interaction concepts via self-compositional learning. In *European Conference on Computer Vision*, pages 461–478. Springer, 2022.

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[13] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022.

[14] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

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

[16] Sanghyun Kim, Deunsol Jung, and Minsu Cho. Relational context learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2925–2934, 2023.

[17] Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*, 2024.

[18] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9579–9589, 2024.

[19] Ting Lei. Exploring conditional multi-modal prompts for zero-shot hoi detection. In *ECCV*, pages 1–19, 2024.

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

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

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[23] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490, 2020.

[24] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20123–20132, 2022.

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

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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

[28] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2034–2044, 2025.

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

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[31] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[32] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021.

[33] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[34] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13617–13626, 2020.

[35] Junfeng Wang, Jiawei Liu, Yongchao Xu, Tao Jiang, Jiangbo Ai, Jin Zhang, Chong Wang, and Qixing Zhang. Learning where to look: Observation policy optimization for thinking with images. *arXiv preprint arXiv:2611.04943*, 2026.

[36] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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

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

[39] Yongchao Xu, Jiawei Liu, Sen Tao, Qiang Zhang, and Zheng-Jun Zha. Hoimamba: Efficient mamba-based disentangled progressive learning for hoi detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8987–8995, 2025.

[40] Yongchao Xu, Jiawei Liu, Junfeng Wang, Sen Tao, Na Jiang, and Zheng-Jun Zha. Learning to diversify and focus: A reinforcement framework for open-vocabulary hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 34673–34682, 2026.

[41] Yongchao Xu, Bowen Ye, Jiefeng Gan, Junkai Ma, Wenzhao Li, Sen Tao, Yi Wei, and Jiawei Liu. Videoevolve: Co-evolving memory and retrieval for long video understanding. *arXiv preprint arXiv:2610.10183*, 2026.

[42] Jie Yang, Bingliang Li, Ailing Zeng, Lei Zhang, and Ruimao Zhang. Open-world human-object interaction detection via multi-modal prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16954–16964, 2024.

[43] Bowen Ye, Yongchao Xu, Zhijian Li, Xiang Yin, Junkai Ma, and Wenzhao Li. Coevo-mem: Co-evolving retrieval policy and memory bank for llm agents. *arXiv preprint arXiv:2608.01739*, 2026.

[44] Bowen Ye, Yongchao Xu, Junkai Ma, Xiang Yin, and Wenzhao Li. Principles that guide, actions that inform: Agent evolution via knowledge abstraction. *arXiv preprint arXiv:2610.06964*, 2026.

[45] Heng Yin, Yuqiang Ren, Ke Yan, Shouhong Ding, and Yongtao Hao. Rod-mllm: Towards more reliable object detection in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14358–14368, 2025.

[46] Zhenlong Yuan, Xiangyan Qu, Jing Tang, Rui Chen, Lei Sun, Ruidong Chen, Hongwei Yu, Chengxuan Qian, Xiangxiang Chu, Shuo Li, et al. What if agents could imagine? reinforcing open-vocabulary hoi comprehension through generation. *arXiv preprint arXiv:2602.11499*, 2026.

[47] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. *Advances in Neural Information Processing Systems*, 34:17209–17220, 2021.

[48] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20104–20112, 2022.

[49] Xiang Zhang, Senyu Li, Ning Shi, Bradley Hauer, Zijun Wu, Grzegorz Kondrak, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. Cross-modal consistency in multimodal large language models. *arXiv preprint arXiv:2411.09273*, 2024.

[50] Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14227–14238, 2024.

[51] Sipeng Zheng, Boshen Xu, and Qin Jin. Open-category human-object interaction pre-training via language modeling framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19392–19402, 2023.

[52] Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. *arXiv preprint arXiv:2410.04780*, 2024.

[53] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A Prompt Templates and Structured Output Schemas

This appendix presents the prompt templates and structured output schemas used in RRL-HOI. All prompts are designed to constrain the MLLM to produce parseable and executable outputs, rather than free-form textual descriptions. This design enables RRL-HOI to transform open-vocabulary HOI detection into a structured generation, reflection, and policy optimization process.

A.1 Initial Triplet Generation Prompt and Schema

Prompt for Initial Triplet Generation. Figure 4 illustrates the prompt used in the first stage, *Initial Triplet Generation via Supervised Fine-Tuning*. This prompt guides the MLLM to directly generate box-level HOI triplets from the input image, including human boxes, object boxes, object categories, interaction verbs, and confidence scores. By enforcing a structured JSON output without relying on external proposals or candidate boxes, this prompt establishes the initial localization-interaction space for subsequent reflection.

Schema for Initial Box-Level HOI Triplets. Figure 5 shows the structured output schema for the initial HOI triplets. The schema explicitly binds each interaction verb to a specific human box and object box, making the prediction directly parseable and preventing the model from producing ungrounded action labels. This representation provides the basic unit for triplet-level verification, revision, and evaluation.

A.2 Structured Triplet Reflection Prompt and Edit Schema

Prompt for Structured Triplet Reflection. Figure 6 presents the prompt used in the second stage, *Structured Triplet Reflection via Edit Instruction Tuning*. Given the image and the serialized initial triplets, the prompt instructs the model to verify whether each triplet is visually grounded and to revise it only through executable edit operations. This design converts reflection from free-form explanation into structured correction, allowing the model to explicitly address inaccurate boxes, wrong human-object bindings, and visually unsupported interaction labels.

Schema for Executable Reflection Edits. Figure 7 defines the structured edit schema used by the reflection policy. Each edit specifies the target triplet, edit type, operation, and operation-specific arguments. This schema ensures that the model's reflection output can be deterministically applied to the initial prediction, enabling verifiable self-correction within the box-level triplet space.

A.3 GRPO Sampling and Inference Prompt

Prompt for GRPO Sampling and Inference-Time Reflection. Figure 8 shows the prompt used in the third stage, *Reflective Policy Optimization via GRPO*, and also during inference. During GRPO training, the prompt is used to sample multiple reflection trajectories for the same initial prediction, enabling group-relative comparison under verifiable rewards. During inference, the same prompt generates the final edit sequence. This shared prompt format aligns reinforcement optimization with inference-time behavior and encourages the reflection policy to favor edits that improve localization, pair binding, and interaction correctness.

B Reward Component Definitions

This section provides the detailed definitions of the reward components used in Eq. (15). Given the initial triplet set $\mathcal{T}^0$, the predicted edit sequence $\mathcal{E}$, and the revised triplet set $\mathcal{T}^1 = \mathrm{Apply}(\mathcal{T}^0, \mathcal{E})$, we compute rewards with respect to the ground-truth triplet set $\mathcal{T}^*$.

The box improvement reward measures whether the revised triplets achieve better human and object localization than the initial prediction:

$$R_{\mathrm{box}} = Q_{\mathrm{box}}(\mathcal{T}^1, \mathcal{T}^*) - Q_{\mathrm{box}}(\mathcal{T}^0, \mathcal{T}^*). \tag{17}$$

Here, Q_{box} evaluates the localization quality of matched human and object boxes, e.g., by the average IoU between predicted boxes and ground-truth boxes.

Table 4: Invalid-output types checked by P_{invalid}. The penalty depends only on whether an edit program can be parsed and deterministically executed, not on whether the interaction is semantically common, rare, or novel.

Invalid-output type	Example
Malformed serialization	Invalid or incomplete JSON
Illegal operation	Operation outside the allowed edit-operation set
Missing required argument	`Adjust_O` without `new_object_box`
Invalid reference	A nonexistent `target_triplet_id`
Malformed argument	Invalid coordinate structure or non-numeric values

The pair improvement reward measures whether the revised prediction produces more accurate human-object associations:

$$R_{\text{pair}} = Q_{\text{pair}}(\mathcal{T}^1, \mathcal{T}^*) - Q_{\text{pair}}(\mathcal{T}^0, \mathcal{T}^*). \tag{18}$$

Here, Q_{pair} evaluates whether the predicted human and object boxes are correctly bound as an interacting pair.

The action improvement reward measures whether the revised triplets contain more accurate interaction verbs. Let $\mathcal{M}$ denote the matched prediction-ground-truth triplet pairs. We define:

$$R_{\text{act}} = \frac{1}{|\mathcal{M}|} \sum_{(i,j)\in\mathcal{M}} \left[\text{F1}(\mathcal{A}_i^1, \mathcal{A}_j^*) - \text{F1}(\mathcal{A}_i^0, \mathcal{A}_j^*)\right], \tag{19}$$

where $\mathcal{A}_i^0$ and $\mathcal{A}_i^1$ are the interaction verb sets before and after reflection, and $\mathcal{A}_j^*$ is the ground-truth verb set. The set-level F1 is computed as:

$$\text{F1}(\mathcal{A}, \mathcal{A}^*) = \frac{2|\mathcal{A} \cap \mathcal{A}^*|}{|\mathcal{A}| + |\mathcal{A}^*| + \epsilon}. \tag{20}$$

The consistency reward encourages the generated edits to be faithfully reflected in the revised triplet set:

$$R_{\text{cons}} = \frac{1}{|\mathcal{E}|} \sum_{e\in\mathcal{E}} C(e, \mathcal{T}^1), \tag{21}$$

where $C(e, \mathcal{T}^1) = 1$ if edit e is successfully applied and reflected in $\mathcal{T}^1$, and 0 otherwise.

In addition, R_{det} measures the final triplet-level HOI detection quality of $\mathcal{T}^1$, considering localization, pair association, and interaction correctness jointly. P_{invalid} penalizes invalid or unparseable edit outputs, such as edits with illegal operation names or missing required arguments. P_{over} penalizes unnecessary over-editing, which discourages the policy from modifying triplets when the initial prediction is already visually well grounded.

What Counts as an Invalid Edit Output. P_{invalid} is purely an *executability* constraint: it checks only serialization validity, reference validity, and edit executability under the edit schema in Appendix A.2. It does *not* rely on a predefined human-object-action compatibility table, a closed HOI vocabulary, or an action-object plausibility whitelist. Consequently, rare, unseen, or compositionally novel interactions are never penalized merely for being novel: a new verb in `Add_A` or `Refine_A` is treated as valid as long as the operation and its required arguments follow the schema. Semantic correctness is instead judged by the action, detection, and improvement rewards after execution, which preserves the separation between syntax-level validity and image-grounded correctness. Table 4 lists the invalid-output types recognized by P_{invalid}. Consistent with Table 3(d), removing P_{invalid} decreases Unseen/Full mAP from 19.94/25.37 to 18.54/23.95, confirming its role as an executability regularizer rather than a semantic restriction on open-vocabulary interactions.

C Experimental Setting

Datasets. We evaluate the proposed RRL-HOI on two public benchmarks, HICO-DET [5] and SWIG-HOI [37]. HICO-DET contains 80 object classes, 117 interaction classes, and 600 HOI

categories. Following previous works [20, 21, 25, 38], we hold out 120 rare HOIs to simulate the open-vocabulary setting. SWIG-HOI naturally contains unseen HOI classes in the test set. At inference time, SWIG-HOI requires a comprehensive evaluation of performance on 1391 unseen, 2841 rare, and 1307 non-rare HOI classes, which puts high demands on the generalization of the OV-HOI detection model.

Evaluation Metric. We use the mean Average Precision (mAP) as the evaluation metric on these datasets. Following prior works [15, 20, 27, 29, 39, 48, 38], a predicted HOI triplet is regarded as a true-positive HOI sample when the Intersection over Union (IoU) between the detected human and object bounding boxes and the corresponding ground-truth boxes is larger than 0.5, and the interaction class prediction is correct simultaneously.

D Additional Implementation Details

For both Qwen3-VL-8B and Qwen2.5-VL-7B, we use the same training and inference recipe unless otherwise specified. LoRA modules are inserted into the language decoder, including the attention projection layers and feed-forward projection layers. We use LoRA rank $r = 64$, scaling factor $\alpha = 128$, and LoRA dropout 0.05. All images are processed with the dynamic-resolution setting of the corresponding Qwen-VL backbone.

Stage 1. The initial triplet generation adapter is trained with the objective in Eq. (6). The loss weights are set to $\lambda_b = 2.0$ and $\lambda_a = 1.0$. We use AdamW with betas $(0.9, 0.999)$, weight decay 0.01, and a cosine learning-rate schedule with a warmup ratio of 0.03.

Stage 2. Initial predictions are generated by the frozen Stage-1 model and then matched to ground-truth triplets using Hungarian matching. The affinity weights in Eq. (7) are set to $\lambda_h = 0.4$, $\lambda_o = 0.4$, and $\lambda_i = 0.2$. The verb confidence threshold γ in Eq. (10) is set to 0.2, and at most the top-5 verbs are kept for each triplet. Box edits are generated when the matched human or object box has an IoU lower than 0.5.

Stage 3. For GRPO training, reflection trajectories are sampled with temperature 0.7 and top-p 0.9. The advantage normalization constant in Eq. (13) is set to 1×10^{-6}. The reward weights in Eq. (15) are set to $\alpha_b = 0.5$, $\alpha_p = 0.5$, $\alpha_a = 1.0$, $\alpha_c = 0.2$, $\beta_i = 1.0$, and $\beta_o = 0.2$. The maximum number of executable edits is limited to 32 per image.

Inference and Post-processing. Invalid JSON outputs and illegal edit operations are discarded. All revised boxes are clipped to the range $[0, 1000]$. Multi-verb triplets are expanded into single-verb detections. Kept verbs inherit their original confidence scores, refined verbs inherit the scores of the replaced verbs, and newly added verbs use the maximum existing verb score of the same triplet or $\gamma = 0.2$ if no previous verb exists. Duplicate detections are removed by triplet-level non-maximum suppression with an IoU threshold of 0.5.

E Additional Ablation Studies

To further analyze the sensitivity of RRL-HOI to key hyperparameters, we conduct additional ablation studies on the SWIG-HOI dataset using Qwen3-VL-8B as the default backbone. We evaluate three groups of hyperparameters: the Stage-1 loss weights, the Stage-2 matching affinity weights, and the Stage-3 reward weights.

E.1 Effect of Stage-1 Loss Weights

Table 5 studies the influence of the loss weights in Eq. (6). Setting $\lambda_b = 2.0$ and $\lambda_a = 1.0$ achieves the best overall performance. Reducing λ_b weakens box-coordinate supervision and leads to worse localization, while increasing λ_a too much over-emphasizes interaction semantics and slightly harms rare-category performance. These results indicate that stronger box supervision is important for establishing reliable initial box-grounded triplets.

Table 5: Ablation study of Stage-1 loss weights on SWIG-HOI.

λ_b	λ_a	Non-rare	Rare	Unseen	Full
0.5	1.0	28.91	23.85	18.31	23.66
1.0	1.0	29.62	24.74	18.96	24.43
2.0	0.5	29.88	24.36	18.72	24.25
1.0	2.0	29.47	24.91	19.10	24.53
3.0	1.0	30.18	25.06	19.43	24.85
2.0	**1.0**	**30.71**	**25.58**	**19.94**	**25.37**

Table 6: Ablation study of Stage-2 matching affinity weights on SWIG-HOI.

λ_h	λ_o	λ_i	Non-rare	Rare	Unseen	Full
0.33	0.33	0.34	29.84	24.69	18.97	24.47
0.50	0.30	0.20	30.16	25.02	19.35	24.81
0.30	0.50	0.20	30.05	24.93	19.28	24.72
0.35	0.35	0.30	29.71	24.78	19.47	24.60
0.45	0.45	0.10	30.22	25.13	19.21	24.84
0.40	**0.40**	**0.20**	**30.71**	**25.58**	**19.94**	**25.37**

E.2 Effect of Stage-2 Matching Affinity Weights

Table 6 evaluates the matching weights in Eq. (7). The setting $\lambda_h = 0.4$, $\lambda_o = 0.4$, and $\lambda_i = 0.2$ performs best. Compared with equal weighting, assigning slightly larger weights to human and object localization produces more reliable edit targets. In contrast, increasing the action weight improves semantic matching in some cases but makes the matching more sensitive to noisy initial verb predictions.

E.3 Effect of Stage-3 Reward Weights

Table 7 studies the sensitivity of the reward weights in Eq. (15). The default setting $\alpha_b = 0.5$, $\alpha_p = 0.5$, $\alpha_a = 1.0$, $\alpha_c = 0.2$, $\beta_i = 1.0$, and $\beta_o = 0.2$ achieves the best results. Reducing α_a weakens interaction-level verification and leads to a clear drop on rare and unseen categories, while further increasing it does not bring additional gains. Increasing α_b or α_p individually, or increasing both of them together, also underperforms the default setting, suggesting that overly strong localization or pair-association rewards may reduce the benefit of semantic refinement. In addition, changing α_c, β_i, or β_o degrades performance, showing that edit consistency, invalid-output penalty, and over-editing penalty are all important for stable policy optimization. Overall, these results indicate that RRL-HOI benefits from a balanced reward design, where interaction correctness receives stronger guidance while localization, pair association, consistency, and penalty terms jointly regularize the reflection policy.

F Limitations and Future Work

Although RRL-HOI achieves strong performance on open-vocabulary HOI detection, it still has several limitations. First, our experiments are conducted on two image-based benchmarks, HICO-DET and SWIG-HOI. While these benchmarks cover diverse human-object interactions, they cannot fully represent all real-world scenarios, such as videos, egocentric views, crowded scenes, low-resolution images, or interactions with severe occlusion. Second, RRL-HOI depends on the visual grounding and instruction-following ability of the underlying MLLM. When the base model fails to perceive small objects, ambiguous contact regions, or fine-grained spatial cues, the reflection policy may still produce incomplete or inaccurate revisions. Third, our reward design relies on ground-truth annotations during training and may not capture all semantic nuances of real-world interactions. Finally, the generate-and-revise inference process introduces additional computation compared with single-pass prediction, although the use of lightweight LoRA adapters avoids full-parameter fine-tuning.

Table 7: Ablation study of Stage-3 reward weights on SWIG-HOI.

α_b	α_p	α_a	α_c	β_i	β_o	Non-rare	Rare	Unseen	Full
0.5	0.5	0.5	0.2	1.0	0.2	29.96	24.32	18.63	24.22
1.0	0.5	1.0	0.2	1.0	0.2	30.18	24.91	19.05	24.68
0.5	1.0	1.0	0.2	1.0	0.2	30.11	24.84	18.98	24.61
1.0	1.0	1.0	0.2	1.0	0.2	30.25	24.88	18.91	24.65
0.5	0.5	1.5	0.2	1.0	0.2	30.22	25.07	19.38	24.86
0.5	0.5	1.0	0.0	1.0	0.2	29.78	24.39	18.73	24.24
0.5	0.5	1.0	0.5	1.0	0.2	30.08	24.96	19.18	24.72
0.5	0.5	1.0	0.2	0.5	0.2	29.94	24.78	19.12	24.58
0.5	0.5	1.0	0.2	1.0	0.0	29.61	24.80	18.91	24.46
0.5	0.5	1.0	0.2	1.0	0.5	30.05	25.02	19.31	24.77
0.5	**0.5**	**1.0**	**0.2**	**1.0**	**0.2**	**30.71**	**25.58**	**19.94**	**25.37**

Future work can improve RRL-HOI in several directions. One direction is to extend reflective revision from static images to video-based HOI detection, where temporal cues can help disambiguate subtle actions and human-object contact. Another direction is to improve uncertainty estimation and confidence calibration, allowing the model to better identify when a prediction should be revised or rejected. It is also promising to design more fine-grained rewards that account for long-tailed interactions, physical plausibility, and compositional generalization. In addition, future work may explore more efficient distillation or pruning strategies to reduce inference cost, as well as broader evaluations across more datasets, domains, and real-world deployment conditions.

G Societal Impacts

RRL-HOI aims to improve open-vocabulary human-object interaction detection by producing more accurate and visually grounded interaction predictions. This capability can have positive societal impacts in applications such as assistive technologies, human-robot collaboration, visual content understanding, and human-centric scene analysis. By reducing the reliance on closed-vocabulary detectors and full-parameter fine-tuning, RRL-HOI may also make open-vocabulary visual understanding more accessible and computationally efficient for research and practical applications.

However, human-object interaction detection also raises potential risks. More accurate human-centric perception systems could be misused for surveillance, behavioral profiling, or monitoring without consent. Prediction errors may also cause harm if the system is used in high-stakes scenarios, especially when the model misinterprets rare actions, ambiguous interactions, or culturally dependent behaviors. In addition, biases in training data or pretrained MLLMs may lead to uneven performance across different environments, activities, or demographic groups. Therefore, RRL-HOI should not be used as a standalone decision-making system in safety-critical or privacy-sensitive applications. We recommend human oversight, privacy-preserving data handling, consent-aware deployment, bias auditing, confidence calibration, and domain-specific validation before applying such systems in real-world settings.

H Isolating the Contribution of GRPO

A central concern raised during review is whether the Stage-3 gain comes from GRPO itself rather than from the stronger MLLM backbone, longer supervised training, or a second generation pass. This appendix provides two controlled analyses that isolate outcome-driven policy optimization: an equal-budget continued-training control and a prediction-transition analysis. Unless otherwise noted, all results are on SWIG-HOI with the Qwen3-VL-8B backbone, under the same data split, prompts, edit space, reward weights, and decoding protocol as the main paper.

H.1 Equal-Budget Continued-Training Control

To exclude the possibility that the improvement merely comes from additional training, we compare GRPO against continued EIT starting from the same Stage-2 checkpoint, under matched data, optimizer updates, effective batch size, and token budget. As shown in Table 8, one additional EIT epoch

Table 8: Equal-budget continued-training control from the same Stage-2 (EIT) checkpoint on SWIG-HOI. One GRPO epoch clearly outperforms one additional EIT epoch under a matched data and compute budget.

Continued training from the same EIT checkpoint	Non-rare	Rare	Unseen	Full
No additional training	27.49	21.85	15.73	21.64
One additional EIT epoch	27.55	21.92	16.02	21.76
One GRPO epoch	**30.71**	**25.58**	**19.94**	**25.37**
GRPO gain over continued EIT	+3.16	+3.66	+3.92	+3.61

Table 9: Prediction-transition analysis on SWIG-HOI. Each initial triplet is tracked by its `triplet_id` and classified before and after reflection. GRPO reduces harmful edits and increases the correction rate over EIT.

Reflection policy	Correct→ Correct	Correct→ Incorrect	Incorrect→ Correct	Incorrect→ Incorrect	New TPs /image
EIT	91.6%	8.4%	25.1%	74.9%	0.28
EIT + GRPO	**95.2%**	**4.8%**	**33.7%**	**66.3%**	**0.41**

brings only $+0.12$ Full mAP ($21.64 \rightarrow 21.76$), whereas one GRPO epoch adds $+3.73$ Full mAP ($\rightarrow 25.37$) and lifts Unseen mAP by a relative 26.8% ($15.73 \rightarrow 19.94$). Thus GRPO remains $+3.61$ Full and $+3.92$ Unseen mAP above equal-budget continued EIT. Moreover, SFT+GRPO *without* EIT reaches only 12.37/18.61 Unseen/Full mAP, confirming that the two stages are complementary rather than redundant: EIT teaches an executable edit language, while GRPO optimizes the executed final prediction.

H.2 Prediction-Transition and Harmful-Edit Analysis

Accuracy alone does not reveal *how* reflection changes predictions. We therefore trace each initial triplet through its `triplet_id` and classify correctness before and after reflection, which directly measures whether reflection damages initially correct predictions. As shown in Table 9, adding GRPO on top of EIT reduces harmful edits (correct $\rightarrow$ incorrect) from 8.4% to 4.8% (a 42.9% relative reduction), raises the correction rate (incorrect $\rightarrow$ correct) from 25.1% to 33.7%, and increases newly created true positives per image from 0.28 to 0.41. This indicates that GRPO learns to prefer edits with positive net outcomes rather than simply producing more edits.

I Backbone Scaling and Adapter Design

Reviewers also asked whether the gains are an artifact of using a larger MLLM, and whether the separate-adapter design is justified. This appendix addresses both questions with a same-family scaling study and a parameter-matched adapter ablation.

I.1 Same-Family Backbone Scaling

We repeat the full three-stage pipeline on Qwen3-VL-2B, 4B, and 8B. For all backbones, the data split, prompts, dynamic-resolution policy, coordinate quantization, LoRA setup, edit space, reward coefficients, and decoding protocol are unchanged. As shown in Table 10, the GRPO-over-EIT gain is consistent across model sizes ($+2.80$/$+3.21$/$+3.73$ Full mAP for 2B/4B/8B), and the improvement of executable editing plus GRPO over the SFT initial prediction holds at every scale. This shows that the benefit comes from the reflective revision mechanism rather than merely from a larger backbone, and the same reward coefficients are reused across all scales without model-specific retuning.

Table 10: Same-family backbone scaling on SWIG-HOI (Unseen/Full mAP). Each backbone uses an identical data split, prompts, resolution policy, quantization, LoRA setup, edit space, reward weights, and decoding protocol.

Backbone	SFT (Unseen/Full)	+EIT (Unseen/Full)	+GRPO (Unseen/Full)	GRPO gain over EIT
Qwen3-VL-2B	3.92 / 10.48	11.74 / 17.51	14.68 / 20.31	+2.94 / +2.80
Qwen3-VL-4B	4.87 / 12.26	13.92 / 19.83	17.46 / 23.04	+3.54 / +3.21
Qwen3-VL-8B	5.69 / 13.39	15.73 / 21.64	**19.94 / 25.37**	+4.21 / +3.73

Table 11: Parameter-matched shared versus separate adapter ablation on SWIG-HOI. At an equal parameter budget (shared rank-128 vs. separate rank-64+64), separate adapters perform consistently better.

Adapter design	Total LoRA params	Unseen	Full
One shared adapter, rank 64	188.7M	17.62	22.94
One shared adapter, rank 128	377.5M	18.43	24.06
Separate adapters, rank 64+64	377.5M	**19.94**	**25.37**

I.2 Shared versus Separate Adapters

RRL-HOI decouples the initial image-to-triplet generator (ϕ_A) from the image-plus-triplet-to-edit reflection policy (ϕ_B) using two separate LoRA adapters. During EIT and GRPO, ϕ_A is frozen, which prevents the initial prediction distribution and the automatically constructed edit targets from drifting together. To verify that this decoupling is beneficial rather than simply a matter of using more parameters, we compare a single shared adapter against separate adapters at a matched parameter budget. As shown in Table 11, a shared rank-128 adapter has the same total LoRA parameters as two separate rank-64 adapters (377.5M), yet the separate design improves Unseen/Full mAP by 1.51/1.31 points. This supports decoupling initial generation from reflection so that reflection-policy optimization does not destabilize the initial prediction distribution.

J Visual Grounding versus Annotation Alignment

A key concern is whether RRL-HOI genuinely reduces visually unsupported interaction predictions or merely conforms to dataset-specific labeling conventions. To disentangle these two effects, we conduct a ground-truth-blind visual-support study in which annotators judge predictions using only visual evidence.

Protocol. We sample 300 SWIG-HOI prediction instances balanced across the non-rare, rare, and unseen splits. Three annotators see only the input image, the predicted boxes, and the predicted interaction; benchmark labels, the model stage, and the before/after order are all hidden. Each prediction is labeled as *Supported*, *Unsupported*, *Contradicted*, or *Ambiguous*; *Unsupported* and *Contradicted* are grouped as visually unsupported, and *Ambiguous* cases are excluded from the main calculation. Inter-annotator agreement is substantial (Fleiss' $\kappa = 0.71$).

Results. As shown in Table 12, visually unsupported predictions drop from 40.7% (initial SFT) to 20.3% (EIT+GRPO), while 93.3% of initially supported predictions are preserved. To further separate visual support from annotation conformity, we examine 90 initial predictions that are absent from the benchmark verb annotations (Table 13). If reflection merely forced predictions toward benchmark labels, visually supported and unsupported unannotated predictions would be removed at similar rates. Instead, RRL-HOI preserves 87.5% of visually supported but unannotated predictions while removing or correcting 76.0% of visually unsupported ones, showing that the model corrects by visual evidence rather than by conforming to labels. We therefore state the claim precisely: RRL-HOI reduces visually unsupported interaction predictions while also improving benchmark-label alignment.

Table 12: Ground-truth-blind visual-support transitions on 300 SWIG-HOI predictions. Annotators see only the image, predicted boxes, and predicted interaction. EIT+GRPO removes visually unsupported predictions while preserving initially supported ones.

Prediction stage	Visually unsupported ↓	Unsupport → support ↑	Support → unsupport ↓	Support preserved ↑
Initial SFT	122/300 (40.7%)	–	–	–
SFT + EIT	86/300 (28.7%)	51/122 (41.8%)	15/178 (8.4%)	163/178 (91.6%)
SFT + EIT + GRPO	**61/300 (20.3%)**	**73/122 (59.8%)**	**12/178 (6.7%)**	**166/178 (93.3%)**

Table 13: Behavior on 90 initial predictions absent from benchmark verb annotations. RRL-HOI preferentially preserves visually supported unannotated predictions and corrects visually unsupported ones, rather than conforming to labels.

Prediction absent from benchmark annotation	Number	After EIT	After EIT+GRPO
Visually supported but unannotated	40	33/40 preserved (82.5%)	35/40 preserved (87.5%)
Visually unsupported and unannotated	50	27/50 removed or corrected (54.0%)	38/50 removed or corrected (76.0%)

```
You are an open-vocabulary human-object interaction detector.
Given an input image, detect all visible human-object interaction triplets.

# Input:
Image: {image}

# Task:
Generate an initial set of box-level HOI triplets. Each triplet should contain:
1. The human bounding box.
2. The object bounding box.
3. The object category.
4. One or more interaction verbs.
5. A confidence score for each interaction verb.

# Output requirements:
1. Return JSON only.
2. Use normalized integer coordinates in [0, 1000].
3. Each box should be represented as [x1, y1, x2, y2].
4. Do not use external proposals or candidate boxes.
5. Do not output free-form explanations.
6. If no valid interaction is visible, return an empty triplet list.

# Output format:
{
  "triplets": [
    {
      "triplet_id": 1,
      "human_box": [x1, y1, x2, y2],
      "object_box": [x1, y1, x2, y2],
      "object_category": "object name",
      "interactions": [
        {
          "verb": "interaction verb",
          "score": 0.0
        }
      ]
    }
  ]
}
```

Figure 4: Prompt design for Stage-1 initial HOI triplet generation.

```
{
  "triplets": [
    {
      "triplet_id": "integer, unique triplet index",
      "human_box": "[x1, y1, x2, y2], normalized integers in [0, 1000]",
      "object_box": "[x1, y1, x2, y2], normalized integers in [0, 1000]",
      "object_category": "open-vocabulary object name",
      "interactions": [
        {
          "verb": "open-vocabulary interaction verb",
          "score": "float in [0, 1]"
        }
      ]
    }
  ]
}
```

Figure 5: Structured JSON schema for initial box-level HOI triplets.

```
You are a structured HOI reflection model.
Given an image and an initial set of HOI triplets, verify whether each triplet is
     visually grounded and revise it when necessary.

# Input:
Image: {image}

Initial triplets:
{initial_triplets}

# Allowed edit types:
1. Box editing:
   - Keep_B: keep the current boxes.
   - Adjust_H: refine the human box.
   - Adjust_O: refine the object box.
   - Replace_H: replace the human box.
   - Replace_O: replace the object box.

2. Pair rebinding:
   - Rebind_H: bind the interaction to another human.
   - Rebind_O: bind the interaction to another object.
   - Swap_P: swap the human-object pairing between triplets.
   - Remove_P: remove an invalid human-object pair.

3. Interaction editing:
   - Keep_A: keep the current interaction verbs.
   - Add_A: add visually supported interaction verbs.
   - Remove_A: remove visually unsupported interaction verbs.
   - Refine_A: replace confusing verbs with more accurate ones.

# Instructions:
1. Check whether the human box covers the interaction subject.
2. Check whether the object box covers the interaction target.
3. Check whether the human-object pair is correctly bound.
4. Check whether each interaction verb is supported by visual evidence.
5. Output executable edits only.
6. Do not output free-form reasoning.
7. Ground-truth annotations are not provided and should not be assumed.

# Output format:
{
  "edits": [
    {
      "edit_id": 1,
      "target_triplet_id": 1,
      "edit_type": "box | pair | action",
      "operation": "operation name",
      "arguments": {}
    }
  ]
}
```

Figure 6: Prompt design for Stage-2 structured triplet reflection.

```
{
  "edits": [
    {
      "edit_id": "integer, unique edit index",
      "target_triplet_id": "integer, the triplet to be edited",
      "edit_type": "box | pair | action",
      "operation": "Keep_B | Adjust_H | Adjust_O | Replace_H | Replace_O | Rebind_H
     | Rebind_O | Swap_P | Remove_P | Keep_A | Add_A | Remove_A | Refine_A",
      "arguments": {
        "new_human_box": "[optional; required for Adjust_H or Replace_H]",
        "new_object_box": "[optional; required for Adjust_O or Replace_O]",
        "new_human_ref": "[optional; required for Rebind_H]",
        "new_object_ref": "[optional; required for Rebind_O]",
        "paired_triplet_id": "[optional; required for Swap_P]",
        "add_verbs": "[optional; required for Add_A]",
        "remove_verbs": "[optional; required for Remove_A]",
        "old_verb": "[optional; required for Refine_A]",
        "new_verb": "[optional; required for Refine_A]"
      }
    }
  ]
}
```

Figure 7: Structured JSON schema for executable triplet-level reflection edits.

```
You are a reflective HOI revision policy.
Given an image and an initial set of HOI triplets, generate a structured edit
     sequence that improves the final HOI prediction.

# Input:
Image: {image}

Initial triplets:
{initial_triplets}

# Objective:
Revise the initial triplets only when the edit is visually justified.
Prefer edits that improve:
1. Human/object localization.
2. Human-object pair binding.
3. Interaction verb correctness.
4. Output validity and edit consistency.

# Constraints:
1. Return JSON only.
2. Use only the allowed edit operations.
3. Do not generate explanations outside the JSON.
4. Avoid unnecessary edits.
5. Avoid creating excessive triplets.
6. All revised boxes must use normalized integer coordinates in [0, 1000].

# Output format:
{
  "edits": [
    {
      "edit_id": 1,
      "target_triplet_id": 1,
      "edit_type": "box | pair | action",
      "operation": "operation name",
      "arguments": {}
    }
  ]
}
```

Figure 8: Prompt design for Stage-3 GRPO sampling and inference-time reflection.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction make claims that accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: This paper discusses the limitations of the work in Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: There are no theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: This work provides all the information needed to reproduce the main experimental results of the paper in the Implementation Details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The data and code will be released upon the publication of our paper.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setup, including data splits, training, and testing details, are provided in the Method and Experiments sections.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We follow the default evaluations protocols in the HOI detection field, which doesn't require error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide them in the implementation details of the main paper and the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This work conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impacts discussion is provided in the Appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not release high-risk models or datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original paper that produced the code package or dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will provide the code to facilitate communication among researchers.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: There is no research with human subjects in this work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We describe the use of LLMs as an important component of the core method.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.