GTR-HOI: GROUND, TRANSPORT, AND REINFORCE FOR WEAKLY SUPERVISED HOI DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Weakly supervised human–object interaction (HOI) detection learns to localize and recognize interactions from image-level labels, without explicit correspondences between labels and human–object pairs. This ambiguity can reinforce incorrect interaction assignments during training. We propose GTR-HOI, a three-stage framework that grounds image-level labels, transports localization supervision, and reinforces interaction predictions. First, a pretrained multimodal large language model (MLLM) generates pseudo-ground truths from human, object, and relation perspectives, with spatial agreement determining their confidence tiers. To account for grounding uncertainty, unbalanced optimal transport transfers localization supervision to detector pairs through soft correspondences. Pair expert routing integrates appearance, spatial, and contextual cues for pair validity prediction and visual HOI scoring. Hierarchical supervision uses consensus assignments with high or medium confidence to train the visual scoring branch and fine-tune the MLLM for semantic HOI scoring. Finally, cross-prompt reinforcement learning uses semantic feedback across four prompt views to refine decisions to accept or reject candidate interactions through weighted proximal policy optimization. Spatial and fixed reference certainty weights regulate these updates to limit the influence of unreliable supervision. Experiments on HICO-DET and V-COCO demonstrate state-of-the-art performance among the compared weakly supervised methods, achieving 36.44 Full mAP and 60.6 Scenario 2 role AP, respectively. The code is provided in the supplementary materials.

1 INTRODUCTION

Human–object interaction (HOI) detection localizes humans and objects and recognizes their relations as $\langle \text{human}, \text{verb}, \text{object} \rangle$ triplets, facilitating human-centric scene understanding. Fully supervised methods require exhaustive annotations for interacting human–object pairs (Chao et al., 2018; Guo et al., 2024; Lei et al., 2023; Kim et al., 2025; Ning et al., 2023), making annotation increasingly costly as the interaction vocabulary and number of pairs grow. Weakly supervised HOI detection (Park et al., 2026; Wan et al., 2023; Kumaraswamy et al., 2021; Kilickaya & Smeulders, 2021) reduces this burden through image-level labels but lacks explicit correspondences between interaction labels and human–object pairs. The central challenge is therefore to recover reliable visual–semantic associations from a large, noisy candidate space without instance-level supervision.

To address this limitation, earlier weakly supervised methods associate image-level labels with candidate human–object pairs through score aggregation, graph-based attribution, or latent alignment (Zhang et al., 2017; Kumaraswamy et al., 2021; Baldassarre et al., 2020; Kilickaya & Smeulders, 2021) (Fig. 1(a)). With advances in vision–language pre-training, more recent methods incorporate transferable semantic knowledge into HOI learning. Weakly HOI-CLIP (Wan et al., 2023) integrates CLIP (Radford et al., 2021) priors into image and human–object pair representations. It uses multiple instance learning (MIL) to aggregate pair predictions for image-level supervision. OpenCat (Zheng et al., 2023) learns open-category interaction knowledge from large-scale image–caption data through language modeling, then matches predicted interactions to detected regions for weakly supervised detection. However, semantic priors alone do not resolve the ambiguity in associating image-level labels with individual human–object pairs. Incorrect pairs may dominate aggregated predictions or receive erroneous interaction assignments, reinforcing spurious associations that can impair both localization and recognition.

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.

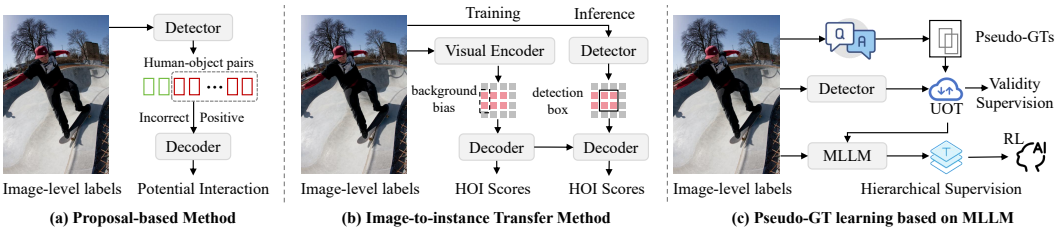


Figure 1: Comparison of weakly supervised HOI detection paradigms. (a) Proposal-based methods associate image-level labels with candidate human–object pairs, potentially reinforcing incorrect assignments during training. (b) Image-to-instance transfer learns HOI representations from image-level labels and transfers them to detected regions at inference, but implicit spatial associations may introduce background bias. (c) GTR-HOI uses confidence-tiered MLLM groundings and UOT for detector-pair validity supervision, followed by hierarchical HOI learning and RL-based refinement.

In contrast, proposal-free methods bypass detector-based pair selection during weakly supervised training (Fig. 1(b)). RegFormer (Park et al., 2026) learns patch-level relational grounding and interactiveness from image-level labels, then converts detector regions into instance-aware queries at inference. This design avoids combinatorial proposal search and reduces reliance on classifier-driven instance selection. However, this transfer relies on spatial associations learned implicitly through patch attention, without explicit localization targets for the human–object pairs corresponding to each interaction. Consequently, reliance on irrelevant background cues may limit the model's ability to capture HOI details, making visually similar interaction categories harder to distinguish.

To overcome these challenges, we propose GTR-HOI, which **G**rounds image-level HOI labels by using a multimodal large language model (MLLM) to localize the corresponding human–object pairs. It then **T**ransports localization supervision to detector pairs through unbalanced optimal transport (UOT) (Chizat et al., 2018) and **R**einforces interaction predictions through cross-prompt reinforcement learning (RL) (Fig. 1(c)). Stage 1 queries the MLLM from human, object, and relation perspectives to localize human–object pairs, with spatial agreement determining high, medium, and low confidence tiers. Stage 2 uses these localization results as pseudo-ground truths (pseudo-GTs) and transfers their supervision to detector pairs through UOT. Pair expert routing (PER) constructs interaction features and learns to predict pair validity from the transferred supervision. Hierarchical supervision then jointly trains the visual and MLLM semantic scoring branches. However, spatial agreement does not guarantee correct interaction assignments, motivating further refinement through semantic feedback. Stage 3 therefore evaluates each candidate interaction across multiple prompt views and uses their agreement to guide proximal policy optimization (PPO) (Schulman et al., 2017).

In summary, our contributions are four-fold: (i) We introduce multi-perspective MLLM grounding to generate HOI-specific pseudo-GTs, using spatial agreement to organize them into confidence tiers for explicit spatial supervision. (ii) We combine PER with UOT to learn detector-pair validity, and introduce hierarchical supervision to jointly train visual and semantic HOI prediction under localization uncertainty. (iii) We develop cross-prompt RL that uses multi-view semantic agreement to guide PPO, refining interaction predictions for each human–object pair. (iv) Extensive experiments on HICO-DET (Chao et al., 2018) and V-COCO (Gupta & Malik, 2015) demonstrate state-of-the-art performance on both benchmarks and validate the effectiveness of our design.

2 Related Work

HOI Detection. Fully supervised HOI approaches have achieved substantial progress using instance-level annotations that specify human and object boxes and their interaction labels (Kim et al., 2021; Tamura et al., 2021; Zhang et al., 2022). However, annotating every interaction category becomes increasingly costly as the vocabulary expands. Zero-shot HOI detection addresses this challenge by extending recognition to interactions absent from the training labels, with pretrained vision–language knowledge facilitating semantic transfer to unseen categories (Liao et al., 2022; Ning et al., 2023; Guo et al., 2024; 2026; Mao et al., 2023). Although this reduces the need for examples of every target interaction, conventional zero-shot approaches still require instance-level

annotations for seen categories. The central difficulty consequently shifts from recognizing interactions on labeled pairs to determining which pairs correspond to each image-level label. *Our work studies this weakly supervised setting, including generalization to unseen interaction compositions.*

Weakly Supervised HOI Detection. Weakly supervised HOI detection uses image-level interaction labels without annotations linking them to human–object pairs, making instance assignment inherently ambiguous. Early methods (Zhang et al., 2017; Baldassarre et al., 2020; Kilickaya & Smeulders, 2021) infer corresponding pairs through multiple instance learning, graph-based attribution, or latent alignment. However, incorrect pair assignments can reinforce selection errors during training. To strengthen interaction representations, Weakly HOI-CLIP (Wan et al., 2023) incorporates CLIP priors into representation learning for images and human–object pairs. Unal & Kovashka (2023) combine caption-guided proposal selection and spatial learning with language priors for interaction plausibility. OpenCat (Zheng et al., 2023) acquires transferable interaction semantics through image–caption pre-training. RegFormer (Park et al., 2026) avoids proposal selection during training by learning proposal-free relational grounding and transferring it to detected instances at inference. These methods improve HOI representations but still face ambiguous human–object associations. *Our method addresses this ambiguity using MLLM-generated pseudo-GTs for explicit spatial supervision, with UOT aligning these groundings with detected human–object pairs.*

MLLM-Based HOI Detection. With advances in MLLMs, recent studies have increasingly explored their application to various aspects of HOI detection, including interaction recognition, structured prediction, and training-free reasoning. DA-HOI (Xuan et al., 2026) decouples detection from interaction recognition, using spatial-aware pooling and deterministic matching to classify human–object pairs. HOI-MLLM (Wu et al., 2026) combines chain-of-thought supervision with GRPO post-training to generate structured HOI predictions. HOI-R1 (Chen et al., 2025) directly predicts HOIs without an external detector, using reasoning distillation and RL to improve classification and localization. AgentHOI (Lei et al., 2026) combines multi-round interaction discovery with descriptive grounding without task-specific training. These studies highlight the potential of MLLMs for interaction reasoning and localization through supervised adaptation or training-free inference. *Our method uses MLLM grounding to convert image-level HOI labels into explicit spatial supervision, followed by hierarchical learning and cross-prompt RL to refine interaction predictions.*

3 METHOD

3.1 PRELIMINARIES

Problem Setting. Following the weakly supervised HOI setting (Wan et al., 2023; Kilickaya & Smeulders, 2021; Park et al., 2026), the training set is $\mathcal{D} = \{(I_n, \mathcal{Y}_n)\}_{n=1}^{N}$, where I_n is an image and $\mathcal{Y}_n$ contains its positive image-level HOI labels. Each label $y = (c_v, c_o)$ combines a verb c_v and an object category c_o, without specifying the corresponding human and object boxes. At inference, an off-the-shelf detector provides instance proposals, and the model predicts a set of scored tuples $(b_{h,j}, b_{o,j}, y, S_{jy})$. Here, j indexes a human–object pair, $b_{h,j}, b_{o,j} \in \mathbb{R}^4$ denote its human and object boxes, and S_{jy} is its final score for label y. Multiple interaction labels may describe the same pair, and each label may apply to multiple pairs. In unseen-composition evaluation, the model must additionally recognize verb–object combinations withheld during training.

Baseline. We build GTR-HOI on DA-HOI (Xuan et al., 2026), which decouples object detection from MLLM-based interaction recognition. Its Spatial-Aware Pooling (SAP) combines regional appearance and pairwise geometry, while one-pass deterministic matching scores object-compatible interaction candidates. However, fine-tuning DA-HOI requires ground-truth human–object boxes and their interaction labels, whereas our weakly supervised setting provides only image-level HOI labels. We therefore query a pretrained MLLM with a single label-conditioned grounding prompt for each positive HOI to generate human and object boxes. Each generated pair is assigned the prompting HOI label and treated as a pseudo-GT. These pseudo-GTs replace annotated pairs for SAP training and are directly fed into the MLLM for interaction classification fine-tuning.

3.2 FRAMEWORK OVERVIEW

As illustrated in Fig. 2, GTR-HOI combines MLLM grounding, spatial supervision transfer, and RL to learn HOI detection from image-level labels. Stage 1 generates pseudo-GTs from human, object,

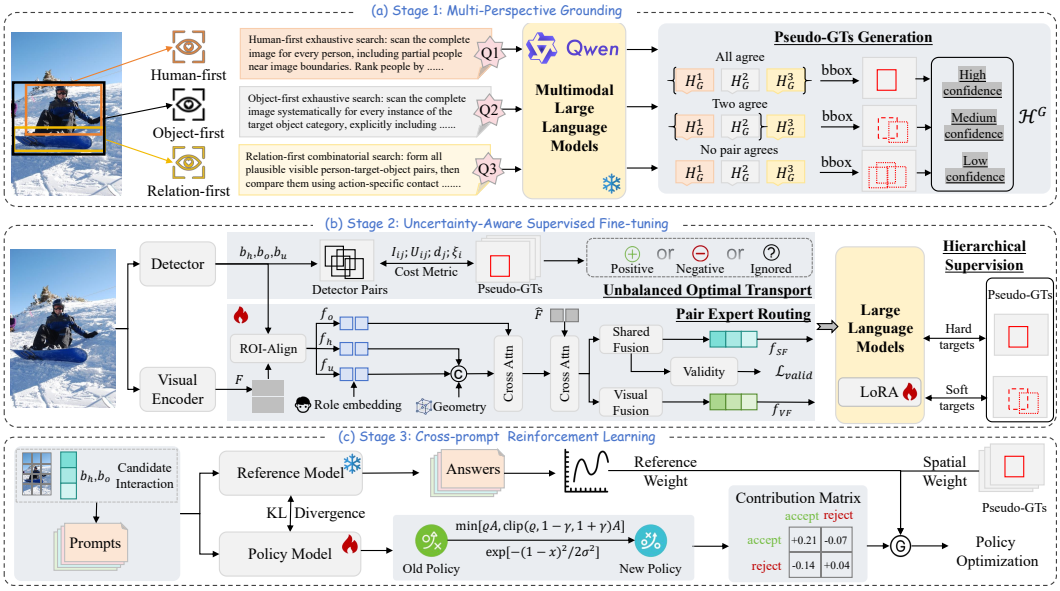


Figure 2: Overview of GTR-HOI. (a) Multi-Perspective Grounding generates pseudo-GTs from human, object, and relation prompts, with spatial agreement determining their confidence tiers. (b) Uncertainty-Aware Supervised Fine-Tuning transfers localization supervision to detector pairs through UOT. PER fuses expert features for validity prediction and visual HOI scoring, while hierarchical supervision guides MLLM fine-tuning. (c) Cross-Prompt RL refines interaction predictions through PPO with multi-view semantic feedback, spatial weighting, and fixed reference certainty.

and relation perspectives, using spatial agreement to define confidence tiers. Stage 2 aligns pseudo-GTs with detector pairs through UOT, providing validity supervision and candidate regions for HOI learning. Hierarchical supervision uses consensus assignments from groups with high or medium confidence to train the visual scoring branch and fine-tune the MLLM for semantic HOI scoring. Stage 3 additionally incorporates eligible medium residuals and low-confidence assignments. Multi-view feedback and PPO with exact action enumeration refine interaction predictions.

3.3 Stage 1: Multi-Perspective Grounding

To obtain pseudo-GTs for the training images, we query Qwen2.5-VL-3B-Instruct (Bai et al., 2025) to localize human–object pairs corresponding to image-level HOI labels. However, a single grounding prompt may produce inaccurate localizations. We therefore query the model from human, object, and relation perspectives, using spatial agreement across these views to assess grounding reliability (Fig. 2(a)). For each image and label y, we define a pseudo-GT group G containing the hypotheses generated by the three prompt views. Each view $p \in \{1, 2, 3\}$ returns at most one raw spatial hypothesis $H_G^p = (\hat{b}_{h,G}^p, \hat{b}_{o,G}^p, \hat{b}_{u,G}^p)$ with a model confidence score. The union box is derived from the human and object boxes. Spatial agreement between two raw hypotheses is measured by

$$\mathrm{Agr}_G(p, p') = \min_{\zeta \in \{h,o,u\}} \mathrm{IoU}(\hat{b}_{\zeta,G}^p, \hat{b}_{\zeta,G}^{p'}), \quad p \neq p'. \tag{1}$$

Two hypotheses agree when $\mathrm{Agr}_G(p, p') \geq \tau_{\text{agree}} = 0.5$. Based on this agreement, we construct the final retained hypothesis set $\mathcal{H}^G$, with $\mathcal{H}_i^G$ denoting its i-th hypothesis. Pairwise agreement among all three hypotheses yields high confidence with one consensus hypothesis. Partial agreement yields medium confidence, retaining the selected pair's consensus and any valid residual hypothesis. No agreement yields low confidence, retaining all valid raw hypotheses.

Within each image and object category, we merge consensus hypotheses from groups with high or medium confidence across HOI labels when the human and object box IoUs both meet the overlap threshold. The merged records share human and object coordinates, while each HOI label retains its own confidence tier. Residual hypotheses from groups with medium confidence and alternative hypotheses from groups with low confidence are excluded from merging across HOI labels.

3.4 STAGE 2: UNCERTAINTY-AWARE SUPERVISED FINE-TUNING

Stage 1 produces pseudo-GTs with varying reliability, so treating them as equally reliable supervision may introduce noise into training. Stage 2 uses UOT to transfer localization supervision to detector pairs, accounting for grounding uncertainty in validity and HOI learning (Fig. 2(b)).

Pair Expert Routing. To integrate appearance, spatial, and contextual cues for pair validity prediction and visual HOI scoring, we introduce Pair Expert Routing (PER). PER applies RoI-Align (He et al., 2017) to the image token grid F from the frozen Qwen visual encoder. Separate projections yield human, object, and union tokens f_h, f_o, and f_u. Mean pooling over human and object tokens produces expert features e_h and e_o. The geometry expert e_g complements appearance features by encoding pairwise spatial information and the object category. To capture relational cues, a query $q_r = g_r([e_h, e_o, e_g])$ attends to union tokens augmented with role embeddings, $\tilde{f}_u = f_u + E_{\text{role}}$. These embeddings distinguish human, object, overlap, and background locations within the union region. The relation feature retrieves global context from projected image tokens $\hat{F} = P_F(F)$:

$$e_r = \text{CrossAttn}(q_r, \tilde{f}_u, \tilde{f}_u), \qquad e_c = \text{CrossAttn}(e_r, \hat{F}, \hat{F}), \tag{2}$$

where CrossAttn denotes cross-attention. The Shared Fusion (SF) and Visual Fusion (VF) modules use the same architecture with separate parameters. Each predicts fusion weights from the five expert features and transforms their weighted sum:

$$f_k = \hat{g}_k \left(\sum_{i=1}^{5} \left[\text{softmax}(g_k(\text{Concat}(E))) \right]_i E_i \right), \qquad k \in \{\text{SF}, \text{VF}\}, \tag{3}$$

where $E = [e_h, e_o, e_r, e_g, e_c] \in \mathbb{R}^{5 \times d_s}$ stacks five expert features. $\text{Concat}(\cdot)$ denotes concatenation, while g_k and $\hat{g}_k$ are learnable mappings. SF produces f_{SF} for validity prediction; after projection, it serves as an interaction feature for the MLLM. VF produces the visual interaction feature f_{VF}. The MLLM contextualizes the interaction token and each candidate interaction text y, yielding f'_{SF} and $\Theta(y)$, respectively. A learned projection P_E maps f_{VF} into the language hidden space. The visual HOI score S^{vis} is obtained by applying a sigmoid to the scaled cosine similarity between $P_E(f_{\text{VF}})$ and $\Theta(y)$. The semantic branch computes S^{sem} analogously using f'_{SF} and $\Theta(y)$.

Unbalanced Optimal Transport. To account for localization uncertainty in Stage 1 pseudo-GTs, we use Unbalanced Optimal Transport (UOT) to transfer localization supervision to detector pairs through soft correspondences. Within group G, we first construct a matching cost matrix C_{ij} between hypotheses i and human–object pairs j produced by the detector:

$$C_{ij} = \nu_p(1 - I_{ij}) + \nu_u(1 - U_{ij}) + \nu_d(1 - d_j) + \nu_s(1 - \xi_i), \tag{4}$$

where $I_{ij} = \min_{\zeta \in \{h,o\}} \text{IoU}(\hat{b}_{\zeta,i}, b_{\zeta,j})$ and $U_{ij} = \text{IoU}(\hat{b}_{u,i}, b_{u,j})$ measure pair and union overlap between pseudo-GT boxes $\hat{b}$ and detector boxes b. The detector prior $d_j = S^{\lambda}_{h,j} S^{\lambda}_{o,j}$ combines human and object detection scores. The reliability $\xi_i = \bar{c}_i \kappa_i$ is specific to hypothesis i: $\bar{c}_i$ averages its supporting responses' grounding confidences, and κ_i is their fraction among the three prompts.

After obtaining the matching cost matrix C_{ij}, we normalize ξ_i to obtain the source marginal m^{s}, while detector confidence d_j defines the target marginal m^{t}. We obtain the soft transport plan by:

$$\Pi^{\star} = \arg\min_{\Pi \geq 0} \left\{ \sum_{ij} \Pi_{ij} \left[C_{ij} + \epsilon(\log \Pi_{ij} - 1) \right] + \rho \left[D_{\text{KL}}(\tilde{m}^{\text{s}} \| m^{\text{s}}) + D_{\text{KL}}(\tilde{m}^{\text{t}} \| m^{\text{t}}) \right] \right\}, \tag{5}$$

where Π_{ij} denotes transferred mass, with actual marginals $\tilde{m}^{\text{s}}_i = \sum_j \Pi_{ij}$ and $\tilde{m}^{\text{t}}_j = \sum_i \Pi_{ij}$. We optimize the nonnegative transport matrix Π using log-domain generalized Sinkhorn iterations (Chizat et al., 2018) to obtain $\Pi^{\star}$. The parameters ϵ and ρ control entropic smoothing and generalized KL penalties, respectively. With received mass $m_j = \sum_i \Pi^{\star}_{ij}$, spatial support is

$$u^{(G)}_j = \min\left(1, \frac{m_j}{m^{\text{t}}_j + \delta}\right) \frac{\sum_i \Pi^{\star}_{ij} \exp(-C_{ij}/T_a)}{m_j + \delta}, \qquad u_j = \max_G u^{(G)}_j. \tag{6}$$

The two factors combine mass support and matching affinity, with temperature $T_a > 0$ and stability constant $\delta > 0$. For each detector pair, maximizing over pseudo-GT groups in the same image with the same object category yields bounded validity support u_j.

Hierarchical Supervision. PER extracts interaction features from detector pairs, while UOT provides positive and negative targets for validity pretraining. The PER features of UOT-selected pairs associated with high or medium confidence consensus hypotheses are then projected into the MLLM. Hierarchical supervision trains both visual and semantic branches using hard positive targets for high-confidence assignments and soft targets for medium-confidence assignments.

3.5 Stage 3: Cross-Prompt Reinforcement Learning

To mitigate the impact of pseudo-GT localization uncertainty on HOI learning, Stage 3 introduces cross-prompt RL (Fig. 2(c)). It uses detector pairs selected through UOT across all confidence tiers. For each pair, we construct four prompt views $p \in \{1, 2, 3, 4\}$: a direct query and queries emphasizing the human, object, and relation, respectively. Semantic feedback across these views guides independent accept or reject decisions for each candidate interaction. For clarity, we omit the indices j, y for a human–object pair j and candidate interaction y throughout this subsection.

Each view averages its visual and semantic HOI scores to obtain an acceptance probability $\pi^p = (S^{\mathrm{vis},p} + S^{\mathrm{sem},p})/2$. The other three views provide feedback:

$$f^{-p} = \frac{1}{3} \sum_{p' \neq p} \pi^{p'}_{\mathrm{old}}, \qquad q_1^{-p} = \frac{\mathcal{G}_\sigma(f^{-p})}{\mathcal{G}_\sigma(f^{-p}) + \mathcal{G}_\sigma(1 - f^{-p})}. \tag{7}$$

Here, $-p$ indicates exclusion of view p, and $\mathcal{G}_\sigma(x) = \exp[-(1-x)^2/(2\sigma^2)]$ is the Gaussian weighting function with width $\sigma > 0$. The weights q_1^{-p} and $q_0^{-p} = 1 - q_1^{-p}$ express support for accept and reject pseudo targets, respectively. old denotes predictions cached at the start of each PPO round.

Action $a = 1$ accepts the candidate interaction, while $a = 0$ rejects it; pseudo target t uses the same encoding. Let $P^p_\theta(a)$ denote the current action distribution, with $P^p_\theta(1) = \pi^p$ and $P^p_\theta(0) = 1 - \pi^p$; $P^p_{\mathrm{old}}(a)$ is its cached counterpart. We assign reward $R_t(a) = 1$ when $a = t$ and 0 otherwise. The advantage is $A^p_t(a) = R_t(a) - P^p_{\mathrm{old}}(t)$, and the probability ratio is $\varrho^p(a) = P^p_\theta(a)/P^p_{\mathrm{old}}(a)$. Old-policy probabilities, feedback weights, and advantages remain fixed during each round's updates. For each pseudo target and action, the clipped PPO surrogate is

$$\Psi^p_{ta} = \min[\varrho^p(a) A^p_t(a), \mathrm{clip}(\varrho^p(a), 1 - \gamma, 1 + \gamma)\, A^p_t(a)], \tag{8}$$

where γ is the PPO clipping width. Weighting each surrogate by the feedback weight and the old-policy action probability gives a 2×2 contribution matrix:

$$M^p_{ta} = q_t^{-p} P^p_{\mathrm{old}}(a) \Psi^p_{ta}, \qquad \mathbf{M}^p = \begin{pmatrix} M^p_{11} & M^p_{10} \\ M^p_{01} & M^p_{00} \end{pmatrix}. \tag{9}$$

Each matrix entry gives a weighted PPO contribution, with rows indexing pseudo targets and columns indexing actions, both ordered as accept and reject. Exact action enumeration and averaging across four views yield $\mathcal{L}_{\mathrm{PPO}} = -\frac{1}{4} \sum_{p=1}^{4} \sum_{t,a \in \{0,1\}} M^p_{ta}$.

3.6 Training and Inference

Training. Stage 1 generates pseudo-GTs using a frozen MLLM. In Stage 2, validity supervision trains PER to determine whether a detector pair is consistent with the pseudo-GT supervision, while hierarchical supervision trains the visual and semantic branches to identify its specific interactions:

$$\mathcal{L}_{\mathrm{stage2}} = \lambda_1 \mathcal{L}_{\mathrm{valid}} + \lambda_2 \mathcal{L}_{\mathrm{high}} + \lambda_3 \mathcal{L}_{\mathrm{medium}}. \tag{10}$$

The validity term uses all confidence tiers during pretraining, followed by hierarchical training of the PER visual HOI scoring branch and the MLLM semantic HOI scoring branch. The high and medium terms supervise consensus assignments with hard and soft targets, respectively.

In Stage 3, inspired by RESTRAIN (Yu et al., 2025), we weight the PPO objective using semantic feedback across prompt views and constrain policy drift through KL regularization. We fine-tune the MLLM by updating only its LoRA (Hu et al., 2022) parameters:

$$\mathcal{L}_{\mathrm{stage3}} = \mathcal{L}_{\mathrm{policy}} + \lambda_{\mathrm{KL}} \mathcal{L}_{\mathrm{KL}}. \tag{11}$$

Here, $\mathcal{L}_{\mathrm{policy}}$ aggregates the $\mathcal{L}_{\mathrm{PPO}}$ using spatial and reference certainty weights. $\mathcal{L}_{\mathrm{KL}}$ penalizes divergence from the frozen Stage 2 reference model, with λ_{KL} controlling the regularization strength.

Table 1: HOI detection results on HICO-DET and V-COCO under different supervision settings.

Method	Detector	Backbone		HICO-DET			V-COCO	
		Vision	Text	Full	Rare	Non-Rare	AP^1_{role}	AP^2_{role}
Fully Supervised								
GEN-VLKT	DETR	RN50	CLIP-B	33.75	29.25	35.10	62.4	64.4
ADA-CM	DETR	CLIP-B	CLIP-B	33.80	31.72	34.42	–	61.4
HOIGen	DETR	CLIP-B	CLIP-B	34.84	34.52	34.94	–	–
HOICLIP	DETR	CLIP-B	CLIP-B	34.69	31.12	35.74	63.5	64.8
DA-HOI	DETR	Qwen2.5-VL	Qwen2.5-VL	44.58	46.17	44.08	69.4	71.7
Training-free								
AgentHOI	–	GPT-4o	GPT-4o	29.61	40.33	26.41	–	–
Weakly Supervised								
Explanation-HOI	Faster R-CNN	RNeXt101	–	10.63	8.71	11.20	–	–
MX-HOI	Faster R-CNN	RN101	–	16.14	12.06	17.50	–	–
OpenCat	Faster R-CNN	RN101	RoBERTa	19.72	14.56	21.01	14.6	15.2
PPR-FCN	Faster R-CNN	CLIP-RN50	–	17.55	15.69	18.41	–	–
Weakly HOI-CLIP	Faster R-CNN	CLIP-RN50	CLIP-RN50	22.89	22.41	23.03	43.0	48.1
RegFormer	Faster R-CNN	DINO-S	CLIP-B	30.06	30.67	29.88	–	56.5
RegFormer	Faster R-CNN	DINO-B	CLIP-B	33.33	35.04	32.82	–	–
GTR-HOI	Faster R-CNN	Qwen2.5-VL	Qwen2.5-VL	**36.07**	**41.45**	**34.46**	**53.1**	**59.9**
AlignFormer	DETR	RN101	–	20.85	18.23	21.64	15.8	16.3
RegFormer	DETR	DINO-S	CLIP-B	30.01	32.05	29.39	–	57.5
RegFormer	DETR	DINO-B	CLIP-B	32.90	35.18	32.21	–	–
GTR-HOI	DETR	Qwen2.5-VL	Qwen2.5-VL	**36.44**	**42.54**	**34.62**	**53.6**	**60.6**

Inference. At inference, detector pairs with validity scores $v_j > \tau_v$ are evaluated using a single classification prompt. The final interaction score combines both HOI branches with validity and detection confidence:

$$S_{jy} = \text{mean}(S^{\text{vis}}_{jy}, S^{\text{sem}}_{jy})\, v_j^{\beta} S^{\lambda}_{h,j} S^{\lambda}_{o,j}. \tag{12}$$

Here, $S_{h,j}$ and $S_{o,j}$ are human and object detection scores; β controls the validity contribution, and λ controls the detection-score contributions.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metrics. We evaluate our method on two widely used public benchmarks for HOI detection, HICO-DET (Chao et al., 2018) and V-COCO (Gupta & Malik, 2015). HICO-DET contains 38,118 training and 9,658 test images, while V-COCO provides 5,400 trainval and 4,946 test images. Following RegFormer (Park et al., 2026), we report results using both DETR (Carion et al., 2020) and Faster R-CNN (Ren et al., 2016) as object detectors. For HICO-DET, we report 600-class mAP on the Full, Rare, and Non-Rare subsets. Zero-shot evaluation adopts RF-UC and NF-UC, each withholding 120 HOI compositions and retaining 480 seen classes. We report Unseen, Seen, and Full mAP, together with the harmonic mean of Unseen and Seen mAP (HM). For V-COCO, we report Scenario 1 and Scenario 2 role AP.

Implementation Details. Stage 1 uses the frozen Qwen2.5-VL-3B-Instruct (Bai et al., 2025) to generate pseudo-GTs. Stage 2 comprises 10 epochs of UOT validity pretraining followed by 5 epochs of semantic fine-tuning. Both phases use AdamW with a learning rate of 10^{-4} and weight decay of 0.02. Semantic fine-tuning uses LoRA (Hu et al., 2022) of rank 8 while keeping the MLLM base parameters frozen. Stage 3 initializes from the Stage 2 checkpoint and updates only the LoRA parameters through weighted PPO across four prompt views. At inference, we apply a pair validity threshold of 0.15, setting $\beta = 1$ for both datasets and $\lambda = 0.5$ and 2.0 for HICO-DET and V-COCO, respectively. All experiments are conducted on two NVIDIA GeForce RTX 5090 D GPUs.

4.2 Comparison with State-of-the-Art Methods

Weakly Supervised HOI detection. Table 1 compares GTR-HOI under different supervision settings. GTR-HOI achieves the best results among the compared weakly supervised methods with

both detector architectures. On HICO-DET, it reaches 36.07 Full mAP with Faster R-CNN (Ren et al., 2016) and 36.44 with DETR (Carion et al., 2020). Compared with RegFormer (Park et al., 2026) using DINO-B (Oquab et al., 2023), these results yield gains of 2.74 and 3.54 mAP on Full, respectively. The Rare gains are larger, at 6.41 and 7.36 mAP, demonstrating stronger performance on classes with limited training examples. Our DETR configuration exceeds fully supervised GEN-VLKT (Liao et al., 2022), HOICLIP (Ning et al., 2023), and HOIGen (Guo et al., 2024) in Full mAP. We compare GTR-HOI with AgentHOI (Lei et al., 2026), another method that applies an MLLM to HOI detection. Using GPT-4o (Achiam et al., 2023) without task-specific training, AgentHOI achieves 29.61, 40.33, and 26.41 mAP on Full, Rare, and Non-Rare, respectively. GTR-HOI with DETR improves upon these results by 6.83, 2.21, and 8.21 mAP, respectively, with the largest gain on Non-Rare classes. On V-COCO, GTR-HOI achieves 53.1/59.9 role AP in Scenarios 1/2 with Faster R-CNN and 53.6/60.6 with DETR. Compared with RegFormer using DINO-S, Scenario 2 gains are 3.4 and 3.1 AP under the respective detectors. With Faster R-CNN, GTR-HOI exceeds Weakly HOI-CLIP (Wan et al., 2023) by 10.1 and 11.8 AP in the two scenarios.

Zero-shot HOI Detection. Table 2 compares zero-shot HOI detection under the RF-UC and NF-UC settings. GTR-HOI consistently outperforms the compared weakly supervised methods across all four metrics in both settings. Specifically, under RF-UC, our method achieves 42.38, 35.59, and 36.95 mAP on Unseen, Seen, and Full, respectively. These results yield improvements of +10.85, +5.34, and +6.45 mAP over RegFormer (Park et al., 2026). Under NF-UC, GTR-HOI achieves 35.20, 35.02, and 35.06 mAP, with corresponding gains of +9.76, +4.16, and +5.28 mAP. The improvements extend to both seen and unseen compositions, resulting in HM scores of 38.69 and 35.11, exceeding RegFormer by 7.81 and 7.22, respectively. In particular, NF-UC achieves nearly equal Unseen and Seen performance, with a gap of only 0.18 mAP. These results demonstrate improved generalization to unseen compositions alongside stronger performance on seen classes. Despite using only image-level HOI supervision, GTR-HOI also compares favorably with several fully supervised methods. Its RF-UC Unseen mAP exceeds all listed fully supervised results, including DA-HOI (Xuan et al., 2026) at 41.79 mAP.

Table 2: Zero-shot HOI detection. RF-UC and NF-UC denote rare-first and non-rare-first unseen compositions.

Method	RF-UC				NF-UC			
	Unseen	Seen	Full	HM	Unseen	Seen	Full	HM
Fully Supervised								
GEN-VLKT	21.36	32.91	30.56	25.91	25.05	23.38	23.71	24.19
ADA-CM	27.63	34.35	33.01	30.63	32.41	31.13	31.39	31.76
HOICLIP	25.53	34.85	32.99	29.47	26.39	28.10	27.75	27.22
CLIP4HOI	28.47	35.48	34.06	31.59	31.44	28.26	28.90	29.77
HOIGen	31.01	34.57	33.86	32.69	33.98	32.86	33.08	33.41
LAIN	31.83	35.06	34.41	33.37	36.41	32.44	33.23	34.31
HypHOI	31.90	37.18	36.12	34.34	37.65	34.30	34.97	35.90
DA-HOI	41.79	44.01	43.56	42.87	43.12	39.63	40.33	41.30
Training-free								
AgentHOI	38.64	-	-	-	28.04	-	-	-
Weakly & Fully Supervised								
OpenCat	21.46	33.86	31.38	26.27	23.25	28.04	27.08	25.42
Weakly Supervised								
RegFormer	31.53	30.25	30.50	30.88	25.44	30.86	29.78	27.89
GTR-HOI	**42.38**	**35.59**	**36.95**	**38.69**	**35.20**	**35.02**	**35.06**	**35.11**

4.3 Qualitative Results

Fig. 3 presents representative HOI predictions on the HICO-DET test set. In (a), both scoring branches assign higher scores to the non-rare interaction *hold teddy bear* and the rare interaction *kiss teddy bear* than to the negative candidate *hug*. In (b), GTR-HOI distinguishes *ride elephant* from *walk* and *hold*, despite their shared human–object pair. The visual score for *ride* reaches 0.7058, compared with 0.1441 for *walk*, with the semantic branch showing the same trend. In (c), the model simultane-



Figure 3: HOI detection results on the HICO-DET test set: (a) scoring rare and non-rare classes; (b) distinguishing subtle interaction differences; and (c) multi-label scoring. Orange, black, and red denote rare classes, non-rare classes, and negative predictions, respectively. Values report validity, visual, and semantic scores, respectively.

ously recognizes *ride*, *sit on*, and *row* for the *human–boat* pair, illustrating recognition of multiple interactions involving the same pair.

4.4 Ablation Studies

Table 3: Component ablations on HICO-DET. Stage 1 varies only pseudo-GTs (H/O/R: human/object/relation perspectives); Stage 2 builds on (e), with (f)–(j) modifying (k); Stage 3 builds on (k), with (l)–(m) modifying (n).

	Variant	Full	Rare	Non-Rare
(a)	Baseline	30.82	39.01	28.38
Stage 1				
(b)	H+O	31.45	39.94	28.91
(c)	3 random templates	32.12	40.20	29.71
(d)	High-confidence only	32.88	39.93	30.78
(e)	Full Stage 1 (H+O+R)	33.31	40.31	31.22
Stage 2				
(f)	Hard IoU	34.66	40.51	32.91
(g)	Balanced OT	35.03	40.67	33.35
(h)	SAP instead of PER	34.98	41.13	33.14
(i)	w/o hierarchical supervision	34.94	41.48	32.98
(j)	All pseudo-GTs	34.89	41.02	33.06
(k)	Full Stage 2	35.58	41.73	33.74
Stage 3				
(l)	4 random templates	36.04	41.96	34.27
(m)	w/o Gaussian	36.10	41.83	34.39
(n)	**GTR-HOI**	**36.44**	**42.54**	**34.62**

Stage 1. Table 3 first evaluates pseudo-GT generation while keeping the remaining baseline components unchanged. Using two grounding prompts in (b) yields a +0.63 mAP improvement on Full over the single-prompt baseline. The complete H+O+R configuration in (e) reaches 33.31 Full mAP, improving over the baseline by +2.49 mAP. Compared with three random templates in (c), the designed grounding perspectives provide a +1.19 mAP improvement on Full, supporting their effectiveness beyond increasing the prompt count. Moreover, retaining all confidence tiers improves Full mAP by +0.43 over using only groups with high confidence in (d), suggesting that uncertain hypotheses also provide useful supervision.

Stage 2. Building on (e), the complete Stage 2 configuration in (k) achieves 35.58 Full mAP, yielding a further +2.27 mAP improvement. UOT outperforms hard IoU matching in (f) and balanced OT in (g) by +0.92 and +0.55 mAP on Full, respectively, supporting soft correspondence with relaxed marginal constraints. Replacing SAP with PER provides a +0.60 mAP improvement across all three splits. Hierarchical supervision also yields a +0.64 Full mAP improvement over nonhierarchical supervision in (i). Variant (j) additionally includes medium residuals and low-confidence assignments in supervised fine-tuning with soft targets, while retaining the consensus targets in (k). Compared with (j), restricting supervised fine-tuning to consensus assignments in (k) yields gains of +0.69, +0.71, and +0.68 mAP on Full, Rare, and Non-Rare, respectively. These results support transferring localization supervision to detector pairs through UOT and combining PER with selective hierarchical supervision to improve HOI recognition under grounding uncertainty.

Stage 3. The complete model in (n) reaches 36.44 Full mAP, improving over Stage 2 by 0.86, 0.81, and 0.88 mAP on Full, Rare, and Non-Rare, respectively. The designed prompt views outperform four random templates in (l) by 0.40 Full mAP and 0.58 Rare mAP. Spatial and semantic Gaussian weighting jointly yield gains of 0.34 and 0.71 mAP on Full and Rare over (m). These results support refining predictions learned from pseudo-GTs through multi-view semantic feedback and uncertainty weighting, building on the spatial supervision established in earlier stages.

5 Conclusion

We presented GTR-HOI, which addresses localization uncertainty and semantic ambiguity in HOI detection using only image-level labels. MLLM grounding from human, object, and relation perspectives generates pseudo-GTs organized by spatial confidence. UOT transfers their spatial support to detector pairs for validity learning, while hierarchical supervision trains visual and semantic HOI scoring through PER and the MLLM. Cross-prompt RL further refines interaction predictions through PPO. Experiments on HICO-DET and V-COCO demonstrate improvements over the compared weakly supervised methods, including gains on rare classes and unseen compositions on HICO-DET. These results highlight the value of combining explicit spatial supervision with semantic feedback for HOI detection without instance-level interaction annotations.

AI USE STATEMENT

We employed a large language model (LLM) as a writing assistant to improve the clarity, readability, and presentation of our paper. Specifically, the LLM was used to polish grammar, refine phrasing, and reorganize text for better readability. All core research ideas, technical designs, experimental implementations, analyses, and conclusions were conceived and developed entirely by the authors. The authors take full responsibility for the content of the paper.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 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.

Federico Baldassarre, Kevin Smith, Josephine Sullivan, and Hossein Azizpour. Explanation-based weakly-supervised learning of visual relations with graph networks. In *Proceedings of the European Conference on Computer Vision*, pp. 612–630, 2020. doi: 10.1007/978-3-030-58604-1_37.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human–object interactions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pp. 381–389, 2018. doi: 10.1109/WACV.2018.00048.

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.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of computation*, 87(314):2563–2609, 2018.

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

Yixin Guo, Yu Liu, Weimin Wang, Yanming Guo, and Qi Jia. Hyphoi: Exploring hierarchical hyperbolic embeddings for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 6517–6527, June 2026.

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

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Mert Kilickaya and Arnold W. M. Smeulders. Human–object interaction detection without alignment supervision. In *Proceedings of the British Machine Vision Conference*, 2021. doi: 10.5244/C.35.12.

Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. HOTR: End-to-end human–object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20190–20200, 2025.

Suresh Kirthi Kumaraswamy, Miaojing Shi, and Ewa Kijak. Detecting human–object interaction with mixed supervision. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pp. 1227–1236, 2021. doi: 10.1109/WACV48630.2021.00127.

Ting Lei, Fabian Caba, Qingchao Chen, Hailin Jin, Yuxin Peng, and Yang Liu. Efficient adaptive human-object interaction detection with concept-guided memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6480–6490, 2023.

Ting Lei, Jialin Liu, Zhu Xu, Yuxin Peng, and Yang Liu. Unleashing multimodal large language models for training-free HOI detection in the wild. *arXiv preprint arXiv:2607.13881*, 2026.

Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. GEN-VLKT: Simplify association and enhance interaction understanding for HOI detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20091–20100, 2022. doi: 10.1109/CVPR52688.2022.01949.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Yunyao Mao, Jiajun Deng, Wengang Zhou, Li Li, Yao Fang, and Houqiang Li. Clip4hoi: towards adapting clip for practical zero-shot hoi detection. *Advances in Neural Information Processing Systems*, 36:45895–45906, 2023.

Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. HOICLIP: Efficient knowledge transfer for HOI detection with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23507–23517, 2023. doi: 10.1109/CVPR52729.2023.02251.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Jihwan Park, Chanhyeong Yang, Jinyoung Park, Taehoon Song, and Hyunwoo J. Kim. RegFormer: Transferable relational grounding for efficient weakly-supervised human–object interaction detection. *arXiv preprint arXiv:2604.00507*, 2026.

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

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. QPIC: Query-based pairwise human–object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10405–10414, 2021. doi: 10.1109/CVPR46437.2021.01027.

Mesut Erhan Unal and Adriana Kovashka. Weakly-supervised HOI detection from interaction labels only and language/vision-language priors. *arXiv preprint arXiv:2303.05546*, 2023. doi: 10.48550/arXiv.2303.05546.

Bo Wan, Yongfei Liu, Desen Zhou, Tinne Tuytelaars, and Xuming He. Weakly-supervised HOI detection via prior-guided bi-level representation learning. In *International Conference on Learning Representations*, 2023.

Eastman Z. Y. Wu, Yali Li, and Shengjin Wang. Towards open-world human–object interaction reasoning with multimodal large language model. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8392–8396, 2026. doi: 10.1109/ICASSP55912.2026.11462048.

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

Zhaoning Yu, Will Su, Leitian Tao, Haozhu Wang, Aashu Singh, Hanchao Yu, Jianyu Wang, Hongyang Gao, Weizhe Yuan, Jason Weston, Ping Yu, and Jing Xu. RESTRAIN: From spurious votes to signals – self-driven RL with self-penalization. *arXiv preprint arXiv:2510.02172*, 2025.

Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human–object interactions with a novel unary-pairwise transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20072–20080, 2022. doi: 10.1109/CVPR52688.2022.01947.

Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. PPR-FCN: Weakly supervised visual relation detection via parallel pairwise R-FCN. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4243–4251, 2017. doi: 10.1109/ICCV.2017.454.

Sipeng Zheng, Boshen Xu, and Qin Jin. Open-category human–object interaction pre-training via language modeling framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19392–19402, 2023. doi: 10.1109/CVPR52729.2023.01858.

APPENDIX

This appendix provides additional methodological details, experimental analyses, and discussion, organized as follows:

- **Section A: Pretrained Models in Zero-Shot HOI Detection.** Evaluation conventions and potential overlap with foundation-model pretraining data.
- **Section B: Choice of the Baseline.** Rationale for adopting DA-HOI and comparison with other MLLM formulations.
- **Section C: Method Details.** Implementation settings, pseudo-GT generation and quality, UOT supervision transfer, prompt templates, and training objectives.
- **Section D: Additional Experiments.** PER expert ablations, matching quality, threshold and cost-weight sensitivity, Gaussian weighting, repeated RL runs, computational cost, and qualitative results.
- **Section E: Discussion and Future Directions.** Current limitations and corresponding directions for future research.

A PRETRAINED MODELS IN ZERO-SHOT HOI DETECTION

Pretrained vision–language models are widely used in zero-shot HOI detection. GEN-VLKT (Liao et al., 2022), HOICLIP (Ning et al., 2023), and HOIGen (Guo et al., 2024) use CLIP (Radford et al., 2021) knowledge to recognize unseen interactions. Recent MLLM methods also build on pretrained knowledge. DA-HOI (Xuan et al., 2026), HOI-R1 (Chen et al., 2025), and HOI-MLLM (Wu et al., 2026) use Qwen2.5-VL-3B-Instruct (Bai et al., 2025), the same backbone adopted by GTR-HOI, while AgentHOI (Lei et al., 2026) employs GPT-4o (Achiam et al., 2023) for training-free HOI detection. Under conventional zero-shot HOI protocols, "unseen" refers to compositions excluded from task-specific supervision, without requiring their absence from foundation-model pretraining. GTR-HOI follows this convention under RF-UC and NF-UC. The 120 held-out compositions in each split are excluded from training assignments, classification candidates, and associated losses; the full candidate vocabulary is restored at inference. Our evaluation therefore assesses the transfer of pretrained knowledge to compositions without task-specific supervision, consistent with prior zero-shot HOI studies. These precedents support our use of a pretrained MLLM, although we cannot fully audit its pretraining corpus or rule out benchmark overlap.

B CHOICE OF THE BASELINE

AgentHOI (Lei et al., 2026) performs interaction discovery and grounding without task-specific training; adapting it as a learning baseline would require defining objectives for image-level supervision. HOI-R1 (Chen et al., 2025) generates bounding boxes and interaction labels as text, using instance annotations for answer supervision and localization rewards. HOI-MLLM (Wu et al., 2026) supports both image-only and detector-conditioned inputs, but formulates prediction as structured text generation. Its training combines chain-of-thought supervision with GRPO using image-level and instance-level rewards. Adapting HOI-R1 or HOI-MLLM would therefore require replacing their instance-level supervision with objectives that account for uncertain pseudo-GTs. We therefore choose DA-HOI (Xuan et al., 2026), which separates object detection from interaction recognition and scores candidate interactions for supplied human–object regions in a single forward pass. Its explicit region inputs and candidate scores provide a direct interface for UOT-based validity learning, hierarchical supervision, and cross-prompt refinement with fixed regions. This allows us to adapt supervision while retaining the interaction recognition formulation, without introducing coordinate-generation targets or localization rewards. The same architecture also accommodates the Faster R-CNN (Ren et al., 2016) and DETR (Carion et al., 2020) configurations evaluated in this work.

C METHOD DETAILS

C.1 IMPLEMENTATION DETAILS

Stage 2. PER uses a 7×7 RoI token grid, projects 2048-dimensional visual features to $d_s = 512$, and encodes geometry in 36 dimensions. Its cross-attention blocks use eight heads and a feed-forward width of 2048, with dropout set to 0.1. The learnable cosine logit scale is initialized to $1/0.07$.

Both trainable stages use BF16 computation, gradient-norm clipping at 1.0, random seed 42, and random horizontal flips with probability 0.5. Validity pretraining and semantic fine-tuning use linear warm-up over 5% and 3% of their optimizer updates, respectively, followed by cosine decay to zero.

The UOT cost weights are $(\nu_p, \nu_u, \nu_d, \nu_s) = (0.55, 0.20, 0.15, 0.10)$. The entropy and marginal-penalty parameters (ϵ, ρ) are $(0.04, 1.0)$, $(0.07, 0.5)$, and $(0.12, 0.25)$ for groups with high, medium, and low confidence, respectively. Affinity computation uses $T_a = 0.35$ and $\delta = 10^{-12}$, while the source-reliability floor is $\xi_{\min} = 0.05$. Sinkhorn (Chizat et al., 2018) optimization uses at most 100 iterations with tolerance 10^{-5}. Validity supervision uses support thresholds $(\tau_+, \tau_-) = (0.3, 0.1)$. Supervised fine-tuning uses only consensus assignments from groups with high or medium confidence; medium residuals and low-confidence assignments remain eligible for Stage 3 RL, subject to the support threshold and pair budget. All confidence tiers contribute to validity pretraining. The loss coefficients in Eq. 10 are $(\lambda_1, \lambda_2, \lambda_3) = (1, 1, 0.75)$.

Stage 3. PPO (Schulman et al., 2017) runs for one epoch using AdamW with learning rate 2×10^{-5} and weight decay 0.02. Linear warm-up covers 5% of optimizer updates, followed by cosine decay to zero. Semantic Gaussian weighting uses width $\sigma = 0.5$, and spatial Gaussian weighting uses $\sigma_g = 0.5$. The clipping width is $\gamma = 0.2$, and each feedback snapshot supports $K_{\text{PPO}} = 2$ optimizer updates. Reference certainty uses floor $u_{\min} = 0.1$ and exponent $\gamma_u = 1$. The KL regularization coefficient is $\lambda_{\text{KL}} = 0.1$. Probabilities are clipped to $[10^{-6}, 1 - 10^{-6}]$ before evaluating policy ratios and Bernoulli KL.

C.2 STAGE 1

Generation and Output. For each image and HOI label, the frozen Qwen2.5-VL-3B-Instruct (Bai et al., 2025) receives the full image and one of three grounding prompts, emphasizing the human, object, or relation. Each independent call requests one human–object pair, its enclosing union, and a self-reported confidence. The complete system message and three user templates appear in Figs. 4–7. The image dimensions, object category, and interaction statement are substituted into each template; the local label ID is always `H0`. This ID is mapped back to the dataset label after parsing. The three prompts are batched for efficiency but do not share responses. No detector proposals or instance annotations are supplied. For HICO-DET (Chao et al., 2018), prompts for *no interaction* explicitly request a visible human–object pair exhibiting no active interaction. These pseudo-GTs are retained, and detector pairs matched to them receive positive validity targets when they satisfy the same positive assignment criterion under UOT. Thus, validity supervision includes pairs associated with *no interaction* and does not exclusively indicate the presence of an active interaction.

Pseudo-GT Localization Quality. Table 4 evaluates the localization quality of pseudo-GT boxes generated by the MLLM in Stage 1 on the HICO-DET (Chao et al., 2018) training split and V-COCO (Gupta & Malik, 2015) trainval split. Groups count retained image–HOI assignments. A match requires both human and object IoU ≥ 0.5 with the same annotated HOI instance. Primary localization precision evaluates the representative pair, whereas bag coverage measures the fraction of groups containing at least one matching hypothesis among the available candidates. The two metrics coincide for high-confidence groups, which contain one consensus pair.

Primary localization precision follows the confidence ordering on both datasets, reaching 71.35% and 69.75% for groups with high confidence on HICO-DET and V-COCO, respectively. Considering alternative hypotheses improves coverage over primary precision for groups with medium and low confidence. The respective absolute gains are 17.3% and 15.8% on HICO-DET, and 11.7% and 8.5% on V-COCO. This indicates that alternatives recover some localizations missed by the representative pair, while spatial agreement distinguishes groups with different localization reliability. Nevertheless, coverage for groups with low confidence remains 53.65% on HICO-DET and 39.94% on V-COCO at IoU 0.5, leaving many groups without a matching hypothesis. These observations

System Message

You are a precise visual grounding annotator for human-object interactions. You receive one image and one image-level interaction label that is known to be positive. Before selecting a pair, exhaustively inspect the full image for every visible person and every instance of the target object, including small, partially occluded, truncated, and image-boundary instances. For the supplied interaction, return the single best-supported visible human-object pair. Do not return null merely because the interaction is subtle, non-contact, or partially occluded. Return null boxes only when no person or no target-object instance can be localized in the image. The label may be marked NO_INTERACTION. For that label, select a clearly visible person and target object that show no direct action, contact, or manipulation. Do not select an actively interacting pair merely because both instances are visible. Return JSON only, with no markdown or explanation. Coordinates must be integer pixels in the original image in [x1, y1, x2, y2] order, where (0, 0) is the top-left. Human and object boxes must be tight. The union box must be the tight enclosing rectangle of the selected human and object boxes. All boxes must stay within the supplied original image bounds. For person-person interactions, human_box and object_box must identify two different people. Return exactly one grounding entry for the supplied interaction ID H0. Use this schema:
{"coordinate_space":"pixel","image_size":[width,height],"groundings":[{"interaction_id":"H0","union_box":[ux1,uy1,ux2,uy2],"human_box":[hx1,hy1,hx2,hy2],"object_box":[ox1,oy1,ox2,oy2],"confidence":0.0}]}
The u*, h*, and o* symbols denote three independently estimated boxes; do not copy one box into another. The human box contains only the person, the object box contains only the target object, and their enclosing rectangle gives the union box. Confidence must be between 0 and 1. Never output any interaction ID other than H0.

Figure 4: System message template.

Human Perspective

Human-first exhaustive search: scan the complete image for every person, including partial people near image boundaries. Rank people by action-specific body pose, hands, gaze, and motion direction, then inspect every target-object instance around each person. Do not stop at the largest or most central person. For NO_INTERACTION, select a visible person whose pose, hands, and gaze show the clearest absence of action toward a visible target object. The image-level label is guaranteed positive. If both a person and a target-object instance are visible, always return the best-supported pair. When several pairs are plausible, prioritize action-specific evidence and correct target-object identity; use visibility only as a final tie-break and never reject a pair solely for occlusion, small size, truncation, or distance from the image center. For NO_INTERACTION, prioritize the clearest absence of an active relation between the selected instances.
Original image size: width=***<W>*** pixels, height=***<H>*** pixels.
Set "coordinate_space" to "pixel" and "image_size" to [***<W>***, ***<H>***].Target object category: ***<OBJECT>***
Positive image-level interaction: H0: ***<MARKER> <STATEMENT>*** (target object category: ***<OBJECT>***)
Silently verify that the selected human and object are distinct visible instances, that the object belongs to the requested category, and that both boxes are tight. For the interaction ID, output its union_box, human_box, and object_box together in the same grounding entry. Do not crop or request another image view. Use null for all three boxes only when no person or no target-object instance is visible.

Figure 5: Human perspective template.

Object Perspective

Scan the complete image systematically for every instance of the target object category, explicitly including small, partially occluded, truncated, and peripheral instances. Verify object identity before comparing all visible people around each instance. Do not default to the largest or most central object. For NO_INTERACTION, select the visible object-person pairing with the clearest absence of contact or manipulation. The image-level label is guaranteed positive. If both a person and a target-object instance are visible, always return the best-supported pair. When several pairs are plausible, prioritize action-specific evidence and correct target-object identity; use visibility only as a final tie-break and never reject a pair solely for occlusion, small size, truncation, or distance from the image center. For NO_INTERACTION, prioritize the clearest absence of an active relation between the selected instances.
Original image size: width=***<W>*** pixels, height=***<H>*** pixels.
Set "coordinate_space" to "pixel" and "image_size" to [***<W>***, ***<H>***].Target object category: ***<OBJECT>***
Positive image-level interaction: H0: ***<MARKER> <STATEMENT>*** (target object category: ***<OBJECT>***)
Silently verify that the selected human and object are distinct visible instances, that the object belongs to the requested category, and that both boxes are tight. For the interaction ID, output its union_box, human_box, and object_box together in the same grounding entry. Do not crop or request another image view. Use null for all three boxes only when no person or no target-object instance is visible.

Figure 6: Object perspective template.

Relation Perspective

Form all plausible visible person-target-object pairs, then compare them using action-specific contact when applicable, relative geometry, body pose, gaze, and semantic compatibility. Non-contact actions do not require visible contact. Check ambiguous and occluded pairs before choosing the strongest relation. For NO_INTERACTION, select the pair with the strongest evidence that no listed action occurs. The image-level label is guaranteed positive. If both a person and a target-object instance are visible, always return the best-supported pair. When several pairs are plausible, prioritize action-specific evidence and correct target-object identity; use visibility only as a final tie-break and never reject a pair solely for occlusion, small size, truncation, or distance from the image center. For NO_INTERACTION, prioritize the clearest absence of an active relation between the selected instances.
Original image size: width=***<W>*** pixels, height=***<H>*** pixels.
Set "coordinate_space" to "pixel" and "image_size" to [***<W>***, ***<H>***].Target object category: ***<OBJECT>***
Positive image-level interaction: H0: ***<MARKER> <STATEMENT>*** (target object category: ***<OBJECT>***)
Silently verify that the selected human and object are distinct visible instances, that the object belongs to the requested category, and that both boxes are tight. For the interaction ID, output its union_box, human_box, and object_box together in the same grounding entry. Do not crop or request another image view. Use null for all three boxes only when no person or no target-object instance is visible.

Figure 7: Relation perspective template.

motivate retaining spatial supervision from groups with low confidence for validity learning while deferring their interaction learning to RL without fixed positive targets.

Table 4: Stage 1 pseudo-GT quality on HICO-DET train and V-COCO trainval at IoU 0.5.

Dataset	Confidence	Groups	Primary localization precision	Bag coverage
HICO-DET	High	49,277	71.35	71.35
	Medium	12,528	42.94	60.19
	Low	4,869	37.81	53.65
V-COCO	High	6,622	69.75	69.75
	Medium	2,370	43.97	55.70
	Low	1,069	31.43	39.94

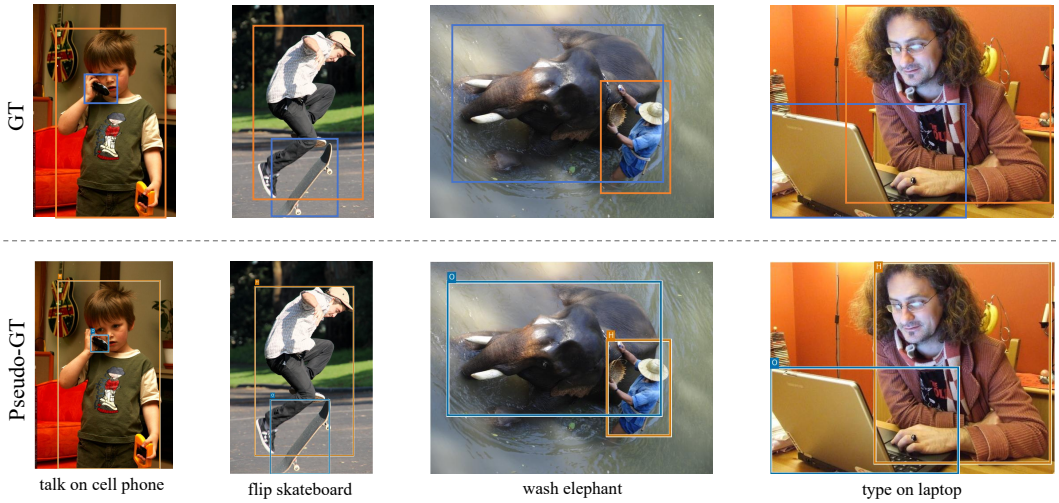


Figure 8: Stage 1 pseudo-GTs with high confidence on HICO-DET. Ground-truth boxes (top) are compared with consensus pairs from three agreeing prompt views (bottom). Orange and blue boxes indicate human and object roles, respectively.

Qualitative Analysis of Pseudo-GTs. Figs. 8, 9, and 10 present selected Stage 1 pseudo-GTs from the HICO-DET training set with high, medium, and low confidence, respectively. Each example depicts a single interaction category, with orange and blue boxes indicating the human and object roles, respectively.

For the high-confidence examples in Fig. 8, the consensus pairs identify the relevant participants across a wide range of object scales, from a cell phone to an elephant. The skateboard and laptop examples likewise show agreement with the annotated participants. Together, these examples illustrate how agreement across the three prompt views can provide explicit spatial supervision.

By contrast, Fig. 9 illustrates two forms of partial agreement underlying medium-confidence pseudo-GTs. For *hold person*, the consensus preserves the annotated participant roles, whereas the residual hypothesis reverses the human and object assignments. For *sit on bench*, the two hypotheses identify similar human regions but differ in the spatial extent of the bench. These examples reveal uncertainty in either participant-role assignment or localization boundaries.

The low-confidence examples in Fig. 10 exhibit more substantial disagreement across prompt views. For *eat hot dog*, some views select food on the plate rather than the food being eaten, while the relation view also mislocalizes the human. For *hold pizza*, the views differ in their selection of the person or pizza region. Despite these discrepancies, the human view for *eat hot dog* and the relation view for *hold pizza* provide plausible localizations of the annotated participants. These observations suggest that low-confidence groups may still contain useful spatial hypotheses, motivating their retention as uncertain spatial evidence for UOT-based validity supervision.

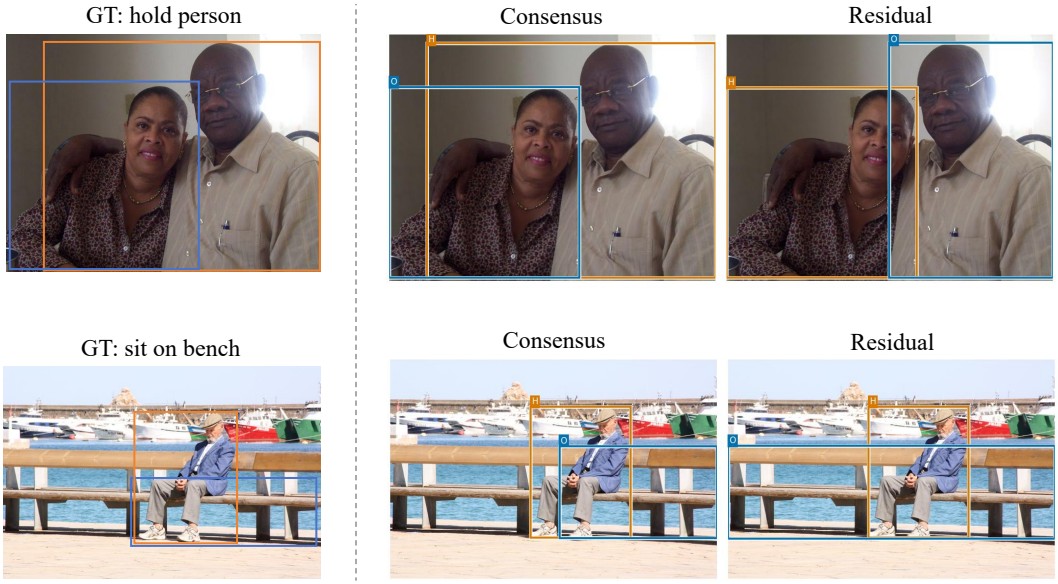


Figure 9: Stage 1 pseudo-GTs with medium confidence on HICO-DET. Each row shows the ground-truth pair (left), the consensus from two agreeing views (center), and the residual hypothesis from the third view (right). Orange and blue boxes indicate human and object roles, respectively.

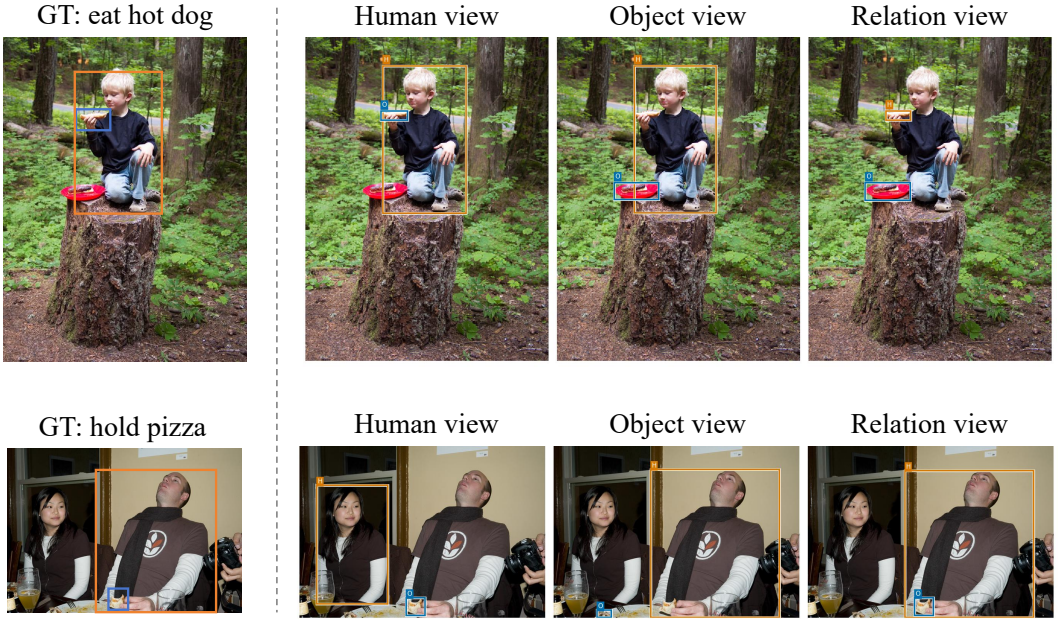


Figure 10: Stage 1 pseudo-GTs with low confidence on HICO-DET. Each row compares the ground-truth pair (left) with hypotheses from the human, object, and relation views. The three hypotheses are retained as alternatives. Orange and blue boxes indicate human and object roles, respectively.

C.3 STAGE 2

Unbalanced Optimal Transport. For each group, we define the normalized reference marginals using the reliability ξ_i and detector prior d_j introduced in the main text:

$$m_i^{\mathrm{s}} = \frac{\max(\xi_i, \xi_{\min})}{\sum_l \max(\xi_l, \xi_{\min})}, \qquad m_j^{\mathrm{t}} = \frac{0.5 + 0.5 d_j}{\sum_l (0.5 + 0.5 d_l)}. \tag{13}$$

The floor $\xi_{\min} = 0.05$ ensures strictly positive source reference masses, while the matching cost retains the original reliability ξ_i. The generalized KL divergence in Eq. 5 is defined as $D_{\mathrm{KL}}(x\|y) = \sum_i [x_i \log(x_i/y_i) - x_i + y_i]$.

Because UOT relaxes the marginal constraints, the received mass m_j alone cannot be interpreted as a validity probability. Eq. 6 therefore combines relative transport mass with matching affinity. The first factor, $\min(1, m_j/(m_j^{\mathrm{t}} + \delta))$, measures transported mass relative to the target reference mass. Clipping at one prevents excess mass from increasing this factor further. The second factor is a numerically stabilized mean of matching affinities, weighted by the transported mass. A high spatial support score therefore requires both sufficient relative mass and favorable matching affinity.

To characterize this score, consider a nonempty group G and define $c_{\min,j} = \min_i C_{ij}$, omitting the group index from costs and marginals. Under $C_{ij} \geq 0$, $\Pi^\star_{ij} \geq 0$, $m_j^{\mathrm{t}} > 0$, $T_a > 0$, and $\delta > 0$, the support satisfies

$$0 \leq u_j^{(G)} \leq \min\left(1, \frac{m_j}{m_j^{\mathrm{t}} + \delta}\right) \frac{m_j}{m_j + \delta} \exp\left(-\frac{c_{\min,j}}{T_a}\right) \leq 1. \tag{14}$$

Proof. Nonnegativity follows from the nonnegative factors in Eq. 6. Since $C_{ij} \geq c_{\min,j}$ and $\sum_i \Pi^\star_{ij} = m_j$,

$$\sum_i \Pi^\star_{ij} \exp\left(-\frac{C_{ij}}{T_a}\right) \leq m_j \exp\left(-\frac{c_{\min,j}}{T_a}\right).$$

Substitution into Eq. 6 yields the upper bound, whose three factors each lie in $[0, 1]$. The bound also holds when $m_j = 0$, in which case the support is zero.

This bound shows that insufficient mass relative to the target reference marginal limits support even when matching affinities are high. Similarly, if all matching costs satisfy $C_{ij} \geq c_0 \geq 0$, then $u_j^{(G)} \leq \exp(-c_0/T_a)$ regardless of the received mass. We obtain u_j by taking the maximum over groups from the same image whose object category matches that of the detector pair. This preserves $u_j \in [0, 1]$, providing a bounded spatial support score for validity supervision.

We assign validity targets by thresholding spatial support u_j, matching cost c_j, and detector confidence d_j, using the same rule for all confidence tiers. Here, $c_j = \min_{G,i} C_{G,ij}$ is the minimum matching cost over source hypotheses from the same image and object category. Pairs without a compatible source have $u_j = 0$ and $c_j = +\infty$ and are ignored. The validity targets are

$$t_j^{\mathrm{valid}} = \begin{cases} 1, & u_j \geq 0.3, \\ 0, & u_j \leq 0.1, \quad 0.65 \leq c_j < +\infty, \quad d_j \geq 0.20, \\ \text{ignore}, & \text{otherwise}. \end{cases} \tag{15}$$

The validity loss is the weighted mean of focal losses over positive and negative pairs, with respective weights of 1 and 0.25. The complete UOT is summarized in Algorithm 1.

Training Objectives. Following DA-HOI (Xuan et al., 2026), we first pretrain pair validity using binary focal loss, denoted by FocalBCE (Lin et al., 2017). For minibatch $\mathcal{B}$, let $\mathcal{J}_I^{\mathrm{valid}}$ contain detector pairs in image I with non-ignored targets from Eq. 15. These targets $t_j^{\mathrm{valid}} \in \{0, 1\}$ are obtained using all pseudo-GT confidence tiers. The validity head predicts v_j from the SF feature $f_{\mathrm{SF},j}$ of pair j, giving

$$\mathcal{L}_{\mathrm{valid}} = \frac{1}{|\mathcal{B}|} \sum_{\substack{I \in \mathcal{B} \\ \mathcal{J}_I^{\mathrm{valid}} \neq \varnothing}} \frac{\sum_{j \in \mathcal{J}_I^{\mathrm{valid}}} \mu_j \, \mathrm{FocalBCE}(v_j, t_j^{\mathrm{valid}})}{\sum_{j \in \mathcal{J}_I^{\mathrm{valid}}} \mu_j}, \tag{16}$$

where $\mu_j = 1$ for positives and 0.25 for negatives. Images without supervised pairs contribute zero. Minimizing $\lambda_1 \mathcal{L}_{\text{valid}}$ with $\lambda_1 = 1$ updates only the PER experts, SF, and validity head.

We then train the visual and semantic branches using assignments retained by Algorithm 1, selecting only those associated with consensus hypotheses of high or medium confidence. For these positive assignments, the targets t_{jy} are 1 and $2/3$ for high and medium confidence, respectively. Assuming complete image-level labels, compatible training classes absent from the image labels receive $t_{jy} = 0$ in each pair's highest supervised tier. For tier $r \in \{\text{high}, \text{medium}\}$, the binary mask m^r_{jy} is one for the corresponding positive and negative entries and zero otherwise. The loss for each tier is

$$\mathcal{L}_r = \frac{1}{\max(1, N_r)} \sum_{j,y} m^r_{jy} \left[\text{FocalBCE}(S^{\text{vis}}_{jy}, t_{jy}) + \text{FocalBCE}(S^{\text{sem}}_{jy}, t_{jy}) \right]. \tag{17}$$

The sums range over selected detector pairs in the minibatch and their compatible training classes. Here, $N_r = \sum_j \mathbf{1}[\sum_y m^r_{jy} > 0]$ counts pairs with supervised entries at tier r. The hierarchical component of Eq. 10 is

$$\mathcal{L}_{\text{sem}} = \lambda_2 \mathcal{L}_{\text{high}} + \lambda_3 \mathcal{L}_{\text{medium}}, \qquad (\lambda_2, \lambda_3) = (1, 0.75). \tag{18}$$

This phase updates only LoRA, VF, the interaction projection, the visual projection P_E, the candidate marker embedding, and the logit scale.

Algorithm 1 UOT

Input: Training images, Stage 1 pseudo-GT groups, and detector pairs.
Output: Pair validity targets (positive, negative, or ignored) and retained UOT assignments with supervision metadata.

for each training image I **do**
  Initialize $u_j \leftarrow 0$ and $c_j \leftarrow +\infty$ for every detector pair j.
  for each pseudo-GT group G in image I **do**
    Use $\mathcal{H}^G$ and detector pairs with the same object category; skip if either set is empty.
    Construct C_{ij}, m^{s}_i, and m^{t}_j using Eqs. 4 and 13; use (ϵ, ρ) for the group's confidence tier.
    Solve Eq. 5 for $\Pi^\star$ using generalized Sinkhorn iterations in the log domain.
    Compute $u^{(G)}_j$ using Eq. 6; update $u_j \leftarrow \max(u_j, u^{(G)}_j)$ and $c_j \leftarrow \min(c_j, \min_i C_{ij})$ for these pairs.
    Retain assignments with $u^{(G)}_j \geq \tau_{\text{sem}}$; attribute each to $\arg\max_i \Pi^\star_{ij}$, favoring consensus hypotheses in ties.
    For retained assignments, compute $c_G = \max_i \kappa_i$ and $c_{Gj} = \frac{\sum_i \Pi^\star_{ij} \kappa_i}{\sum_i \Pi^\star_{ij}}$.
    Store retained assignments with detector coordinates and supervision metadata.
  end for
  Assign fixed validity targets to all detector pairs using Eq. 15.
end for

C.4 Stage 3

Classification Prompt Templates. The four prompt views evaluate the same human–object pair from direct, human, object, and relation perspectives, indexed by $p = \{1, 2, 3, 4\}$, respectively. They share the image, pair coordinates, projected interaction token, and object-compatible candidate classes. Each uses the system message `You are a helpful assistant.` and the following user-message template:

```
[IMAGE]
The coordinates of the person are [HUMAN_BOX].
The coordinates of the object are [OBJECT_BOX].
The features of the interaction are [INTERACTION_FEATURE].
[VIEW_INSTRUCTION]: [CANDIDATE_LIST].
```

Here, `[IMAGE]` represents the full-image visual tokens, and the box placeholders specify integer coordinates (x_1, y_1, x_2, y_2) in the resized image. The placeholder `[INTERACTION_FEATURE]` is replaced by `<|box_start|>`, enclosed by the vision boundary tokens. The input embedding

at this marker is set to a projection of the SF feature f_{SF}. The placeholder `[CANDIDATE_LIST]` expands to the object-compatible interaction phrases, each followed by `<|box_end|>`. The contextualized hidden state at each candidate marker provides the feature $\Theta(y)$ for the preceding interaction phrase y. The four replacements for `[VIEW_INSTRUCTION]` are:

Direct view ($p = 1$):

```
Select all correct interactions between the given person and
object from the list:
```

Human perspective ($p = 2$):

```
Focus on the specified person's overall body posture, arm
and hand positions, and gaze direction when assessing the
given person-object pair.  Examine how these human-centered
cues relate to the target object, considering both physical
contact and non-contact interactions such as directed
attention.  Base your judgment on the visible evidence
rather than the object's presence or category alone.  Select
all supported interactions from the following list:
```

Object perspective ($p = 3$):

```
Focus on the target object's visible state, orientation,
and position relative to the specified person.  Inspect its
relevant parts, contact points, and signs of being held,
supported, worn, or manipulated.  Use these object-centered
cues to assess the candidate interactions, including those
that do not require direct physical contact.  Do not infer
an action solely from the object's typical function.  Select
all supported interactions from the following list:
```

Relation perspective ($p = 4$):

```
Focus on the spatial relationship between the specified
person and object, including their relative positions,
distance, orientation, and contact configuration.  Consider
how these pairwise cues support or contradict each candidate
interaction, using the surrounding context to clarify
ambiguous relationships rather than as sufficient evidence
on its own.  Distinguish meaningful interaction from
simple co-occurrence, and consider both contact-based and
non-contact actions.  Select all supported interactions from
the following list:
```

Training Objectives. Stage 3 takes detector human–object pairs selected through UOT as input. For minibatch $\mathcal{B}$, let $\mathcal{G}_{\mathcal{B}}$ contain the groups with at least one selected pair. For each group G, y_G denotes its HOI label and $\mathcal{J}_G$ contains its selected detector pairs. Algorithm 1 provides the group and pair prompt support values c_G and c_{Gj}. Using the Gaussian weighting function defined in Section 3.5, we obtain a group weight and normalized weights within each group:

$$w_G = \mathcal{G}_{\sigma_g}(c_G), \qquad \omega_{Gj} = \frac{\mathcal{G}_{\sigma_g}(c_{Gj})}{\sum_{l \in \mathcal{J}_G} \mathcal{G}_{\sigma_g}(c_{Gl})}, \tag{19}$$

where σ_g is the spatial Gaussian width. We write π^p_{jy} for the acceptance probability π^p in Section 3.5, with pair and class indices restored. Its counterpart from the frozen Stage 2 reference is $\pi^{\mathrm{ref},p}_{jy}$. Reference predictions computed once on fixed inputs give the certainty weight:

$$\bar{\pi}^{\mathrm{ref}}_{jy} = \frac{1}{4}\sum_{p=1}^{4} \pi^{\mathrm{ref},p}_{jy}, \qquad u^{\mathrm{ref}}_{jy} = u_{\min} + (1 - u_{\min}) \left|2\bar{\pi}^{\mathrm{ref}}_{jy} - 1\right|^{\gamma_u}, \tag{20}$$

where $u_{\min}$ sets the weight floor and γ_u controls its dependence on certainty. Confident acceptance and rejection both receive larger weights. The spatial and certainty weights remain fixed throughout

Stage 3. Let $\mathcal{L}_{\mathrm{PPO}}(j, y_G)$ denote the PPO loss from Section 3.5 evaluated for pair j and label y_G. The weighted policy loss is

$$\mathcal{L}_{\mathrm{policy}} = \frac{1}{|\mathcal{G}_{\mathcal{B}}|} \sum_{G \in \mathcal{G}_{\mathcal{B}}} w_G \sum_{j \in \mathcal{J}_G} \omega_{Gj} u^{\mathrm{ref}}_{jy_G} \mathcal{L}_{\mathrm{PPO}}(j, y_G). \tag{21}$$

For KL regularization, let $\mathcal{J} = \bigcup_{G \in \mathcal{G}_{\mathcal{B}}} \mathcal{J}_G$ contain the distinct selected pairs. For pair j with object category $c_{o,j}$, $\mathcal{C}_{\mathrm{tr}}(c_{o,j})$ denotes the compatible training HOI classes. The reference and current policy receive identical inputs. We average their Bernoulli KL divergence over the selected pairs, compatible classes, and four views:

$$\begin{aligned} \mathcal{L}_{\mathrm{KL}} &= \frac{1}{4|\mathcal{J}|} \sum_{j \in \mathcal{J}} \sum_{p=1}^{4} \frac{1}{|\mathcal{C}_{\mathrm{tr}}(c_{o,j})|} \sum_{y \in \mathcal{C}_{\mathrm{tr}}(c_{o,j})} D_{\mathrm{B}}(\pi^{\mathrm{ref},p}_{jy} \| \pi^{p}_{jy}), \\ D_{\mathrm{B}}(r\|s) &= r \log \frac{r}{s} + (1-r) \log \frac{1-r}{1-s}, \end{aligned} \tag{22}$$

where $D_{\mathrm{B}}(r\|s)$ is the Bernoulli KL divergence for acceptance probabilities r, s. Algorithm 2 summarizes Cross-prompt RL.

Algorithm 2 Cross-prompt RL

Input: Stage 2 checkpoint, detector-pair assignments retained by Algorithm 1, four prompt views, and K_{PPO} updates per round.
Output: Updated MLLM LoRA parameters.

```
Initialize the policy and frozen reference from the Stage 2 checkpoint.
Compute fixed weights w_G, ω_Gj, u^ref_jy using Eqs. 19 and 20.
for each PPO round do
    Load minibatch B; keep its images, pair regions, prompts, and candidate orders fixed for this round.
    For each assignment (j, y_G) and view p, cache P^p_old(a), q_t^{-p}, and A^p_t(a) from Section 3.5, omitting
  pair and class indices; keep these quantities fixed.
    Evaluate the reference on these inputs for KL regularization.
    repeat K_PPO times
        Compute current acceptance probabilities for all compatible training classes under all four views.
        Obtain L_PPO(j, y_G) for each selected assignment by exact enumeration, following Section 3.5.
        Compute L_policy and L_KL using Eqs. 21 and 22.
        Update LoRA parameters using L_stage3 in Eq. 11.
    end repeat
end for
```

D Additional Experiments

D.1 Pair Expert Routing

Table 5 evaluates the contribution of each expert to PER fusion, with the complete configuration (e) achieving the best results. Jointly excluding the human and object experts e_h, e_o in (a) reduces Full performance by 0.12 mAP, while removing the relation expert e_r in (b) yields a 0.18 mAP decrease. Excluding the geometry expert e_g in (c) reduces Rare performance by 0.41 mAP, compared with 0.05 mAP on Non-Rare, suggesting greater sensitivity to spatial information for rare interactions. Removing the context expert e_c in (d) causes the largest decreases: 0.39, 0.71, and 0.30 mAP on Full, Rare, and Non-Rare, respectively. These results highlight the contribution of contextual information and support combining appearance, relation, geometry, and context features. The consistently larger decreases on Rare classes suggest that the detailed and discriminative representations provided by the combined experts are particularly beneficial when training examples are limited.

D.2 Unbalanced Optimal Transport

Comparison of Pseudo-GT Matching Methods. Table 6 compares the validity supervision obtained by matching pseudo-GTs to detected human–object pairs. Incomplete or inaccurate pseudo-

Table 5: Expert ablation in PER on HICO-DET.

	Expert features					HICO-DET		
	e_h	e_o	e_r	e_g	e_c	Full	Rare	Non-Rare
(a)	–	–	✓	✓	✓	36.32	42.32	34.53
(b)	✓	✓	–	✓	✓	36.26	42.24	34.47
(c)	✓	✓	✓	–	✓	36.31	42.13	34.57
(d)	✓	✓	✓	✓	–	36.05	41.83	34.32
(e)	✓	✓	✓	✓	✓	36.44	42.54	34.62

GTs may leave genuine interactions unmatched, so weak matching support alone does not establish a negative label. The ignored category accommodates this uncertainty by withholding supervision when the available evidence is insufficient for either a positive or a negative assignment. UOT achieves the highest positive share and precision, providing more positive assignments with greater reliability. Compared with hard IoU matching, it improves positive precision from 72.25% to 80.30% and negative precision from 90.94% to 94.37%. Although balanced OT achieves the highest negative precision of 95.87%, it assigns only 6.49% of pairs as negatives and ignores 81.55%, substantially limiting supervision coverage. UOT increases the negative share to 39.35% while maintaining 94.37% precision and reduces the ignored share to 47.91%. These results indicate that UOT provides broader supervision than balanced OT and more accurate assignments than hard IoU matching, while retaining the flexibility to exclude uncertain pairs from validity learning.

Table 6: Comparison of pseudo-GT matching methods. Both OT methods use $(\tau_+, \tau_-) = (0.30, 0.10)$. Share denotes the percentage of all evaluated detector pairs assigned as positive, negative, or ignored. Positive (negative) precision is the percentage of pairs assigned as positive (negative) that are confirmed as such by comparison with GT annotations.

	Positive		Negative		Ignored
Method	Share	Prec.↑	Share	Prec.↑	Share
Hard IoU	10.56	72.25	89.44	90.94	0.00
Balanced OT	11.96	68.11	6.49	95.87	81.55
UOT	12.74	80.30	39.35	94.37	47.91

Table 7: Ablation of thresholds for human-object pair validity supervision on HICO-DET.

	τ_+	τ_-	Full	Rare	Non-Rare
(a)	0.30	0.100	36.44	42.54	34.62
(b)	0.25	0.100	36.12	42.68	34.17
(c)	0.35	0.100	36.13	42.01	34.37
(d)	0.30	0.050	36.30	42.56	34.43
(e)	0.30	0.075	36.34	42.47	34.51
(f)	0.30	0.125	36.37	42.11	34.65
(g)	0.30	0.150	36.36	41.89	34.71

Thresholds for Human–Object Pair Validity Supervision. Table 7 examines the positive and negative assignment thresholds used for pair validity supervision. The reference setting $(\tau_+, \tau_-) = (0.30, 0.10)$ in (a) achieves the highest Full mAP of 36.44. Lowering τ_+ to 0.25 in (b) yields the highest Rare mAP of 42.68, but reduces Full and Non-Rare performance to 36.12 and 34.17 mAP, respectively. Increasing τ_+ to 0.35 in (c) lowers performance, suggesting that stricter positive selection may exclude useful supervision. With $\tau_+ = 0.30$, the smaller negative thresholds in (d) and (e) achieve 36.30 and 36.34 Full mAP, respectively, below the reference configuration. Increasing τ_- beyond 0.10 in (f) and (g) improves Non-Rare performance to 34.65 and 34.71 mAP, but lowers Rare performance to 42.11 and 41.89 mAP, respectively. This divergence suggests that broader negative supervision may benefit frequent interactions while adversely affecting rare interactions.

These results support the reference thresholds for overall detection performance, while showing that more permissive assignment does not benefit Rare and Non-Rare classes uniformly.

Sensitivity to UOT Cost Weights. Table 8 presents a comparison of the relative weights assigned to pair overlap, union overlap, detector confidence, and source reliability. The reference configuration (a) yields the highest performance. Equal weighting in (b) reduces Full, Rare, and Non-Rare performance by 0.46, 0.92, and 0.32 mAP, respectively, suggesting a benefit from emphasizing pair overlap. Adjusting the union-overlap weight in (c) and (d) lowers Full performance to 36.32 and 36.18 mAP, respectively. Changes to the detector-confidence weight in (e) and (f) produce smaller Full reductions of 0.11 and 0.16 mAP, although the larger weight in (f) reduces Rare performance by 0.43 mAP. For source reliability, reducing its weight in (g) lowers Full mAP by 0.21, whereas increasing it in (h) yields results closest to the reference, with a 0.06 mAP decrease.

Table 8: UOT cost-weight sensitivity on HICO-DET with DETR.

	ν_p	ν_u	ν_d	ν_s	Full	Rare	Non-Rare
(a)	0.55	0.20	0.15	0.10	36.44	42.54	34.62
(b)	0.25	0.25	0.25	0.25	35.98	41.62	34.30
(c)	0.60	0.15	0.15	0.10	36.32	42.20	34.57
(d)	0.50	0.25	0.15	0.10	36.18	42.36	34.34
(e)	0.60	0.20	0.10	0.10	36.33	42.44	34.51
(f)	0.50	0.20	0.20	0.10	36.28	42.11	34.54
(g)	0.60	0.20	0.15	0.05	36.23	42.13	34.47
(h)	0.50	0.20	0.15	0.15	36.38	42.42	34.57

D.3 Cross-Prompt Reinforcement Learning

Gaussian Weighting. Table 9 separates the three Gaussian weighting operations combined in the main ablation. Removing semantic Gaussian weighting sets $q_1^{-p} = f^{-p}$ and $q_0^{-p} = 1 - f^{-p}$. Removing Gaussian weighting across groups sets $w_G = 1$, while removing it within each group sets $\omega_{Gj} = 1/|\mathcal{J}_G|$. Each variant changes only one operation, retaining the other two, fixed reference certainty, PPO clipping, and KL regularization. The complete removal of Gaussian weighting is already reported in Table 3.

Table 9: Separate ablations of Gaussian weighting on HICO-DET with DETR.

	Gaussian weighting			HICO-DET		
	q_t^{-p}	w_G	ω_{Gj}	Full	Rare	Non-Rare
(a)	–	✓	✓	36.19	42.19	34.40
(b)	✓	–	✓	36.27	42.22	34.49
(c)	✓	✓	–	36.34	42.38	34.54
(d)	✓	✓	✓	36.44	42.54	34.62

Table 9 examines the three Gaussian weighting operations used in RL refinement. The complete configuration (d) achieves the best performance. Removing Gaussian weighting from semantic feedback in (a) produces the largest reductions, lowering Full, Rare, and Non-Rare performance by 0.25, 0.35, and 0.22 mAP, respectively. This suggests that weighting semantic feedback helps refine interaction predictions. Removing Gaussian weighting across groups through w_G in (b) reduces Full and Rare performance by 0.17 and 0.32 mAP, respectively. Removing Gaussian weighting within groups through ω_{Gj} in (c) causes smaller decreases of 0.10 and 0.16 mAP. These comparisons support accounting for spatial uncertainty both across groups and among their assigned detector pairs. Rare classes exhibit the largest decreases in every variant, suggesting that uncertainty weighting is particularly beneficial when training examples are limited.

Stability Across Repeated Runs. Table 10 summarizes five repeated Stage 3 training runs under the same experimental configuration. The mean Full, Rare, and Non-Rare performance reaches 36.34, 42.55, and 34.49 mAP, with sample standard deviations of 0.07, 0.08, and 0.10 mAP, respectively.

Full performance ranges from 36.26 mAP in (b) to 36.44 mAP in (a), a difference of 0.18 mAP. The similarly small standard deviations on Rare and Non-Rare indicate limited variability. These results suggest that Stage 3 produces consistent detection performance across repeated training under the evaluated configuration.

Table 10: Results of five repeated Stage 3 training runs on HICO-DET with DETR. Rows (a)–(e) denote individual runs. The final row reports the mean and sample standard deviation.

	Full	Rare	Non-Rare
(a)	36.44	42.54	34.62
(b)	36.26	42.61	34.36
(c)	36.31	42.65	34.41
(d)	36.38	42.52	34.54
(e)	36.33	42.44	34.50
Mean $\pm$ std.	36.34$\pm$0.07	42.55$\pm$0.08	34.49$\pm$0.10

D.4 Computational Cost

Table 11 reports computational costs on a single NVIDIA RTX 5090 GPU, excluding the external detector. Validity pretraining requires the least GPU memory at 4.71 GiB, with 17.419M trainable parameters. Supervised fine-tuning and cross-prompt RL each use 3780.864M total parameters, but update only 8.556M and 2.507M parameters, respectively. Stage 3 thus updates approximately 0.07% of the model parameters through LoRA. Despite this smaller trainable parameter count, its peak memory reaches 20.39 GiB, compared with 15.52 GiB for supervised fine-tuning, indicating the additional memory demands of cross-prompt refinement. At inference, the model requires 13.28 GiB of peak memory and processes each image in 66.06 ms, corresponding to 15.14 FPS.

Table 11: Computational cost on one NVIDIA RTX 5090 GPU. Total parameter counts include frozen parameters. The external detector is excluded from all measurements.

Stage / setting	Total params. (M)	Trainable params. (M)	Peak GPU memory (GiB)	Inference time (ms/image)	FPS
Stage 2: Validity pretraining	686.104	17.419	4.71	–	–
Stage 2: Supervised fine-tuning	3780.864	8.556	15.52	–	–
Stage 3: Cross-prompt RL	3780.864	2.507	20.39	–	–
Inference	3780.864	–	13.28	66.06	15.14

D.5 Additional Qualitative Results

Fig. 11 illustrates successful predictions and remaining recognition errors. The cell phone and cup examples demonstrate recognition of multiple valid interactions for the same pair, while the second column shows positive interactions ranked above negative candidates. The third column correctly predicts *sit on* for two different people paired with the same bench, illustrating the ability to recognize multiple interacting pairs within one image.

The last column highlights difficulties in distinguishing closely related actions. In the upper example, the pair receives a validity score of 0.8903, but the visual and semantic scores for the correct interaction *carry* are only 0.2054 and 0.1840. Limited visibility of the racket and ambiguous pose cues may favor recognition of *hold* while providing weaker evidence for the co-occurring interaction *carry*. In the lower example, the annotated negative *hold* receives scores of 0.5100 and 0.5131, exceeding those for *carry*. This confusion may arise from overlapping visual and semantic cues, since carrying a person also involves physical support and close contact. These examples suggest that identifying an interacting pair is easier than resolving its specific actions when interaction categories share similar cues. Such ambiguity can affect both scoring branches, so their agreement does not necessarily ensure a correct prediction.

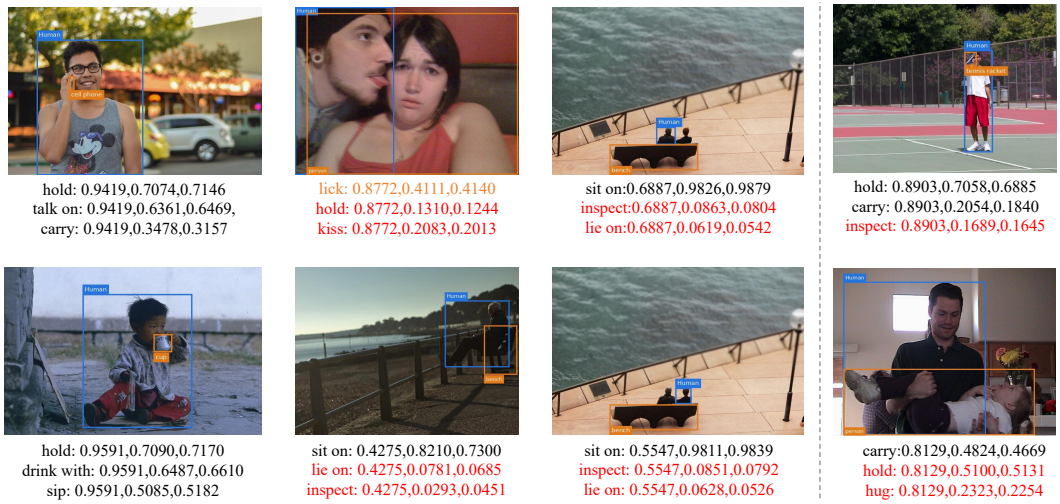


Figure 11: Additional qualitative HOI results. The first three columns show successful examples, while the last column illustrates underestimated scores for *carry* (top) and overestimated scores for *hold* (bottom). Blue and orange boxes denote humans and objects, respectively. Interaction labels in orange, black, and red indicate rare positives, non-rare positives, and negative categories. Each score triplet reports validity, visual, and semantic scores, respectively.

E Discussion and Future Directions

GTR-HOI remains dependent on the quality of MLLM grounding and detector proposals. Although spatial agreement helps assess grounding reliability, it does not guarantee correct localization (Table 4). Exhaustively localizing all interacting human–object pairs in an image also remains challenging for the MLLM, particularly when multiple pairs share the same interaction label. Each grounding prompt currently returns at most one hypothesis, and different views may identify the same pair, leaving other valid pairs without corresponding pseudo labels. Such omissions lead to incomplete supervision at the instance level, even when the relevant image-level labels are available. UOT transfers the available supervision to detector pairs, while Stage 3 refines predictions on fixed regions and cannot recover pairs absent from the detector proposals. Future work could explore iterative grounding to discover overlooked pairs and proposal refinement guided by image-level labels, improving supervision coverage without requiring annotated interaction boxes.