ImagineAgent: What if Agents Could Imagine? Reinforcing Open-Vocabulary HOI Comprehension through Generation

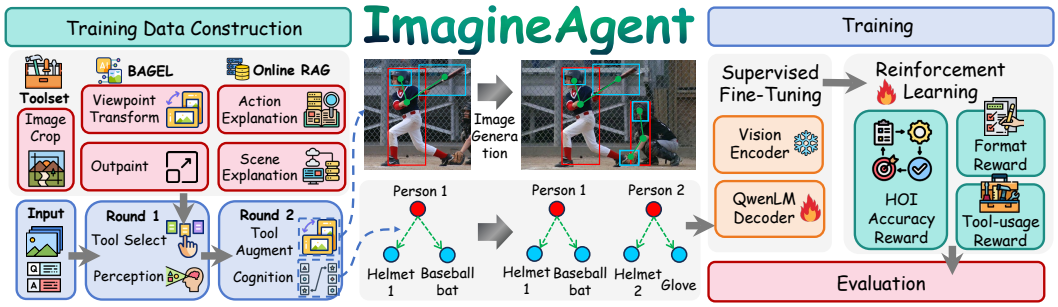
Zhenlong Yuan[1] * **Yue Wang**[2] * **Dapeng Zhang**[3] * **Kejin Cui**[4]
Rui Chen[1] **Jing Tang**[1] **Lei Sun**[1] ‡ **Hongwei Yu**[1]
Chengxuan Qian[1] **Xiangxiang Chu**[1] **Shuo Li**[5] **Yuyin Zhou**[6] §

[1] AMAP, Alibaba Group [2] Meta Platforms [3] National University of Singapore
[4] Independent Researcher [5] Case Western Reserve University [6] UC Santa Cruz

* *Equal contribution* ‡ *Project Lead* § *Corresponding Author*

Abstract

Multimodal Large Language Models have shown promising capabilities in bridging visual and textual reasoning, yet their reasoning capabilities in Open-Vocabulary Human-Object Interaction (OV-HOI) are limited by cross-modal hallucinations and limited viewpoints of images. To address this, we propose **ImagineAgent**, an agentic framework that integrates cognitive mapping, tool-augmented reinforcement learning (RL), and generative world modeling for robust OV-HOI understanding. Specifically, we first propose an innovative CoT dataset named hicodet-6K for supervised fine-tuning (SFT), which effectively bridges the perception-to-cognition gap by structuring perceived entities into interaction pairs for comprehensive predictions. Subsequently, we develop a multimodal tool library integrating online retrieval, image cropping, and generative modeling, enabling the agent to dynamically augment reasoning with domain-specific tools to resolve visual-semantic ambiguities and hallucinations during inference. Moreover, we incorporate a generative model to reconstruct alternative viewpoints, enabling the agent to "imagine" under limited viewpoints. Finally, we propose a composite reward mechanism to jointly optimize prediction accuracy and tool efficiency. Evaluations on both SWIG-HOI and HICO-DET datasets demonstrate that our method achieves state-of-the-art performance while requiring merely 36.7% of the training data compared to existing methods, validating our robustness, empirical effectiveness and efficiency.



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

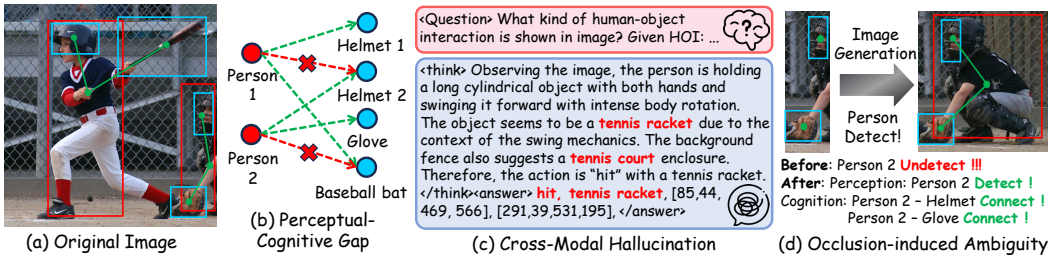


Figure 1: **Three limitations in OV-HOI tasks.** Given (a) an original image, MLLMs may suffer from (b) perceptual-cognitive gap, where they perceive individual entities but fail to form a coherent understanding, leading to flawed cognition (e.g., linking Person 1 to Helmet 2 and Glove they are not wearing). Another critical issue is (c) cross-modal hallucination, where models generate plausible-sounding but visually unsupported interactions by over-relying on textual priors (e.g., mistaking a baseball bat for a tennis racket). Finally, (d) occlusion-induced ambiguity arises when key visual information is missing, causing the model fail to detect entities and their connections.

1 Introduction

Human-Object Interaction (HOI) detection [14] aims to identify and localize interacting human and object pairs within an image, while also recognizing their semantic relationships as *<human, object, action>* triplets. This fine-grained understanding of human-centric activities is foundational for next-generation applications in areas such as robot manipulation, assistive technologies, and intelligent surveillance [34, 22, 49]. To truly unlock this potential, a critical frontier is extending HOI detection beyond a predefined, closed set of categories to an open-vocabulary (OV) setting [46, 26, 17]. This capability would enable the system to generalize to detect any conceivable interaction, even those unseen during training [25, 23], thereby enabling true adaptability in the unpredictable real world.

Recent breakthroughs in Multimodal Large Language Models (MLLMs) [40, 20, 32, 9, 47] have revolutionized both visual perception and reasoning [42]. By leveraging the vast world knowledge embedded in their underlying language models and grounding it in visual evidence, MLLMs interpret nuanced visual details and their semantic implications with high fidelity. Models like Qwen-VL series [2, 1] have demonstrated exceptional performance across a spectrum of challenging tasks, including visual dialogue, few-shot learning, and generating coherent narratives from visual inputs.

Despite their capabilities, MLLMs still exhibit three fundamental limitations when applied to OV-HOI tasks: ❶ **Perceptual-Cognitive Gap**: As illustrated in Fig. 1 (b), while MLLMs detect all persons and objects individually, they often fail to establish semantically coherent physical relationships between them [29]. This stems from a lack of structured cognitive mapping that connects visual perception to plausible interaction hypotheses, which may cause incomplete or fragmented reasoning despite accurate detection. ❷ **Cross-Modal Hallucination**: As shown in Fig. 1 (c), MLLMs frequently generate plausible-sounding but visually unsupported interactions, driven by textual priors rather than visual cues [19]. This cross-modal hallucination arises from imprecise alignment between high-level semantic concepts and low-level visual features, especially under distribution shifts in open-vocabulary settings. ❸ **Occlusion-induced Ambiguity**: As depicted in Fig. 1 (d), real-world scenes frequently encounter partial occlusions or suboptimal viewpoints, which truncate critical visual features for interaction verification. Without mechanisms to recover limited visual information, MLLMs may resort to biased textual priors, causing incorrect predictions in ambiguous scenarios.

To address the above-mentioned challenges, we argue that an effective OV-HOI agent requires three synergistic capabilities: ❶ **Cognitive Mapping**, which adopts a structured reasoning framework that explicitly models the cognitive mapping between detected entities and possible actions. By bridging the gap from perception to cognition, this structure enables the model to construct comprehensive interaction hypotheses with physical plausibility. ❷ **Tool Augmenting**. The agent learns to dynamically invoke domain-specific tools for gathering targeted visual-semantic evidence during reasoning. Such an interactive workflow transforms the static foundation model into an interactive agent, effectively reinforcing cross-modal alignment while suppressing hallucinations. ❸ **Generative World Modeling**, which adopts diffusion models to reconstruct plausible alternative views under occlusion, effectively expanding its observational horizon. This generative prior acts as a world model that empowers the agent to reasonably **"imagine"** through ambiguity or limited viewpoints.

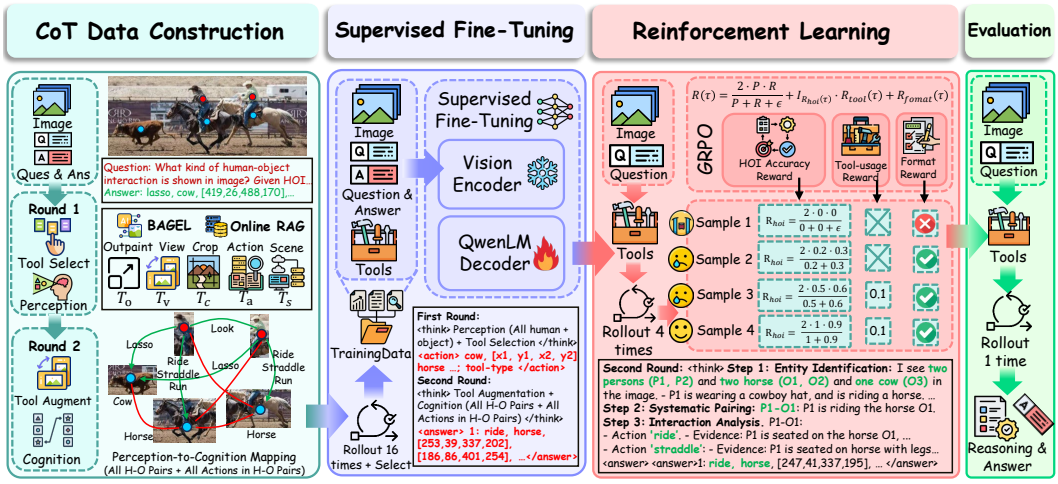


Figure 2: **Pipeline of our method.** (i) We first construct a high-quality dataset of structured reasoning chains named *hicodet-6K* by performing multiple rollouts with a powerful base model. Each data sample encapsulates a two-round reasoning process: (1) perception and tool selection, and (2) tool augmentation and cognition. (ii) Subsequently, *hicodet-6K* dataset is adopted for SFT to initialize the agent's policy, teaching it the fundamental structure of reasoning. (iii) Following this, the model's policy is further refined through RL using the GRPO algorithm, where the agent learns to make optimal decisions by performing rollouts and receiving feedback from a composite reward that balance HOI predictions, structural coherence and tool efficiency, thus achieving robust inference.

Therefore, we introduce ImagineAgent, a novel framework that integrates cognitive reasoning, generative imagination, and **tool-augmented reinforcement learning (RL)** for robust open-vocabulary HOI detection. Specifically, to bridge the perception-to-cognition gap, our method first constructs ***hicodet-6K* dataset** for **supervised fine-tuning (SFT)**. *hicodet-6K* is the first chain-of-thought (CoT) dataset in the Open-Vocabulary Human-Object Interaction (OV-HOI) field that simulates both reasoning and cognitive mapping by structuring perceived entities into interaction pairs for comprehensive predictions. Subsequently, we develop a tool-augmented framework tailored to group relative policy optimization (GRPO) in RL stage. During the inference process, our method dynamically invokes domain-specific tools like online retrieval and image cropping for cross-modal interleaving.

Moreover, to resolve occlusion-induced ambiguity, we further incorporate a generative model as an auxiliary tool to reconstruct alternative viewpoints of the original image, effectively addressing the ambiguities of limitations of its initial perspective. Furthermore, we propose a composite reward mechanism that jointly optimizes both the accuracy of HOI predictions and the efficiency of tool utilization. In summary, our contributions are four-fold:

- **Tool-Augmented Agentic RL.** ImagineAgent introduces a novel framework that unifies both cognitive mapping reasoning and generative modeling with tool-augmentation reinforcement learning, thereby enabling robust OV-HOI through dynamic, visually grounded reasoning.
- **CoT Cognitive Mapping.** We propose *hicodet-6K* dataset that simulates both reasoning and cognitive mapping by structuring perceived entities into interaction pairs for comprehensive predictions, thus effectively bridging the gap between perception and cognition.
- **Multi-Tool Integration.** We design a multimodal tool library integrating online retrieval, image cropping and generative modeling, which dynamically augments reasoning with domain-specific tools to resolve both cross-modal hallucinations and occlusion-induced image ambiguity.
- **Empirical Effectiveness.** Experimental results on both SWIG-HOI [45] and HICO-DET [5] datasets demonstrate that our proposed method achieves state-of-the-art performance in OV-HOI detection, requiring merely approximately 36.7% of the training data compared to existing methods.

2 Methodology

Overview. We propose ImagineAgent, a novel framework that integrates cognitive reasoning with generative imagination and tool-augmented RL for robust OV-HOI, as shown in Fig. 2. Section 2.1

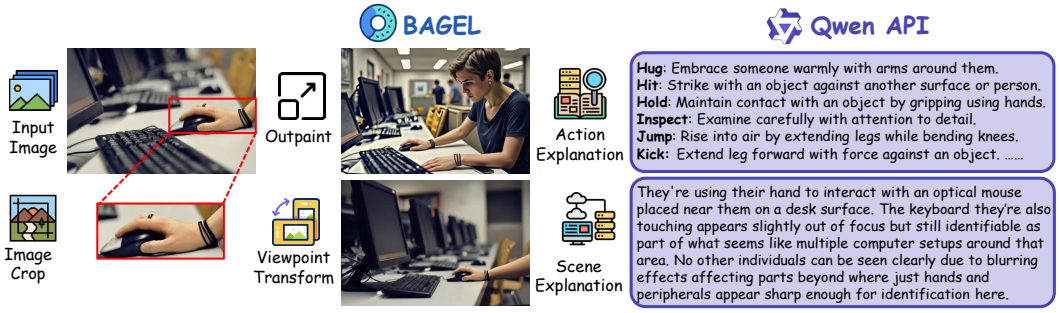


Figure 3: **Tool Library.** To resolve visual and semantic ambiguities of the image, our framework equips the agent with diverse tools. It leverages the BAGEL model for generative imagination via Outpaint and Viewpoint Transform, the Qwen API for Online RAG through Action Explanation and Scene Explanation, and a standard Image Crop tool to extract focused fine-grained interaction details.

presents the problem formulation and our workflow. Section 2.2 details the **Training Data Construction**, which synthesizes structured reasoning chain essential for tool invocation and interaction detection. Section 2.3 describes the **Agentic Supervised Fine-tuning**, where the Qwen3-VL base model is pretrained to initialize the agent's policy for tool invocation. Section 2.4 outlines the **Agentic Reinforcement Learning**, which employs the GRPO algorithm to optimize agent policy via a composite reward that balances prediction accuracy and tool efficiency.

2.1 Problem Formulation & Workflow Design

The goal in OV-HOI is to identify and classify comprehensive interactions from a vocabulary unseen during training. The input is a single image I and a query Q, while the output is a set of interaction triplets $\mathcal{H} = \{\langle h_k, o_k, v_k \rangle\}_{k=1}^{K}$, where h_k is a human, o_k is an object, and v_k is their interaction verb. We deconstruct the problem into a two-stage process:

- **Perception.** This stage involves identifying and localizing all potential human and object instances within the image I to produce sets of localized human instances $\mathcal{I}_H = \{(h_i, b_i)\}_{i=1}^{N}$ and object instances $\mathcal{I}_O = \{(o_j, b_j)\}_{j=1}^{M}$, where b denotes the entity's bounding box.
- **Cognition**: This stage involves reasoning over the perceived instances to correctly pair entities and identify their interaction. It pairs instances (h_i, o_j) from $\mathcal{I}_H \times \mathcal{I}_O$ and predicts all plausible interaction verb v_k, thereby forming each valuable triplet $\langle h_i, o_j, v_k \rangle \in \mathcal{H}$ as final prediction.

To achieve this goal, we design following two-stage workflow to integrates domain-specific tools:

Stage 1: Perception & Tool Selection. Given an image I and query Q, our model identifies sets of localized human instances $\mathcal{I}_H$ and object instances $\mathcal{I}_O$. Concurrently, it constructs a cognitive map of potential interaction pairs (h_i, o_j). For each potential interaction that exhibits ambiguity or insufficient evidence, our agentic framework formulates a plan by selecting an appropriate toolset from tool library $\mathcal{T}$, which comprises an image cropping tool ($\mathcal{T}_c$), retrieval-augmentation tools for scene and action explanation ($\mathcal{T}_s$ and $\mathcal{T}_a$), and generative tools for viewpoint transformation and outpainting ($\mathcal{T}_v$ and $\mathcal{T}_o$). The policy network π_θ determines the tool call $C \in \mathcal{T}$ based on the initial visual context and the specific interaction hypothesis being evaluated.

Stage 2: Cognition & Tool Augmentation. The second stage executes the selected tool call C and integrates its output to transition from perception to cognition. The tool outputs are not merely concatenated but are adopted to dynamically enhancing agent's scenario understanding. A detailed view from the cropping tool ($\mathcal{T}_c$) allows for focused attention, while contextual knowledge from retrieval tools ($\mathcal{T}_s, \mathcal{T}_a$) enriches the agent's semantic representation of the query. Moreover, an enhanced image I' that resolves prior visual ambiguities from generative tools ($\mathcal{T}_v, \mathcal{T}_o$) provides the agent reasons over an *imagined reality*. The final prediction for an interaction triplet $\langle h, o, v \rangle \in \mathcal{H}$ is then inferred by conditioning the model π_θ on these augmented inputs:

$$p(\langle h, o, v \rangle | I', Q') = \pi_\theta(I', Q'), \tag{1}$$

where $I' = I \oplus I_{\text{aug}}$ and $Q' = Q \oplus E_{\text{aug}}$, $\oplus$ denotes modality-specific augmentation, and I_{aug} and E_{aug} respectively represent the augmented visual and textual evidence, while π_θ maintains the same policy network of MLLMs parameterized by weights θ.

Tool Library. As shown in Fig. 3, our framework integrates a diverse tool library to facilitate a multi-stage reasoning process. ❶ **Image Crop**: to ensure detailed evidence is not overlooked, this tool allows the agent to scrutinize fine-grained interaction details within high-resolution sub-regions. ❷ **Outpaint & Viewpoint Transform**: to resolve occlusion-induced ambiguity, we equip the agent with generative imagination via BAGEL [10], allowing it to reconstruct missing context or synthesize alternative views. ❸ **Action & Scene Explanation**: to mitigate cross-modal hallucination, we leverage the Qwen API for online RAG, which injects precise semantic definitions and contextual information, anchoring predictions in external knowledge. Details are provided in the Appendix B.

2.2 Training Data Construction

A fundamental challenge when applying existing MLLMs to OV-HOI tasks is the absence of explicit reasoning chains. Standard pipelines provide static labels but lack intermediate cognitive steps: **what** tool to use, **which** human-object pairs to link, and **how** plausible verbs are identified. To address this, we devise a process to synthesize a high-fidelity structured reasoning chains for constructing our proposed *hicodet-6K* dataset, which creates explicit examples to subtly resolve the above-mentioned cognitive steps, thus effectively bridge the gap between perception and cognition during inference.

Data Collection. Our data generation begins by leveraging the Qwen3-VL-32B [1] to explore the HICO-DET training dataset. For each image, we perform 16 independent policy rollouts to generate a diverse set of potential reasoning pathways. An instance is considered solvable if at least one rollout produces a triplet set that matches the training annotations under the same matching protocol used for reward computation, including both semantic consistency and spatial IoU constraints.

Data Assessment. To ensure the quality and reliability of the synthesized data, we introduce a rigorous assessment stage. Specifically, we employ the Qwen-VL-Max model as an expert validator to meticulously review and score each reasoning chain. Chains containing factual errors, logical inconsistencies, or hallucinated reasoning are systematically discarded. This validation procedure yields approximately 6,000 high-fidelity reasoning chains for SFT. The remaining 8,000 trajectories, which were confirmed as successful but not selected in SFT, are then served as the training data for RL stage, providing a wider range of successful behaviors for policy exploration. To prevent data leakage, hicodet-6K and all RL trajectories are constructed only from the HICO-DET training set.

Generating Pipeline. Our generating pipeline implements a three-step reasoning process. First, based on the initial perception stage, the agent identifies all human instances $\mathcal{I}_H$ and object instances $\mathcal{I}_O$. Second, it constructs a comprehensive hypothesis space by performing an exhaustive pairing of all perceived entities, creating a set of candidate human-object pairs $(h_i, o_j) \in \mathcal{I}_H \times \mathcal{I}_O$. Finally, for each pair, the agent evaluates a constrained set of valid verbs v_k, scoring each of them based on visual evidence and contextual cues. Actions with high scores deemed valid and selected to constitute all validated interaction triplets $\mathcal{H} = \{\langle h_k, o_k, v_k \rangle\}_{k=1}^{K}$. This structured approach ensures comprehensive consideration of all potential HOIs, effectively bridging perception-to-cognition gap.

2.3 Agentic Supervised Fine-Tuning

After constructing the training data, we introduce a **cold-start phase** that prioritizes supervised fine-tuning (SFT) before applying reinforcement learning (RL). Inspiring from R1-Zero [12], we initially attempt direct RL optimization to train our method. However, preliminary experiments reveal a progressive decline in the frequency of tool invocation during policy rollouts. This behavior likely arises from a distributional discrepancy in target domain between tool-enhanced visual features and model's pretraining data. Therefore, we opt to perform SFT before RL.

Specifically, We formalize each training instance as $\mathcal{W} = (\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{O})$, where $\mathcal{X}$ denotes the input modality, $\mathcal{Y}$ represents task instructions, $\mathcal{Z} = \{z_1, \ldots, z_T\}$ captures reasoning steps, and $\mathcal{O}$ is the target output. The objective minimizes the negative log-likelihood of reasoning process:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{\mathcal{W}\sim\mathcal{D}} \left[\sum_{t=1}^{T} \log p_\theta(z_t \mid \mathcal{X}, \mathcal{Y}, z_{<t}) \right], \tag{2}$$

where $p_\theta(\cdot)$ denotes conditional probability distribution. This formulation enables reasoning chains are aligned with both task instructions and GT during pre-training.

2.4 Agentic Reinforcement Learning

Group Relative Policy Optimization (GRPO). We adopt the GRPO algorithm [41] for policy optimization. GRPO innovatively employs a groupwise comparison framework to evaluate candidate responses. Specifically, for each query q paired with its ground-truth solution a from dataset D, the algorithm generates a set of rollout trajectories $\{o_1, o_2, \ldots, o_G\}$ based on the previous policy $\pi_{\theta_{\text{old}}}$. The policy π_θ is then refined through the optimization of this objective function:

$$\mathcal{L}_{GRPO}(\theta) = -\mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(O|q)} \left[\frac{1}{G}\sum_{i=1}^{G} \left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\epsilon, 1+\epsilon \right) A_i \right) - \beta \mathbb{D}_{KL}(\pi_\theta \| \pi_{ref}) \right) \right], \tag{3}$$

$$\mathbb{D}_{KL}(\pi_\theta||\pi_{ref}) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1, \tag{4}$$

where β is adopted to balance the trade-off between exploration and stability during optimization. Then the advantage estimator A_i is calculated using normalized rewards from the trajectory group:

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

Each trajectory o_i receives a composite reward through rule-based verification designed as follows.

Reward Design. Effective reward functions should balance precision & recall of predictions, structural coherence, and tool efficiency. To this end, we design a composite reward that integrates three components: tool-usage reward R_{tool}, format reward R_{format} and hoi-based accuracy reward R_{hoi}. The tool-usage reward is defined as $R_{\text{tool}}(\tau) = 1$ when at least one valid tool invocation is executed and the generated answer obtains a positive HOI reward, otherwise $R_{\text{tool}}(\tau) = 0$. $R_{\text{format}}(\tau) = 1$ if the trajectory τ follows the required structured reasoning format; otherwise $R_{\text{format}}(\tau) = 0$.

Differently, the hoi-based accuracy reward R_{hoi} quantifies the similarity between the predicted triplets $\mathcal{H}_p = \{\langle h_{p_i}, o_{p_i}, v_{p_i}\rangle\}_{i=1}^{N_p}$ against the ground-truth triplets $\mathcal{H}_g = \{\langle h_{g_i}, o_{g_i}, v_{g_i}\rangle\}_{i=1}^{N_g}$. This requires solving a complex set-to-set matching problem, for which we first compute pairwise affinity score s_{ij} between each predicted triplet p_i and ground-truth triplet g_j. A non-zero score is assigned only if the pair (p_i, g_j) passes both semantic and spatial verification checks. For semantic verification, we encode the verb and object strings using the CLIP ViT-L/14 text encoder and compute cosine similarity in the resulting text embedding space. For spatial verification, we use the human and object bounding boxes of the predicted and ground-truth triplets. The matching conditions are defined as:

$$\begin{aligned} \cos(v_{p_i}, v_{g_j}) > \delta \quad \wedge \quad \cos(o_{p_i}, o_{g_j}) > \delta, \\ \text{IoU}(b_{h_{p_i}}, b_{h_{g_j}}) > \eta \quad \wedge \quad \text{IoU}(b_{o_{p_i}}, b_{o_{g_j}}) > \eta, \end{aligned} \tag{6}$$

where $\cos(\cdot,\cdot)$ is the cosine similarity, with $\delta = 0.8$ and $\eta = 0.5$. The affinity score s_{ij} is defined as 1 if Eqn. 6 satisfy, otherwise $s_{ij} = 0$. With the $N_p \times N_g$ affinity score matrix $S = [s_{ij}]$, we formulate it as a linear assignment problem and employ Hungarian algorithm to find the optimal one-to-one matching. The algorithm operates on a cost matrix C, where $c_{ij} = 1 - s_{ij}$, and finds the assignment M^* that minimizes the total cost:

$$M^* = \underset{M\in\Pi(N_p,N_g)}{\arg\min} \sum_{(i,j)\in M} c_{ij}, \tag{7}$$

where $\Pi(N_p, N_g)$ is the set of all possible one-to-one matchings. From this optimal matching, we identify the number of true positives (TP) as the count of matched pairs with a score greater than zero, i.e., $\text{TP} = |\{(i,j) \in M^* \mid s_{ij} > 0\}|$. The final accuracy reward is then calculated as the F1-score, which harmonizes precision (P) and recall (R):

$$P = \frac{\text{TP}}{N_p}, \quad R = \frac{\text{TP}}{N_g}, \quad R_{\text{hoi}} = \frac{2\cdot P\cdot R}{P+R+\epsilon}, \tag{8}$$

where ϵ is a small constant to prevent division by zero. Then the final reward is formally defined as:

$$R(\tau) = R_{\text{hoi}} + R_{\text{format}}(\tau) + \mathbb{I}_{R_{\text{hoi}}(\tau)>0} \cdot R_{\text{tool}}(\tau), \tag{9}$$

where $\mathbb{I}_{R_{\text{hoi}}(\tau)>0}$ is an indicator function. Such composite reward not only maximizes a balanced accuracy metric but also achieves structural coherence and tool-usage efficiency.

3 Experiment

Datasets. We perform our experiments on two datasets: HICO-DET [5] and SWIG-HOI [45]. HICO-DET includes 600 interaction types made from 117 human actions and 80 objects. Following previous studies [15, 46], we exclude 120 rare interactions to test our model in a zero-shot detection setting. SWIG-HOI provides a wide range of interactions with 400 actions and 1,000 objects. This test set naturally contains many unseen combinations. It has about 14K images and 5.5K interaction types, including 1.8K interactions that are not present during training.

Evaluation Metrics. Following prior work [46, 26, 27], we evaluate our model using mean Average Precision (mAP). A prediction is a true positive if it satisfies: (1) The Intersection over Union (IoU) for both the human and object bounding boxes are greater than 0.5. (2) The predicted verb (interaction) label and corresponding object label are both correct.

Implementation Details. We perform experiments with Qwen3-VL-8B as our base model. According to Sec. 2.2, we respectively select approximately 6,000 and 8,000 training samples for both SFT and RL. The training was conducted on 8 NVIDIA H20 GPUs (96 GB VRAM) with batch size of 8. The training process took around 40 hours for 1,000 iterations (1 epoch), with 4 rollouts per sample and a learning rate of 5e-7. More details are in Appendix A.1.2.

3.1 Comparison with SOTA Methods

Results on HICO-DET. We evaluate our method on the HICO-DET dataset in Tab. 1. Following INP-CC [27], we categorize existing methods into two groups for fair assessment. Note that the best and second best performances are highlighted. As can be seen, while zero-shot methods like HOICLIP [38] and CLIP4HOI [37] have shown strong results, they inevitably rely on detector backbones pre-trained on large-scale object detection datasets like COCO. This pre-training can provide an unfair advantage, as both HICO-DET and COCO datasets share a significant portion of their object categories, limiting the assessment of true generalization in an open-world settings.

Table 1: Comparison of our model with existing methods on HICO-DET under the simulated zero-shot setting.

Method	Pretrained Detector	Unseen	Seen	Full
Zero-shot Methods	-	-	-	-
FCL [16]	✓	13.16	24.23	22.01
GEN-VLKT [31]	✓	21.36	32.91	30.56
HOICLIP [38]	✓	23.48	34.47	32.26
CLIP4HOI [37]	✓	28.47	35.48	34.08
HOIGen [13]	✓	31.01	34.57	33.86
Open-vocabulary Methods	-	-	-	-
THID [46]	✗	15.53	24.32	22.38
CMD-SE [26]	✗	16.70	23.95	22.35
INP-CC [27]	✗	17.38	24.74	23.12
Qwen2.5-VL-7B	✗	16.94	13.43	14.13
Qwen3-VL-8B (Base)	✗	19.41	18.24	18.47
ImagineAgent (Ours)	✗	37.85	36.49	36.76

Open-vocabulary methods offer a more rigorous test of generalization by forgoing such detection pre-training. ImagineAgent achieves **36.76%** mAP on the Full split, a remarkable **13.64% mAP** improvement over the prior state-of-the-art, INP-CC [27]. This significant leap demonstrates the powerful reasoning and ambiguity-resolution capabilities of our agentic framework.

Results on SWIG-HOI. We further evaluate our method on the SWIG-HOI dataset. As shown in Tab. 2, our method consistently surpasses all previous methods, achieving a Full mAP of **20.84%**, outperforming INP-CC by **4.10%** mAP. The strength of our approach is further evidenced by its performance on novel and infrequent interactions, with a **14.70%** mAP on the Unseen split and **21.16%** on the Rare split. Such results validate our robustness, which excels at detecting interactions in open-vocabulary long-tail scenes.

Table 2: Comparison of our model with existing methods on SWIG-HOI under simulated zero-shot setting.

Method	Non-rare	Rare	Unseen	Full
QPIC [44]	16.95	10.84	6.21	11.12
GEN-VLKT [31]	20.91	10.41	-	10.87
MP-HOI [49]	20.28	14.78	-	12.61
THID [46]	17.67	12.82	10.04	13.26
CMD-SE [26]	21.46	14.64	10.70	15.26
INP-CC [27]	22.84	16.74	11.02	16.74
Qwen2.5-VL-7B	1.55	3.54	5.82	3.61
Qwen3-VL-8B (Base)	3.09	7.05	11.61	7.25
ImagineAgent (Ours)	26.88	21.16	14.70	20.84

3.2 Ablation Study

Effect of Training Stages. We analyze each training stage's contribution in Tab. 3. The baseline model (row a) struggles with the HOI task. SFT alone (row b) significantly boosts performance to **27.23%** mAP by teaching model structured reasoning. Conversely, RL-only training (row c) shows limited improvement, highlighting the challenge of exploring the vast policy space without a strong prior. Differently, combining SFT and RL (row d) achieves the highest performance at **36.76%** mAP. This confirms that SFT provides a crucial cognitive foundation for subsequent RL refinement.

Table 3: Ablation of training stages on HICO-DET dataset.

Method	Training Steps	Unseen	Seen	Full
(a) Qwen3-VL-8B	0 steps	19.41	18.24	18.47
(b) w/ SFT (*hicodet-6K*)	750 steps	28.24	26.98	27.23
(c) w/ RL (GRPO)	1,000 steps	25.33	24.34	24.54
ImagineAgent (Ours)	1,750 steps	37.85	36.49	36.76

Effect of Tool-Usage. We investigate each tool's contribution in Tab. 4, where disabling any single tool causes performance degradation. Removing generative tools like viewpoint transformation (d) and outpainting (e) hinders reasoning under occlusion, causing drops to **35.48%** and **35.61%** mAP. Excluding retrieval-based explanations (b, c) impairs semantic disambiguation. Disabling image cropping (a) leads to a significant drop to **35.69%** mAP, indicating that fine-grained analysis is indispensable. The full framework's superior performance validates our multi-tool integration strategy.

Table 4: Ablation of tool-usage on HICO-DET.

Method	Unseen	Seen	Full
(a) w/o image crop	37.69	35.19	35.69
(b) w/o action explanation	37.37	34.17	34.81
(c) w/o scene explanation	37.31	35.96	36.23
(d) w/o view transform	37.82	34.90	35.48
(e) w/o outpaint	37.58	35.10	35.61
ImagineAgent (Ours)	37.85	36.49	36.76

Effect of Reward Functions. Our composite reward's design is pivotal for guiding the agent's policy. As shown in Tab. 5, removing the core HOI accuracy reward (a) causes a catastrophic drop to **32.62%** mAP. Simplifying the accuracy metric by removing the Hungarian algorithm (c) or verb similarity check (d) also degrades performance, confirming that our nuanced F1-score is essential for evaluating complex scenes. Finally, excluding the tool-usage reward (e) drops mAP to **35.66%**, showing its importance in encouraging efficient tool invocation. Each component is thus crucial for steering the agent towards accurate and efficient behavior.

Table 5: Ablation of reward func. on HICO-DET.

Method	Unseen	Seen	Full
(a) w/o HOI Reward	33.75	32.34	32.62
(b) w/o Precision & Recall	34.56	33.42	33.65
(c) w/o Hungarian	35.62	34.53	34.75
(d) w/o Verb Similarity	37.25	36.03	36.27
(e) w/o Tool Reward	36.62	35.42	35.66
ImagineAgent (Ours)	37.85	36.49	36.76

Effect of Generative Models. We analyze the impact of applying different generative models in Tab. 6. Our investigation reveals that while more advanced models such as FLUX.1 Kontext (a) and Qwen-Image (b) can yield marginal performance improvements, the former reaching **37.23%** mAP while come at a greater computational cost. Differently, BAGEL (c), provides a compelling balance between performance and processing speed, makes it a more practical choice for our workflow.

Table 6: Effect of different generative models on HICO-DET.

Editing Model	Unseen	Seen	Full
(a) FLUX.1 Kontext [24]	38.34	36.95	37.23
(b) Qwen-Image [48]	38.19	36.83	37.10
(c) BAGEL [10] (**Ours**)	37.85	36.49	36.76

3.3 Qualitative Results

Our method utilizes outpainting, view transformation, and cropping to respectively enrich context, resolve occlusion, and focus on critical details (see Appendix A.4 for more cases). This enhanced perception directly translates to superior cognitive reasoning, as shown in Fig. 4, where our method accurately deduces a comprehensive set of correct interactions. To mitigate potential hallucinations from generative images, we append designed prompt instructions to govern the use of synthesized images (see Appendix B.3 for details).

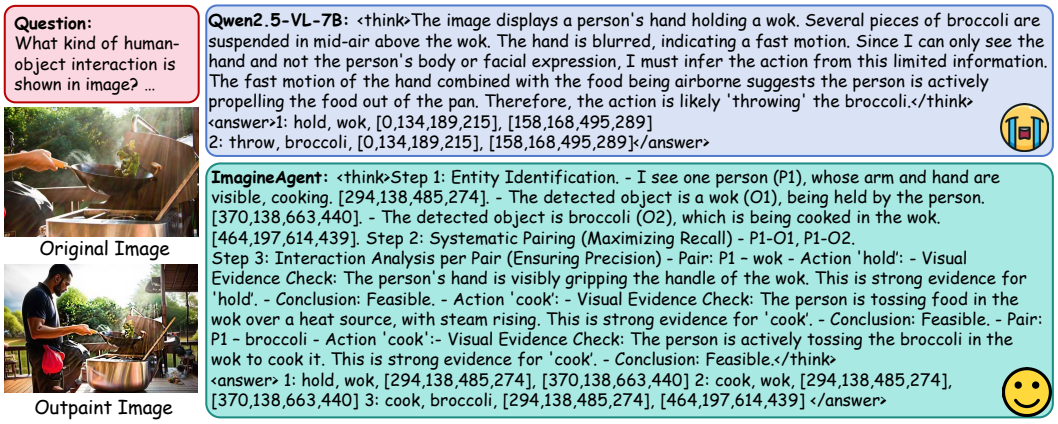


Figure 4: **Case Study between Qwen2.5-VL-7B and our method.** Our method can accurately identify comprehensive HOI predictions, while Qwen2.5-VL-7B misclassifies the interaction labels.

4 Related Work

Open-Vocabulary HOI Detection. Previous methods for open-vocabulary Human-Object Interaction (HOI) detection primarily leverage the cross-modal alignment capability of CLIP [31, 38, 37], and utilize DETR [4] as a critical component to perform object detection in the first place [23]. With the advances of Large Language Models (LLMs), recent literature [25–27] proposed to improve the fine-grained understanding of diverse HOI concepts through prompting LLMs. Furthermore, MLLMs [2, 1] show great potential in open-vocabulary HOI detection in an end-to-end manner with strong visual grounding ability and world knowledge. Thus, we build **ImagineAgent** upon MLLMs and further achieve tool-augmented RL for robust open-vocabulary inference.

Multimodal LLMs Reasoning. Recent advances in Large Language Models (LLMs) have demonstrated that RL-based post-training can significantly enhance reasoning capabilities [8, 50, 6], as exemplified by OpenAI-o1 [21] and DeepSeek-R1 [12]. These paradigms have been extended to Multimodal Large Language Models (MLLMs) for tasks like mathematical VQA [39], image segmentation [35], visual grounding [3], and video understanding [11]. However, existing methods struggle with long-sequence hallucination [7] and limited cross-modal interaction [28]. To address this, we propose multimodal CoT reasoning via tool-augmented RL to explicitly reduce cross-modal hallucination.

Tool-Augmented Agentic System. Recent advancements in MLLMs have shown that external tools can enhance multimodal reasoning. Early works like FAST [43] and MVoT [30] introduce visual evidence into reasoning, forming multimodal CoT for image tasks. LLaVA-Plus [33] pioneered training strategies for tool use, while VPD [18] leveraged program-derived data to transfer tool skills. Recent methods like TACO [36] and PyVision [52] expand tool use with RL. Video-STAR [51] further explores tool-augmented RL for video action recognition. Our framework unifies cognitive mapping reasoning and generative modeling with tool-augmented RL for robust OV-HOI prediction.

5 Conclusion

In this work, we introduced **ImagineAgent**, a novel agentic framework designed to enhance OV-HOI understanding through structured reasoning and multimodal augmentation. Our approach leverages four key components: (i) The *hicodet-6K* CoT dataset for SFT, bridging the perception-to-cognition gap by structuring perceived entities into interaction pairs; (ii) A multimodal tool library integrating online retrieval, image cropping, and generative modeling, enabling dynamic reasoning with domain-specific tools to resolve ambiguities and hallucinations; (iii) A generative model that reconstructs alternative viewpoints, allowing the agent to "imagine" beyond limited perspectives; and (iv) A composite reward mechanism that jointly optimizes prediction accuracy and tool efficiency. Extensive evaluations on the SWIG-HOI and HICO-DET datasets validate our SOTA performance while using merely 36.7% of the training data compared to existing methods. Future work will explore expanding the tool library and applying imagination-driven reasoning to broader complex visual tasks.

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, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[3] Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*, 2025.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 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] Rui Chen, Lei Sun, Jing Tang, Geng Li, and Xiangxiang Chu. Finger: Content aware fine-grained evaluation with reasoning for ai-generated videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3517–3526, 2025.

[7] Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, et al. Scaling RL to long videos. *arXiv preprint arXiv:2507.07966*, 2025.

[8] Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. Gpg: A simple and strong reinforcement learning baseline for model reasoning. In *International Conference on Learning Representations*, volume 2026, pages 59637–59659, 2026.

[9] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[10] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Shi Guang, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *CoRR*, abs/2505.14683, 2025.

[11] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

[12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, and et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

[13] Yixin Guo, Yu Liu, Jianghao Li, Weimin Wang, and Qi Jia. Unseen no more: Unlocking the potential of clip for generative zero-shot HOI detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1711–1720, 2024.

[14] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

[15] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *European Conference on Computer Vision*, pages 584–600. Springer, 2020.

[16] 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.

[17] Yupeng Hu, Changxing Ding, Chang Sun, Shaoli Huang, and Xiangmin Xu. Bilateral collaboration with large vision-language models for open vocabulary human-object interaction detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20126–20136, 2025.

[18] Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9590–9601, 2024.

[19] Zhe Huang, Hao Wen, Aiming Hao, Bingze Song, Meiqi Wu, Jiahong Wu, Xiangxiang Chu, Sheng Lu, and Haoqian Wang. Taming hallucinations: Boosting mllms' video understanding via counterfactual video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2026.

[20] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[21] 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.

[22] 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.

[23] Sanghyun Kim, Deunsol Jung, and Minsu Cho. Locality-aware zero-shot human-object interaction detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20190–20200, 2025.

[24] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 kontext: Flow matching for in-context image generation and editing in latent space. *CoRR*, abs/2506.15742, 2025.

[25] Qinqian Lei, Bo Wang, and Robby Tan. Ez-HOI: VLM adaptation via guided prompt learning for zero-shot hoi detection. *Advances in Neural Information Processing Systems*, 37:55831–55857, 2024.

[26] 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.

[27] Ting Lei, Shaofeng Yin, Qingchao Chen, Yuxin Peng, and Yang Liu. Open-vocabulary HOI detection with interaction-aware prompt and concept calibration. In *ICCV*. IEEE, 2025.

[28] Bozhou Li, Xinda Xue, Sihan Yang, Yang Shi, Xinlong Chen, Yushuo Guan, Yuanxing Zhang, and Wentao Zhang. The unseen bias: How norm discrepancy in pre-norm mllms leads to visual information loss. In *International Conference on Learning Representations*, volume 2026, pages 1687–1709, 2026.

[29] Bozhou Li, Jiahang Zhang, Yue Ding, Yushuo Guan, Bohan Zeng, Yiyan Ji, Xinlong Chen, Yang Shi, Yifan Dai, Yuran Wang, et al. Human-centric image captioning with subject-centered spatial understanding. *arXiv preprint arXiv:2609.08300*, 2026.

[30] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025.

[31] 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.

[32] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[33] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. LLaVA-Plus: Learning to use tools for creating multimodal agents, 2023. URL https://arxiv.org/abs/2311.05437.

[34] Xinpeng Liu, Yong-Lu Li, Xiaoqian Wu, Yu-Wing Tai, Cewu Lu, and Chi-Keung Tang. Interactiveness field in human-object interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20113–20122, 2022.

[35] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025.

[36] Zixian Ma, Jianguo Zhang, Zhiwei Liu, Jieyu Zhang, Juntao Tan, Manli Shu, Juan Carlos Niebles, Shelby Heinecke, Huan Wang, Caiming Xiong, et al. Taco: Learning multi-modal action models with synthetic chains-of-thought-and-action. *arXiv preprint arXiv:2412.05479*, 2024.

[37] Yunyao Mao, Jiajun Deng, Wengang Zhou, Li Li, Yao Fang, and Houqiang Li. CLIP4HOI: towards adapting CLIP for practical zero-shot HOI detection. *Advances in Neural Information Processing Systems*, 36:45895–45906, 2023.

[38] 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.

[39] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based RL. *arXiv preprint arXiv:2503.07536*, 2025.

[40] 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.

[41] 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.

[42] Zihan Su, Hongyang Wei, Kangrui Cen, Yong Wang, Guanhua Chen, Chun Yuan, and Xiangxiang Chu. Generation enhances understanding in unified multimodal models via multi-representation generation. *arXiv preprint arXiv:2601.21406*, 2026.

[43] Guangyan Sun, Mingyu Jin, Zhenting Wang, Cheng-Long Wang, Siqi Ma, Qifan Wang, Tong Geng, Ying Nian Wu, Yongfeng Zhang, and Dongfang Liu. Visual agents as fast and slow thinkers. In *The Thirteenth International Conference on Learning Representations*, 2025.

[44] 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.

[45] 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.

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

[47] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl 3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[48] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Shengming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report. *CoRR*, abs/2508.02324, 2025.

[49] 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.

[50] Zhenlong Yuan, Chengxuan Qian, Jing Tang, Rui Chen, Zijian Song, Lei Sun, Xiangxiang Chu, Yujun Cai, Dapeng Zhang, and Shuo Li. Autodrive-r^2: Incentivizing reasoning and self-reflection capacity for vla model in autonomous driving. In *International Conference on Learning Representations*, volume 2026, pages 97934–97955, 2026.

[51] Zhenlong Yuan, Xiangyan Qu, Chengxuan Qian, Rui Chen, Jing Tang, Lei Sun, Xiangxiang Chu, Dapeng Zhang, Yiwei Wang, Yujun Cai, et al. Video-star: Reinforcing open-vocabulary action recognition with tools. In *International Conference on Learning Representations*, volume 2026, pages 51445–51468, 2026.

[52] Shitian Zhao, Haoquan Zhang, Shaoheng Lin, Ming Li, Qilong Wu, Kaipeng Zhang, and Chen Wei. PyVision: Agentic vision with dynamic tooling. *arXiv preprint arXiv:2507.07998*, 2025.

Appendix

In this appendix, we provide more experimental details, tool library details, and prompt details of our proposed method. Specific detailed contents are as follows:

A Experiment **14**

A.1 Full Experimental Setup (Reproducibility) 14

A.1.1 Datasets and evaluation protocols 14

A.1.2 Implementation Details and Hyperparameters 15

A.2 More Ablation Studies 15

A.3 Data Construction Details (Reasoning Chains) 16

A.3.1 Generation Prompts 16

A.3.2 Expert Validation and Filtering 17

A.3.3 Examples of Synthesized Reasoning Chains 17

A.4 More Case Studies 17

B Tool Library Details **18**

B.1 Visual Refinement Tools 18

B.2 Online Retrieval-Augmentation Tools 19

B.3 Generative Imagination Tools 19

C Discussion on Generative Hallucination **20**

D Prompts **21**

E Limitation **22**

A Experiment

A.1 Full Experimental Setup (Reproducibility)

A.1.1 Datasets and evaluation protocols

HICO-DET We conduct experiments on HICO-DET, a standard benchmark for HOI detection with 600 interaction categories composed of 117 verbs and 80 object classes and use the official test set for evaluation. Following prior zero-shot HOI protocols (CMD-SE [26] and INP-CC [27]), 120 rare interaction categories are excluded from training to simulate the zero-shot setting, where these categories are treated as Unseen, and the remaining categories are treated as Seen. Results on Unseen, Seen, and Full splits are reported.

SWIG-HOI We also evaluate on SWIG-HOI, a large-vocabulary HOI benchmark built upon the publicly available SWiG dataset to study large-vocabulary and open-vocabulary HOI detection. SWIG-HOI contains more than 400 action (verb) categories and 1,000 object categories, substantially expanding the label space compared to earlier datasets such as HICO-DET. In our setting, Unseen refers to out-of-vocabulary action types that do not appear in the training set, and we report results on the standard Non-rare, Rare, Unseen, and Full splits.

Metric. We follow prior work and evaluate using mean Average Precision (mAP) computed by the official evaluation scripts, with no additional filtering.

True positive criterion. A predicted HOI triplet is considered correct if it satisfies both: (i) IoU > 0.5 for the predicted human bounding box and its matched ground-truth human box, and IoU > 0.5 for the predicted object bounding box and its matched ground-truth object box. (ii) the predicted interaction label matches the ground truth.

A.1.2 Implementation Details and Hyperparameters

Base Model. We choose Qwen3-VL-8B as our backbone and fine-tune the model using full-parameter finetuning.

Input Configuration. For each sample, the maximum image resolution is set to 401408 pixels, and the number of image tokens is capped at 512. The maximum text sequence length is 8192 tokens, and the maximum generation length during rollouts is 4096 tokens. We use resize and center crop for image preprocessing.

Optimization. We train the model with Adam optimizer using a learning rate of 1e-6 for SFT Stage, and 5e-7 for RL Stage with weight decay 0.01, and gradient clipping with max norm 5. All experiments are conducted with BF16 mixed precision, and the random seed is fixed to 1234 for reproducibility.

Batching and Training Budget. We use a per-device batch size of 1 on 8 GPUs, with gradient accumulation steps 1 for both SFT and RL. The resulting global batch size is therefore $\mathbf{8} = \mathbf{1} \times \mathbf{8} \times \mathbf{1}$ for SFT, and $\mathbf{8} = \mathbf{1} \times \mathbf{8} \times \mathbf{1}$ for RL.

SFT Stage. In supervised fine-tuning (SFT), we train for 1 epoch and use 5999 trajectories.

RL Stage (GRPO). In agentic reinforcement learning, we optimize the policy using GRPO for approximately 1,000 steps. For each training query, we sample 4 rollouts. We use KL regularization coefficient $\beta = 0.04$. The reference policy π_{ref} is chosen as SFT model, and the old policy $\pi_{\theta_{\text{old}}}$ is updated every 1 steps.

Hardware. Training is performed on eight **Nvidia H20** GPUs with **96GB** memory for each GPU. The total training time is approximately **2.5 hours** for SFT and **40 hours** for RL.

A.2 More Ablation Studies

Comparison with Frontier MLLMs. To further evaluate whether general-purpose frontier MLLMs can directly solve fine-grained OV-HOI reasoning, we compare ImagineAgent with several strong proprietary models, including GPT-4o, Gemini 3.0 Pro, and GPT-5. As shown in Tab. 7, although these frontier models possess strong native multimodal reasoning abilities, their performance remains limited on fine-grained HOI comprehension. In particular, GPT-5 achieves only 23.18% Full mAP, which is still substantially lower than ImagineAgent with Qwen3-VL-8B as the backbone.

This result suggests that parameter scaling alone is insufficient for resolving the severe occlusions, spatial ambiguities, and human-object pairing challenges in OV-HOI. In contrast, our specialized tool-augmented reasoning framework explicitly decomposes the task into perception, tool selection, and interaction-aware cognition, thereby providing more reliable visual grounding and structured reasoning. These results further validate the necessity of specialized agentic reasoning for HOI tasks.

Table 7: Comparison with frontier MLLMs on HICO-DET under the simulated zero-shot setting.

Method	Unseen	Seen	Full
GPT-4o	22.15	21.45	21.78
Gemini 3.0 Pro	22.45	21.75	22.06
GPT-5	23.45	22.85	23.18
ImagineAgent (Ours)	33.55	32.40	32.84

Prevalence of Hallucination and Visual Ambiguity. We further analyze how often the challenges illustrated in Fig. 1 occur in real OV-HOI evaluation scenarios. Specifically, we evaluate Qwen-VL-Max on the full HICO-DET test set containing 9,546 images and regard a prediction as hallucinated if the model fails to cover all ground-truth action and object categories in its output set. In addition, we randomly sample 600 images from HICO-DET and manually categorize whether each image contains significant object occlusion or severe viewpoint limitation.

As shown in Tab. 8, hallucination occurs in 38.7% of the full test set, indicating that native MLLM reasoning frequently fails to fully ground HOI predictions in visual evidence. Meanwhile, 12.3% of the sampled images contain significant object occlusion, and 19.7% suffer from severe viewpoint

limitations. These observations confirm that both cross-modal hallucination and visual ambiguity are prevalent in OV-HOI, motivating the need for tool-augmented reasoning and generative imagination.

Table 8: Statistics of hallucination and visual ambiguity on HICO-DET.

Metric	Num. Images	Rate
MLLM Hallucination	9,546	38.7%
Object Occlusion	600	12.3%
Viewpoint Limitation	600	19.7%

Error Analysis. To better understand the remaining failure cases of ImagineAgent, we randomly sample 600 examples from HICO-DET and manually categorize the major error types. As reported in Tab. 9, incomplete HOI matching accounts for the largest proportion of errors, reaching 39.7%. This indicates that the model still struggles to exhaustively identify all valid human-object interaction pairs, especially in crowded or visually complex scenes. Object classification and action classification errors account for 25.5% and 23.2%, respectively, suggesting that fine-grained semantic discrimination remains challenging. Grounding errors account for 20.8%, reflecting the difficulty of accurately localizing interacting humans and objects under occlusion or small-object scenarios. These findings reveal that although ImagineAgent significantly improves OV-HOI reasoning, its remaining errors are mainly caused by incomplete interaction coverage and fine-grained semantic confusion.

Table 9: Error analysis on 600 randomly sampled HICO-DET examples.

Error Type	Proportion
Action classification error	23.2%
Grounding error	20.8%
Object classification error	25.5%
Incomplete HOI matching	39.7%

A.3 Data Construction Details (Reasoning Chains)

This section details the reproducible pipeline used to synthesize structured reasoning chains for HOI comprehension on HICO-DET. Our construction is implemented as a **two-turn agentic prompting** process with optional auxiliary images and optional textual knowledge, followed by trajectory packaging into training-ready JSON.

A.3.1 Generation Prompts

Prompt design (two-turn decomposition). We generate reasoning chains using a two-turn chat template with **Qwen3-VL-32B-Instruct** (served by vLLM). Turn-1 performs *object set selection* and *tool selection*, and Turn-2 performs *exhaustive HOI enumeration* under strict constraints and outputs HOI triplets with bounding boxes.

Turn-1 (Perception & Tool Decision). Given the input image, the model is instructed to: (i) detect all visible objects from the predefined HICO object list. (ii) decide which tools to use among {*image_description*, *action_description*, *outpaint*, *viewpoint_transform*, *image_crop*}. The output format is enforced as: *object_1, [ox1,oy1,ox2,oy2], object_2, [ox1,oy1,ox2,oy2], ... ; tool_1, tool_2, ...* wrapped by <*think*> and <*answer*> tags. For training-data synthesis only, we additionally provide *ground-truth objects extracted from the HOI annotation string* inside the prompt to bias the chain toward GT-aligned decisions (used for CoT supervision rather than evaluation).

Turn-2 (Constrained HOI Chain-of-Thought). Turn-2 takes the original image (always) plus up to two optional auxiliary images (outpaint and/or viewpoint transform) depending on Turn-1 tool decisions. It also injects: (i) **DETECTED OBJECTS** from Turn-1. (ii) a per-object **VALID ACTIONS** constraint list derived from HICO interactions (i.e., allowed verbs for each object class). (iii) optional textual evidence, including an image-level description and/or action definitions if the corresponding tools were selected.

Inside <*think*>, the model must follow a fixed structure: (1) entity listing with person/object boxes, (2) exhaustive person–object pairing, (3) per-pair verb feasibility checks restricted to valid verbs, (4)

final review and compilation. Inside <*answer*>, the model must emit HOIs as a comma-separated list of records: *idx: action, object, [px1,py1,px2,py2], [ox1,oy1,ox2,oy2]*. For synthesis, we include the ground-truth HOI string in the prompt as a reference target (the model is instructed to build a reasoning bridge that leads to the GT).

Tool library and how it is realized in our code. Unlike an online tool-execution agent, our synthesis pipeline implements tool augmentation *by conditional evidence injection*:

- ***image_description***: if selected in Turn-1, we load a precomputed image-level analysis from *train_all_description.json* and append it into the Turn-2 prompt.
- ***action_description***: if selected, we load per-action textual definitions from *hico_actions_rephase_20.json*. To avoid leaking irrelevant knowledge, we only include definitions for actions appearing in the valid-action constraint set of the current sample.
- ***outpaint* / *viewpoint_transform***: We explicitly instruct the model that auxiliary images are *only* for disambiguating actions, and that all bounding box coordinates must be taken from the original image.
- ***image_crop***: if selected, we extract high-resolution sub-regions centered on specific human-object pairs from the original image and inject them as auxiliary visual evidence into the Turn-2 prompt.

Independent rollouts and decoding. In the released synthesis code, we generate **one** trajectory per image-query pair with stochastic decoding: temperature 0.8, and max generation length 4096 tokens for both turns.

Randomness control. The synthesis code relies on the underlying sampler randomness from vLLM at the process level and does not explicitly set per-sample random seeds. All other decoding hyperparameters are kept fixed across samples to ensure reproducibility given the same runtime environment and model weights.

A.3.2 Expert Validation and Filtering

Trajectory packaging (what is saved). For each HICO training instance, we store: *detected_objects*, *selected_tools*, *first_turn_output*, and *second_turn_output*. We also store the exact textual prompt strings for both turns to enable prompt-level reproduction.

Expert validation. In our overall framework, we employ **Qwen-VL-Max** as an expert validator to score and filter synthesized chains. Concretely, the validator checks whether the generated chain is (i) visually grounded, (ii) logically consistent with the constrained action set and the claimed evidence, and (iii) format-compliant with the required <*think*>/<*answer*> schema. Chains that fail validation (e.g., hallucinated entities/interactions, invalid format, or ungrounded claims) are removed before forming the final SFT/RL corpora.

Hard constraints used in construction (rule-based). Even before expert scoring, we apply strict rule constraints embedded in the prompts and enforced by parsing:

- **Closed-set object constraint:** all objects must come from the HICO object vocabulary.
- **Per-object verb constraint:** verbs must come from the HICO interaction-derived mapping *object_to_verbs*.
- **Bounding box constraint:** all coordinates must correspond to the *original image*, and auxiliary images are prohibited from providing coordinates.
- **Format constraint:** outputs must contain parsable <*answer*> blocks, otherwise the sample is marked invalid.

A.3.3 Examples of Synthesized Reasoning Chains

Each synthesized sample contains: (i) the original image (and optional auxiliary images), (ii) Turn-1 object/tool decisions, (iii) Turn-2 structured reasoning in <*think*> and HOI triplets in <*answer*>.

A.4 More Case Studies

As shown in Fig. 5, our method utilizes outpainting, view transformation, and cropping to respectively enrich context, resolve occlusion, and focus on critical details. As shown in Fig. 6 and Fig. 7, we present more case studies between Qwen2.5-VL-7B and our model on the HICO-DET dataset.

These examples highlight the baseline model's susceptibility to cross-modal hallucination, where it generates plausible but incorrect interactions. In contrast, ImagineAgent employs a structured cognitive process, systematically evaluating entity pairs against visual evidence enriched by tools like Outpaint. This allows it to discard flawed hypotheses and accurately identify a comprehensive set of correct interactions, demonstrating superior grounded reasoning.

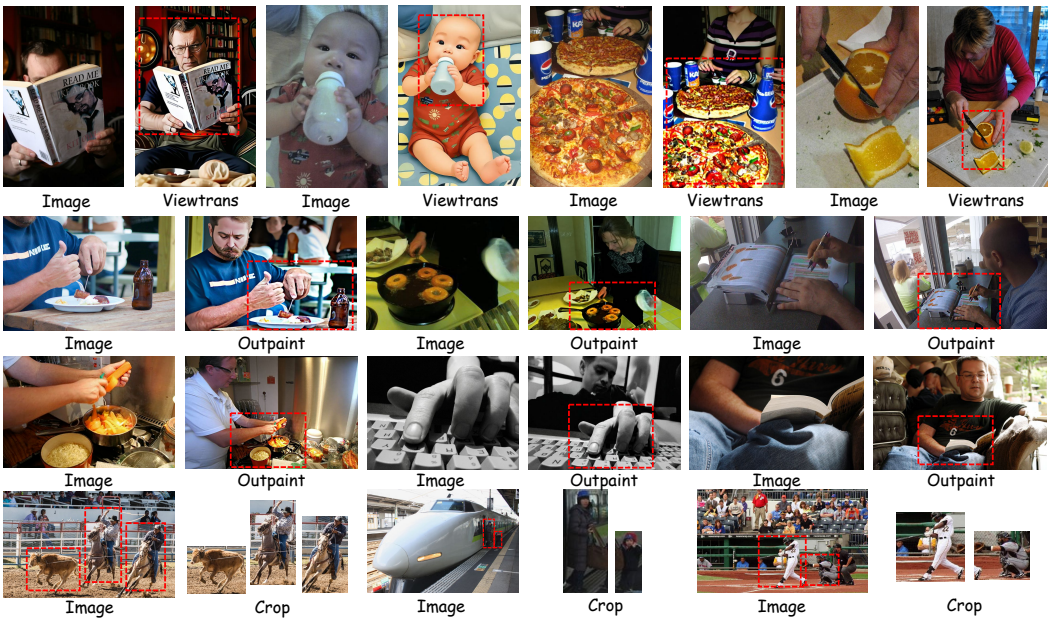


Figure 5: **Qualitative results of ImagineAgent's generative imagination.** Respectively, **Image Outpaint** is used to extend the scene's context, providing a more holistic view of the environment that aids in understanding the overall activity. To overcome severe occlusion, the agent employs **Image View Transformation** to synthesize a novel perspective of the interaction. Furthermore, **Image Crop** is utilized for a focused analysis of fine-grained interactions, scrutinizing critical details for detection.

B Tool Library Details

This section provides a detailed specification of the multimodal tools integrated into our ImagineAgent framework. These tools are designed to solve ambiguities encountered during HOI prediction and are categorized into visual refinement, generative imagination, and online retrieval-augmentation tools.

Table 10: Implementation details of the tool library.

Tool	Input	Output	Failure Handling
Image Crop	predicted human/object boxes	cropped image patch	ignored if boxes are invalid
Action Explanation	candidate action phrase	textual action definition	fallback to original query
Scene Explanation	image and scene query	textual scene context	fallback to original query
Viewpoint Transform	image and editing prompt	transformed image	used only as auxiliary evidence
Outpaint	image and editing prompt	expanded image	used only as auxiliary evidence

B.1 Visual Refinement Tools

To ensure that fine-grained interaction details are not overlooked in complex scenes, our framework incorporates an **Image Crop** tool (T_c). Its primary function is to allow the agent to isolate and scrutinize high-resolution sub-regions of the input image. This tool is dynamically invoked when the initial perception suggests a potential interaction, but the visual evidence is too small or subtle to be conclusively identified from a global view. By extracting a focused patch centered on a candidate human-object pair, the agent can perform a more detailed analysis of subtle cues, such as hand-object contact or the specific orientation of a tool. This capability is critical for accurately distinguishing

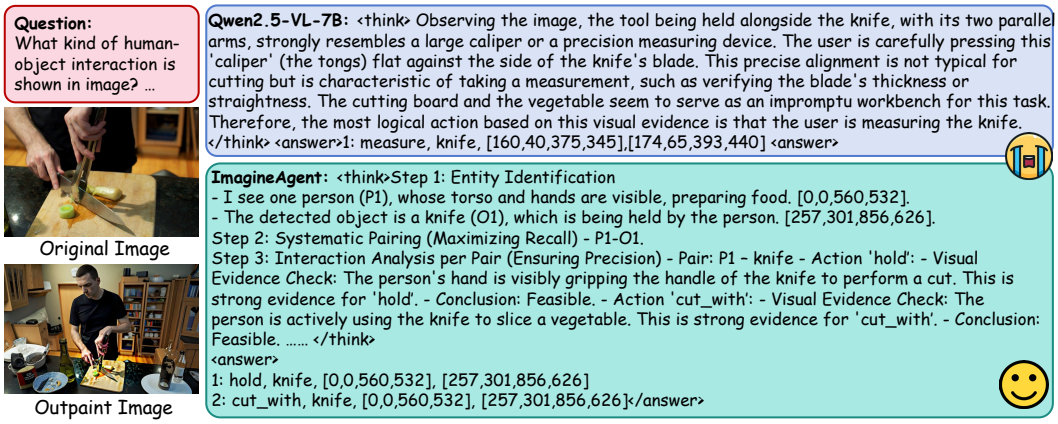


Figure 6: **Case Study between Qwen2.5-VL-7B and our method.** Our method can accurately identify comprehensive HOI predictions, while Qwen2.5-VL-7B misclassifies the interaction labels.

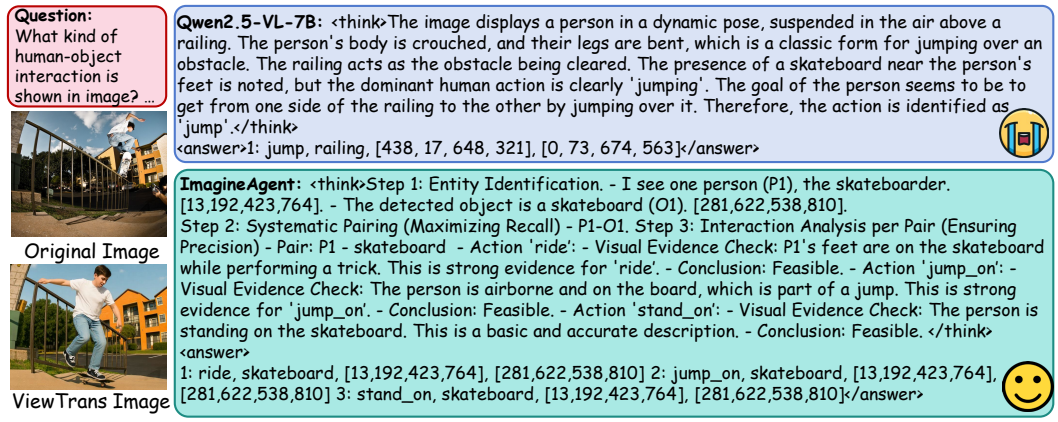


Figure 7: **Case Study between Qwen2.5-VL-7B and our method.** Our method can accurately identify comprehensive HOI predictions, while Qwen2.5-VL-7B misclassifies the interaction labels.

between similar interactions and directly mitigates failures arising from the perceptual-cognitive gap, where global context is insufficient for precise grounding.

B.2 Online Retrieval-Augmentation Tools

To mitigate cross-modal hallucinations, we leverage the **Qwen API** as an online Retrieval-Augmented Generation (RAG) tool. This component anchors the agent's reasoning in external, factual knowledge, preventing over-reliance on learned textual priors. It serves two key functions: **Action Explanation** (T_a) and **Scene Explanation** (T_s). When faced with semantically similar verbs, the agent can invoke Action Explanation to retrieve precise definitions and distinguishing characteristics, thus enabling a more nuanced classification. Concurrently, the Scene Explanation function provides high-level contextual descriptions of the entire image, which helps ground the agent's understanding of the overall environment and the plausibility of certain interactions within it. By dynamically injecting this external knowledge, the framework ensures that predictions are semantically consistent with both the visual evidence and real-world knowledge.

B.3 Generative Imagination Tools

To address challenges of occlusion-induced ambiguity, our agent is endowed with generative imagination capabilities through the **BAGEL** model [10]. This toolset operates in two distinct modes: **Outpaint** (T_o) and **Viewpoint Transform** (T_v). The Outpaint function is employed when key contextual elements are cut off by the image frame, allowing the agent to reconstruct a more complete scene

and better infer the overall activity. The Viewpoint Transform function is selected when an object or part of an interaction is occluded from the current perspective. By synthesizing a plausible alternative view, the agent can "see" behind obstructions, effectively filling in missing visual information. This process provides the agent with an imagined reality upon which to base its reasoning, transforming an incomplete observation into a more comprehensive one and enabling robust decision-making under visual uncertainty. Moreover, to control for potential generative artifacts and ensure the reliability of the agent's reasoning, we embed a strict protocol within its prompt, governing the use of supplementary images. The complete instructions are detailed in Fig. 8.

2.3. SUPPLEMENTARY IMAGE DESCRIPTIONS
- Primary Goal: Use SUPPLEMENTARY images only when the ORIGINAL image is ambiguous in identifying the action.
- Viewpoint Image: Use only if the original has a blocked view and the viewpoint image provides a clear, alternative angle of the same scene.
- Outpaint Image: Use only if the original cuts off an interaction and the outpaint image provides a valid, logical completion.
- NEVER use extra images for bounding boxes. All coordinates must come from the original image.
- IGNORE extra images if the action is already clear or if they are poor quality (distorted, irrelevant).
- Final Judgment: If an extra image creates more confusion than it solves, discard it. When in doubt, the original image wins.

Figure 8: Detailed prompt of supplementary image descriptions.

C Discussion on Generative Hallucination

ImagineAgent leverages generative tools such as Viewpoint Transform and Outpainting to recover occluded or ambiguous interactions. While these tools improve the semantic clarity of interactions, they inherently carry a risk of producing hallucinated visual content. It is therefore important to contextualize this risk within our task objective: OV-HOI detection fundamentally aims to recognize semantic ⟨human, object, action⟩ triplets rather than to reconstruct precise physical grounding or millimeter-level contact points. Consequently, as long as the core interaction semantics are preserved, positional or geometric deviations in generated views do not compromise recognition accuracy.

This rationale is illustrated in Fig. 7. The original image suffers from severe occlusion of the skateboard, which causes the baseline Qwen2.5-VL-7B to hallucinate incorrect actions. By contrast, the ViewTrans output reconstructs a clearer perspective of the person-skateboard interaction. For HOI detection, the decisive evidence lies in identifying the "skateboarding" relational pattern rather than verifying exact foot-to-surface coordinates. Additional results are provided in Fig. 4 and Fig. 6.

To quantify the impact of generative hallucinations, we conducted two complementary analyses:

1. **Empirical Evaluation of Hallucination Frequency.** Using 9,546 HICO-DET test images, we define hallucination conservatively: a prediction is considered hallucinated if the model's output set of actions and objects does not fully cover the ground-truth categories. Under this metric, 38.7% of images exhibit some form of hallucination, with the majority (29.4%) arising from incomplete interaction coverage and only 13.9% due to incorrect predictions.
2. **Semantic Fidelity Verification of Generated Views.** We randomly sampled 600 generated images and evaluated them using both human annotators and an automated verification pipeline based on Qwen-VL-Max. Results indicate a human validation accuracy of 86.7% and a model-based validation accuracy of 88.0%, confirming that the majority of generated views maintain high semantic fidelity and provide actionable evidence for interaction recognition.

Our pipeline further mitigates hallucination risk through three key design choices:

- *Interaction Filtering*: The agent first detects candidate objects and pre-filters only those with clear interaction potential, ensuring distant or non-interacting objects do not enter the reasoning space.
- *Targeted Generative Reasoning*: Generative tools are applied exclusively to pre-filtered interaction pairs, serving as auxiliary signals to recover occluded or ambiguous cues without influencing unrelated objects.
- *Reward-Guided Tool Usage*: The agent's composite reward credits tool usage only when it improves HOI predictions. Invalid or unnecessary tool invocations are penalized, preventing over-reliance on potentially hallucinated content.

Despite these safeguards, we acknowledge that generative bias remains an intrinsic limitation of diffusion-based image manipulation. Extreme occlusions, crowded scenes, or severe motion blur may

You are an expert in detecting objects in images for Human-Object Interaction analysis.

TASK:
1. Detect ALL objects from the PREDEFINED OBJECTS that are present.
2. Decide whether it is necessary to call any external tools to improve the final HOI predictions.

PREDEFINED OBJECTS:
{OBJECTS_INFO}

Available tools:
1. "image_description" - If the image is unclear, blurry, or has complex lighting conditions, request detailed image analysis to better understand the scene
2. "action_description" - Get detailed definitions of action categories to distinguish similar actions (e.g., "hold" vs "carry", "cut" vs "cut_with")
3. "outpaint" - Select this if the frame cuts off the person, showing just a part of them (e.g., hands in the kitchen, feet on a skateboard) instead of the complete scene.
4. "viewpoint_transform" - Select this if the person is occluded or indistinct from the current angle.
5. "image_crop" – Select this if specific regions of interest need closer examination to detect fine-grained interactions.

{OBJECT_DESCRIPTIONS}

Instructions:
- For object detection: Detect ALL objects from the PREDEFINED LIST that are VISIBLE in the image, regardless of whether they are involved in interactions
- IMPORTANT: Only detect "person" as an OBJECT (not as subject) if there are clear interactions where a person is the target/recipient of an action. For example:
  * CORRECT: Detect "person" if someone is hugging another person, or touching another person
  * INCORRECT: Do NOT detect "person" if someone is holding an object, sitting on a chair, cooking, using a computer, etc. (person is the subject/doer, not the object/target)
- For tool selection: Carefully evaluate which tools would provide the most benefit for accurate HOI detection
- Detect ONLY objects from the predefined list
- Return results in the specified format, no extra text

Please think step by step about the image content, your detection and tool selection.
Provide your detailed reasoning between the <think> </think> tags, and then give your final results in the <answer> </answer> tags.

Output Format:
<think>step-by-step reasoning process</think>
<answer>
Format: object_1, [ox1,oy1,ox2,oy2], object_2, [ox1,oy1,ox2,oy2], object_3, [ox1,oy1,ox2,oy2], ... ; tool_1, tool_2, tool_3, ...
Examples:
boat, [299,235,385,395], car, [320,306,359,349], person, [148,345,376,414]; object_description, action_description
bench, [39,239,192,452], horse, [79,25,571,411]; object_description, outpaint
chair, [320,306,359,349], table, [148,345,376,414]; object_description, action_description, viewpoint_transform
</answer>

Figure 9: Detailed prompt template in the first round.

still yield unreliable synthesized cues. Nevertheless, empirical evidence indicates that the semantic recovery provided by generative modules consistently outweighs geometric deviations. This is further supported by the ablation study in Tab. 4 , where disabling the Viewpoint Transform module leads to a consistent performance drop, confirming its effectiveness in resolving visual ambiguities.

In conclusion, while hallucination risk cannot be completely eliminated, our pipeline design, empirical validation, and reward-guided constraints collectively ensure that ImagineAgent relies predominantly on semantically meaningful evidence. This establishes both the robustness and reliability of generative-assisted perception for OV-HOI detection, while transparently acknowledging the remaining limitations for future work.

D Prompts

To generate the high-quality, structured reasoning chains essential for our training pipeline, we employ a meticulously designed two-turn prompting strategy. As shown in Fig. 9, in the first turn, the model is tasked with initial perception and strategic planning. It is prompted to identify all relevant objects within the image from a predefined vocabulary and to select appropriate tools from a library including image cropping, online retrieval for scene and action explanations, and generative modeling including outpainting and viewpoint transformation.

As shown in Fig. 10, the second turn focuses on detailed cognitive reasoning and evidence integration. It takes the set of detected objects and any augmented visual or textual data from the deployed tools to generate a constrained, step-by-step analysis. This involves systematically pairing all human and object instances, evaluating a predefined set of valid verbs for each pair, and grounding each potential interaction in direct visual evidence.

You are a world-class AI expert in Human-Object Interaction (HOI) detection.
Your primary goal is to achieve both **maximum recall** (finding every possible interaction) and **perfect precision** (ensuring every detected interaction is accurate and valid).

 # 1. CONTEXT & DATA
{IMAGE}
1.1. Detected Objects
Here is the definitive list of objects you must consider.
DETECTED OBJECTS: {DETECTED_OBJECTS}
1.2. Valid Actions per Object
These are the only actions you are allowed to use for each object class.
VALID ACTIONS: {OBJECT_VERB_CONSTRAINTS}
1.3. Additional Information (Optional)
{ACTION_DESCRIPTIONS}
{IMAGE_DESCRIPTIONS}
{DETAILED_IMAGE_INSTRUCTIONS}

 # 2. CORE TASK & RULES
Your mission is to exhaustively identify and list ALL valid human-object interactions present in the ORIGINAL image, following these strict rules:
1. **Be Comprehensive (For Recall)**: You must systematically check every person against every object from the `DETECTED OBJECTS` list. Do not miss any potential interaction pair.
2. **Strictly Adhere to Lists (For Precision):**
   - You MUST ONLY use objects from the `DETECTED OBJECTS` list.
   - You MUST ONLY use actions from the `VALID ACTIONS` list for the corresponding object.
3. **Ground in Visual Evidence (For Precision):** Every reported interaction must be directly supported by clear visual cues in the image (e.g., hand contact, body posture, gaze direction, spatial proximity).
4. **Precise Formatting:**
   - Use precise pixel coordinates for all bounding boxes.
   - Format each finding as: `number: action, object, [person_x1,person_y1,person_x2,person_y2], [object_x1,object_y1,object_x2,object_y2]`
   - Do not add any extra text or explanations outside the designated tags.

 # 3. YOUR THOUGHT PROCESS (Mandatory)
You must follow this structured reasoning process inside the <think> tags.
<think>
Step 1: Entity Identification
- I will first identify and list all persons (P1, P2, ...) and all relevant objects from the `DETECTED OBJECTS` list (O1, O2, ...) with their bounding boxes.
Step 2: Systematic Pairing (Maximizing Recall)
- I will now create a list of all possible Person-Object pairs for analysis (e.g., P1-O1, P1-O2, P2-O1, ...). This ensures an exhaustive search.
Step 3: Interaction Analysis per Pair (Ensuring Precision)
- For each pair, I will analyze all `VALID ACTIONS` for that specific object.
- For each action, I will perform a detailed visual evidence check.
- Example for a single pair (P1-O1 where O1 is a 'cup'):
  - **Pair:** P1 - O1 (cup)
  - **Action 'hold':**
    - **Visual Evidence Check:** I see P1's hand is wrapped around the cup O1.
    - **Conclusion:** Feasible.
  - **Action 'drink_from':**
    - **Visual Evidence Check:** The cup O1 is near P1's mouth, and P1's head is tilted back. This confirms the 'drink_from' action.
    - **Conclusion:** Feasible.
  - **Action 'sit_on':**
    - **Visual Evidence Check:** A person cannot sit on a cup. This is physically impossible and not supported by the image.
    - **Conclusion:** Not Feasible.
Step 4: Final Compilation and Review
- Based on the 'Feasible' conclusions from Step 3, I will compile the final list of interactions.
- I will do a final review to ensure I have not missed any interactions and that every reported interaction strictly follows all rules and is backed by evidence. </think>
<answer>
Format: number: action, object, [px1,py1,px2,py2], [ox1,oy1,ox2,oy2]
Example: 1: sit_on, bench, [320,306,359,349], [148,345,376,414], 2: hold, horse, [226,18,340,210], [174,65,393,440]
</answer>

Figure 10: Detailed prompt template in the second round.

E Limitation

Despite its effectiveness, ImagineAgent still has several limitations. First, our framework relies on external tools, including retrieval, image cropping, and generative image editing, which inevitably introduces additional inference cost compared with a single-pass MLLM baseline. Although the tool-usage reward encourages effective tool invocation, reducing unnecessary tool calls and further improving inference efficiency remain important directions for future work. Second, our current experiments are conducted on HICO-DET and SWIG-HOI, and the proposed training data construction process is tailored to image-based OV-HOI detection. Extending the framework to more complex scenarios, such as video-based HOI, embodied interaction understanding, and real-time robotic perception, may require additional temporal reasoning modules and more efficient tool scheduling strategies. Finally, our reward design relies on rule-based matching with predefined thresholds for semantic similarity and spatial IoU, more adaptive reward formulations may further improve robustness across datasets with different annotation styles.

Licenses

All existing datasets and models used in this paper are credited to their original creators. We follow the licenses and terms of use specified in the corresponding original papers, official repositories, or official model cards. These assets are used only for research purposes, including OV-HOI evaluation, reasoning-chain construction, semantic matching, and auxiliary tool-augmented inference.

Impact Statement

This work aims to improve open-vocabulary human-object interaction understanding, which may benefit assistive perception, robotics, and embodied AI systems. However, HOI recognition also has potential risks in surveillance or activity-monitoring applications, especially when used without consent or human oversight. In addition, the generative tools used in our framework, such as Outpaint and Viewpoint Transform, may have dual-use concerns if misused to synthesize misleading visual evidence. Our method is intended for research use, and deployment in sensitive scenarios should follow appropriate ethical, legal, and safety constraints.

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: We emphasize the contributions and scope in the Introduction.

   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: The limitation of the proposed algorithm has been discussed in the supplementary material.

   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 is no theoretical result.

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: We provide comprehensive implementation details both in main paper and in supplementary material.

   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: As we promised, 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 detailed, are provided in 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 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 implementation details of main paper and supplementary material.

   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 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: We provide a substantive impact statement in the supplementary material, discussing both positive applications of OV-HOI understanding, such as assistive perception and robotics, and potential risks, including surveillance misuse, erroneous human-object interaction predictions, and dual-use concerns introduced by generative tools such as Outpaint and Viewpoint Transform.

    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: The proposed method uses pre-trained models. This proposed methods is safe under the safeguards of adopted pre-trained models.

    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 credit the original creators of all existing datasets, models, and tools used in this work, including HICO-DET, SWIG-HOI, Qwen-VL/Qwen3-VL, CLIP, and BAGEL. We follow the licenses and terms of use specified in their original papers, official repositories, or official model cards, and use these assets only for research purposes.

    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: There is no new assets released in this work.

    Justification: This work introduces *hicodet-6K*, a curated reasoning dataset for OV-HOI supervised fine-tuning. We document its construction pipeline, filtering and validation procedure, and data usage in Appendix. The current submission does not publicly release the dataset for anonymity, and we will clarify the release status upon publication.

    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: [N/A]

    Justification: LLM does not impact the core methodology, scientific rigorousness, or originality of the research.

    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.