HARNESSING MULTIMODAL LARGE LANGUAGE MODELS FOR TRAINING-FREE HUMAN-OBJECT INTERACTION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Human-object interaction (HOI) detection aims to localize human-object pairs and recognize their interactions. Traditional supervised methods perform strongly but rely on task-specific training. Recent multimodal large language models (MLLMs) offer a promising route to training-free HOI detection through their broad visual-semantic knowledge and versatile perceptual and reasoning capabilities. However, existing approaches largely invoke these capabilities through loosely coordinated inference stages. This fragmented execution restricts the role of interaction hypotheses in guiding visual exploration, leaving key participants overlooked and local ambiguities unresolved. Furthermore, propagating early semantic assumptions through subsequent visual grounding and relation prediction induces self-reinforcing semantic circularity. To resolve these challenges, we propose HarnessHOI, a training-free framework that transforms passive MLLM inference into an active interaction-centric harness. Specifically, we introduce an interaction-guided perception mechanism that projects emerging interaction hypotheses back into the visual space to discover missing participants and refine ambiguous evidence through targeted observation. Furthermore, a relation-agnostic geometric adjudication module reconciles multi-source evidence to establish a unified spatial basis for grounded interaction reasoning across multiple actions and semantic roles. Extensive experiments on HICO-DET and V-COCO demonstrate that HarnessHOI achieves state-of-the-art performance among training-free methods, confirming the effectiveness of the proposed harness for complex interaction understanding. Code will be released upon publication.

1 INTRODUCTION

Human-object interaction (HOI) detection aims to localize human-object pairs and recognize the interactions between them, typically represented as $\langle \text{human}, \text{action}, \text{object} \rangle$ triplets (Chao et al., 2018; Gupta & Malik, 2015). As a foundational task for scene understanding, it supports applications ranging from visual surveillance (Dogariu et al., 2020; Li et al., 2024) to embodied AI (Mascaro et al., 2023; Bahl et al., 2023). Traditional supervised methods have achieved strong performance by learning task-specific interaction representations and prediction functions from HOI annotations (Gkioxari et al., 2018; Liao et al., 2022). However, their reliance on extensive annotations and dedicated training limits their scalability to the compositional and long-tailed interaction space encountered in real-world scenes (Hou et al., 2020; 2021).

Recent vision-language models and multimodal large language models (MLLMs) have substantially expanded their ability to visual understanding and visual-semantic knowledge (Radford et al., 2021; Bai et al., 2025). Initial approaches integrate this knowledge into HOI detectors through feature alignment, prompt adaptation, or supervised optimization (Lei et al., 2024b;a). Recent work further develops part-aware interaction representations (Tonini et al., 2026) and unifies HOI detection with generation (Yang et al., 2026). Although effective, these approaches still require specific training, reintroducing the annotation and optimization burden that foundation models are expected to alleviate. Training-free paradigms therefore directly exploit frozen pretrained models for interaction grounding and prediction. Retrieval-based approaches score candidate interactions against visual-semantic knowledge (Tonini et al., 2025), while the variant formulates recognition as candidate-conditioned semantic assessment (Xuan et al., 2026). Recent agentic methods further revisit candidate interactions and recover visual entities through multi-stage semantic reasoning (Lei et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

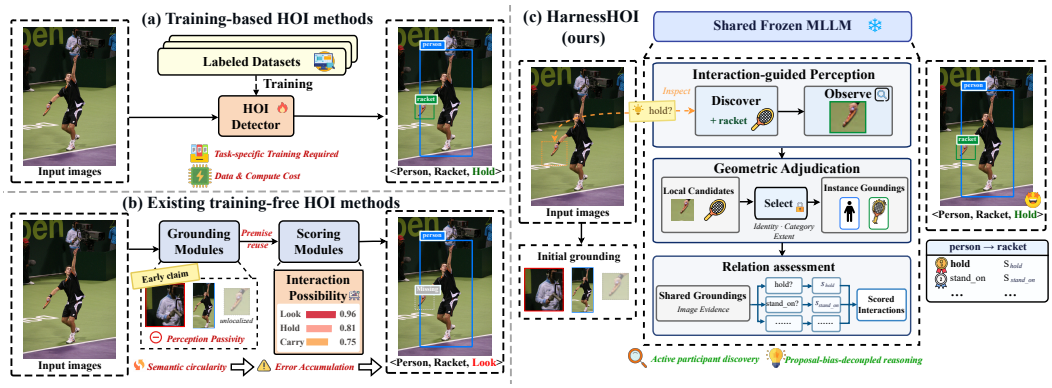

Figure 1: Comparison of HOI methods. (a) Training-based HOI methods learn from annotated interactions but require task-specific training. (b) Existing training-free methods rely on loosely coordinated perception and reasoning modules, resulting in passive visual exploration and semantic circularity. (c) HarnessHOI (ours) introduces interaction-centric harness to actively search and select visual evidence, and establish reliable grounding for final HOI predictions.

2026). Despite this progress, existing approaches leave semantic reasoning and visual perception through a separate process, leaving perception passive to the emerging interaction hypotheses.

Human-object relations are fundamentally constrained by both relational semantics and fine-grained visual evidence distributed across the scene (Wang et al., 2019; Wan et al., 2019; Zhang et al., 2023).An emerging interaction hypothesis provides cues to recover overlooked participants or resolve local visual ambiguity. Conversely, grounded visual evidence provides necessary spatial constraints to verify candidate relations. Existing methods fail to regulate this reciprocal dependency, resulting in two coupled limitations. First, visual perception remains static when emerging interaction hypotheses are not systematically projected back into the image space to guide observation. This creates *perceptual passivity*, leaving missing entities and ambiguous regions unresolved. Second, directly propagating semantic proposals through visual grounding and subsequent relation evaluation induces *semantic circularity*, where initial semantic assumptions bias the visual interpretation and produce self-reinforcing predictions and spurious relational judgments. The central challenge is therefore to harness the complementary semantic and perceptual capabilities of MLLMs within a unified framework, enabling interaction hypotheses to actively guide visual exploration while grounding the resulting judgments in instance-level visual evidence.

To address these limitations, we propose **HarnessHOI**, a training-free framework that transforms MLLM interpretation into an active interaction-centric harness paradigm(Figure 1). To overcome perceptual passivity, HarnessHOI incorporates *interaction-guided perception* that projects emerging hypotheses back into image space, actively discovering interaction participants and requesting targeted observations for ambiguous evidence. To mitigate semantic circularity, *relation-agnostic geometric adjudication* reconciles the resulting localization candidates under persistent instance identities, establishing a unified visual grounding according to participant evidence. Grounded interaction reasoning then assesses multiple actions and semantic roles using these shared groundings and the acquired visual evidence. Together, these mechanisms establish an interaction-centric harness that coordinates interaction-guided evidence acquisition with relation-agnostic visual grounding, enabling coherent and reliable training-free HOI prediction with frozen MLLMs.

Our main contributions are summarized as follows:

- We present **HarnessHOI**, a training-free framework that harnesses complementary perceptual and semantic capabilities of frozen MLLMs for HOI detection, moving beyond isolated semantic scoring or loosely coupled multi-stage reasoning.
- We introduce interaction-guided perception, relation-agnostic geometric adjudication, and grounded interaction reasoning to connect hypothesis-driven visual exploration with shared participant groundings and evidence-grounded relation assessment, addressing perceptual passivity and semantic circularity.
- Extensive experiments on HICO-DET and V-COCO demonstrate that HarnessHOI achieves state-of-the-art performance among training-free methods, validating the effectiveness of the proposed interaction-centric harness for human-object interaction detection.

2 Related Work

2.1 Traditional Human-Object Interaction Detection

Traditional human-object interaction (HOI) detection frameworks primarily rely on fully supervised paradigms, which are broadly grouped into two-stage pipelines and single-stage set-prediction architectures (Zhang et al., 2021a; 2022b). Two-stage methods detect human and object proposals prior to classifying interaction predicates (Gao et al., 2018; Li et al., 2019; Zhang et al., 2021b; 2022a). Conversely, single-stage transformer architectures utilize learnable queries to directly decode relation triplets from global visual contexts (Kim et al., 2021; Zou et al., 2021; Chen et al., 2021; Tamura et al., 2021). To improve generalization across unseen combinations, recent work incorporates pretrained vision-language representations by aligning visual features with text embeddings (Yuan et al., 2022; 2023; Ning et al., 2023; Mao et al., 2023). Although these supervised methods achieve competitive performance on closed-world benchmarks, they remain inherently constrained by task-specific parameter optimization and heavy reliance on exhaustive triplet annotations, impeding their scalability to open-world scenarios.

2.2 Foundation Models for HOI Detection

The evolution of vision-language models and multimodal large language models offers expansive visual-semantic knowledge, prompting a shift toward parameter-efficient and training-free HOI detection. Early adaptation strategies align interaction representations with pretrained visual-semantic knowledge (Wu et al., 2023; Cao et al., 2023; Lei et al., 2025a). Recent methods learn part-aware representations (Tonini et al., 2026), distill MLLM knowledge into HOI detectors (Gao et al., 2025), or unify detection and generation (Yang et al., 2026). Supervised fine-tuning and reinforcement learning further adapt MLLMs for interaction reasoning (Wu et al., 2026b). Although effective, these approaches still require specific optimization. Training-free methods therefore directly exploit frozen foundation models for interaction prediction. Visual-semantic retrieval approaches calculate similarity scores between representations and text embeddings (Tonini et al., 2025). The variant formulates HOI detection as candidate-conditioned generation (Xuan et al., 2026). Recent agentic frameworks introduce multi-agent and multi-turn reasoning to propose and revisit candidate interactions prior to visual grounding (Wang et al., 2026; Lei et al., 2026). Despite these advancements, current training-free paradigms treat semantic reasoning and visual grounding as isolated steps. As a result, visual perception remains fundamentally passive to initial semantic hypotheses, leaving models susceptible to perceptual omission and semantic circularity.

2.3 MLLM Harnessing for Visual Reasoning

Active perception principles advocate dynamically adapting visual observation based on task demands rather than relying on static image features Bajcsy (1988). Recent vision-language research operationalizes active perception by incorporating selective region re-encoding or localized attention mechanisms during MLLM inference Zhang et al. (2025a); Shalam et al. (2026). In parallel, agentic workflows decompose complex visual tasks into sequential reasoning and grounding operations Lei et al. (2026). However, existing active harnessing strategies are primarily designed for free-form visual question answering or coarse region grounding, lacking effective and controllable mechanisms to manage complex relational understanding inherent to HOI detection. Applying generic workflows directly to HOI detection allows unconstrained semantic assumptions to distort visual search. HarnessHOI overcomes these limitations by unifying interaction-guided perception with relational reasoning, enabling a controlled harnessing paradigm that prevents semantic bias while maintaining rigorous spatial and relational grounding.

3 Method

3.1 Interaction-Centric HOI Harness

HarnessHOI introduces an interaction-centric harness in which semantic hypotheses guide participant acquisition and the acquired evidence supports grounded relation assessment (Figure 2). Shared participant identities connect these operations across multiple actions and semantic roles, allowing evidence acquired under one hypothesis to support further interaction judgments.

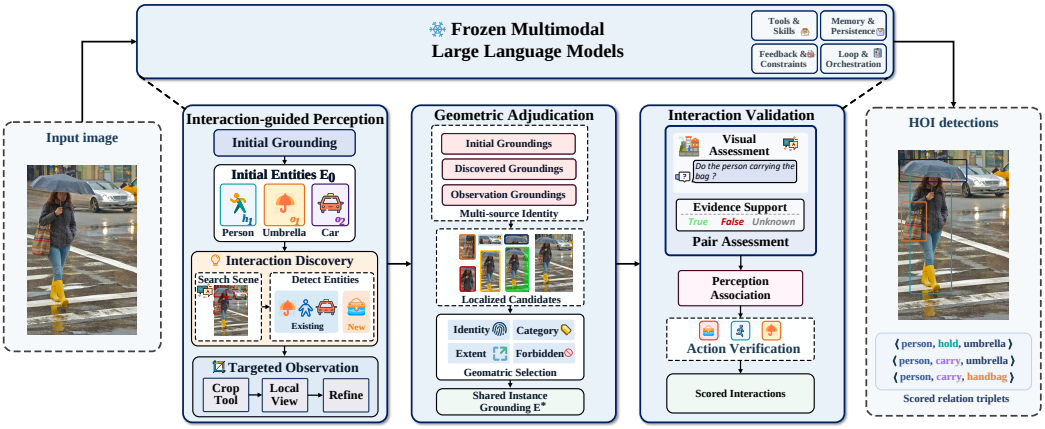

Figure 2: **Overview of HarnessHOI.** Interaction-guided perception uses interaction hypotheses to discover participants and acquire targeted observations. Geometrics adjudication reconciles multi-source identities to establish shared instance-level groundings. Interaction validation assesses evidence support, associates instances with semantic roles, and verifies candidate actions to produce scored HOI triplets. The frozen MLLM orchestrates all components using external tools, persistent memory, feedback constraints, and iterative loops.

Problem formulation. Given an image I and an interaction taxonomy $\mathcal{T}$, our goal is to predict a scored set of grounded interactions $\widehat{\mathcal{Y}} = \{(k, s(k))\}$. Each entity $e_i = (i, b_i, c_i) \in \mathcal{E}$ comprises an image-level identity, a normalized bounding box, and a category. An interaction $k = (h, a, o)$ associates a human $h \in \mathcal{H} \subseteq \mathcal{E}$ with an action a and a target $o \in \mathcal{E}$ under the semantics of $\mathcal{T}$. The score $s(k)$ expresses visual support for the interaction between the localized participants. We denote the MLLM by F_θ, with θ fixed throughout inference.

HarnessHOI overview. As illustrated in Figure 2, HarnessHOI consists of interaction-guided perception, relation-agnostic geometric adjudication, and grounded interaction reasoning. Specifically, we use the interaction hypotheses generated by F_θ to guide participant discovery and targeted observation, associating the resulting localization candidates and their source views with persistent instance identities. We then apply relation-agnostic geometric adjudication to select a consistent grounding for each participant according to its identity, category, and visible extent. Based on these shared groundings and the acquired image evidence, grounded interaction reasoning evaluates multiple actions and semantic roles through visual question answering. The same frozen MLLM performs all three stages, with its answer distributions determining geometric selection and interaction scores.

3.2 Interaction-Guided Perception

The semantics of an interaction provide cues to its participating objects and their spatial relation to the human. Interaction-guided perception exploits these cues to recover overlooked participants and obtain local observations for precise instance grounding.

Hypothesis-directed discovery. Starting from an initial entity set $\mathcal{E}_0$ localized by F_θ in I, we expand the participant set by querying the model about the visible interactions of each human in $\mathcal{H}_0 \subseteq \mathcal{E}_0$. Conditioned on the full image, initial entities, and interaction semantics in $\mathcal{T}$, F_θ jointly generates interaction hypotheses, localizations of newly discovered participants, and observation requests:

$$\big(\mathcal{P}_D, \Delta\mathcal{E}_D, \mathcal{R}\big) = D_\theta\big(I, \mathcal{E}_0, \mathcal{T}\big), \tag{1}$$

where D_θ denotes the discovery operation of F_θ. The positive interaction hypotheses define human-target associations $\mathcal{P}_D$ over the expanded participant set $\mathcal{E}_0 \cup \Delta\mathcal{E}_D$, where $\Delta\mathcal{E}_D$ contains participants newly localized in the full image. For hypotheses involving unlocalized targets or ambiguous instance boundaries, the observation requests $\mathcal{R}$ specify the relevant image regions and participants for further visual inspection.

Targeted visual observation. Targeted observation complements global interaction context with local visual evidence of participant identity and extent. Using high-resolution crops acquired from I according to $\mathcal{R}$, F_θ refines existing localizations and grounds additional participants, conditioned on the requested target description and local instance context. To integrate observations across image scales, we associate candidate localizations with persistent instance identities and express their boxes in image coordinates. The corresponding human-target associations form $\mathcal{P}_O$, connecting humans to the existing targets addressed by their observations and to newly localized participants.

Across initial localization, discovery, and local observation, the shared instance representation organizes visual evidence under N participant identities. Each identity i has a candidate localization set $\mathcal{B}_i = \{(b_{i,m}, c_{i,m}, v_{i,m})\}_{m=1}^{M_i}$. Here $b_{i,m}$ and $c_{i,m}$ specify the box and category of candidate m, and $v_{i,m}$ identifies its source view. The global image and acquired crops form the shared view set $\mathcal{V}$. The candidate sets $\{\mathcal{B}_i\}_{i=1}^{N}$ preserve alternative spatial interpretations of each participant for geometric adjudication, while $\mathcal{V}$ provides the global context and local detail for grounded interaction reasoning. The observation policy and coordinate handling are detailed in Appendix A.

3.3 Relation-Agnostic Geometric Adjudication

We introduce relation-agnostic geometric adjudication to establish a consistent localization for each participant across its interactions. For each instance, we compare the candidate localizations against the shared views $\mathcal{V}$ according to its identity, category, and complete visible extent. This instance-level criterion decouples geometric selection from support for individual actions or roles.

Let x denote a visual question together with its instance context and image evidence, and $\mathcal{L}_x$ the candidate answer set. We define the normalized support of F_θ for an answer $\ell \in \mathcal{L}_x$ as

$$q_\theta(\ell \mid x) = \frac{\exp\big(z_\ell(x)\big)}{\sum_{j \in \mathcal{L}_x} \exp\big(z_j(x)\big)}, \tag{2}$$

where $z_\ell(x)$ is the output logit of F_θ for answer label ℓ. For an instance with $M_i > 1$, x_i^{geo} contains the comparison question, candidate localizations $\mathcal{B}_i$, and shared views $\mathcal{V}$. Each answer label $\ell_{i,m}$ identifies a candidate m, and we select the candidate with the greatest visual support:

$$m_i^* = \underset{m=1,\dots,M_i}{\arg\max}\ q_\theta\big(\ell_{i,m} \mid x_i^{\mathrm{geo}}\big). \tag{3}$$

The selected candidate defines $(b_i^*, c_i^*) = (b_{i,m_i^*}, c_{i,m_i^*})$; identities with one candidate retain that localization. Collecting these selections yields the grounded entity set $\mathcal{E}^* = \{(i, b_i^*, c_i^*)\}_{i=1}^{N}$. All relations involving an entity in $\mathcal{E}^*$ use its selected box and category, so hypotheses concerning the same participant are evaluated with consistent spatial grounding.

3.4 Grounded Interaction Reasoning

Given the shared instance groundings $\mathcal{E}^*$, we formulate visual questions for candidate interactions using the action or role definitions in $\mathcal{T}$. The frozen MLLM F_θ evaluates each question with the global image and relevant local views from $\mathcal{V}$. Pair, action, agent, and binding questions use the answer set $\mathcal{L}_{\mathrm{rel}} = \{\mathrm{false}, \mathrm{true}, \mathrm{unknown}\}$, representing rejection, support, and unresolved visual evidence, respectively, with normalized support given by Equation 2.

Candidate pair construction. The relation candidates draw on both the associations established during perception and the visual evidence across the expanded participant set. Let $\mathcal{H}^* \subseteq \mathcal{E}^*$ denote the localized humans and $\mathcal{P} = \{(h, o) \in \mathcal{H}^* \times \mathcal{E}^* : h \neq o\}$ the human-entity pairs. For each pair, F_θ evaluates whether any positive action or physical role binding in $\mathcal{T}$ is visually supported, yielding $q_{\mathrm{F}}^{h,o}$, $q_{\mathrm{T}}^{h,o}$, and $q_{\mathrm{U}}^{h,o}$. These assessments extend the discovery and observation associations to define the candidate pair set:

$$\mathcal{K} = \mathcal{P}_D \cup \mathcal{P}_O \cup \big\{(h, o) \in \mathcal{P} \;\big|\; q_{\mathrm{F}}^{h,o} \leq \max\big(q_{\mathrm{T}}^{h,o}, q_{\mathrm{U}}^{h,o}\big)\big\}. \tag{4}$$

Multi-label action verification. To recognize multiple interactions involving the same participants, we evaluate each pair in $\mathcal{K}$ over its category-compatible action set. The action domain extends beyond the hypotheses used for participant discovery:

$$\mathcal{Q}_+ = \big\{(h, a, o) \;\big|\; (h, o) \in \mathcal{K},\ a \in \mathcal{A}^+(c_o^*)\big\}, \tag{5}$$

where $\mathcal{A}^+(c)$ is the set of positive actions associated with category c in $\mathcal{T}$. For each $(h, a, o) \in \mathcal{Q}_+$, $x^{\text{rel}}_{h,a,o}$ denotes the visual question context for verifying action a between the selected instances h, o. The interaction score is

$$s(h, a, o) = q_\theta\big(\text{true} \mid x^{\text{rel}}_{h,a,o}\big). \tag{6}$$

Separate normalization over $\mathcal{L}_{\text{rel}}$ for each action preserves the multi-label nature of HOI.

Compositional role verification. In semantic-role prediction, action occurrence and the visibility of its target are distinct aspects of the interaction. We extend the prediction to $k = (h, a, r, o)$, where r specifies the semantic role and $o \in \mathcal{E}^* \cup \{\varnothing\}$ accommodates physical and null targets. To distinguish target visibility from target identity, we decompose role prediction into action support, conditional target visibility, and instance-specific binding. For a given interaction, we suppress the arguments of the component supports and write

$$\begin{aligned} q_A &= q_\theta\big(\text{true} \mid x^{\text{agent}}_{h,a}\big), \\ q_\sigma &= q_\theta\big(\sigma \mid x^{\text{role}}_{h,a,r}\big), \\ q_B &= q_\theta\big(\text{true} \mid x^{\text{bind}}_{h,a,r,o}\big), \end{aligned} \tag{7}$$

where $x^{\text{agent}}_{h,a}$ denotes the context for verifying whether human h performs action a. Target visibility is assessed in $x^{\text{role}}_{h,a,r}$ conditional on action occurrence, and the binding of o to role r is assessed in $x^{\text{bind}}_{h,a,r,o}$ conditional on both action occurrence and target visibility. Action and binding questions use $\mathcal{L}_{\text{rel}}$, while the role-state answer σ ranges over $\mathcal{L}_{\text{role}} = \{\text{missing}, \text{visible}, \text{unknown}\}$, whose outcomes denote the absence of a visible, localizable target, the presence of such a target, and unresolved visual evidence, respectively.

Action and role-state factors cover every human in $\mathcal{H}^*$ and all corresponding actions and roles in $\mathcal{T}$, while physical binding factors are evaluated for pairs in $\mathcal{K}$. Writing q_M and q_V for the missing and visible supports, we obtain scores for person-level actions $k = (h, a)$, physical role targets, and null targets:

$$s_{\text{agent}} = q_A, \qquad s_{\text{physical}} = q_A\, q_V\, q_B, \qquad s_{\text{NULL}} = q_A\, q_M. \tag{8}$$

We apply the action-triplet formulation to HICO-DET and the semantic-role formulation to V-COCO. For HICO-DET, we additionally evaluate the no_interaction predicate over all pairs in $\mathcal{P}$, using its independently assessed true support as the detection score.

The resulting prediction set $\widehat{\mathcal{Y}}$ couples each interaction score $s(k)$ with the selected participant boxes from $\mathcal{E}^*$. Participants discovered and observations acquired under individual interaction hypotheses thus contribute to action and role verification across the expanded scene representation. Further details on inference, implementation, and question semantics are provided in Appendices A, B.3, and G.

4 Experiments

4.1 Experimental Setup

Benchmarks and metrics. We evaluate HarnessHOI on the complete HICO-DET (Chao et al., 2018) and V-COCO (Gupta & Malik, 2015) test sets, containing 9,658 and 4,946 images, respectively. HICO-DET reports Full, Rare, and Non-Rare mAP over 600 HOI categories under Default and Known Object evaluation. Open-vocabulary evaluation follows RF-UC, NF-UC, UO, and UV. V-COCO reports role AP over 25 semantic roles under scenarios S1 and S2. Both benchmarks require participant matching at IoU 0.5. Detailed metric definitions and evaluation protocols appear in Appendix B.

Implementation details. All perceptual and reasoning operations use a single frozen Qwen3.5-9B (Qwen Team, 2026), with at most two local observations per image. The harness receives the image and public interaction vocabulary. Prompts and inference policy are selected using the official training sets, following the protocol in Appendix B. Inference settings, implementation, and visual questions appear in Appendices A, B.3, and G.

Table 1: Comparison with state-of-the-art methods under regular settings using predicted boxes. HICO-DET reports Default and Known Object mAP (%); V-COCO reports role AP (%) under S1 and S2, with external entries following each source's role definitions. HarnessHOI uses the 25-role definition. Within each method group, the best results are bold; the strongest previous training-free results are underlined. Dashes denote unavailable entries.

		HICO-DET Default			HICO-DET Known Object			V-COCO	
Method	Model	Full	Rare	Non-Rare	Full	Rare	Non-Rare	S1	S2
Task-optimized methods									
QPIC (Tamura et al., 2021)	R101	29.90	23.92	31.69	32.38	26.06	34.27	58.30	60.70
UPT (Zhang et al., 2022a)	R101-DC5	32.62	28.62	33.81	36.08	31.41	37.47	61.30	67.10
GEN-VLKT (Liao et al., 2022)	R101/CLIP-B	34.95	31.18	36.08	38.22	34.36	39.37	63.58	65.93
PViC (Zhang et al., 2023)	Swin-L/H-DETR	44.32	44.61	44.24	47.81	48.38	47.64	64.10	70.20
ADA-CM (FT) (Lei et al., 2023)	R50/CLIP-L	38.40	37.52	38.66	–	–	–	58.57	63.97
HORP (Geng et al., 2025)	Swin-L/CLIP-B/32	47.53	46.81	47.74	51.24	50.78	**51.38**	68.90	71.10
HOI-IDiff (Hui et al., 2025)	DDETR	47.71	48.36	47.52	50.56	51.95	50.14	73.40	76.10
InCoM-Net (Seo et al., 2026)	R50/CLIP-L	43.96	45.61	43.46	45.59	45.92	45.49	**73.60**	75.40
InstructHOI (Luo et al., 2025)	R101/ViT-L	47.68	47.97	47.59	49.89	50.92	49.58	72.40	76.10
HOI-R1 (Chen et al., 2025)	Rex-Omni-3B	25.24	22.57	26.04	27.00	23.21	28.14	–	–
DA-HOI (FT, G-DINO) (Xuan et al., 2026)	G-DINO/Qwen2.5-VL-3B	46.21	47.54	45.81	–	–	–	70.40	72.50
LINK (Wu et al., 2026a)	Swin-L/CLIP-L	**49.06**	**53.63**	47.60	**51.34**	**56.29**	49.86	–	69.20
FineHOI (Tonini et al., 2026)	DETR/DINOtxt-L/16	39.13	39.91	38.89	–	–	–	–	66.50
UniHOI (Yang et al., 2026)	VQGAN/Llama3-8B	48.16	50.74	**48.12**	**51.34**	53.72	50.33	72.91	**77.45**
Training-free methods									
BLIP-2 analysis (Ai, 2024)	G-DINO/BLIP-2-H@364	19.71	24.07	18.41	–	–	–	–	–
ADA-CM (TF) (Lei et al., 2023)	R50/CLIP-B	25.19	27.24	24.58	–	–	–	39.09	43.93
DYSCO (Tonini et al., 2025)	R50/LongCLIP-L	29.00	33.63	27.62	–	–	–	42.80	47.82
AgentHOI (Wang et al.) (Wang et al., 2026)	YOLOv8/BLIP/Qwen2.5	13.26	–	–	–	–	–	–	–
AgentHOI (Lei et al.) (Lei et al., 2026)	G-DINO/GPT-4o	29.61	40.33	26.41	–	–	–	–	–
HarnessHOI	Qwen3.5-9B	**33.20**	**41.63**	**30.68**	**36.85**	**44.20**	**34.65**	**44.65**	**49.70**

4.2 Comparison with State-of-the-Art Methods

Regular settings. Table 1 compares HarnessHOI with representative task-optimized and training-free HOI detectors. HarnessHOI establishes state-of-the-art training-free performance on both benchmarks. On HICO-DET, it achieves 33.20% Default Full mAP, surpassing AgentHOI (Lei et al., 2026) by 3.59 points and the training-free variant of DA-HOI (Xuan et al., 2026) by 1.70 points, and reaches 36.85% under Known Object evaluation. It also obtains the best training-free Rare and Non-Rare results, with 41.63% Rare mAP exceeding AgentHOI by 1.30 points. The consistent gains across frequency groups and evaluation settings demonstrate the effectiveness of the HOI harness in recognizing common and infrequent interactions.

On V-COCO, HarnessHOI obtains 44.65%/49.70% role AP under S1/S2, surpassing DYSCO (Tonini et al., 2025) by 1.85/1.88 points. The gains under both scenarios demonstrate the effectiveness of the harness in recognizing actions and identifying their corresponding role targets. Extended results, including a common-evaluator comparison with the released predictions of AgentHOI, are provided in Appendix C.

Open-vocabulary settings. Table 2 reports results under RF-UC, NF-UC, UO, and UV using the same HarnessHOI predictions. The method achieves 40.15%, 33.64%, 35.45%, and 33.87% Unseen mAP, respectively, attaining the best training-free results in all four partitions. Improvements over AgentHOI (Lei et al., 2026) reach 1.75 points on UO and 4.02 points on UV. On NF-UC, HarnessHOI surpasses the strongest previous training-free result, ADA-CM (Lei et al., 2023), by 3.53 points. It also obtains the strongest reported training-free Seen results across all four settings. The simultaneous gains on Unseen and Seen classes demonstrate the effectiveness of the harness across interaction compositions, object categories, and actions. Complete results appear in Appendix C.

4.3 Ablation Studies

We conduct ablation studies to evaluate the overall HOI harness and the contributions of its perceptual and reasoning components. The component studies use the same frozen Qwen3.5-9B and

Table 2: Open-vocabulary HOI detection on HICO-DET using predicted boxes. RF-UC, NF-UC, UO, and UV each report Unseen, Seen, and Full mAP (%). HarnessHOI uses the same predictions across all four partitions. Within each method group, the best results are bold; the strongest previous training-free results are underlined. Dashes denote unavailable entries.

	RF-UC			NF-UC			UO			UV		
Method	Unseen	Seen	Full	Unseen	Seen	Full	Unseen	Seen	Full	Unseen	Seen	Full
Task-optimized methods												
GEN-VLKT (Liao et al., 2022)	21.36	32.91	30.56	25.05	23.38	23.71	10.51	28.92	25.63	20.96	30.23	28.74
HOICLIP (Ning et al., 2023)	25.53	34.85	32.99	26.39	28.10	27.75	16.20	30.99	28.53	24.30	32.19	31.09
CLIP4HOI (Mao et al., 2023)	28.47	35.48	34.08	31.44	28.26	28.90	31.79	32.73	32.58	26.02	31.14	30.42
ADA-CM (FT) (Lei et al., 2023)	27.63	34.35	33.01	32.41	31.13	31.39	–	–	–	–	–	–
CMMP (ViT-L) (Lei et al., 2024b)	35.98	37.42	37.13	33.52	35.53	35.13	39.67	36.15	36.74	30.84	37.28	36.38
LAIN (ViT-L) (Kim et al., 2025)	36.57	38.54	38.13	37.52	35.90	36.22	40.78	36.96	37.60	32.05	38.04	37.20
BC-HOI (BLIP-2) (Hu et al., 2025)	42.31	40.67	40.99	33.01	37.24	36.40	19.94	37.03	34.18	31.18	41.31	39.89
InstructHOI (Luo et al., 2025)	36.82	44.86	43.25	36.42	38.82	38.34	39.92	37.05	37.53	31.64	39.17	38.12
LINK (ViT-L) (Wu et al., 2026a)	38.51	40.33	39.97	34.63	35.67	35.43	38.24	37.61	37.91	27.22	38.36	36.88
FineHOI (Tonini et al., 2026)	37.79	38.55	38.39	38.11	36.30	36.58	40.13	35.92	36.62	35.58	38.51	38.10
DA-HOI (FT, DETR) (Xuan et al., 2026)	41.79	44.01	43.56	**43.12**	39.63	40.33	48.67	42.58	43.60	**36.89**	43.84	42.88
DA-HOI (FT, G-DINO) (Xuan et al., 2026)	**43.30**	**45.19**	**44.81**	41.52	**41.51**	**41.51**	**50.15**	**44.31**	**45.28**	36.88	**45.66**	**44.43**
Training-free methods												
BLIP-2 analysis (Ai, 2024)	23.96	18.65	19.71	19.97	19.65	19.71	22.71	19.12	19.71	21.96	19.35	19.71
ADA-CM (TF) (Lei et al., 2023)	26.83	24.54	25.00	30.11	23.16	24.55	–	–	–	–	–	–
DYSCO (Tonini et al., 2025)	30.36	23.96	25.24	27.56	24.58	25.18	–	–	–	–	–	–
AgentHOI (Lei et al.) (Lei et al., 2026)	38.64	–	–	28.04	–	–	33.70	–	–	29.85	–	–
HarnessHOI	**40.15**	**31.46**	**33.20**	**33.64**	**33.09**	**33.20**	**35.45**	**32.75**	**33.20**	**33.87**	**33.09**	**33.20**

Table 3: Configuration and component analysis of the HOI harness. D: inclusion of discovered participants; O: targeted observation; G: geometric adjudication; V: visual relation verification. ✓ denotes retained, × removed, and – not applied. HICO-DET reports Default Full mAP; V-COCO reports S1/S2 role AP (%). Best results are bold.

Variant	Harness configuration				HICO-DET	V-COCO	
	D	O	G	V	Full	S1	S2
Direct HOI generation	–	–	–	–	15.62	16.41	21.20
Initial-instance baseline	–	–	–	✓	31.45	42.90	47.20
w/o discovered participants	×	✓	✓	✓	32.55	44.02	48.93
w/o targeted observation	✓	×	–	✓	32.80	44.33	49.21
First-candidate geometry	✓	✓	×	✓	32.95	44.38	49.48
HarnessHOI	✓	✓	✓	✓	**33.20**	**44.65**	**49.70**

report HICO-DET Default Full mAP and V-COCO S1/S2 role AP on the complete test sets. Detailed settings and complete results are provided in Appendix D.

Effectiveness of the HOI harness. Table 3 compares the complete HOI harness with direct generation and an initial-instance baseline that verifies relations among initially localized participants. On HICO-DET, the initial-instance baseline achieves 31.45% mAP compared with 15.62% for direct generation, and the complete harness further reaches 33.20%. On V-COCO, the complete harness improves the initial-instance baseline from 42.90%/47.20% to 44.65%/49.70% AP. These results demonstrate that grounded verification strengthens interaction recognition, while interaction-guided perception and geometric adjudication further improve detection through participant discovery and shared grounding. Candidate-ranking analysis and backbone generalization with InternVL3-8B and Qwen2.5-VL-7B are presented in Appendices E.4 and D.6.

Interaction-guided perception. As shown in Table 3, participant discovery improves V-COCO S2 AP by 0.77 points. Under the same generation budget, interaction-guided discovery outperforms repeated participant localization by 0.75 points in HICO-DET mAP and 0.78/0.82 points in V-COCO AP (Table 4). With region counts and sizes fixed, targeted observation improves over non-directed observation by 0.50 and 0.58/0.62 points, respectively. Interaction semantics therefore

Table 4: Controlled alternatives to the harness design. Each alternative replaces the corresponding design choice within the complete harness. HICO-DET reports Default Full mAP; V-COCO reports S1/S2 role AP (%). The comparison conditions are specified in Appendix D.

Design factor	Alternative	HICO-DET	V-COCO	
		Full	S1	S2
Participant discovery	Repeated localization	32.45	43.87	48.88
Local observation	Non-directed regions	32.70	44.07	49.08
Geometric adjudication	Relation-conditioned	32.79	44.18	49.16
Relation verification	Proposal-conditioned	32.50	43.92	48.84
HarnessHOI		33.20	44.65	49.70

provide effective guidance for recovering overlooked participants and observing relevant regions. The coverage analysis in Appendix D.3 further shows how recovered participants contribute to interaction recognition.

Geometric adjudication and interaction reasoning. Table 4 evaluates relation-agnostic geometric adjudication and grounded interaction reasoning. With candidate boxes and image views fixed, selecting shared groundings according to participant identity, category, and visible extent improves over relation-conditioned adjudication by 0.41 points in HICO-DET mAP and 0.47/0.54 points in V-COCO AP. On identical relation candidates, verification using the selected participants and acquired visual evidence further improves over proposal-conditioned questions by 0.70 and 0.73/0.86 points. Together, these results validate the separation between hypothesis-guided visual exploration and evidence-grounded interaction assessment.

Action and role prediction. The analyses in Appendix D.5 show that combining perceptual associations with visual pair assessment improves detection, while verifying compatible labels recovers multiple actions involving the same participants. Role prediction further benefits from separately assessing action occurrence, target visibility, and role binding. On identical role candidates, compositional verification improves S1/S2 AP from 41.10%/46.30% to 44.65%/49.70%. With the final true-positive counts unchanged, this improvement reflects more accurate ranking of role predictions through the composition of action and target evidence.

Inference computation. Table 33 compares detection performance and computational cost using the same frozen MLLM. Compared with direct generation using eight independent responses per image, HarnessHOI improves HICO-DET Full mAP by 13.36 points, while reducing the measured cost from 19.95 to 13.26 GPU-worker seconds per image. On V-COCO, the gains reach 22.44/23.20 points in S1/S2 AP, with measured costs of 22.07 and 20.94 GPU-worker seconds per image for HarnessHOI and repeated generation, respectively. The substantial detection gains at comparable computational cost demonstrate the advantage of coordinating visual exploration and grounded reasoning within the harness. Further computation analyses and hardware measurements are provided in Appendix E.

5 Conclusion

We presented HarnessHOI, a training-free framework that harnesses frozen MLLMs for active interaction-guided perceptual reasoning. Its HOI harness connects interaction understanding with visual exploration, enabling the model to discover overlooked participants and inspect ambiguous regions. Relation-agnostic geometric adjudication reconciles the acquired localization candidates, and grounded interaction reasoning uses the selected instance groundings to verify multiple actions and semantic roles. Extensive experiments on HICO-DET and V-COCO show that HarnessHOI achieves state-of-the-art performance among training-free methods, demonstrating the value of coordinating visual exploration and relation assessment through shared instance groundings.

AI USE STATEMENT

In the preparation of this work, generative AI tools assisted with the purpose of language polishing and grammar correction to improve clarity and readability. We confirm that all core research ideas, the conceptual framework, experimental design, data analysis are entirely the original work of the authors. The authors take full and sole responsibility for the accuracy, integrity, and all claims made within this work.

ETHICS STATEMENT

Our work adheres to the ICLR Code of Ethics. The research is conducted exclusively on publicly available, standard academic benchmarks (HICO-DET and V-COCO), which are widely used for evaluating human-object interaction detection algorithms. Our primary motivation is to advance the field for societal benefit, such as assistive robotics and accessible visual understanding. We believe that our framework supports the responsible development of AI by making the visual basis of interaction predictions available for inspection. Specifically, the harness associates discovered participants, targeted observations, and relation assessments with shared groundings, allowing predictions to be examined against the image evidence and corresponding visual questions. Responsible applications should respect privacy, consent, and the original data providers' terms, while accounting for biases in pretrained models and benchmark coverage. Human oversight should be appropriate to the consequences of an incorrect prediction. We advocate for the responsible development and deployment of this technology, ensuring its application aligns with ethical principles and contributes positively to human well-being.

REPRODUCIBILITY STATEMENT

We are fully committed to ensuring the complete reproducibility of our research. To this end, we provide a detailed description of our methodology in Section 3, supplemented by the complete inference procedure in Appendix A and the prompt instructions and visual questions in Appendix G. All crucial implementation details, including the frozen Qwen3.5-9B backbone, image processing, inference settings, and evaluation protocols, are clearly documented in Section 4.1 and Appendix B. Our experiments are built on standard public benchmarks and common libraries, including PyTorch, Hugging Face Transformers, and vLLM, to ensure broad accessibility. Extended comparisons and ablations specify their baseline formulations and controlled variables in Appendices C and D, while Appendix E details the computation measurements and independent-run evaluation. We also pledge to release our source code, inference configurations, image lists, seed assignments, and evaluation scripts upon the publication of this paper, which we believe will allow the community to fully replicate our findings and extend our work.

REFERENCES

Chaoyi Ai. An analysis of HOI: Using a training-free method with multimodal visual foundation models when only the test set is available, without the training set. *arXiv preprint arXiv:2408.05772*, 2024.

Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *CVPR*, pp. 13778–13790, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

R. Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988. doi: 10.1109/5.5968.

Ping Cao, Yepeng Tang, Chunjie Zhang, Xiaolong Zheng, Chao Liang, Yunchao Wei, and Yao Zhao. Visual relation diffusion for human-object interaction detection. In *ICCV*, pp. 23551–23560, 2025.

Yichao Cao, Qingfei Tang, Xiu Su, Song Chen, Shan You, Xiaobo Lu, and Chang Xu. Detecting any human-object interaction relationship: Universal HOI detector with spatial prompt learning on foundation models. In *NeurIPS*, volume 36, pp. 739–751. Curran Associates, Inc., 2023. doi: 10.52202/075280-0034.

Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, pp. 381–389, 2018. doi: 10.1109/wacv.2018.00048.

Junwen Chen and Keiji Yanai. HOIBlender: Blending lightweight detection with vision-language priors for efficient human-object interaction detection. *arXiv preprint arXiv:2609.23431*, 2026.

Junwen Chen, Peilin Xiong, and Keiji Yanai. HOI-R1: Exploring the potential of multimodal large language models for human-object interaction detection. *arXiv preprint arXiv:2510.05609*, 2025.

Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating HOI detection as adaptive set prediction. In *CVPR*, pp. 9004–9013, 2021.

Mihai Dogariu, Liviu-Daniel Stefan, Mihai Gabriel Constantin, and Bogdan Ionescu. Human-object interaction: Application to abandoned luggage detection in video surveillance scenarios. In *COMM*, pp. 157–160, 2020. doi: 10.1109/COMM48946.2020.9141973.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010. doi: 10.1007/s11263-009-0275-4.

Yaowu Fan, Tao Han, Dazhao Du, Andy J. Ma, and Jia Wan. Train the agent, not the expert: Learning to harness heterogeneous experts for multi-turn visual reasoning. *arXiv preprint arXiv:2605.29894*, 2026.

Chen Gao, Yuliang Zou, and Jia-Bin Huang. iCAN: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018.

Jianjun Gao, Chen Cai, Ruoyu Wang, Wenyang Liu, Kim-Hui Yap, Kratika Garg, and Boon Siew Han. CL-HOI: Cross-level human-object interaction distillation from multimodal large language models. *Knowledge-Based Systems*, 320:113561, 2025. doi: 10.1016/j.knosys.2025.113561.

Pei Geng, Jian Yang, and Shanshan Zhang. HORP: Human-object relation priors guided HOI detection. In *CVPR*, pp. 25325–25335, 2025.

Georgia Gkioxari, Ross Girshick, Piotr Dollar, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, pp. 8359–8367. IEEE, June 2018. doi: 10.1109/CVPR.2018.00872.

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

Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, pp. 584–600, 2020.

Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *CVPR*, pp. 14646–14655, 2021.

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 *ICCV*, pp. 20126–20136, 2025. doi: 10.1109/ICCV51701.2025.01872.

Xiaofei Hui, Haoxuan Qu, Hossein Rahmani, and Jun Liu. An image-like diffusion method for human-object interaction detection. In *CVPR*, pp. 14002–14012, 2025.

Mingda Jia, Liming Zhao, Ge Li, and Yun Zheng. ContextHOI: Spatial context learning for human-object interaction detection. In *AAAI*, volume 39, pp. 3931–3939, 2025a. doi: 10.1609/aaai.v39i4.32411.

Mingda Jia, Liming Zhao, Ge Li, and Yun Zheng. Orchestrating the symphony of prompt distribution learning for human-object interaction detection. In *AAAI*, volume 39, pp. 3940–3948, 2025b. doi: 10.1609/aaai.v39i4.32412.

Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. HOTR: End-to-end human-object interaction detection with transformers. In *CVPR*, pp. 74–83, 2021. doi: 10.1109/CVPR46437.2021.00014.

Sanghyun Kim, Deunsol Jung, and Minsu Cho. Locality-aware zero-shot human-object interaction detection. In *CVPR*, pp. 20190–20200. IEEE, June 2025. doi: 10.1109/CVPR52734.2025.01880.

Qinqian Lei, Bo Wang, and Robby T. Tan. EZ-HOI: VLM adaptation via guided prompt learning for zero-shot HOI detection. In *NeurIPS*, volume 37, pp. 55831–55857. Curran Associates, Inc., 2024a. doi: 10.52202/079017-1775.

Qinqian Lei, Bo Wang, and Robby T. Tan. HOLa: Zero-shot HOI detection with low-rank decomposed VLM feature adaptation. In *ICCV*, pp. 1825–1835, 2025a. doi: 10.1109/ICCV51701.2025.00178.

Ting Lei, Fabian Caba, Qingchao Chen, Hailin Jin, Yuxin Peng, and Yang Liu. Efficient adaptive human-object interaction detection with concept-guided memory. In *ICCV*, pp. 6457–6467. IEEE, October 2023. doi: 10.1109/ICCV51070.2023.00596.

Ting Lei, Shaofeng Yin, Yuxin Peng, and Yang Liu. Exploring conditional multi-modal prompts for zero-shot HOI detection. In *ECCV*, pp. 1–19. Springer, 2024b. doi: 10.1007/978-3-031-73007-8_1.

Ting Lei, Shaofeng Yin, Qingchao Chen, Yuxin Peng, and Yang Liu. Open-vocabulary HOI detection with interaction-aware prompt and concept calibration. In *ICCV*, pp. 23945–23957, 2025b. doi: 10.1109/iccv51701.2025.02220.

Ting Lei, Jialin Liu, Zhu Xu, Yuxin Peng, and Yang Liu. Unleashing multimodal large language models for training-free HOI detection in the wild. In *ECCV*, 2026.

Huadong Li, Ying Wei, Shuailei Ma, Mingyu Chen, and Ge Li. Ripple transformer: A human-object interaction backbone and a new prediction strategy for smart surveillance devices. *IEEE Transactions on Consumer Electronics*, 70(1):2257–2268, February 2024. doi: 10.1109/TCE.2024.3371451.

Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *CVPR*, pp. 3580–3589. IEEE, June 2019. doi: 10.1109/CVPR.2019.00370.

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 *CVPR*, pp. 20091–20100. IEEE, June 2022. doi: 10.1109/CVPR52688.2022.01949.

Xin Lin, Chong Shi, Zuopeng Yang, Haojin Tang, and Zhili Zhou. SGC-Net: Stratified granular comparison network for open-vocabulary HOI detection. In *CVPR*, pp. 4539–4549, 2025.

Jinguo Luo, Weihong Ren, Quanlong Zheng, Yanhao Zhang, Zhenlong Yuan, Zhiyong Wang, Haonan Lu, and Honghai Liu. InstructHOI: Context-aware instruction for multi-modal reasoning in human-object interaction detection. In *NeurIPS*, volume 38, pp. 142558–142577. Curran Associates, Inc., 2025. doi: 10.52202/085713-4772.

Yunyao Mao, Jiajun Deng, Wengang Zhou, Li Li, Yao Fang, and Houqiang Li. CLIP4HOI: Towards adapting CLIP for practical zero-shot HOI detection. In *NeurIPS*, volume 36 of *NeurIPS 2023*, pp. 45895–45906. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-1989.

Esteve Valls Mascaro, Daniel Sliwowski, and Dongheui Lee. HOI4ABOT: Human-object interaction anticipation for human intention reading collaborative roBOTs. In *CoRL*, volume 229 of *Proceedings of Machine Learning Research*, pp. 1111–1130. PMLR, 2023.

Minh Anh Nguyen, Quang Huy Tran, Bao Ngoc Le, SuiYang Guang, Tuan Kiet Pham, and Linh Chi Vo. ScriptHOI: Learning scripted state transitions for open-vocabulary human-object interaction detection. *arXiv preprint arXiv:2605.05057*, 2026.

Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. HOICLIP: Efficient knowledge transfer for HOI detection with vision-language models. In *CVPR*, pp. 23507–23517. IEEE, June 2023. doi: 10.1109/CVPR52729.2023.02251.

Qwen Team. Qwen3.5-9B model card. Hugging Face model card, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Sandipan Sarma, Agney Talwarr, and Arijit Sur. Funnel-HOI: Top-down perception for zero-shot HOI detection. *Machine Vision and Applications*, 37(3):59, 2026. doi: 10.1007/s00138-026-01818-9.

Soo Won Seo, KyungChae Lee, Hyungchan Cho, Taein Son, Nam Ik Cho, and Jun Won Choi. Mining instance-centric vision-language contexts for human-object interaction detection. In *CVPR*, pp. 40418–40427, June 2026.

Daniel Shalam, Emanuel Ben Baruch, Avi Ben Cohen, and Tal Remez. Propose and attend: Training-free MLLM grounding confidence via multi-token localized attention. *arXiv preprint arXiv:2607.05978*, 2026.

Chang Sun, Dongliang Liao, and Changxing Ding. Streamlined open-vocabulary human-object interaction detection. In *CVPR*, pp. 20324–20333, 2026.

Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, pp. 10405–10414. IEEE, June 2021. doi: 10.1109/CVPR46437.2021.01027.

Francesco Tonini, Lorenzo Vaquero, Alessandro Conti, Cigdem Beyan, and Elisa Ricci. Dynamic scoring with enhanced semantics for training-free human–object interaction detection. In *ACM MM*, pp. 2801–2810, 2025. doi: 10.1145/3746027.3754770.

Francesco Tonini, Lorenzo Vaquero, Mohammad Mahdi Derakhshani, Cees Snoek, Elisa Ricci, and Cigdem Beyan. FineHOI: Part-aware dense representations for zero-shot human-object interaction detection. In *ACM MM*, 2026. doi: 10.1145/3767308.3836016.

Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *ICCV*, pp. 9469–9478, 2019.

Guangzhi Wang, Yangyang Guo, Ziwei Xu, and Mohan Kankanhalli. Bilateral adaptation for human-object interaction detection with occlusion-robustness. In *CVPR*, pp. 27970–27980, 2024.

Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. Deep contextual attention for human-object interaction detection. In *ICCV*, pp. 5693–5701. IEEE, October 2019. doi: 10.1109/ICCV.2019.00579.

Yuxiao Wang, Yu Lei, and Qi Liu. AgentHOI: Towards multi-agent collaborative human-object interaction detection. *IEEE Internet of Things Journal*, 13(18):41354–41361, 2026. doi: 10.1109/JIOT.2026.3685409.

Eastman Z. Y. Wu, Ya-Li Li, Yuan Wang, and Shengjin Wang. LINK: Learning instance-level knowledge from vision-language models for human-object interaction detection. In *ICLR*, 2026a.

Eastman Z. Y. Wu, Yali Li, and Shengjin Wang. Towards open-world human-object interaction reasoning with multimodal large language model. In *ICASSP*, pp. 8392–8396, 2026b. doi: 10.1109/ICASSP55912.2026.11462048.

Mingrui Wu, Jiaxin Gu, Yunhang Shen, Mingbao Lin, Chao Chen, and Xiaoshuai Sun. End-to-end zero-shot HOI detection via vision and language knowledge distillation. In *AAAI*, pp. 2839–2846, 2023.

Yongchao Xu, Jiawei Liu, Junfeng Wang, Sen Tao, Na Jiang, and Zheng-Jun Zha. Learning to diversify and focus: A reinforcement framework for open-vocabulary HOI detection. In *CVPR*, pp. 34673–34682, 2026.

Shiyu Xuan, Dongkai Wang, Zechao Li, and Jinhui Tang. Zero-shot HOI detection with MLLM-based detector-agnostic interaction recognition. In *ICLR*, 2026.

Bin Yang, Yulin Zhang, Hong-Yu Zhou, and Sibei Yang. No more sibling rivalry: Debiasing human-object interaction detection. In *ICCV*, pp. 22707–22717, 2025a. doi: 10.1109/ICCV51701.2025.02108.

Chanhyeong Yang, Taehoon Song, Jihwan Park, and Hyunwoo J. Kim. Visual diversity and region-aware prompt learning for zero-shot HOI detection. In *NeurIPS*, volume 38, pp. 119981–120010. Curran Associates, Inc., 2025b. doi: 10.52202/085713-4006.

Panqi Yang, Haodong Jing, Nanning Zheng, and Yongqiang Ma. UniHOI: Unified human-object interaction understanding via unified token space. In *AAAI*, volume 40, pp. 11640–11648, 2026. doi: 10.1609/aaai.v40i14.38148.

Hangjie Yuan, Jianwen Jiang, Samuel Albanie, Tao Feng, Ziyuan Huang, Dong Ni, and Mingqian Tang. RLIP: Relational language-image pre-training for human-object interaction detection. In *NeurIPS*, volume 35, pp. 37416–37431. Curran Associates, Inc., 2022. doi: 10.52202/068431-2712.

Hangjie Yuan, Shiwei Zhang, Xiang Wang, Samuel Albanie, Yining Pan, Tao Feng, Jianwen Jiang, Dong Ni, Yingya Zhang, and Deli Zhao. RLIPv2: Fast scaling of relational language–image pre-training. In *ICCV*, pp. 21649–21661, 2023.

Zhenlong Yuan, Yue Wang, Dapeng Zhang, Kejin Cui, Rui Chen, Jing Tang, Lei Sun, Hongwei Yu, Chengxuan Qian, Xiangxiang Chu, Shuo Li, and Yuyin Zhou. What if agents could imagine? reinforcing open-vocabulary HOI comprehension through generation. *arXiv preprint arXiv:2602.11499*, 2026.

Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage HOI detection. In *NeurIPS*, volume 34, pp. 17209–17220, 2021a.

Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *ICCV*, pp. 13319–13327, 2021b.

Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *CVPR*, pp. 20072–20080. IEEE, June 2022a. doi: 10.1109/CVPR52688.2022.01947.

Frederic Z. Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring predicate visual context in detecting of human-object interactions. In *ICCV*, pp. 10377–10387. IEEE, October 2023. doi: 10.1109/ICCV51070.2023.00955.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *ICLR*, pp. 68194–68213, 2025a.

Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *CVPR*, pp. 19526–19535, 2022b. doi: 10.1109/CVPR52688.2022.01894.

Zhenhao Zhang, Hanqing Wang, Xiangyu Zeng, Ziyu Cheng, Jiaxin Liu, Haoyu Yan, Zhirui Liu, Kaiyang Ji, Tianxiang Gui, Ke Hu, Kangyi Chen, Yahao Fan, and Mokai Pan. HOID-R1: Reinforcement learning for open-world human-object interaction detection reasoning with multimodal large language model. *arXiv preprint arXiv:2508.11350*, 2025b.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, and Jian Sun. End-to-end human object interaction detection with HOI transformer. In *CVPR*, pp. 11820–11829, 2021. doi: 10.1109/CVPR46437.2021.01165.

A Additional Details of HarnessHOI

A.1 Shared Participants and Interaction Evidence

Table 5 summarizes how the MLLM performs interaction-guided perception, relation-agnostic geometric adjudication, and grounded interaction reasoning. Throughout these stages, each participant retains a common identity that links its localization candidates to the interactions involving it.

Participant identity. Each entity e_i has a persistent identity and a candidate set $\mathcal{B}_i$ containing its alternative boxes, categories, and source views. Discovery identifies additional participants and relations among existing participants, while observation supplies local grounding candidates. Geometric adjudication selects the box and category shared by all relations involving the entity.

Initial localization and discovery associate exact category-and-box matches with the same identity. Discovery outputs are merged in human-group order, matching each new localization against the accumulated participant set. Other localized participants receive new identities, while observation refinements retain the identity of the designated participant.

Evidence and associations. Discovery associations $\mathcal{P}_D$ connect humans to the targets of positive interaction hypotheses. Observation associations $\mathcal{P}_O$ connect the inspected human to the requested target and newly localized participants. Both contribute candidate pairs for relation verification, where each action or role receives its own visual assessment.

Table 5: MLLM operations within the three stages of HarnessHOI. Perceptual generation constructs participants and observation requests; visual question answering selects shared groundings and evaluates candidate relations.

Operation	Visual and semantic context	MLLM operation	Shared outcome
Interaction-guided perception			
Initial localization	Full image; object vocabulary	Generate participants in interaction groups	Initial entities and humans
Interaction discovery	Full image; initial entities; designated humans; public relations	Generate hypotheses, additional participants, and observation requests	Expanded entities and discovery associations
Targeted observation	One requested crop; target description; local instance context	Generate instance updates and newly visible participants	Candidate geometry and observation associations
Relation-agnostic geometric adjudication			
Grounding selection	Full image and acquired crops; candidate boxes for one identity	Answer which candidate matches the visible instance	One grounding per identity
Grounded interaction reasoning			
Relation assessment	Full image and a relevant crop; selected participants; public relation definition	Answer pair, action, agent, visibility, or binding questions	Relation supports and final detections

A.2 Interaction-Guided Perception

Initial localization and grouped discovery. Initial localization groups each person with the identifiable targets involved in their interactions. We then divide the initially localized humans into groups of at most four for discovery. Conditioned on the full image and initial entity set, each group is examined for additional interactions and participants. Newly discovered humans participate in subsequent geometric and relational assessment, while discovery groups are defined by the initial human set.

A hypothesis specifies the human, action, and target identities, together with semantic roles for V-COCO. Each person may participate in multiple actions with multiple targets. Existing targets are referenced by identity, and new participants are localized by category and complete visible extent. A hypothesis with unresolved target localization can request a local observation, specifying the human, target description, action and role, and region to inspect.

Observation selection. The harness executes at most two observation requests per image. Requests concerning an unlocalized target take precedence over requests to complete an existing participant's boundary. Within each priority, different humans receive their first observation before a second request for the same human is considered. Requests for the same known target are deduplicated by identity; unresolved targets are distinguished by their associated human and description. Each discovery group may propose two requests, while the image-level budget governs execution after merging.

Given the target description and instances within the requested region, the MLLM refines eligible groundings and localizes additional visible participants. A single observation can therefore provide localization candidates for several participants in the same crop.

Local-to-global geometry. For a crop with image-pixel boundaries $[L, T, R, B]$, let (x, y) be a point normalized to $[0, 1000]$ within the crop. Its image coordinate is

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} L \\ T \end{bmatrix} + \frac{1}{1000} \begin{bmatrix} R-L & 0 \\ 0 & B-T \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{9}$$

Here (x', y') is a full-image pixel position. For an image of size $W \times H$, $(1000x'/W, 1000y'/H)$ gives its normalized coordinates. Both box corners are transformed to this full-image representation before updating the candidate set. An existing identity is eligible for refinement when all its candidate boxes lie within the crop; boundary-crossing instances provide contextual references. Localized boxes cover the visible extent of each instance.

Observation associations. Let h_ρ denote the human associated with observation ρ, t_ρ its known target when available, and $\mathcal{E}_\rho^{\text{new}}$ the newly localized identities. For the acquired observations $\mathcal{R}_{\text{obs}}$, the corresponding relation associations are

$$\mathcal{P}_O = \bigcup_{\rho \in \mathcal{R}_{\text{obs}}} \big\{(h_\rho, o) : o \in \mathcal{E}_\rho^{\text{new}} \cup \{t_\rho\},\ o \neq h_\rho\big\}, \tag{10}$$

where $\{t_\rho\}$ is omitted for an unlocalized target. These associations connect the requested target and newly localized participants to the human whose interaction prompted the observation, forming candidate pairs for grounded relation verification.

A.3 Relation-Agnostic Geometric Adjudication

One grounding per identity. For an identity with multiple grounding candidates, the MLLM compares their category and box descriptions against the full image and acquired crops. Each candidate is assigned an answer letter, and the candidate with the highest answer support is selected. The judgment considers the participant's identity, category, and complete visible extent to establish a shared grounding across actions and roles. Identities with one candidate retain their original grounding.

The selected geometry is shared by every relation involving the participant: for relations d_1 and d_2 containing identity i, $b_i(d_1) = b_i(d_2) = b_i^*$, where $b_i^* = b_{i,m_i^*}$. With at most two observations, an identity has at most three candidates: its initial or discovery grounding and up to two local revisions.

A.4 Grounded Interaction Reasoning

Question construction. The question context comprises the selected participants' identities, categories, and image-coordinate boxes, the public action or role definition, and the corresponding image views. Pair questions assess the presence of an interaction, and action questions verify a specific predicate. Agent, role-state, and binding questions respectively assess action occurrence, target visibility, and role identity. The prompt formulations appear in Appendix G.

Visual context. Localization and discovery use the full image, local observation uses the requested crop, and geometric adjudication receives all acquired views. Relation questions use the full image and, when available, one local view associated with the queried participants. The source view v_{i,m_i^*} of a selected grounding takes priority, with the most recent observation chosen when both participants have local source views. If neither selected grounding has a local source view, the crop is taken from the most recent compatible observation request.

A compatible request concerns the same human and agrees with the question's specified action and target, or leaves those attributes unspecified. For role questions, it concerns the queried role or general participant localization. Agent questions admit requests for any role, pair questions use general participant requests, and no-interaction questions impose no action or role restriction. The full image provides the visual context when no compatible local view is available.

Answer support. Answer supports are obtained by normalizing the raw first-token logits according to Equation 2. Pair, action, agent, and binding questions use {`false`, `true`, `unknown`}; role-state questions use {`missing`, `visible`, `unknown`}; geometry questions use the candidate letters. Each label corresponds to one token. These normalized supports determine the selected geometry and final detection scores, with the generation temperature applied to perceptual generation alone.

Candidate retention and multi-label prediction. Discovery and observation associations complement visual pair assessment in Equation 4, with $\mathcal{P}_D \cup \mathcal{P}_O \subseteq \mathcal{K}$. Each retained HICO-DET pair is evaluated over all positive actions compatible with its target category, using a separate answer distribution for each action. The `no_interaction` label is evaluated independently over all directed pairs.

Physical and missing role targets. V-COCO uses the agent support q_A, role-state supports q_V and q_M, and binding support q_B defined in the main text. Physical candidates share the action and visibility factors and receive individual binding assessments, giving $s_{\text{physical}} = q_A q_V q_B$. The NULL alternative combines action and missing-target support as $s_{\text{NULL}} = q_A q_M$. Since visible, missing, and unknown states share one answer distribution, $q_V + q_M \leq 1$; the combined score of any one physical candidate and its NULL alternative is bounded by q_A.

A.5 Complete Inference Procedure

Table 6 summarizes the complete inference procedure, from participant localization and observation to grounded interaction prediction.

Table 6: Inference procedure for one image. All generation and question-answering operations use the same frozen MLLM.

Step	Operation
1	Localize initial entities $\mathcal{E}_0$ and identify the initial humans $\mathcal{H}_0$. Initialize each entity's geometry candidates with its returned box.
2	Partition $\mathcal{H}_0$ into groups of at most four. For each group, discover interactions and participants using the full image and $\mathcal{E}_0$. Associate the resulting hypotheses and observation requests with the expanded participant set.
3	Select at most two image-level observation requests. Localize participants in each requested crop, map their boxes to image coordinates, and update geometry candidates and observation associations.
4	For each multi-candidate identity, ask the geometry question over the full image and acquired crops. Select one grounding and construct the shared entity state $\mathcal{E}^*$.
5	Assess all directed human-entity pairs. Form $\mathcal{K}$ from pair support, discovery associations, and observation associations. For HICO-DET, independently assess no-interaction for all pairs.
6	For HICO-DET, evaluate every compatible positive action on each retained pair through a visual question. For V-COCO, assess each human's actions and role states, followed by role bindings for retained pairs.
7	Compute the HICO-DET and V-COCO scores using Equations 6 and 8. Form the final detections from the selected boxes and continuous scores, retaining maximum support for exact duplicates.

B Experimental Setup and Evaluation Protocols

B.1 Detection and Semantic-Role Evaluation

HICO-DET. We evaluate all 9,658 official test images over 600 human-object interaction classes, comprising 520 positive-interaction classes and 80 `no_interaction` classes. Each prediction consists of a human box, an object box, an interaction class, and a confidence score. A detection is correct when its class matches the annotation and both boxes reach IoU 0.5. Predictions are matched in descending score order within each class, with at most one true positive per annotation. Rare contains 138 classes with fewer than ten training instances, and Non-Rare contains the remaining 462 classes.

Default evaluation includes all test images. Known Object evaluates each interaction class on images containing an annotated instance of its object category. Both settings use the same predictions, with Known Object restricting the images considered when assessing false positives.

Average precision. HICO-DET uses the VOC 2007 eleven-point definition (Everingham et al., 2010). For class c, let $p_c(r)$ denote precision at an attained recall value r, and let $\mathcal{R}_{\mathrm{AP}} = \{0, 0.1, \ldots, 1\}$. Then

$$\mathrm{AP}_c = \frac{1}{11} \sum_{t \in \mathcal{R}_{\mathrm{AP}}} \max_{r \geq t} p_c(r), \qquad \mathrm{mAP}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathrm{AP}_c, \tag{11}$$

where a recall threshold with no qualifying prediction contributes zero. Full mAP averages all 600 classes. We report AP as percentages and compute differences before rounding.

V-COCO. We evaluate all 4,946 official test images. Agent evaluation measures whether the localized person performs one of 26 actions. Role evaluation additionally identifies the target fulfilling the corresponding semantic role and averages over 25 action-role categories, including `point`. Actions such as `hit`, `eat`, and `cut` distinguish an object from an instrument. Conversely, `stand`, `walk`, `run`, and `smile` have agent annotations without a target role.

For a role with an annotated physical target, both the person and role target must match at IoU 0.5. When the role target is annotated as missing, Scenario 1 (S1) requires a matching person and a NULL target prediction. Scenario 2 (S2) relaxes target localization for that missing-role annotation. Both scenarios follow the benchmark's treatment of incompletely annotated people and integrate the precision-recall envelope.

V-COCO reporting conventions. The primary S1/S2 metric for HarnessHOI, direct-generation baselines, and ablations is the 25-role mean, including `point`. Appendix C.5 additionally gives the 24-role mean excluding `point` and the 26-action agent mean. Published comparison results retain their sources' role definitions and missing-target conventions; the S1/S2 columns follow the scenario designations in those sources.

Coordinate conventions. Predicted boxes use coordinates normalized within the input view and are mapped to image pixels for evaluation. HICO-DET then shifts all four zero-based coordinates by one and computes IoU with inclusive pixel extents, following its annotation convention. Crop coordinates are mapped to the full image as specified in Appendix A. All official test images are evaluated, including empty outputs and valid detections from partially completed inference.

B.2 Class Partitions and Development Settings

Open-vocabulary partitions. We report rare-first unseen combinations (RF-UC), non-rare-first unseen combinations (NF-UC), unseen objects (UO), and unseen verbs (UV). The Unseen subsets contain 120, 120, 100, and 84 classes, respectively; their Seen complements contain 480, 480, 500, and 516 classes. We use the class masks released with GEN-VLKT (Liao et al., 2022). Each partition aggregates the same class AP values from the same predictions. In particular,

$$\mathrm{mAP}_{\mathrm{Full}} = \frac{|\mathcal{C}_U|\mathrm{mAP}_U + |\mathcal{C}_S|\mathrm{mAP}_S}{600}, \tag{12}$$

where $\mathcal{C}_U$ and $\mathcal{C}_S$ are the benchmark's Unseen and Seen class sets. HarnessHOI uses the full public interaction vocabulary and the same frozen model and inference policy across these partitions. The class masks are applied to the resulting AP values, so the Full result is shared by all four settings.

Development protocol. Prompts and inference policy are developed using evaluation feedback from fixed subsets of the official train set: 128 images for optimization and 256 for confirmation, with four additional images for interface checks. Development feedback also includes train-image inspection and evaluation on the complete train set. The final configuration is evaluated on every image in the official test sets.

The optimization and confirmation sets are sampled without replacement from disjoint source subsets, excluding the interface images. Sampling uses fixed seeds, and each image retains its index in the official train list for generation across both development-subset and complete-train-set runs. The image lists and seed assignments will accompany the released evaluation configurations.

Model inputs. The model remains frozen throughout development and evaluation. Its inference inputs comprise the image, generated participants, and public action and role definitions. Target-benchmark annotations are used for development feedback, evaluation, and the ground-truth-based analyses described below; they are not supplied to the model at inference.

B.3 Model and Inference Configuration

We use frozen Qwen3.5-9B for participant localization, interaction discovery, local observation, and grounded question answering. Its native visual encoder processes both the full image and local crops. Table 7 summarizes the inference settings.

Table 7: Inference configuration. Generation settings apply to localization, discovery, and observation; visual question answering reads raw first-token logits over its answer set.

Setting	Value
Model and visual inputs	
Backbone	Frozen Qwen3.5-9B; shared weights across stages
Precision	BF16 weights and KV cache
Maximum context	16,384 tokens
Full-image processor	Roughly 3 MP minimum and 17 MP maximum, depending on aspect ratio
Crop processor	Roughly 0.8 MP
Generation and visual question answering	
Initial localization output	At most 4,096 tokens
Discovery / observation output	At most 2,048 tokens per response
Generation sampling	Temperature 0.7; top-p 0.8; top-k 20; min-p 0
Generation penalties	Presence 1.5; frequency 0; repetition 1
Question answering	One output token; raw logits for the designated labels
Generation seeds	Base seed 66; seeds 67 and 68 for independent runs; fixed image and stage offsets
Harness settings	
Discovery group	At most four initially localized humans
Observation budget	At most two requests per image
Visual context	At most the full image and two distinct crops
Software	Torch 2.10.0; vLLM 0.19.1; Transformers 5.15.1; xgrammar 0.2.3

Image views. The image processor preserves aspect ratio and aligns the input dimensions to the model's visual grid. Local crops are extracted from the original image and encoded by the same backbone.

Generation and scoring. Perceptual generation produces participants, interaction hypotheses, and localization updates. Visual question answering uses one output token with decoding temperature

zero; answer support is computed from the raw logits before output-grammar modification. Independent generation runs preserve the image and stage seed offsets and vary the base seed as specified in Appendix E.5.

Detection output. We retain continuous detection scores and merge exact duplicate predictions by maximum support. No score threshold or non-maximum suppression is applied. V-COCO predictions include agent actions, physical-role targets, and NULL alternatives, scored according to Equation 8.

B.4 Baseline Inference

We compare three inference strategies using frozen Qwen3.5-9B on the complete test sets. Direct generation predicts participants and relations in one response, with a shared person box across actions. The initial-instance baseline applies visual pair assessment and relation verification to the initially localized participants. HarnessHOI additionally performs interaction-guided discovery, observation, and geometric adjudication. Direct generation uses the same full-image processing, sampling settings, and initial output budget.

Direct-generation confidence. For each generated interaction record, let $\mathcal{J}$ contain the token positions overlapping its emitted interaction label. The direct-generation score is the geometric mean of the corresponding token probabilities:

$$s_{\text{gen}} = \exp\left(\frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}}\log p_\theta(y_j \mid I, y_{<j})\right). \tag{13}$$

We use token probabilities before output-grammar modification. HICO-DET scores the emitted verb-object label; V-COCO scores the action label and assigns this score to its agent prediction and explicitly generated physical or NULL roles.

B.5 Evaluation Coverage

Evaluation includes all 9,658 HICO-DET and 4,946 V-COCO test images. Valid predictions from partially completed inference are evaluated normally. Images with no predictions, numbering 98 and 12 on the two benchmarks, respectively, remain in the evaluation population.

B.6 Evaluation Implementations

We use a common evaluator for HarnessHOI, its ablations, and the released predictions re-evaluated in Tables 13 and 14. Tables 8 and 9 compare this implementation with the original HICO-DET code under Octave and the original V-COCO evaluator on the same predictions, including direct generation and the principal perceptual and relational ablations.

On HICO-DET, the original code under Octave gives 32.97% Default and 36.61% Known Object Full mAP, compared with 33.20% and 36.85% from the common evaluator. The difference follows from the recall-threshold sequences constructed by the two implementations, which sample the precision envelope at slightly different boundaries in eleven-point AP. V-COCO role scores agree at the reported precision.

C Extended Benchmark Results

C.1 Regular and Open-Vocabulary Comparisons

Tables 10 and 11 provide the complete regular and open-vocabulary comparisons corresponding to Tables 1 and 2. They include additional backbone configurations, MLLM-based methods, retrieval baselines, and the training-free variant of DA-HOI.

Table 8: Evaluation implementations on identical HICO-DET predictions (Full mAP, %). Common denotes the evaluator used throughout this paper; Original denotes the benchmark code under Octave.

	Default		Known Object	
Method	Common	Original	Common	Original
Direct generation	15.62	15.50	17.27	17.11
Initial-instance baseline	31.45	31.23	35.05	34.82
w/o targeted observation	32.80	32.56	36.39	36.15
First-candidate geometry	32.95	32.71	36.55	36.30
HarnessHOI	33.20	32.97	36.85	36.61

Table 9: Evaluation implementations on identical V-COCO predictions (role AP, %). Common denotes the evaluator used in this paper; Original denotes the benchmark implementation. Both average the same 25 roles.

	S1		S2	
Method	Common	Original	Common	Original
Direct generation	16.41	16.41	21.20	21.20
Joint role question	41.10	41.10	46.30	46.30
HarnessHOI	44.65	44.65	49.70	49.70

DA-HOI and vision-language baselines. The fine-tuned DA-HOI (Xuan et al., 2026) comparisons include both DETR and GroundingDINO configurations under regular and open-vocabulary evaluation. Its training-free variant combines R50-DETR proposals with deterministic MLLM generation. DYSCO (Tonini et al., 2025) uses LongCLIP-L for regular evaluation and CLIP ViT-B/16 for open-vocabulary evaluation.

Other comparison configurations. GEN-VLKT uses its large and small configurations for regular and open-vocabulary evaluation, respectively. VDRP follows its default ViT-B configuration for regular evaluation and its ViT-L configuration for open-vocabulary evaluation. The UO result of EZ-HOI uses its HICO-DET-pretrained detector. The CLIP baseline follows LAIN (Kim et al., 2025); direct Qwen inference and LongCLIP retrieval follow HOI-R1 and DYSCO. LINK's V-COCO S2 result follows its ICLR version. V-COCO reporting conventions are specified in Appendix B.1.

Table 10: Complete regular-setting comparison corresponding to Table 1, including additional model configurations and training-free baselines. HICO-DET reports Default and Known Object mAP (%); V-COCO reports S1/S2 role AP (%). HarnessHOI uses the 25-role definition. Within each method group, the best results are bold; the strongest previous training-free results are underlined. Dashes denote unavailable entries.

		HICO-DET Default			HICO-DET Known Object			V-COCO	
Method	Model	Full	Rare	Non-Rare	Full	Rare	Non-Rare	S1	S2
Task-optimized methods									
QPIC (Tamura et al., 2021)	R101	29.90	23.92	31.69	32.38	26.06	34.27	58.30	60.70
UPT (Zhang et al., 2022a)	R101-DC5	32.62	28.62	33.81	36.08	31.41	37.47	61.30	67.10
GEN-VLKT (Liao et al., 2022)	R101/CLIP-B	34.95	31.18	36.08	38.22	34.36	39.37	63.58	65.93
HOICLIP (Ning et al., 2023)	R50/CLIP-B	34.69	31.12	35.74	37.61	34.47	38.54	63.50	64.80
PViC (Zhang et al., 2023)	Swin-L/H-DETR	44.32	44.61	44.24	47.81	48.38	47.64	64.10	70.20
CLIP4HOI (Mao et al., 2023)	R50/CLIP-B	35.33	33.95	35.74	37.19	35.27	37.77	–	66.30
UniHOI-s (Cao et al., 2023)	R50/BLIP-2	40.06	39.91	40.11	42.20	42.60	42.08	65.58	68.27
CMMP (Lei et al., 2024b)	R50/CLIP-L	38.14	37.75	38.25	–	–	–	–	64.00
BCOM (Wang et al., 2024)	R50/CLIP-L	39.34	39.90	39.17	42.24	42.86	42.05	65.80	69.90
LAIN (Kim et al., 2025)	R50/CLIP-B	36.02	35.70	36.11	–	–	–	–	65.10
HOLa (Lei et al., 2025a)	R50/CLIP-L	39.05	38.66	39.17	–	–	–	60.30	66.00
ADA-CM (FT) (Lei et al., 2023)	R50/CLIP-L	38.40	37.52	38.66	–	–	–	58.57	63.97
VRDiff (Cao et al., 2025)	R50/CLIP-L	41.40	41.69	41.31	44.25	44.65	44.13	–	–
HORP (Geng et al., 2025)	Swin-L/CLIP-B/32	47.53	46.81	47.74	51.24	50.78	**51.38**	68.90	71.10
VDRP (Yang et al., 2025b)	R50/CLIP-B	39.07	39.08	39.06	–	–	–	60.60	66.20
ContextHOI (Jia et al., 2025a)	R101/CLIP-L	42.09	42.41	41.99	–	–	–	67.30	–
InterProDa (Jia et al., 2025b)	R50/CLIP-L	42.67	45.21	41.92	–	–	–	67.60	–
BC-HOI (Hu et al., 2025)	R50/BLIP-2	43.01	45.76	42.18	–	–	–	68.20	70.61
NMSR (Yang et al., 2025a)	R50/BLIP-2	43.98	43.27	44.00	45.90	45.29	46.32	71.10	73.30
HOI-IDiff (Hui et al., 2025)	DDETR	47.71	48.36	47.52	50.56	51.95	50.14	73.40	76.10
InCoM-Net (Seo et al., 2026)	R50/CLIP-L	43.96	45.61	43.46	45.59	45.92	45.49	**73.60**	75.40
SL-HOI (Sun et al., 2026)	DINOv3-L	45.05	47.71	44.25	–	–	–	–	–
InstructHOI (Luo et al., 2025)	R101/ViT-L	47.68	47.97	47.59	49.89	50.92	49.58	72.40	76.10
HOI-R1 (Chen et al., 2025)	Qwen2.5-VL-3B	18.33	16.03	19.02	19.83	16.25	20.90	–	–
HOI-R1 (Chen et al., 2025)	Qwen3-VL-4B	23.35	23.58	23.28	25.12	24.11	25.42	–	–
HOI-R1 (Chen et al., 2025)	Rex-Omni-3B	25.24	22.57	26.04	27.00	23.21	28.14	–	–
DA-HOI (FT, DETR) (Xuan et al., 2026)	R50/Qwen2.5-VL-3B	44.58	46.17	44.08	–	–	–	69.40	71.70
DA-HOI (FT, G-DINO) (Xuan et al., 2026)	G-DINO/Qwen2.5-VL-3B	46.21	47.54	45.81	–	–	–	70.40	72.50
LINK (Wu et al., 2026a)	Swin-L/CLIP-L	**49.06**	**53.63**	47.60	**51.34**	**56.29**	49.86	–	69.20
FineHOI (Tonini et al., 2026)	DETR/DINOtxt-L/16	39.13	39.91	38.89	–	–	–	–	66.50
UniHOI (Yang et al., 2026)	VQGAN/Llama3-8B	48.16	50.74	**48.12**	**51.34**	53.72	50.33	72.91	**77.45**
Funnel-HOI (Sarma et al., 2026)	R50/CLIP-B/32	34.36	31.07	35.33	37.24	33.76	38.27	61.60	64.60
ScriptHOI (Nguyen et al., 2026)	–	37.10	37.80	36.90	40.00	41.40	39.60	67.80	69.40
HOIBlender-2XL (Chen & Yanai, 2026)	DINOv2-B/BLIP-2	44.49	43.21	44.86	47.45	47.76	47.36	–	–
Training-free methods									
BLIP-2 analysis (Ai, 2024)	G-DINO/BLIP-2-H@364	19.71	24.07	18.41	–	–	–	–	–
Qwen2.5-VL direct (Chen et al., 2025)	72B-AWQ	20.71	29.62	18.05	22.93	31.87	20.26	–	–
Qwen3-VL direct (Chen et al., 2025)	8B	17.41	21.96	16.05	19.17	23.39	17.91	–	–
LongCLIP retrieval (Tonini et al., 2025)	R50/LongCLIP-L	23.13	31.32	20.68	–	–	–	40.13	45.11
ADA-CM (TF) (Lei et al., 2023)	R50/CLIP-B	25.19	27.24	24.58	–	–	–	39.09	43.93
DYSCO (Tonini et al., 2025)	R50/LongCLIP-L	29.00	33.63	27.62	–	–	–	42.80	47.82
AgentHOI (Wang et al.) (Wang et al., 2026)	YOLOv8/BLIP/Qwen2.5	13.26	–	–	–	–	–	–	–
AgentHOI (Lei et al.) (Lei et al., 2026)	G-DINO/GPT-4o	29.61	40.33	26.41	–	–	–	–	–
DA-HOI (TF) (Xuan et al., 2026)	R50/Qwen2.5-VL-3B	31.50	–	–	–	–	–	–	–
HarnessHOI	Qwen3.5-9B	**33.20**	**41.63**	**30.68**	**36.85**	**44.20**	**34.65**	**44.65**	**49.70**

Table 11: Complete open-vocabulary comparison corresponding to Table 2, including additional methods and training-free baselines. Each partition reports Unseen, Seen, and Full mAP (%) using predicted boxes. HarnessHOI uses the same predictions across all four partitions. Within each method group, the best results are bold; the strongest previous training-free results are underlined. Dashes denote unavailable entries. CLIP baseline results are reported by LAIN.

	RF-UC			NF-UC			UO			UV		
Method	Unseen	Seen	Full	Unseen	Seen	Full	Unseen	Seen	Full	Unseen	Seen	Full
Task-optimized methods												
GEN-VLKT (Liao et al., 2022)	21.36	32.91	30.56	25.05	23.38	23.71	10.51	28.92	25.63	20.96	30.23	28.74
HOICLIP (Ning et al., 2023)	25.53	34.85	32.99	26.39	28.10	27.75	16.20	30.99	28.53	24.30	32.19	31.09
CLIP4HOI (Mao et al., 2023)	28.47	35.48	34.08	31.44	28.26	28.90	31.79	32.73	32.58	26.02	31.14	30.42
ADA-CM (FT) (Lei et al., 2023)	27.63	34.35	33.01	32.41	31.13	31.39	–	–	–	–	–	–
BCOM (Wang et al., 2024)	28.52	35.04	33.74	33.12	31.76	32.03	–	–	–	–	–	–
CMMP (ViT-L) (Lei et al., 2024b)	35.98	37.42	37.13	33.52	35.53	35.13	39.67	36.15	36.74	30.84	37.28	36.38
InterProDa (Jia et al., 2025b)	36.38	40.88	39.58	33.64	36.47	35.50	–	–	–	–	–	–
INP-CC (Lei et al., 2025b)	17.38	24.74	23.12	–	–	–	–	–	–	–	–	–
SGC-Net (Lin et al., 2025)	23.27	28.34	27.22	–	–	–	–	–	–	–	–	–
VDRP (ViT-L) (Yang et al., 2025b)	36.72	38.48	38.13	37.48	36.21	36.46	39.36	37.50	37.81	31.16	38.16	37.18
LAIN (ViT-L) (Kim et al., 2025)	36.57	38.54	38.13	37.52	35.90	36.22	40.78	36.96	37.60	32.05	38.04	37.20
EZ-HOI (ViT-L) (Lei et al., 2024a)	34.24	37.35	36.73	36.33	34.47	34.84	38.17	36.02	36.38	28.82	38.15	36.84
UniHOI-s (BLIP-2) (Cao et al., 2023)	28.68	33.16	32.27	28.45	32.63	31.79	19.72	34.76	31.56	26.05	36.78	34.68
HOLa (ViT-B) (Lei et al., 2025a)	30.61	35.08	34.19	35.25	31.64	32.36	36.45	33.02	33.59	27.91	35.09	34.09
NMSR (CLIP) (Yang et al., 2025a)	35.48	–	–	36.12	–	–	31.01	–	–	30.47	–	–
VRDiff (ViT-L) (Cao et al., 2025)	38.92	40.83	40.45	39.33	39.03	39.09	41.52	39.66	39.97	32.14	40.61	39.43
BC-HOI (BLIP-2) (Hu et al., 2025)	42.31	40.67	40.99	33.01	37.24	36.40	19.94	37.03	34.18	31.18	41.31	39.89
SD-IF (Xu et al., 2026)	28.01	30.09	29.63	–	–	–	–	–	–	–	–	–
InCoM-Net (Seo et al., 2026)	37.69	41.46	40.71	39.45	38.42	38.62	–	–	–	–	–	–
InstructHOI (Luo et al., 2025)	36.82	44.86	43.25	36.42	38.82	38.34	39.92	37.05	37.53	31.64	39.17	38.12
LINK (ViT-L) (Wu et al., 2026a)	38.51	40.33	39.97	34.63	35.67	35.43	38.24	37.61	37.91	27.22	38.36	36.88
FineHOI (Tonini et al., 2026)	37.79	38.55	38.39	38.11	36.30	36.58	40.13	35.92	36.62	35.58	38.51	38.10
Funnel-HOI (Sarma et al., 2026)	25.90	34.43	32.70	26.63	28.32	28.00	15.28	31.50	28.66	26.10	32.56	31.66
ImagineAgent (Yuan et al., 2026)	37.85	36.49	36.76	–	–	–	–	–	–	–	–	–
SL-HOI (Sun et al., 2026)	40.53	42.99	42.49	–	–	–	–	–	–	–	–	–
DA-HOI (FT, DETR) (Xuan et al., 2026)	41.79	44.01	43.56	**43.12**	39.63	40.33	48.67	42.58	43.60	**36.89**	43.84	42.88
DA-HOI (FT, G-DINO) (Xuan et al., 2026)	**43.30**	**45.19**	**44.81**	41.52	**41.51**	**41.51**	**50.15**	**44.31**	**45.28**	36.88	**45.66**	**44.43**
Training-free methods												
CLIP baseline (Kim et al., 2025)	28.79	22.00	23.36	28.52	22.06	23.36	28.66	22.29	23.36	26.16	22.90	23.36
BLIP-2 analysis (Ai, 2024)	23.96	18.65	19.71	19.97	19.65	19.71	22.71	19.12	19.71	21.96	19.35	19.71
ADA-CM (TF) (Lei et al., 2023)	26.83	24.54	25.00	30.11	23.16	24.55	–	–	–	–	–	–
DYSCO (Tonini et al., 2025)	30.36	23.96	25.24	27.56	24.58	25.18	–	–	–	–	–	–
AgentHOI (Lei et al.) (Lei et al., 2026)	38.64	–	–	28.04	–	–	33.70	–	–	29.85	–	–
DA-HOI (TF) (Xuan et al., 2026)	–	–	31.50	–	–	31.50	–	–	31.50	–	–	31.50
HarnessHOI	**40.15**	**31.46**	**33.20**	**33.64**	**33.09**	**33.20**	**35.45**	**32.75**	**33.20**	**33.87**	**33.09**	**33.20**

C.2 Models and Inference Resources

Table 12 compares the models, inference strategies, and additional resources used by related methods. HarnessHOI uses one frozen MLLM for participant acquisition, geometric adjudication, and relation verification. The public vocabulary and development protocol are specified in Appendix B.

Table 12: Resource and organization comparison. VisHarness provides a related visual-reasoning formulation; the remaining systems address HOI detection.

System	Capability source	Task organization	Additional resources
HOI-R1	SFT+GRPO-tuned MLLM	Joint interaction reasoning and localization	HOI instruction and reward supervision
DA-HOI (FT)	Adapted Qwen2.5-VL-3B	Spatial-aware pooling and deterministic matching	DETR or GroundingDINO; HOI supervision
DA-HOI (TF)	Frozen Qwen2.5-VL-3B	Candidate-conditioned deterministic generation	R50 DETR proposals
AgentHOI (Wang et al.)	BLIP and Qwen2.5	Hypothesis proposal, multi-agent debate, and consensus	YOLOv8
AgentHOI (Lei et al.)	GPT-4o or Qwen2.5-VL-72B	Recognition, re-mining, reassignment, grounding	GroundingDINO
DYSCO	CLIP/LongCLIP	Visual-semantic registry and scoring	R50 detector; target-domain visual support
VisHarness	Trained visual agent	Learned multi-turn expert selection	Visual experts; policy training
HarnessHOI	One frozen Qwen3.5-9B	Interaction-guided perception, geometric adjudication, and grounded reasoning	No auxiliary models or retrieved visual support sets

Agent-based HOI inference. AgentHOI by Wang et al. (Wang et al., 2026) combines hypothesis proposal with multi-agent debate, whereas AgentHOI by Lei et al. (Lei et al., 2026) combines multi-round interaction reasoning with instance-specific visual grounding. Table 10 includes both methods under regular evaluation, while Table 11 includes the open-vocabulary results reported by Lei et al. Wang et al. also report 14.58 mAP on V-COCO without an S1/S2 designation.

Task-optimized configurations. The extended tables include the learned interaction representations of FineHOI (Tonini et al., 2026), SGC-Net (Lin et al., 2025), Funnel-HOI (Sarma et al., 2026), ScriptHOI (Nguyen et al., 2026), and HOIBlender (Chen & Yanai, 2026), with their reported backbone configurations. UniHOI by Yang et al. (Yang et al., 2026) jointly learns detection and generation in a unified token space; UniHOI-s (Cao et al., 2023) denotes the earlier BLIP-2-based detector. Table 10 also includes the three SFT+GRPO backbones reported by HOI-R1 (Chen et al., 2025). MLLM distillation in CL-HOI (Gao et al., 2025) and generative evaluation in HOI-MLLM (Wu et al., 2026b) are presented with their respective metrics in Appendix C.6.

Tool-augmented interaction reasoning. ImagineAgent (Yuan et al., 2026) learns interaction reasoning through supervised fine-tuning and tool-augmented reinforcement learning, combining image cropping, retrieval, and generative modeling to augment its observations. Related visual-agent work learns to select heterogeneous visual experts across reasoning rounds (Fan et al., 2026). HarnessHOI uses the same frozen MLLM for visual exploration and relation verification, with geometric adjudication establishing the participant groundings shared across actions and roles.

C.3 COMMON-EVALUATOR COMPARISON

Tables 13 and 14 evaluate the released predictions of AgentHOI by Lei et al. and HarnessHOI with identical matching and AP computation, complementing the source-reported comparisons above. We evaluate the released AgentHOI (Lei et al., 2026) predictions with the authors' default score completion, sigmoid, box exponent 0.3, and threshold 0.1.[1] The release covers 9,546 test images, including 23 without boxes. The remaining 112 official test images receive empty predictions. Both methods are evaluated over the complete test set using identical inclusive-IoU, matching, and sorting conventions.

Table 13: Common-evaluator comparison (AP, %; differences in points, computed before rounding). AgentHOI (Lei et al.) uses GPT-4o and GroundingDINO; HarnessHOI uses Qwen3.5-9B.

HICO-DET metric	AgentHOI (Lei et al.)	HarnessHOI	Difference
Default Full	29.68	33.20	3.52
Default Rare	40.49	41.63	1.14
Default Non-Rare	26.45	30.68	4.23
Known Object Full	33.71	36.85	3.14
Known Object Rare	43.39	44.20	0.81
Known Object Non-Rare	30.82	34.65	3.83

Table 14: HICO-DET class-partition analysis (Default AP, %). The same HarnessHOI predictions are evaluated under four standard class partitions. Published and common-evaluator results of AgentHOI (Lei et al.) are shown separately. Bold identifies HarnessHOI's Unseen results.

	RF-UC	NF-UC	UO	UV
AgentHOI (Lei et al.), reported Unseen	38.64	28.04	33.70	29.85
AgentHOI (Lei et al.), common evaluator Unseen	39.11	28.56	33.97	30.13
HarnessHOI, Unseen	**40.15**	**33.64**	**35.45**	**33.87**
HarnessHOI, Seen	31.46	33.09	32.75	33.09
HarnessHOI, Full	33.20	33.20	33.20	33.20

HarnessHOI consistently outperforms AgentHOI under the same evaluator, with improvements on both Rare and Non-Rare classes under Default and Known Object evaluation. The gains also extend to Unseen classes in all four open-vocabulary settings (Table 14).

C.4 HICO-DET INTERACTION CLASSES

Positive interactions and no interaction. Table 15 reports results for the 520 positive-interaction and 80 no-interaction classes. Default mAP is 36.84% and 9.55%, respectively, giving a class-weighted Full mAP of 33.20%. The corresponding Known Object results are 39.79% and 17.73%. Positive actions and no-interaction labels are assessed through separate visual questions.

Frequency and evaluation domain. Rare and Non-Rare mAP average the classes within each frequency group, whereas coverage and true-positive counts aggregate individual annotations. Default and Known Object evaluation yield the same 22,161 true positives. Their AP differs because Known Object restricts the images considered for each class and therefore changes the false-positive distribution.

C.5 V-COCO ACTIONS AND SEMANTIC ROLES

Table 16 reports S1/S2 AP and matching statistics for all 25 roles. Of the 12,423 physical-role and 3,969 missing-role annotations, the predictions match 9,728 and 3,854, respectively. S1 and S2

[1] AgentHOI release identifier: `5599af3794f6b6006be156da353e54f9ca64f486`.

Table 15: HICO-DET interaction-type and frequency breakdowns. The upper and lower blocks each partition all 600 classes. GT and final TP counts are shared by Default and Known Object evaluation; AP is reported in percent.

Class group	Classes	GT	TP	Default	Known Object
Positive interactions	520	29,110	19,965	36.84	39.79
No interaction	80	4,295	2,196	9.55	17.73
Rare	138	981	598	41.63	44.20
Non-Rare	462	32,424	21,563	30.68	34.65

produce the same final true-positive counts in these results, while their missing-role criteria change the matching order and precision-recall curves. S2 exceeds S1 for every role, with small gaps for roles with few missing annotations such as surf, skateboard, and drink, and large gaps for roles with many missing annotations such as lay and eat/instr. The corresponding 24-role means, excluding `point`, are 46.09% and 51.07%; the 26-action agent mean is 75.58%.

Table 16: Complete V-COCO role results. GT and TP are annotation and final true-positive counts; physical and missing targets are separated. TP counts coincide under S1 and S2, while their score-ordered precision–recall curves differ. The last row sums counts and averages AP over 25 roles.

	GT		TP		AP (%)	
Action / role	Physical	Missing	Physical	Missing	S1	S2
hold / obj	2,733	875	2,234	854	42.95	52.39
sit / instr	1,089	827	747	786	38.93	47.88
ride / instr	505	51	428	47	63.13	63.40
look / obj	2,752	595	1,944	583	26.27	29.18
hit / instr	315	34	286	34	43.63	46.52
hit / obj	226	123	187	121	57.19	57.68
eat / obj	334	187	271	181	46.80	54.71
eat / instr	100	421	64	411	37.20	48.48
jump / instr	402	233	371	228	62.40	66.18
lay / instr	230	157	136	152	18.74	43.50
talk_on_phone / instr	246	39	176	39	45.71	49.00
carry / obj	433	39	347	36	26.17	27.29
throw / obj	214	30	194	30	50.01	52.25
catch / obj	201	45	187	44	32.42	32.70
cut / instr	230	39	139	39	18.86	21.49
cut / obj	203	66	159	64	49.20	58.92
work_on_computer / instr	375	35	304	33	50.37	53.68
ski / instr	360	64	249	64	39.58	46.82
surf / instr	464	22	408	22	73.10	73.60
skateboard / instr	402	15	378	15	80.28	80.78
drink / instr	78	4	71	4	40.99	41.49
kick / obj	163	17	131	17	46.78	47.28
point / instr	21	10	13	10	10.05	16.97
read / obj	86	25	71	24	43.83	58.55
snowboard / instr	261	16	233	16	71.64	71.84
Total / mean	12,423	3,969	9,728	3,854	44.65	49.70

Agent-level action recognition. Table 17 reports all 26 agent actions, including the four without target roles. HarnessHOI obtains 75.58% mean agent AP by assessing each action on the localized person. For role-bearing actions, the same support is combined with target visibility and binding to produce role predictions.

Table 17: V-COCO agent results for all 26 actions (AP, %). Agent detection assesses action occurrence and person localization. The mean AP is 75.58%.

Action	GT	TP	AP	Action	GT	TP	AP
hold	3,608	3,533	85.42	catch	246	243	59.84
stand	4,118	4,013	78.45	cut	269	260	83.41
sit	1,916	1,838	84.87	run	687	681	80.02
ride	556	523	80.74	work_on_computer	410	393	80.02
walk	597	566	76.98	ski	424	420	88.57
look	3,347	3,276	43.27	surf	486	484	92.99
hit	349	344	81.31	skateboard	417	417	94.04
eat	521	508	82.90	smile	1,415	1,392	84.45
jump	635	626	84.40	drink	82	82	58.05
lay	387	366	78.22	kick	180	180	65.29
talk_on_phone	285	285	90.84	point	31	29	28.09
carry	472	456	50.55	read	111	108	75.03
throw	244	242	68.87	snowboard	277	274	88.36

C.6 Results under Related Evaluation Protocols

We summarize related results on MLLM distillation, generative HOI prediction, and instruction-conditioned detection under their respective evaluation protocols.

Cross-level distillation results. Table 18 presents CL-HOI under full supervision and MLLM distillation, including its ResNet-50 and Swin-L configurations. The reported metrics are HICO-DET Full mAP and the authors' V-COCO Role AP. The distillation setting transfers image-level interaction knowledge from the MLLM to an instance-level detector.

Table 18: CL-HOI (Gao et al., 2025) results under full supervision and MLLM distillation. Values follow the authors' HICO-DET Full mAP and V-COCO Role AP reporting (%).

Supervision	Backbone	HICO-DET Full mAP	V-COCO Role AP
Full	R50+Transformer	34.10	65.44
MLLM	R50+Transformer	15.62	26.74
MLLM	Swin-L+Transformer	17.55	36.63

Generative interaction results. Table 19 reports HOI-MLLM with Qwen2.5-VL-3B using mean F1, precision, and recall. Its end-to-end configuration jointly predicts participants and interactions, the R50-DETR configuration supplies predicted object boxes, and Oracle supplies ground-truth boxes. These settings examine interaction reasoning with different sources of participant localization.

Instruction-conditioned detection. HOID-R1 by Zhang et al. (Zhang et al., 2025b) evaluates fine-grained, precise, and open-vocabulary instructions on a reannotated HICO-DET benchmark (Table 20). Its Seen and Unseen groups refer to image subsets. Human and object mIoU, action accuracy, and mAP measure localization and interaction recognition under these instruction settings.

Table 19: HOI-MLLM (Wu et al., 2026b) results with Qwen2.5-VL-3B under its mean F1, precision, and recall evaluation (%). End-to-end jointly predicts objects and interactions; R50-DETR supplies predicted object boxes; Oracle supplies ground-truth boxes.

	mF1			mPrec	mRec
Setting	Full	Rare	Non-Rare	Full	Full
HICO-DET					
End-to-end	24.28	25.75	23.27	26.72	26.26
R50-DETR	30.32	27.12	31.68	35.71	36.21
Oracle	50.61	51.16	49.90	51.37	56.47
V-COCO					
End-to-end	51.69	–	–	61.57	45.81
R50-DETR	69.28	–	–	74.44	65.56
Oracle	81.94	–	–	85.32	79.35

Table 20: HOID-R1 (Zhang et al., 2025b) results on its reannotated HICO-DET benchmark with fine-grained, precise, and open-vocabulary instructions. Seen and Unseen refer to image subsets. H-mIoU, O-mIoU, and A-Acc measure human localization, object localization, and action accuracy; mAP is reported in percent.

Annotation	Images	H-mIoU	O-mIoU	A-Acc	mAP
Fine-grained	Seen	0.72 ± 0.026	0.73 ± 0.028	0.79 ± 0.031	55.07 ± 2.11
	Unseen	0.71 ± 0.025	0.70 ± 0.027	0.75 ± 0.030	52.98 ± 2.18
Precise	Seen	0.70 ± 0.025	0.72 ± 0.027	0.76 ± 0.030	53.81 ± 2.13
	Unseen	0.71 ± 0.026	0.69 ± 0.025	0.72 ± 0.029	51.19 ± 2.21
Open-vocabulary	Seen	0.69 ± 0.024	0.71 ± 0.027	0.75 ± 0.032	51.68 ± 2.36
	Unseen	0.68 ± 0.023	0.68 ± 0.022	0.72 ± 0.030	50.03 ± 2.14

D Harness Design Analysis

We examine how interaction-guided perception, geometric adjudication, and grounded interaction reasoning contribute to HOI detection. Table 21 summarizes the experimental conditions used to evaluate each component.

D.1 Effectiveness of the Complete Harness

Table 3 compares direct generation, the initial-instance baseline, and the complete harness under the scoring rules in Appendix B.4. Grounded verification over the initial participants substantially improves detection over direct generation. Interaction-guided perception and geometric adjudication further improve results on both benchmarks.

D.2 Complete Ablation Results

Tables 22 and 23 report the complete HICO-DET and V-COCO metrics for the ablations examined in this section.

Table 24 reports the complete ablations under RF-UC, NF-UC, UO, and UV. Each variant is evaluated across the four class partitions using its predictions from regular evaluation. HarnessHOI obtains Unseen mAP of 40.15%, 33.64%, 35.45%, and 33.87%, respectively. Its advantage over direct generation extends across combinations, object categories, and actions. The pair and label ablations examine the contribution of candidate selection and compatible-action verification within these class groups.

Table 21: Design of the controlled comparisons on the complete official test sets. Each comparison varies the indicated aspect of inference while retaining the common conditions listed below.

Comparison	Aspect examined	Common conditions
Overall inference	Direct generation, initial-instance baseline, and the complete harness	Backbone, full-image processing, public vocabulary, and evaluation
Perceptual components	Discovery participants, local observation, and geometric selection	Common initial localization; relation questions evaluated on each resulting participant set
Interaction guidance	Discovery cues, observation location, geometry criterion, or proposal context	Inputs and budgets specified for each replacement in Appendix D
Relation candidates	Pair sources and labels on the proposed pairs	Selected participants, image views, and relation supports
Role verification	Joint questions and compositional confidence	Physical and NULL candidates and participant geometry; joint questions use the corresponding binding or role-state views
Repeated generation	Response count and duplicate fusion	Ordered response prefixes and original generation scores
Generation stability	Generation seed	Backbone, inference policy, and evaluation population
Backbone generality	Different frozen MLLM backbones	Harness configuration, public vocabulary, seed 66, and evaluation

Table 22: Complete HICO-DET harness comparisons under Default and Known Object evaluation (mAP, %). All variants use the same backbone and evaluator.

	Default			Known Object		
Variant	Full	Rare	Non-Rare	Full	Rare	Non-Rare
HarnessHOI	33.20	41.63	30.68	36.85	44.20	34.65
Direct HOI generation	15.62	22.13	13.68	17.27	23.25	15.49
Initial-instance baseline	31.45	39.79	28.96	35.05	41.30	33.18
w/o discovered participants	32.55	40.91	30.05	36.15	42.42	34.28
w/o targeted observation	32.80	41.24	30.28	36.39	42.62	34.53
First-candidate geometry	32.95	41.30	30.46	36.55	42.89	34.65
Proposed pairs, proposed labels	18.35	21.08	17.53	20.39	22.13	19.87
Proposed pairs, all compatible labels	29.85	34.19	28.55	33.09	36.03	32.21
Visual pair assessment only	32.30	40.58	29.83	35.85	41.90	34.04
Discovery and observation only	28.09	32.41	26.80	31.20	34.26	30.28
Binary relation readout	33.05	41.37	30.56	36.70	42.89	34.85
Repeated participant localization	32.45	40.84	29.95	36.05	42.36	34.16
Non-directed observation	32.70	41.17	30.17	36.30	42.49	34.45
Relation-conditioned adjudication	32.79	41.11	30.30	36.38	42.62	34.51
Proposal-conditioned verification	32.50	40.84	30.01	36.09	42.36	34.22

D.3 Interaction-Guided Perception

Interaction-guided perception uses action and role hypotheses to discover relevant participants and select regions for closer observation. We first evaluate participant discovery and local observation, then compare them with repeated localization and non-directed observation to examine the contribution of interaction guidance.

Participant discovery and targeted observation. The discovered-participant ablation excludes identities introduced by discovery and performs observation and relation assessment over the remaining participants. The observation ablation retains global discovery and its single localization per instance, omitting local observation and the corresponding geometric adjudication. These component configurations are reported in Table 3, with full metrics in Tables 22 and 23.

Table 23: Complete V-COCO harness comparisons corresponding to Table 22, reporting role AP over 25 roles under S1 and S2 (%).

Variant	S1	S2
HarnessHOI	44.65	49.70
Direct HOI generation	16.41	21.20
Initial-instance baseline	42.90	47.20
w/o discovered participants	44.02	48.93
w/o targeted observation	44.33	49.21
First-candidate geometry	44.38	49.48
Proposed pairs, proposed labels	36.43	39.81
Proposed pairs, all compatible labels	42.27	47.14
Visual pair assessment only	43.23	47.77
Discovery and observation only	42.48	47.19
Binary relation readout	44.52	49.57
w/o action factor	5.83	6.57
w/o visibility factor	44.13	49.55
w/o NULL predictions	39.20	49.43
Joint action-and-role question	41.10	46.30
Repeated participant localization	43.87	48.88
Non-directed observation	44.07	49.08
Relation-conditioned adjudication	44.18	49.16
Proposal-conditioned verification	43.92	48.84

Table 24: Complete HICO-DET class-partition ablations (Default mAP, %). All partitions use the same predictions for each variant; Full values appear in Table 22.

	RF-UC		NF-UC	
Variant	Unseen	Seen	Unseen	Seen
HarnessHOI	40.15	31.46	33.64	33.09
Direct generation	21.82	14.07	12.80	16.33
Initial-instance baseline	38.33	29.73	31.59	31.42
w/o discovered participants	39.17	30.89	33.26	32.37
w/o targeted observation	40.00	31.00	32.88	32.78
First-candidate geometry	39.85	31.23	33.48	32.82
Proposed pairs, proposed labels	23.33	17.10	18.41	18.33
Proposed pairs, all compatible labels	36.52	28.18	30.61	29.66
Visual pair assessment only	39.39	30.53	32.50	32.25
Discovery and observation only	35.08	26.34	28.71	27.94
Binary relation readout	40.00	31.31	33.48	32.94

	UO		UV	
Variant	Unseen	Seen	Unseen	Seen
HarnessHOI	35.45	32.75	33.87	33.09
Direct generation	17.09	15.33	19.26	15.03
Initial-instance baseline	33.91	30.96	32.03	31.36
w/o discovered participants	34.55	32.15	33.33	32.42
w/o targeted observation	35.27	32.31	33.12	32.75
First-candidate geometry	34.91	32.56	33.33	32.89
Proposed pairs, proposed labels	20.73	17.87	18.61	18.31
Proposed pairs, all compatible labels	31.91	29.44	30.63	29.72
Visual pair assessment only	34.82	31.80	32.79	32.22
Discovery and observation only	30.27	27.65	29.11	27.92
Binary relation readout	35.18	32.62	33.66	32.95

Guidance for participant discovery. The repeated-localization control receives the same full image, initial participant identities, and designated human groups as interaction-guided discovery.

Its question directly requests additional visible participants associated with those humans. Both conditions use the same generation budget and the observation requests produced by interaction-guided discovery, followed by the same observation, geometric adjudication, and relation verification. Interaction-guided discovery improves detection on both benchmarks (Table 4), supporting the use of interaction semantics to direct participant search.

Selection of local observations. To evaluate region selection, the non-directed control retains the discovery output, target descriptions, observation count, and crop dimensions. Each crop center is sampled uniformly from valid image positions using a fixed seed. Both conditions use the same image processing and localization instructions, followed by geometric and relational assessment on their acquired observations. The gains from targeted observation in Table 4 demonstrate the value of selecting regions according to the interaction hypothesis.

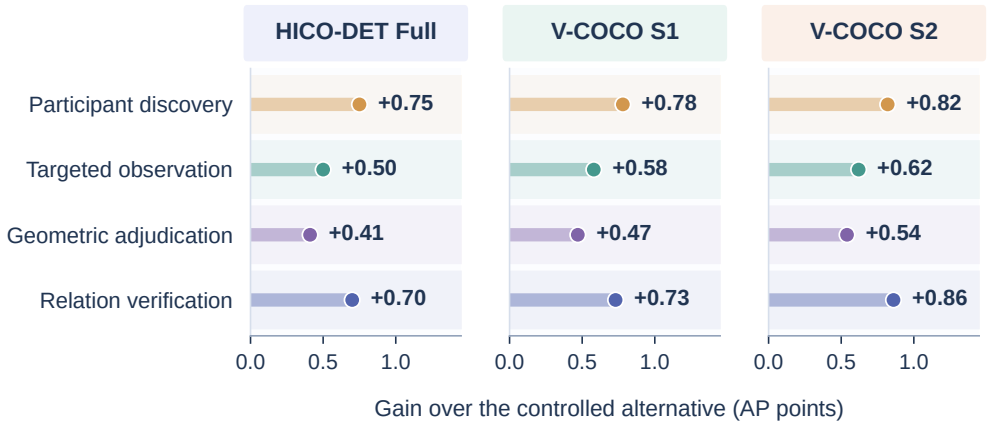

Figure 3: **Interaction guidance and visual judgment.** Each point reports the AP gain of the complete harness over the corresponding controlled alternative in Table 4: repeated localization, non-directed observation, relation-conditioned adjudication, and proposal-conditioned verification. The three panels use the same scale.

Stage-wise participant coverage. We measure grounding coverage and relation-candidate coverage to evaluate participant acquisition at successive stages. Grounding coverage requires participant boxes that match an annotation at IoU 0.5, while candidate coverage additionally requires the corresponding action or role label. Each annotation is counted once.

Initial localization and discovery are evaluated on their respective participant sets. The all-localization condition uses the accumulated grounding candidates, and the selected-grounding condition uses one adjudicated candidate per identity. Physical interactions require matching human and target instances with distinct identities; HICO-DET also requires the target category to match. Missing roles require a matching person, together with the appropriate NULL prediction for candidate coverage.

As shown in Table 25, discovery adds matching groundings for 571 HICO-DET interactions and 217 V-COCO roles. Local observation extends coverage by a further 12 and 23 annotations, respectively. The final selected groundings cover 22,895 HICO-DET interactions and 14,473 V-COCO roles, providing the participant pairs for subsequent verification.

Interaction configurations. We group the interaction annotations by participant and action structure. The multiple-human-box group contains images with at least two distinct annotated human boxes. The HICO-DET multiple-target group contains annotations whose action has at least two distinct physical targets within the image. For V-COCO, the corresponding multiple-target-state group treats each distinct physical target and NULL as target states, while reporting physical-role annotations. Small targets occupy at most 1% of the image area. The multiple-action group contains

Table 25: Coverage of annotated interactions through participant construction and relation assessment. Grounding coverage requires matching participant boxes at IoU 0.5; the final candidate domain additionally requires the action or role label. V-COCO includes 3,969 missing-role annotations, for which participant grounding requires the person alone.

Participant and relation domain	HICO-DET	V-COCO
Initial participant groundings	22,314	14,248
After interaction discovery	22,885	14,465
All localization candidates	22,897	14,488
Selected shared groundings	22,895	14,473
Final action or role candidates	22,232	13,628
Annotated interactions or roles	33,405	16,392

human-target pairs annotated with at least two actions, including NULL targets in V-COCO. Groups may overlap, and their sizes count annotations.

Missing-role annotations form a separate V-COCO group and contribute to the full totals. Their grounding and candidate coverage follow the person and NULL criteria above. The small-target group includes physical targets alone.

Table 26: Interaction coverage in the final action or role candidate domain. A ground-truth annotation is covered when at least one candidate has the required participant localization and interaction label. The initial-instance baseline verifies relations over the initially localized participants. Small targets occupy at most 1% of the image area. Scene groups can overlap; missing-role coverage requires the person and the corresponding NULL role candidate.

Interaction group	Annotations	Initial-instance baseline	HarnessHOI
HICO-DET			
All interactions	33,405	21,354	22,232
Multiple annotated human boxes	30,441	19,498	20,288
Multiple targets for the same action	11,549	6,172	6,437
Small targets	5,960	2,184	2,356
V-COCO			
All roles	16,392	12,703	13,628
Multiple annotated human boxes	8,183	5,909	6,553
Multiple target states per action	4,468	3,026	3,375
Small targets	4,343	2,761	3,208
Multiple actions per human–target state	10,072	8,375	8,909
Missing role targets	3,969	3,863	3,863

Coverage in complex scenes. The full harness covers 878 additional HICO-DET interactions and 925 additional V-COCO roles over the initial-instance baseline (Table 26 and Figure 4). These gains include 172 and 447 annotations with small targets, respectively, and 534 V-COCO roles involving multiple actions. Table 27 reports the corresponding stage-wise coverage. For V-COCO small targets, coverage increases from 3,407 after initial localization to 3,518 after discovery and 3,529 after observation, with 3,208 annotations represented in the final relation candidates.

The gains on small targets demonstrate the effectiveness of interaction-guided perception in recovering visually inconspicuous participants. Improvements on multiple-action pairs further show that the acquired groundings support several interactions involving the same entities.

Table 27: Stage-wise coverage within complex-scene groups. Initial, Discovery, and Observed count annotations with matching participant geometry in the cumulative candidate sets; Selected uses adjudicated boxes; Relations additionally requires the correct action or role candidate. Counts use IoU 0.5 and overlapping annotation-defined groups.

Group	GT	Initial	Discovery	Observed	Selected	Relations
HICO-DET						
All interactions	33,405	22,314	22,885	22,897	22,895	22,232
Multiple annotated human boxes	30,441	20,323	20,848	20,858	20,856	20,288
Multiple targets per action	11,549	6,578	6,728	6,733	6,731	6,437
Small targets	5,960	2,387	2,489	2,493	2,492	2,356
V-COCO						
All roles	16,392	14,248	14,465	14,488	14,473	13,628
Missing roles	3,969	3,863	3,863	3,863	3,863	3,863
Multiple actions per human–target state	10,072	8,939	9,086	9,102	9,088	8,909
Multiple annotated human boxes	8,183	7,100	7,213	7,219	7,214	6,553
Multiple target states per action	4,468	3,663	3,756	3,761	3,759	3,375
Small targets	4,343	3,407	3,518	3,529	3,522	3,208

D.4 Relation-Agnostic Geometric Adjudication

We evaluate geometric adjudication by retaining discovery and observation and replacing visual selection with the first box associated with each identity (Table 3). This comparison examines the contribution of selecting a shared participant grounding from the acquired localization candidates.

On HICO-DET and V-COCO, 479 and 424 identities adopt an observation-derived grounding, respectively. The selected grounding is shared across all relations involving the participant. The improvement over first-candidate geometry in Table 3 confirms the effectiveness of visual adjudication for interaction grounding.

Criterion for shared grounding. Relation-conditioned and relation-agnostic adjudication receive identical candidate boxes and image views for each instance. Relation-conditioned adjudication adds the earliest discovery or observation association involving that instance and asks which candidate best localizes the participant in that interaction. Relation-agnostic adjudication uses the identity, category, and visible extent criterion in Section 3.3. Instances without a recorded association use the common identity-based criterion. Both conditions select one shared grounding per instance and apply the same subsequent relation verification.

With candidate boxes and image views fixed, relation-agnostic adjudication improves detection on both benchmarks (Table 4). The results support selecting shared groundings according to participant identity, category, and visible extent, independently of the proposed relation.

D.5 Grounded Interaction Reasoning

We examine grounded interaction reasoning through complementary pair selection, multi-label verification, and compositional semantic-role assessment. These comparisons evaluate which participant pairs and relations are considered and how their visual evidence determines the final predictions.

Context for relation verification. The proposal-conditioned control appends the original discovery proposals associated with the queried participants to the final action, agent, role-state, and binding questions. The harness condition uses the selected participants, public relation definitions, and image evidence. Pair retention, candidate relations, boxes, views, answer labels, and score composition are identical. Both conditions assess the full candidate domain, including relations absent

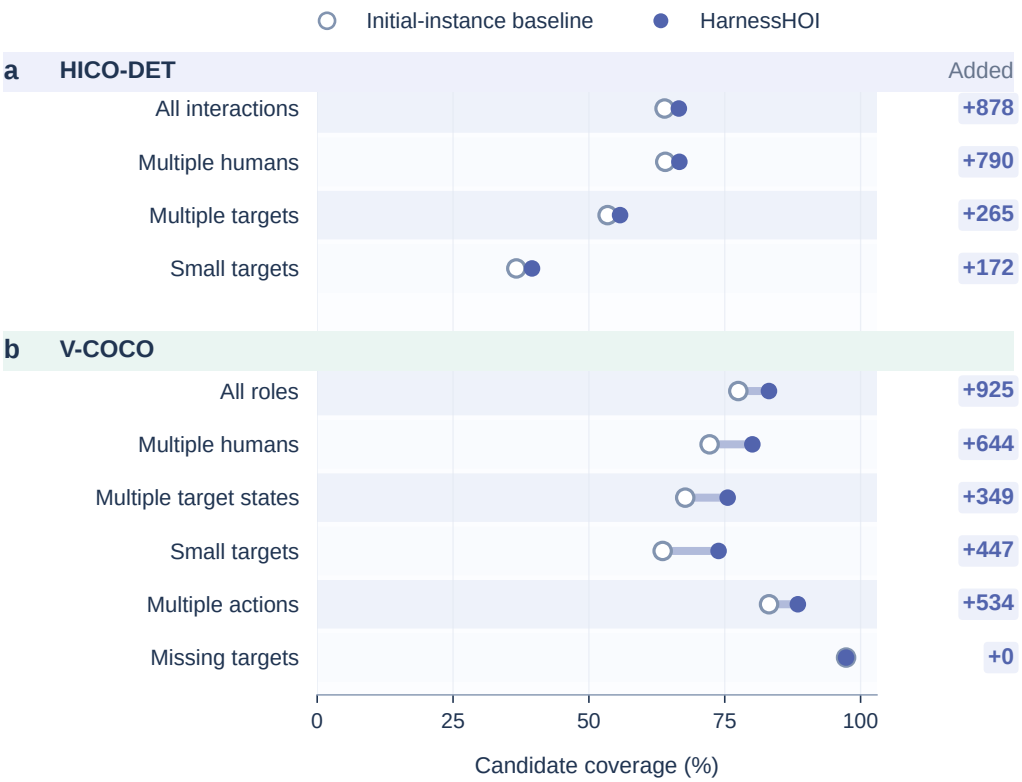

Figure 4: **Candidate coverage across interaction configurations.** Open and filled markers compare the initial-instance baseline and the complete harness. Coverage is normalized by the number of annotations in each group; the right column reports additional covered annotations. Groups can overlap. The underlying counts and group definitions appear in Table 26.

from the discovery proposals. Grounded verification improves detection over proposal-conditioned questions on these identical candidates (Table 4), demonstrating the benefit of assessing relations directly from the acquired visual evidence.

Table 28: Analysis of grounded interaction reasoning. HICO-DET reports Default Full mAP; V-COCO reports S1/S2 role AP (%). Panel (a) varies pair retention with fixed visual assessments; panel (b) varies the retained labels on identical proposed pairs with fixed scores; panel (c) compares joint and compositional verification on the same role candidates, with each joint question receiving the views of its corresponding binding or role-state question. Best results within each panel are bold.

	HICO-DET	V-COCO	
Variant	Full	S1	S2
(a) Sources of relation candidates			
Discovery and observation	28.09	42.48	47.19
Visual pair assessment	32.30	43.23	47.77
Combined	**33.20**	**44.65**	**49.70**
(b) Labels retained on identical proposed pairs			
Proposed labels	18.35	36.43	39.81
All compatible labels	**29.85**	**42.27**	**47.14**
(c) Semantic-role verification			
Joint action-and-role question	–	41.10	46.30
Compositional verification	–	**44.65**	**49.70**

Complementary relation candidates. Table 28(a) compares visual pair assessment, discovery and observation associations $\mathcal{P}_D \cup \mathcal{P}_O$, and their union $\mathcal{K}$. Each configuration retains the original relation scores on its selected pairs, together with HICO-DET no-interaction and V-COCO agent and NULL predictions. Combining the two sources achieves the best results on both benchmarks, improving HICO-DET mAP by 5.11 points over discovery and observation associations and V-COCO AP by 1.42/1.93 points over visual pair assessment. The gains demonstrate that associations established during perception complement the relations identified by visual pair assessment.

Coverage and selective verification. The final candidates cover 22,232 HICO-DET interactions and 13,628 V-COCO roles, retaining 97.1% and 94.2% of the annotations covered by the selected groundings. Positive-action and role-binding verification uses 39,791 of 416,280 human-entity pairs on HICO-DET and 17,804 of 251,376 on V-COCO. Thus, pair selection reduces the respective domains by 90.4% and 92.9% while retaining high coverage of the grounded interactions.

Coverage measures the existence of matching candidates. Detection evaluation additionally assigns each prediction to its best-matching annotation and removes duplicates in score order. A prediction overlapping several annotations contributes at most one true positive, so final matching counts can differ from candidate coverage. The ranking bounds in Appendix E.4 follow the evaluator's assignment rule.

Multi-label interaction verification. We evaluate label expansion on the pairs represented by explicitly proposed and scored relations, denoted by $\mathcal{P}_{\text{prop}}$. The proposed-label condition retains these relations; the compatible-label condition includes all assessed category-compatible positive actions in HICO-DET or all 25 action-role categories in V-COCO on the same pairs. An observation association enters $\mathcal{P}_O$, whereas membership in $\mathcal{P}_{\text{prop}}$ requires a scored relation proposal. Geometry, image evidence, scores, and all no-interaction, agent, and NULL predictions remain fixed.

As shown in Table 28(b), label expansion increases HICO-DET mAP from 18.35% to 29.85% and V-COCO AP from 36.43%/39.81% to 42.27%/47.14%. The gains demonstrate the importance of verifying multiple compatible relations over shared participants, extending prediction beyond the labels proposed during discovery.

Answer support and role factors. Role visibility and binding are assessed conditionally on the action, while q_A supplies the evidence that the person performs it. For a given person, action, and role, physical candidates share $q_A q_V$ and are ordered by their binding support q_B. Across people and images, action and visibility also contribute to the confidence ranking used for role AP. Table 23 examines these factors and the treatment of unresolved and missing targets.

The binary-readout ablation removes unknown support and renormalizes false/true or missing/visible within each final question, preserving pair retention and score composition. The action-factor ablation uses $q_V q_B$ for physical targets and q_M for NULL targets. The visibility-factor ablation uses $q_A q_B$ for physical targets and preserves NULL scores. Both retain the original agent predictions. The NULL ablation removes missing-target predictions while retaining physical-role and agent outputs. All four conditions are derived from the full method's answer supports.

Compositional role verification. We compare compositional verification with a joint action-and-role question on identical participants, geometry, and role candidates. Each physical-target joint question uses the same full-image and crop inputs as its corresponding binding question; each NULL joint question uses those of its corresponding role-state question. The joint question evaluates action occurrence together with the target's role; its NULL formulation evaluates action occurrence together with the absence of a visible role target. True-label support directly scores each candidate, with agent predictions unchanged.

Compositional verification improves V-COCO S1/S2 AP from 41.10%/46.30% to 44.65%/49.70% (Table 28(c)), with gains on 15 of 25 roles in S1 and 20 of 25 in S2. The final true-positive counts remain unchanged, showing that the improvement comes from more accurate ranking of role predictions. Image-paired bootstrap with 1,000 resamples gives 95% percentile intervals of [3.02, 4.10] and [2.85, 3.96] points for the S1 and S2 gains.

Table 29: Joint and compositional role verification on identical candidates (AP, %). Δ denotes compositional minus joint AP in points. Both formulations have the same final true-positive count for every role.

	S1			S2		
Action / role	Joint	Compositional	Δ	Joint	Compositional	Δ
hold / obj	46.08	42.95	-3.13	52.08	52.39	+0.31
sit / instr	38.97	38.93	-0.04	45.02	47.88	+2.86
ride / instr	56.50	63.13	+6.63	57.05	63.40	+6.35
look / obj	31.49	26.27	-5.22	35.52	29.18	-6.34
hit / instr	27.68	43.63	+15.95	30.64	46.52	+15.88
hit / obj	30.73	57.19	+26.46	31.48	57.68	+26.20
eat / obj	45.98	46.80	+0.82	54.14	54.71	+0.57
eat / instr	10.91	37.20	+26.29	53.88	48.48	-5.40
jump / instr	57.87	62.40	+4.53	64.18	66.18	+2.00
lay / instr	22.76	18.74	-4.02	40.67	43.50	+2.83
talk_on_phone / instr	46.79	45.71	-1.08	49.43	49.00	-0.43
carry / obj	25.85	26.17	+0.32	26.31	27.29	+0.98
throw / obj	46.92	50.01	+3.09	48.35	52.25	+3.90
catch / obj	31.30	32.42	+1.12	32.69	32.70	+0.01
cut / instr	7.89	18.86	+10.97	9.43	21.49	+12.06
cut / obj	49.12	49.20	+0.08	56.87	58.92	+2.05
work_on_computer / instr	50.41	50.37	-0.04	52.07	53.68	+1.61
ski / instr	39.84	39.58	-0.26	43.22	46.82	+3.60
surf / instr	67.33	73.10	+5.77	67.90	73.60	+5.70
skateboard / instr	81.07	80.28	-0.79	81.51	80.78	-0.73
drink / instr	41.58	40.99	-0.59	42.14	41.49	-0.65
kick / obj	45.16	46.78	+1.62	45.61	47.28	+1.67
point / instr	7.13	10.05	+2.92	11.40	16.97	+5.57
read / obj	47.62	43.83	-3.79	54.86	58.55	+3.69
snowboard / instr	70.49	71.64	+1.15	71.05	71.84	+0.79
Mean	41.10	44.65	+3.55	46.30	49.70	+3.40

Role-specific confidence composition. Table 29 and Figure 5 report the comparison for all 25 roles. Compositional verification improves both roles of `hit` and the instrument role of `cut` under S1 and S2, with a 26.29-point gain for `eat/instr` under S1. The gains demonstrate the effectiveness of combining shared action evidence with the visibility and binding evidence of each role target.

Image-paired resampling. For the joint and compositional role comparison, we draw 1,000 paired bootstrap samples of the test images with replacement (Efron, 1979), using seed 20260920. Each sample contains the original number of images and assigns identical image multiplicities to both methods' predictions and annotations. We recompute AP under the original matching rules and class denominator, assigning zero to classes absent from a sample. The 2.5th and 97.5th percentiles of the paired AP differences form the confidence interval for each scenario.

Missing role targets. Explicit missing-role assessment represents actions whose targets are not visible in the image. Removing NULL predictions reduces S1 AP from 44.65% to 39.20%, while the complete method obtains 49.70% under S2 (Table 23). This comparison demonstrates the importance of distinguishing a missing role target from an absent action.

D.6 Backbone Generality

We apply the identical harness configuration, public vocabulary, and evaluator to Qwen2.5-VL-7B (Bai et al., 2025) and InternVL3-8B (Zhu et al., 2025) on the complete test sets under generation seed 66. Table 30 reports the complete-harness results for each backbone. InternVL3-8B obtains 32.11% HICO-DET Full mAP and 43.60%/48.40% V-COCO S1/S2 AP, while Qwen2.5-VL-7B obtains 31.61% and 43.10%/47.90%, respectively. The same three-stage harness supports

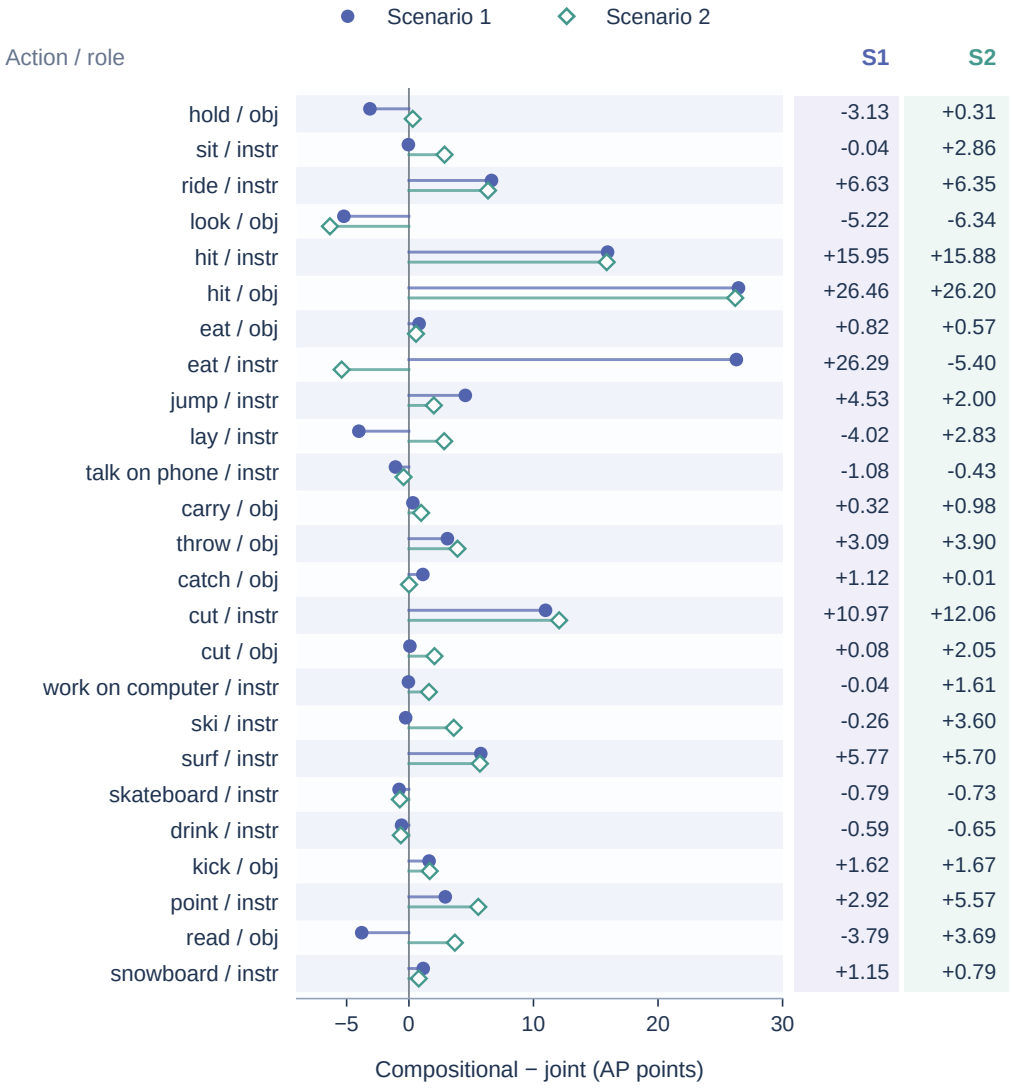

Figure 5: **Compositional verification across semantic roles.** AP differences between compositional and joint verification for all 25 V-COCO roles on the same candidates. Each joint question receives the views of its corresponding binding or role-state question. Filled circles and open diamonds denote S1 and S2 on a common scale. Values in the right columns correspond to Table 29.

Table 30: Backbone generality on the complete test sets under generation seed 66. Each backbone uses the identical harness configuration, public vocabulary, and evaluator. HICO-DET reports Default Full mAP; V-COCO reports S1/S2 role AP (%).

Backbone	HICO-DET Full	V-COCO S1	V-COCO S2
Qwen3.5-9B	33.20	44.65	49.70
InternVL3-8B	32.11	43.60	48.40
Qwen2.5-VL-7B	31.61	43.10	47.90

action-triplet and semantic-role prediction across these backbones, demonstrating its applicability to different MLLMs.

E Inference Computation and Stability

We analyze the computational cost and generation stability of HarnessHOI. Comparisons with repeated direct generation examine the benefit of coordinating perception and reasoning, while ground-truth-assisted ranking separates candidate quality from confidence estimation.

E.1 Computational Structure

Perceptual generation comprises participant localization, interaction discovery, and local observation. Discovery groups share the initially localized participants and can be processed concurrently. Let H_0 be the number of initially localized humans and R the number of local observations. With four humans per discovery group, the number of generative responses is

$$N_{\text{gen}} = 1 + \left\lceil \frac{H_0}{4} \right\rceil + R, \qquad 0 \le R \le 2. \tag{14}$$

For C distinct crop windows, the visual context comprises $1+C$ views, with $C \le R$. Their encoded representations are shared across the associated generation and verification requests.

Candidate verification evaluates independent questions through first-token answer scores. Geometric adjudication selects the participant groundings used in pair assessment and action or role verification. Independent questions within each stage are processed with continuous batching. Let $H = |\mathcal{H}^*|$, $N = |\mathcal{E}^*|$, $P = H(N-1)$, $K = |\mathcal{K}|$, and let G count identities with multiple geometry candidates. The number of scoring requests on HICO-DET is

$$N_{\text{score}}^{\text{HICO}} = G + 2P + \sum_{(h,o)\in\mathcal{K}} |\mathcal{A}^+(c_o^*)|. \tag{15}$$

The two pair terms correspond to interaction assessment and no-interaction assessment. Positive-action verification covers the category-compatible actions of each retained pair. On V-COCO, the scoring workload is

$$N_{\text{score}}^{\text{VCOCO}} = G + 51H + P + 25K. \tag{16}$$

Agent-action and role-state assessment contribute $26H$ and $25H$ requests, respectively; each retained pair contributes 25 role-binding requests. In both benchmarks, the verification workload depends on the number of grounded participants and retained relation candidates.

E.2 Measured Computation

Table 31 reports the per-image computation and hardware cost of the harness. Perceptual generation requires an average of 2.51 and 2.78 responses per image on HICO-DET and V-COCO, respectively, with a median of two on both benchmarks. Candidate verification processes independent scoring questions in batches. Table 32 reports their mean, median, and 95th percentile for each type of question, reflecting the variation in candidate-set size across images.

Hardware cost. Hardware cost is measured independently as amortized GPU-worker time, computed by summing the recorded processing time across GPU workers and dividing by the number of images. The measurement includes input preparation, model inference, and result collection, with all recorded inference work counted once; model loading and evaluation are excluded. On HICO-DET and V-COCO, the measured costs are 13.26 and 22.07 GPU-worker seconds per image, respectively.

Execution settings. Hardware measurements use frozen Qwen3.5-9B in BF16 on NVIDIA A6000 GPUs, with one model replica per GPU, four active images per replica, and up to 32 concurrent requests per GPU. The harness measurement uses four GPUs. Continuous batching schedules generation and scoring requests, while shared visual representations and prefix caching reuse the common image and prompt context. The cache hit rates in Table 31 measure the fraction of scoring input tokens served from the prefix cache.

Table 31: Per-image inference workload and hardware cost. Generative responses and independent scoring requests are reported separately. Output tokens refer to perceptual generation; P95 denotes the 95th percentile over images.

Quantity	HICO-DET	V-COCO
Perceptual generation		
Generative responses (mean)	2.51	2.78
Generative responses (median)	2	2
Generative responses (P95)	5	6
Output tokens (mean)	563.29	615.26
Local observations (mean)	0.19	0.40
Distinct image views (mean)	1.18	1.38
Candidate verification		
Scoring requests (mean)	119.81	311.27
Scoring requests (median)	28	158
Scoring requests (P95)	470	1,011
Hardware measurements		
GPU-worker seconds / image	13.26	22.07
Scoring input cache hits (%)	88.43	89.93

Table 32: Mean independent scoring requests per image. Each request evaluates a candidate question through first-token answer scores, with independent requests processed in continuous batches.

Scoring task	HICO-DET	V-COCO
Geometry selection	0.17	0.33
Pair assessment	43.10	50.82
No-interaction assessment	43.10	–
Positive-action verification	33.44	–
Agent assessment	–	86.73
Role visibility	–	83.40
Role binding	–	89.99

Ablation computation. Pair, label, and factor ablations use the same relation supports to evaluate alternative candidate sets and score compositions. The V-COCO joint action-and-role control evaluates new questions with the view assignment specified in Appendix D.5. Perceptual ablations use the participants and views produced by each configuration. Shared computations are counted once when measuring a complete inference configuration.

E.3 Allocating Inference Computation

We compare the harness with repeated direct generation to examine how coordinated perception and reasoning affect detection performance and computational cost. Both methods use the same frozen backbone, full-image processing, sampling settings, and initial output budget. Repeated generation independently predicts participants and relations in each response, whereas the harness uses interaction hypotheses to acquire visual evidence for grounded verification.

Repeated generation and fusion. We combine the first $K \in \{1, 2, 4, 8\}$ responses in a predetermined order using two fusion rules. Exact-match fusion retains the maximum score among predictions with identical categories, roles, target states, and boxes. Pair-IoU suppression removes lower-scored duplicates at threshold 0.5 within the same category, role, and target state. The hardware cost at each K sums the recorded GPU-worker processing time of the constituent generation runs per image. Tables 33 and 34 report both rules on the same response sets.

Performance and computation. Figure 6 presents detection performance alongside hardware cost. With pair-IoU suppression, increasing the number of responses from one to eight raises HICO-

Table 33: Detection performance and measured hardware cost. HICO-DET reports Default Full mAP and V-COCO reports S1/S2 role AP (%). Cost is amortized GPU-worker seconds per image. Repeated generation uses pair-IoU suppression; Table 34 compares both fusion rules.

	HICO-DET		V-COCO		
Method	Full	Cost	S1	S2	Cost
Direct generation, $K = 1$	15.61	2.50	16.38	21.18	2.60
Direct generation, $K = 2$	17.62	4.98	19.39	24.19	5.22
Direct generation, $K = 4$	19.04	9.97	20.86	25.54	10.50
Direct generation, $K = 8$	19.84	19.95	22.21	26.50	20.94
HarnessHOI	33.20	13.26	44.65	49.70	22.07

Table 34: Prediction fusion for repeated direct generation (AP, %). Exact-match fusion retains the maximum score among identical detections. Pair-IoU suppression uses threshold 0.5 within matching categories, roles, and target states. Both rules use the same fixed response prefixes.

Fusion	Responses	HICO-DET Full	V-COCO S1	V-COCO S2
Exact-match fusion	1	15.62	16.41	21.20
	2	14.53	13.31	16.55
	4	12.78	9.43	11.61
	8	10.87	6.51	7.84
Pair-IoU suppression	1	15.61	16.38	21.18
	2	17.62	19.39	24.19
	4	19.04	20.86	25.54
	8	19.84	22.21	26.50

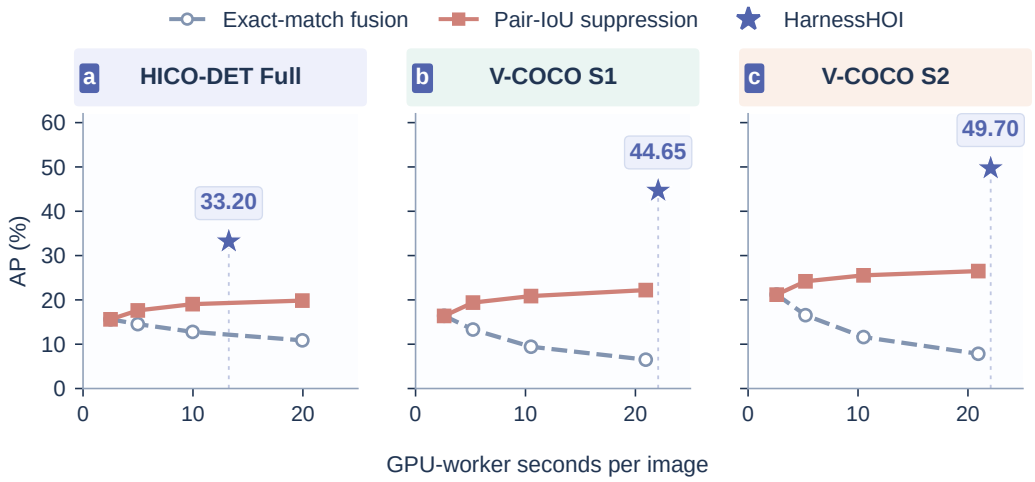

Figure 6: **Detection performance and hardware cost.** Direct-generation curves use $K = 1, 2, 4, 8$ responses in increasing computation order, with exact-match fusion or pair-IoU suppression. Stars indicate the complete harness. Cost is amortized GPU-worker time per image; values are reported in Tables 33 and 34.

DET mAP from 15.61% to 19.84% and V-COCO AP from 16.38%/21.18% to 22.21%/26.50%. The measured eight-response costs are 19.95 and 20.94 GPU-worker seconds per image, respectively. HarnessHOI achieves 33.20% and 44.65%/49.70%, with measured costs of 13.26 and 22.07 seconds per image. The gains over repeated generation demonstrate the effectiveness of coordinating visual exploration and relation verification within the harness.

Fusion and candidate construction. Repeated sampling produces additional interaction candidates and overlapping detections with slightly different boxes. Exact-match fusion retains these duplicates, and AP decreases as more responses are combined. Pair-IoU suppression improves AP by removing overlapping detections from the same responses. The following analysis evaluates candidate quality through the best AP attainable by reordering the predicted boxes and labels.

E.4 Fixed-Candidate Ranking Analysis

We evaluate the quality of the generated candidates by fixing their boxes and labels and using ground-truth annotations to determine the optimal ranking.

Single-response candidates. Table 35 compares the harness with ground-truth-assisted ranking of a single direct-generation response. The resulting ranking upper bounds are 21.58% HICO-DET mAP and 30.69%/35.01% V-COCO AP. HarnessHOI exceeds these bounds by 11.62 and 13.96/14.69 points, respectively, demonstrating that its advantage extends beyond confidence estimation to the participants and relations available for prediction.

Table 35: Ground-truth-assisted ranking of fixed direct-generation predictions. GT ranking retains the generated boxes and labels and optimally orders the candidates using annotations. HICO-DET reports Default Full mAP; V-COCO reports S1/S2 role AP (%).

Method	HICO-DET Full	V-COCO S1	V-COCO S2
Direct HOI generation	15.62	16.41	21.20
Direct HOI generation, GT ranking	21.58	30.69	35.01
HarnessHOI	33.20	44.65	49.70

Interaction candidates beyond repeated generation. With eight responses, exact-match fusion reaches ranking upper bounds of 30.50% HICO-DET mAP and 41.90%/46.40% V-COCO AP. HarnessHOI exceeds these bounds with its predicted scores by 2.70 and 2.75/3.30 points, respectively, and also exceeds the pair-IoU bounds of 30.09% and 41.40%/45.90%. The advantage persists even when the repeated-generation predictions are ranked optimally, demonstrating the importance of interaction-guided participant discovery and relation construction for HOI detection.

Table 36: Ground-truth ranking upper bounds for fixed candidates from repeated direct generation (AP, %). Each row preserves the corresponding fused boxes, labels, and candidate count, and uses ground truth to place correct detections before false positives. HICO-DET averages all 600 classes. The last row uses the harness's predicted scores.

Candidate construction	Responses	HICO-DET Full	V-COCO S1	V-COCO S2
Exact-match fusion	1	21.58	30.69	35.01
	2	24.30	34.80	38.90
	4	27.80	39.20	43.10
	8	30.50	41.90	46.40
Pair-IoU suppression	1	21.55	30.58	34.90
	2	24.09	34.50	38.60
	4	27.39	38.80	42.70
	8	30.09	41.40	45.90
HarnessHOI (predicted scores)	–	33.20	44.65	49.70

Ground-truth ranking. For each response count and fusion rule, we compute the maximum AP obtainable by reordering the fixed predictions. Ground-truth matching identifies one correct prediction per covered annotation and ranks these predictions before false positives. Boxes, labels, and candidate counts remain unchanged, and HICO-DET averages over all 600 classes.

Ranking bound under fixed matching. The evaluators assign each prediction to its best-matching ground-truth instance before duplicate assignment: HICO-DET uses participant-pair overlap, while V-COCO first selects the best-matching person and then assesses its role target. A duplicate does not switch to a second-best annotation. For fixed predicted boxes and labels, the set of annotations with an eligible matching prediction is therefore fixed. Let M_c be the number of such annotations for class c and $G_c > 0$ its ground-truth count. Placing one eligible prediction for each of these annotations before false positives attains precision one up to recall M_c/G_c. Under the respective AP definitions, the class-wise bounds are

$$U_c^{\mathrm{HICO}} = \frac{\mathbf{1}[M_c > 0]}{11} \sum_{t \in \mathcal{R}_{\mathrm{AP}}} \mathbf{1}\left[t \le \frac{M_c}{G_c}\right], \qquad U_c^{\mathrm{VCOCO}} = \frac{M_c}{G_c}. \tag{17}$$

Here $\mathcal{R}_{\mathrm{AP}}$ is the HICO-DET evaluator's recall-threshold sequence. V-COCO uses the matching rule of each role scenario, with S1 and S2 ranked separately. Averaging the class-wise bounds gives the results in Tables 35 and 36.

E.5 Generation Stability

We evaluate each method with generation seeds 66, 67, and 68, keeping the backbone, inference policy, and complete test set fixed. For the harness, the seed controls participant localization, discovery, and observation, while grounded verification uses the same first-token scoring rule. Table 37 reports the individual runs, and Table 38 gives the HICO-DET results under Default, Known Object, Rare, and Non-Rare evaluation.

Table 37: Run-to-run results under generation seeds 66, 67, and 68. The backbone, inference policy, and official test population are fixed for each method. HICO-DET reports Default Full mAP and V-COCO reports S1/S2 role AP (%). SD is the sample standard deviation across three runs.

Method	Seed	HICO-DET Full	V-COCO S1	V-COCO S2
Direct generation	66	15.62	16.41	21.20
	67	15.81	16.89	21.93
	68	15.55	16.58	21.74
	Mean ± SD	15.66 ± 0.13	16.63 ± 0.24	21.62 ± 0.38
HarnessHOI	66	33.20	44.65	49.70
	67	32.79	44.80	49.75
	68	32.86	44.50	49.55
	Mean ± SD	32.95 ± 0.22	44.65 ± 0.15	49.67 ± 0.10

Table 38: HICO-DET run-to-run results by evaluation setting and class frequency (mAP, %). All three seeds evaluate the complete official test set.

	Default			Known Object		
Seed	Full	Rare	Non-Rare	Full	Rare	Non-Rare
66	33.20	41.63	30.68	36.85	44.20	34.65
67	32.79	41.04	30.32	36.39	43.74	34.20
68	32.86	41.17	30.38	36.50	43.87	34.30
Mean	32.95	41.28	30.46	36.58	43.94	34.38
SD	0.22	0.31	0.19	0.24	0.24	0.24

Stability across benchmarks. Across three HICO-DET runs, HarnessHOI obtains 32.95% mean Default Full mAP with a sample standard deviation of 0.22 points, compared with 15.66% $\pm$ 0.13 for direct generation. Its Default Rare and Non-Rare means are 41.28% and 30.46%, and its Known Object Full mean is 36.58%. The gains remain consistent across seeds, evaluation settings, and frequency groups. V-COCO exhibits the same advantage in all three runs under both role-evaluation scenarios (Table 37).

F Qualitative Results

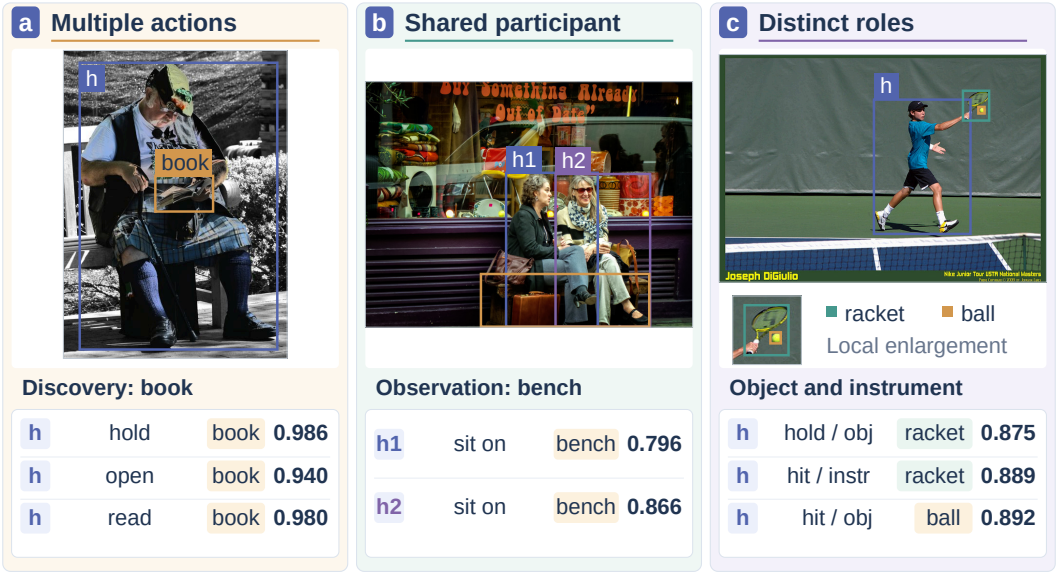

Figure 8: **Interaction-guided perception and grounded reasoning.** (a) Participant discovery supports multiple actions on a shared person-book pair. (b) Targeted observation recovers a bench shared by two people. (c) Compositional verification assigns the object and instrument of a tennis action to the ball and racket. Boxes and labels indicate the predicted participants and relations.

Figure 8 illustrates how the harness coordinates participant discovery, targeted observation, and grounded reasoning. Interaction hypotheses guide the recovery of the book and shared bench. The recovered book supports both *hold* and *read*, and a common bench grounding supports the sitting relations of two people. In the tennis scene, the racket participates in both *hold* and *hit*, while role verification distinguishes the racket as the instrument and the ball as the object of hitting. These examples complement the coverage analysis in Appendix D.3 by showing how the acquired participants support multiple actions and semantic roles.

Figures 9, 10, and 11 provide detailed examples of participant recovery, pair selection, and semantic-role assignment, with the corresponding visual-question scores reported below.

F.1 Participant Discovery and Targeted Observation

Recovering a book. In HICO-DET image `00000116`, interaction discovery recovers a book omitted from initial localization and associates it with the person's holding and reading actions. The two actions are then verified independently using the same person-book groundings. Table 39 reports the localization overlaps and relation scores.

A bench shared by two people. In HICO-DET image `00001195`, the sitting hypotheses of two people direct observation to a shared crop containing the supporting bench. Observation recovers the bench, and both sitting relations use its selected grounding. One person-bench pair is retained through the observation association and the other through visual pair assessment, illustrating their complementary contributions to relation coverage.

The same observation supplies an alternative suitcase box. Comparing both candidates against the acquired views, geometric adjudication selects the original box as the shared grounding for subsequent relation verification.

A small food target. In V-COCO image `18193`, the eating hypothesis directs observation to a small sandwich omitted from the initial participant set. The recovered sandwich is verified as the target of both holding and eating, showing how observation guided by one interaction can support another action involving the same participant. Table 41 gives the corresponding action, visibility, and binding factors.

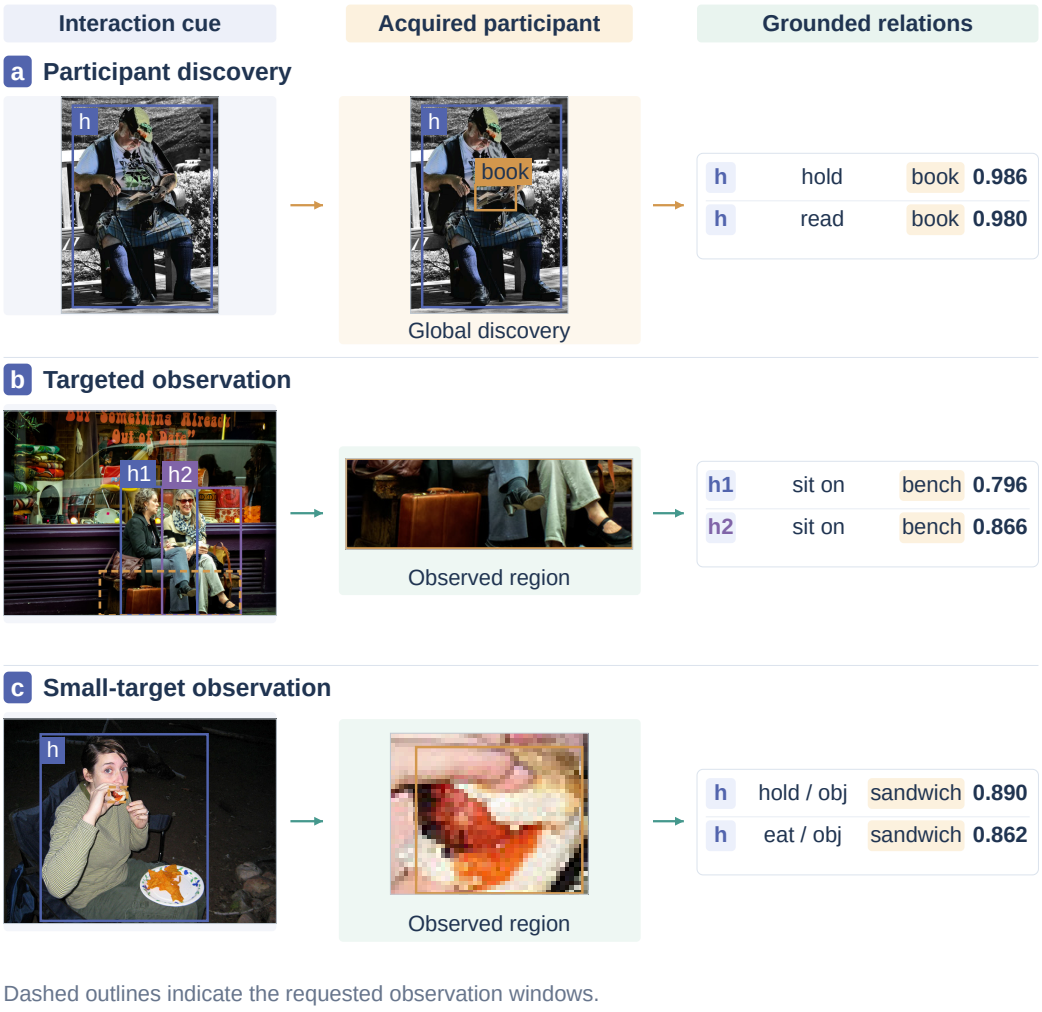

Figure 9: **Interaction-guided participant acquisition.** Selected relations from the book, shared-bench, and sandwich examples. The book is introduced through global discovery. Two sitting hypotheses request the same local window, where observation introduces a bench used by both people. In the eating example, observation localizes the small food target, supporting both holding and eating. Each sequence connects the interaction cue, the acquired participant, and its final grounded relation questions.

Grounded multi-label verification. For the person-book pair in HICO-DET image `00000116`, the selected boxes, full image, and public action definition specify the visual question: whether the person holds the book. The MLLM assesses this question over {`false`, `true`, `unknown`}. Normalizing the first-token logits by Equation 2 gives a holding score of 0.9859. A separate reading question over the same participants gives 0.9796 (Table 39). Independent action assessments there-

Table 39: Selected interaction groundings and scores in the qualitative cases. Each IoU is computed against the ground-truth boxes of the listed interaction annotation. Repeated targets share the same selected geometry across their relations.

Case	Action and target role	Person IoU	Target IoU	Score
Book	hold book	0.9419	0.6543	0.9859
	read book	0.9545	0.7631	0.9796
Shared bench	sit_on bench	0.9497	0.8209	0.7959
	sit_on bench	0.8886	0.8209	0.8658
Sandwich	hold / obj	0.9659	0.8844	0.8896
	eat / obj	0.9659	0.8844	0.8616
Tennis	hold / obj	0.9615	0.9477	0.8748
	hit / instr	0.9615	0.9477	0.8889
	hit / obj	0.9615	0.8294	0.8924
Cutting	hold / obj	0.9845	0.9382	0.7955
	cut / instr	0.9845	0.9382	0.9443
	cut / obj	0.9845	0.8939	0.9445

fore retain both interactions on the shared groundings. Table 41 presents the corresponding factor composition for V-COCO roles.

F.2 Pair Retention in a Distant Gaze Relation

Distant gaze. V-COCO image `2473` contains an annotated gaze relation between two people. The observer and target boxes match their annotations at IoU 0.9560 and 0.9517. Although the agent and role-state questions favor looking and a visible target, the pair question assigns its highest support to `false` (Table 40). Neither discovery nor observation associates this pair. Equation 4 therefore excludes it from the binding domain, and no physical `look/obj` prediction is produced for that pair.

Table 40: Visual-question supports for the gaze example. Agent and role-state questions assess the human's action and target visibility; the pair decision determines whether the specific target reaches binding verification.

Question	False / missing	True / visible	Unknown
Observer performs `look`	0.2644	0.7187	0.0169
Visibility of `look/obj`	0.4069	0.5224	0.0707
Observer-target pair	0.7546	0.2162	0.0293

In this example, pair selection excludes the gaze relation despite accurate participant localization. In the shared-bench example, observation associations and visual pair assessment retain different person-bench pairs, illustrating how their combination extends the relations available for verification.

F.3 Shared Geometry and Distinct Semantic Roles

Object and instrument in hitting. In V-COCO image `579060`, initial localization identifies the person, racket, and ball. The racket is both a held object and the instrument of hitting, while the ball fulfills the object role of hitting. The two hitting roles share action evidence, with separate visibility and binding questions identifying their respective targets (Table 41).

Object and instrument in cutting. V-COCO image `468608` similarly distinguishes a knife as the cutting instrument and a cake as the object being cut. Both are available from initial localization, and their role predictions combine shared action evidence with target-specific assessments. The knife also participates in holding, illustrating the use of one grounding across different action-role assignments.

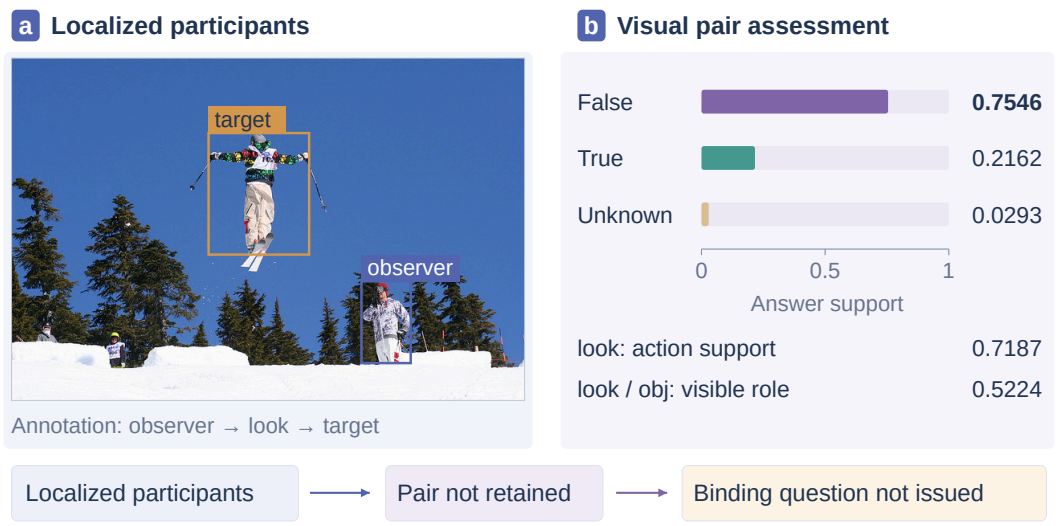

Figure 10: **Pair retention and relation coverage.** Both participants match their annotated boxes, but visual pair assessment assigns its highest support to *false*. Neither discovery nor observation associates them, so the pair is excluded from subsequent *look*-role binding verification. Table 40 reports the corresponding question supports.

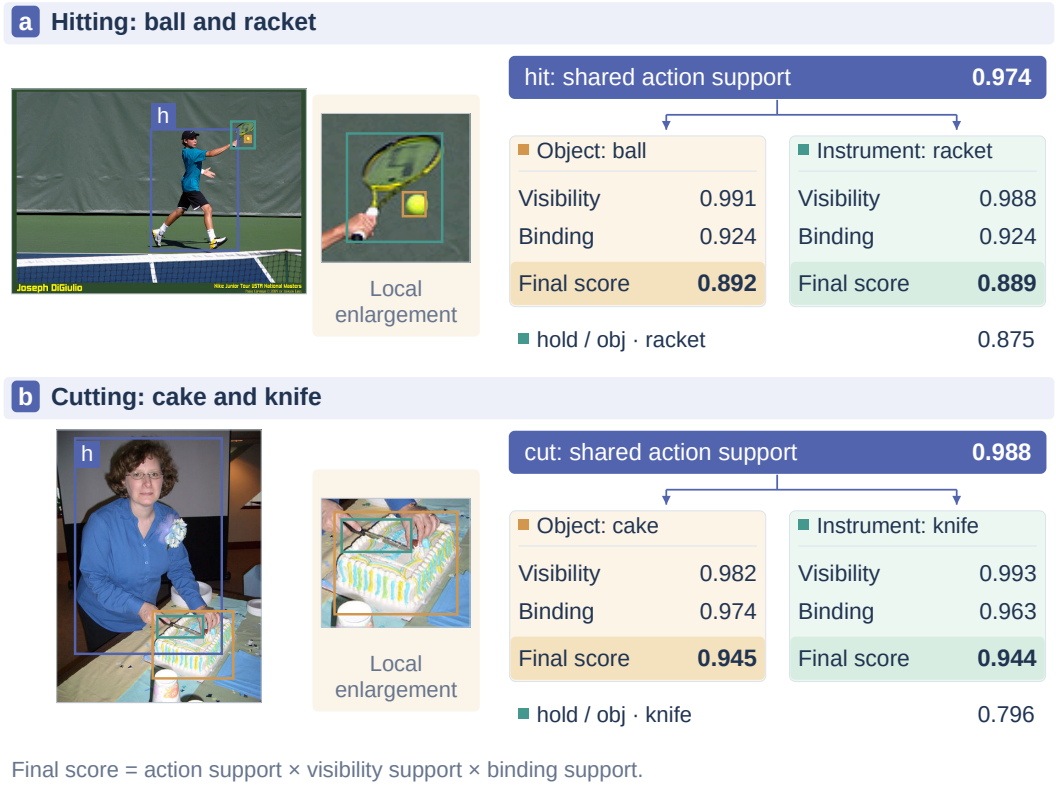

Figure 11: **Compositional role verification.** Selected predictions from the tennis and cutting examples. The racket and knife fulfill instrument roles, whereas the ball and cake fulfill object roles. Each action contributes a shared agent support, combined with role-specific visibility and binding supports. Holding and the instrument role can share the same target grounding.

G PROMPTS AND VISUAL QUESTIONS

We provide the prompt instructions used for perception and relation verification. Each prompt incorporates the relevant participants, image views, and public interaction definitions.

G.1 PARTICIPANT LOCALIZATION

Initial localization receives the full image and public object vocabulary and returns categories, boxes normalized to $[0, 1000]$, and appearance references:

> Localize the people and identifiable targets involved in the scene's interactions. Organize the output by interaction group: list one person and all associated targets immediately after that person, then proceed to the next person. Cover holding, using, riding, sitting or leaning, wearing, carrying, caring, and person-to-person interactions. Include identifiable nearby objects that may support no-interaction assessment, while prioritizing interaction participants over unrelated background objects.
>
> Give one box for each complete visible instance and include each instance once. A person's box should cover the visible person; an object's box should cover the object itself. Describe each reference through visible appearance and position. Return the instance category, normalized box, and reference in the requested format.

G.2 INTERACTION DISCOVERY AND OBSERVATION PLANNING

Discovery receives the full image, initial entities, designated human group, and public interaction definitions. The prompt requests hypotheses, additional participants, and observation regions:

> Inspect every designated human. Give concise visible evidence and identify the interactions involving that person. A person can perform several actions and interact with several targets. Use existing instance identities when their targets are already localized. Add an omitted, visually identifiable participant with its category, full-image box, and appearance reference. For role-based actions, distinguish the target fulfilling each object or instrument role.
>
> Examine held, worn, ridden, supporting, and distant visible targets. When the target cannot yet be localized reliably, retain its description and formulate a question for closer observation. Propose up to two regions, giving priority to unlocalized targets and then incomplete instance boundaries. Each request identifies the human, possible action and role, known target when available, target description, localization question, image region, and priority. The region should support localization of the complete visible instance. When the existing view resolves participant localization, return no observation request.

G.3 TARGETED OBSERVATION

Observation receives a crop, the target description, and the local instances eligible for refinement. Output coordinates are normalized within the crop:

> Use this crop and target description to localize the relevant participants. Identify changes to the eligible instances and any additional visible participants. For an existing instance, retain its identity and report the changed category or box. Instances crossing the crop boundary provide context; update complete-instance boxes for eligible identities whose extents lie within the view.
>
> Describe newly identified participants with complete visible-instance boxes and categories. Preserve the distinction between a person and the object they hold or use. Localize visible extent without extrapolating occluded boundaries. Return instance updates and new participants. If the observation provides no reliable localization change, return empty changes.

Returned boxes are mapped to the full image by Equation 9.

G.4 Geometric Adjudication

The geometry question receives the candidate category-and-box descriptions, full image, and acquired crops. Each candidate is assigned an answer letter:

> Compare the candidate descriptions of this instance against the image evidence. Select the candidate that best matches its identity, category, and complete visible boundary. For a person, assess the full visible person rather than a hand or torso. For an object, assess the extent of that object rather than its contact region with another participant. Judge localization correspondence; the candidate order does not indicate quality. Answer with the letter of the selected candidate.

Candidate supports are normalized by Equation 2, and the grounding is selected by Equation 3.

G.5 Grounded Relation Questions

Shared instruction. Relation questions receive the selected participants, action or role definition, and corresponding image views. They share the following instruction:

> Assess the specified instances from the image evidence. Use the given boxes to identify the person and target, and verify that those instances correspond to the visual question. Distinguish this person's action from another person's action and this target from another instance of its category. An interaction requires visual evidence beyond proximity or the object's typical use. Multiple actions can hold for the same participants. Apply the supplied action and role meanings and answer with one word from the designated answer set.

Table 42 lists the question types and answer sets.

Table 41: Compositional role scores in the qualitative examples. q_A, q_V, and q_B denote action, visible-role, and binding support. The displayed score is computed from the unrounded factors.

Case	Action / role	q_A	q_V	q_B	$q_A q_V q_B$
Sandwich	hold / obj	0.9770	0.9769	0.9320	0.8896
	eat / obj	0.9890	0.9525	0.9146	0.8616
Tennis	hold / obj	0.9399	0.9669	0.9627	0.8748
	hit / instr	0.9740	0.9876	0.9241	0.8889
	hit / obj	0.9740	0.9914	0.9241	0.8924
Cutting	hold / obj	0.9323	0.9234	0.9241	0.7955
	cut / instr	0.9875	0.9933	0.9626	0.9443
	cut / obj	0.9875	0.9820	0.9740	0.9445

Pair assessment. The pair question asks whether any legal positive interaction is visually supported. Its answers denote support for at least one relation (`true`), rejection of all such relations (`false`), or unresolved evidence (`unknown`). Pair retention follows Equation 4, combining this assessment with discovery and observation associations.

No-interaction and action assessment. No-interaction assessment verifies the participants' identity and geometry together with the absence of legal positive interactions. Each positive-action question evaluates a specific predicate under its supplied definition: for example, `hold` concerns bodily or manual support of the target, and `read` concerns engagement with its textual content. Each question produces a separate false/true/unknown distribution.

Role visibility and binding. Conditional on the action, `visible` denotes a localizable role target in the image, including one absent from the entity set; `missing` denotes the absence of such a

target; and `unknown` denotes unresolved visibility. Binding assesses whether the specified instance fulfills the object or instrument role.

G.6 Public Role Definitions

Table 42: Semantic decisions in grounded question answering. Answer supports are obtained from first-token logits normalized within the designated set.

Question	Visual judgment	Answer set
Pair	Whether any legal positive action or physical-role binding connects the specified human and target, including distant and non-contact relations	false / true / unknown
No interaction	Whether both instances and their boxes are valid and no legal positive interaction holds between them	false / true / unknown
Action	Whether the specified participant pair and action jointly have visual support	false / true / unknown
Agent	Whether the specified human performs the action, independently of target visibility	false / true / unknown
Role state	Given the action, whether its role has a localizable visible target, has a missing target, or remains unresolved	missing / visible / unknown
Binding	Given the action and visible role, whether this particular target fulfills that role	false / true / unknown

Table 42 gives the complete target meanings used in V-COCO questions. The role keys follow the benchmark representation. In particular, `point/instr` denotes the target indicated by the pointing gesture. Action definitions complement these role meanings by specifying the person's activity: holding a phone and talking on a phone are assessed separately; standing on a board does not alone establish riding or jumping; and skiing, snowboarding, skateboarding, and surfing retain their distinct activity meanings. HICO-DET questions follow the public verb-object taxonomy, with definitions distinguishing predicates such as `cut` and `cut_with`, `sit_at` and `sit_on`, and `hold` and `carry`.