# HEIR: LEARNING HUMAN-ENTITY INTERACTIONS WITH FUNCTIONAL ROLES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Understanding human–entity interactions requires recovering each person–action event's participants, roles, and shared identities. This structure can support embodied agents by clarifying who acts on which entities and how, informing anticipation and coordination in shared environments. Standard HOI metrics score individual links, leaving complete event composition undermeasured. We introduce HEIR (Human–Entity Interactions with Functional Roles), an image benchmark for complete grounded participant–role sets across object, interpersonal, and self-directed interactions. It contains 18,730 images, six roles, 105 actions, and 437 nouns, with shared entities, role changes, and repeated fillers; 51.6% of images contain multiple actors and 62.1% contain multiple actions. HEIR pairs relation AP with complete-set AP and structural evaluation. We also introduce CoRISP (Compositional Role-aware Interaction Set Prediction), which uses shared entity identities to combine role-conditioned evidence and predict normalized participant–role sets. Cardinality and role-multiplicity potentials couple assignments through event size and role composition, with exact per-event normalization. Across 16 baselines, relation and complete-event rankings diverge even after aligning action weights. CoRISP leads the evaluated systems on repeated-role events and shared-participant images in HEIR by 2.87 and 3.82 Set mAP points, respectively. On V-COCO, CoRISP achieves 73.72/76.23 role AP and 61.06/68.59 complete-set AP on two-slot actions under Scenarios 1/2. These results show the value of learning and evaluating event composition alongside individual relations. We will release the code and benchmark data.

## 1 INTRODUCTION

A child pedals a bicycle as a parent runs alongside, steadying the saddle. A detector may recognize the child riding the bicycle yet fail to recover the event structure that tells us whether the child is riding unaided: the bicycle supports the child in the ride and is the target of the parent's stabilizing action. Understanding the scene therefore requires more than detecting plausible person–action–object relations. A model must recover the complete set of grounded participants in each person's action, assign each participant its functional role, and preserve entity identity across concurrent events. For embodied AI (Chang et al., 2025; Yuan et al., 2024; Li et al., 2023a), this role- and identity-aware grounding can inform action planning: a robot that knows who is acting on which entity—and whether that entity functions as a target, instrument, or support—can better anticipate human activity and coordinate its actions with people in shared environments. *Can a vision model do this reliably in a single image, including when several participants share a role, people act on one another, or actions are self-directed?*

Most human–object interaction evaluations emphasize localized person–action–object relations scored individually (Gkioxari et al., 2018; Chao et al., 2018). Such scores can reward a correctly detected relation without showing whether all participants have been assembled into the right event or whether shared entities have been tracked consistently across events. We address this gap with HEIR, a benchmark that evaluates complete participant–role sets for each person–action event, and CoRISP, which predicts a normalized distribution over participant–role sets while preserving entity identities across actions.



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.

The benchmarks that shaped image-based interaction research were designed for complementary tasks. V-COCO grounds objects in action-specific roles (Gupta & Malik, 2015); HICO-DET localizes person–verb–object triplets over the COCO object vocabulary (Chao et al., 2018); and SWiG grounds semantic arguments of a salient activity (Pratt et al., 2020). A correct local prediction can receive credit while another actor's relation to the same entity is missed. Existing benchmarks (Gupta & Malik, 2015; Chao et al., 2018) penalize that omission when it is annotated, but credit for the first relation does not establish that the concurrent actions have been recovered together. Meanwhile, work on social interaction and open-vocabulary HOI has brought new settings and vocabularies into focus (Wei et al., 2024; Lei et al., 2025). Large pretrained vision–language models can recognize a wider range of interactions, but CrossHOI-Bench finds that they still struggle to assign concurrent actions to the correct person (Lei et al., 2026). Recognition of an action or noun therefore does not establish the composition of the event in which it appears.

We introduce HEIR as a benchmark for reconstructing interaction composition in a single image. Its unit is the person–action event: a complete, variable-sized set of localized participants, each assigned a functional role, with entity identity preserved wherever it recurs. A shared ontology spans six roles, *i.e.*, target, instrument, support, source, destination, and constraint, across 105 actions and 437 nouns from everyday and specialized settings. Participants include objects, other people, and localized body regions in self-directed actions; entities may recur across events in different roles, and several entities may fill the same role within one event. By bringing concurrent actors, shared entities, role changes, and repeated fillers into one annotation protocol, HEIR makes these sources of compositional complexity directly measurable. Across its 18,730 images, 51.6% contain at least two acting people, 62.1% contain at least two actions, and 4,178 person–action events contain multiple fillers for a role. HEIR measures relation and complete-event AP, exposing composition errors in multi-participant events, shared-entity scenes, and repeated-role events. Together, these design choices pose a focused question: can a model recover each event's complete role structure while binding shared participants consistently across the image?

On HEIR, standard HOI baselines can be evaluated on role-qualified relations, but relation AP does not reveal whether all participants and roles have been assembled into the correct event. For these methods (Liao et al., 2022; Yuan et al., 2023; Kim et al., 2025; Lei et al., 2025; Sun et al., 2026), this distinction in evaluation is especially consequential for shared entities and repeated role fillers, where individually plausible links can still leave an event incomplete or bind its participants inconsistently. At the modeling level, contextual relation scores do not by themselves constrain event size or role multiplicity. This motivates coupling participant assignments through a normalized set distribution while preserving the identities of entities shared across actions.

We develop CoRISP to predict this set-level target directly. It uses image-level entity identities to exchange role-specific summaries across three linked neighborhoods: participants within an event, actions involving the same person–entity pair, and events sharing an entity. For each person–action event, CoRISP defines a normalized distribution over assignments of candidate entities to functional roles. Cardinality and role-multiplicity potentials couple participant assignments through event size and repeated role fillers. A dynamic program computes the exact partition function and role marginals for each event over retained proposals and fixed support. The resulting model aligns its output with HEIR's evaluation target while representing uncertainty over participants and set composition.

Our contributions are threefold:

1. **An event-complete benchmark.** HEIR jointly evaluates complete grounded participant–role sets for each annotated person–action event across object, interpersonal, and self-directed interactions. Its annotations preserve entity identity across concurrent events, even when an entity changes roles, and allow multiple participants to fill the same role. HEIR thus makes shared-entity reasoning and event composition explicit, measurable challenges.

2. **A composition-aware evaluation.** HEIR pairs role-qualified relation AP with complete-set AP and reports results for single- and multi-participant events, shared-entity images, and repeated role fillers. This evaluation distinguishes detecting individual relations from correctly binding participants to actors and roles and recovering each event as a complete set.

3. **A normalized model of interaction sets.** CoRISP predicts a normalized distribution over each event's complete participant–role assignments. It uses shared entity identities to aggregate role-specific evidence within an event, across actions involving the same person–entity pair, and across events that share an entity. Cardinality and role-multiplicity potentials couple assignments through event size and repeated role fillers. A dynamic program computes exact per-event normalization and role marginals over retained proposals and fixed support.

## 2 Related Work

**Human–Object Interaction Detection.**

Human–object interaction (HOI) detection spans appearance-, geometry-, and graph-based recognition of localized human–verb–object relations (Gkioxari et al., 2018; Qi et al., 2018; Ulutan et al., 2020; Gao et al., 2020) as well as query-based interaction prediction (HOTR, QPIC) and cascaded detection and interaction decoders (CDN) (Kim et al., 2021; Tamura et al., 2021; Zhang et al., 2021). STIP, MUREN, and PViC strengthen interaction proposals and relational or predicate-specific context (Zhang et al., 2022; Kim et al., 2023; Zhang et al., 2023). SWiG-HOI broadens object coverage (Wang et al., 2021); vision–language and open-vocabulary methods extend the interaction vocabulary (Liao et al., 2022; Yuan et al., 2023; Kim et al., 2025; Lei et al., 2025; Sun et al., 2026), while RoHOI evaluates robustness to image corruptions (Wen et al., 2025). Across these advances, standard HOI protocols chiefly score localized relations, leaving unclear whether every participant in each person–action event has been recovered, assigned the right role, and tracked consistently across concurrent interactions. HEIR makes complete grounded participant–role sets the unit of an image-level benchmark, spanning object, interpersonal, and self-directed actions while preserving shared identities and allowing repeated role fillers. CoRISP directly models this target through role-preserving cross-event context and a normalized set distribution with cardinality and role-multiplicity potentials. A dynamic program computes exact per-event normalization over retained candidates and fixed support.

**Grounded Roles and Interaction Composition.** Prior work has advanced grounded interaction understanding along complementary directions. V-COCO (Gupta & Malik, 2015) annotates action-specific participant roles, while SWiG (Pratt et al., 2020) grounds the semantic arguments of a salient activity and evaluates whether its role values and boxes are jointly recovered. NVI-DET (Wei et al., 2024) models nonverbal behavior among individuals and groups, and GroupHOI (Hong et al., 2025) uses geometric and semantic group context to improve pairwise HOI prediction. HOI-M3 (Zhang et al., 2024) captures multi-person, multi-object interactions in 3D sequences for reconstruction and generation. CrossHOI-Bench (Lei et al., 2026) evaluates multi-person attribution through multiple-answer and multiple-choice questions, including image-wide settings; Luo et al. (2026) jointly study pair-centric detection and anticipation in video. We use still images to isolate within-scene participant–role composition from temporal tracking and anticipation. Building on these complementary efforts, HEIR centers evaluation on each grounded person–action event in an image: it scores the complete participant–role set, allows multiple fillers per role, and preserves entity identity across concurrent events, including when an entity's role changes.

## 3 The HEIR Benchmark

**Grounded events.** HEIR represents each annotated person–action event as a participant–role set in a still image. Entity $j$ has one image-level identity, noun $n_j$, and box $b_j$, reused across events. For person $h$, action $v$, and directed relations $\mathcal{Q}_I$, the event is $S_{h,v} = \{(j, r) : (h, j, v, r) \in \mathcal{Q}_I\}$. Participants include objects, other people, and separately localized body regions in self-directed actions. Each participant has one role per event; multiple entities may fill the same role, and an entity may change roles across events.

**Functional roles.** The six roles describe an entity's function in an event: `target` is acted on or attended to; `instrument` is used to perform the action; `support` bears weight or supports posture; `source` is an origin; `destination` is an endpoint; and `constraint` restricts movement or state. Role annotation follows the depicted action and the participant's function. Definitions and annotation rules appear in Appendix A.

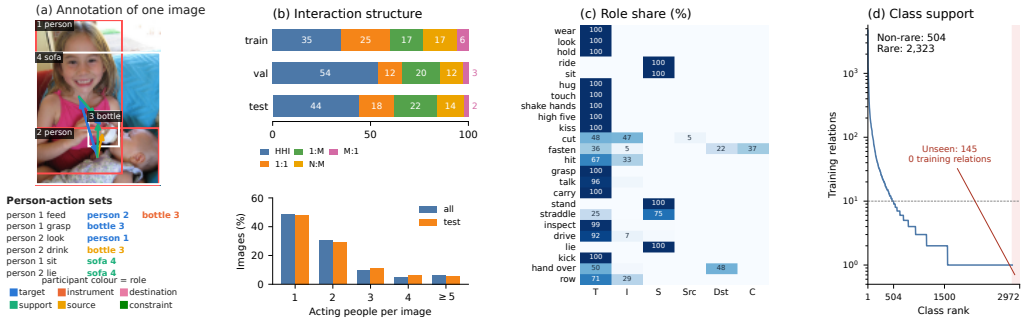



Figure 1: **HEIR structure.** (a) Shared participants across events. (b) Interaction structure (top): HHI denotes images with at least one human–human relation; other images are grouped by acting-person count versus distinct participant count. Bottom: actor counts. All split images are included. (c) Role shares for the 24 most frequent actions. (d) Training support of 2,972 observed classes, including 145 unseen classes with zero training relations.

**Construction and review.** We selected 18,730 of 220,362 images and frames from 29 sources (Appendix Table 4). Screening removes duplicates and prioritizes visible, localizable interactions with semantic and structural diversity. Every released image has complete human-corrected interaction annotations, covering participant identities, boxes, actions, and functional roles. Appendix A details model assistance, human review.

**Interaction structure.** HEIR contains 78,345 boxes and 68,803 relations over 105 actions, 437 nouns, and six roles. Images with multiple actors and multiple actions comprise 51.6% and 62.1%, respectively. Cross-actor sharing occurs in 3,797 images, and 4,178 events have repeated role fillers. Figure 1 illustrates these structures. The train/validation/test split contains 15,158/615/2,957 images; all primary test results use the complete test split without class exclusions.

**Evaluation.** Role mAP averages AP over test-supported $(v, n, r)$ classes. A correct relation requires the action, noun, and role, and actor and participant IoU $\geq 0.5$, with one-use matching. Complete-set AP instead requires the entire event. Let $\phi_I$ be a noun-compatible, one-to-one entity correspondence at IoU $\geq$ 0.5, shared across an image and constructed from boxes, noun labels, and prediction confidence. A predicted event is correct only if its actor and all participants have matches, its actor and action identify an unmatched annotated event, and

$$\{(\phi_I(\hat{j}), \hat{r}) : (\hat{j}, \hat{r}) \in \hat{S}_{\hat{h},\hat{v}}\} = S_{\phi_I(\hat{h}),\hat{v}}. \tag{1}$$

Missing, additional, wrongly labeled, or unmatched participants make the set incorrect. Set mAP averages all-point interpolated AP over test-supported actions. Correspondence, tied scores, duplicates, and structural strata are specified in Appendix C.

## 4 CoRISP: Grounded Interaction Composition

Probabilistic set prediction models membership and cardinality (Rezatofighi et al., 2017). CoRISP uses participant–role assignments both to organize visual evidence and to define a normalized event distribution (Figure 2). In event $(h, v)$, each candidate takes state 0 (unselected) or one role in $\mathcal{R}_v$; several candidates may share a role. Visual–text features initialize event states $\mathbf{e}_{h,v}$, shared entity states $\mathbf{u}_j$, and pair–action features $\mathbf{x}_{hj,v}$. Entities retain their identities across events, which exchange evidence but have separate normalizers.

### 4.1 Role-Conditioned Interaction Field

For a visible candidate $i = (h, j, v)$, a softmax $p_i$ over state 0 and applicable roles gives interaction confidence and conditional role context:

$$q_i = \sum_{r \in \mathcal{R}_v} p_i(r), \qquad \bar{\mathbf{t}}_i = \frac{\sum_{r \in \mathcal{R}_v} p_i(r)\mathbf{t}_r}{\max(q_i, \epsilon)}. \tag{2}$$

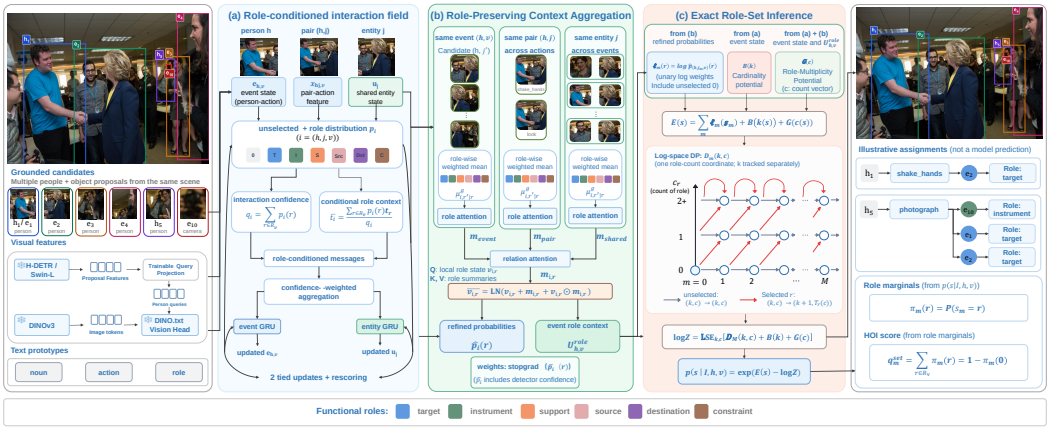



Figure 2: **CoRISP overview.** (a) Role-conditioned recurrent updates. (b) Role-preserving aggregation over event, pair, and shared-entity neighborhoods. (c) Exact set normalization with cardinality and role-multiplicity potentials. Marginals score individual relations; joint probabilities score complete assignments. The displayed assignments illustrate the output structure.

Here $\mathbf{t}_r$ is a projected role prototype and $\epsilon$ a numerical floor. Interaction confidence $q_i$ weights each message, while $\bar{\mathbf{t}}_i$ specifies its role content. Messages combine this context with pair features and visual evidence. Confidence-weighted means feed event and entity GRUs (Cho et al., 2014). Two tied updates followed by rescoring let role estimates change the features used for subsequent interaction decisions. Actions remain multilabel.

### 4.2 Role-Preserving Context Aggregation

Each pair–action index receives three kinds of context: co-participants in event $(h, v)$, other actions involving pair $(h, j)$, and other events involving entity $j$. These neighborhoods connect event composition, action-dependent roles, and shared participants. Detached positive-role probabilities weight candidates within each role; the query's same-role contribution is removed. Pooling within roles keeps evidence for different participant functions separate. Attention reads these summaries and fuses the three neighborhood messages into $\mathbf{m}_{i,r}$.

The local role feature $\mathbf{v}_{i,r}$ combines pair evidence with noun and action–role anchors. Binding it to the fused message gives $\bar{\mathbf{v}}_{i,r} = \mathrm{LN}(\mathbf{v}_{i,r} + \mathbf{m}_{i,r} + \mathbf{v}_{i,r} \odot \mathbf{m}_{i,r})$, where LN is layer normalization (Ba et al., 2016) and $\odot$ is elementwise multiplication. A bounded residual $\delta_{i,r}$ refines detector-weighted probabilities $\widetilde{p}_i$:

$$\widehat{p}_i(r) = \frac{\widetilde{p}_i(r) e^{\delta_{i,r}}}{\sum_{r' \in \{0\} \cup \mathcal{R}_v} \widetilde{p}_i(r') e^{\delta_{i,r'}}}, \qquad \delta_{i,0} = 0. \tag{3}$$

Role-wise pooling of the same bound features, combined with event and action–role features, gives $\mathbf{U}^{\mathrm{role}}_{h,v}$ for set composition. Appendix B gives the pooling and potential parameterizations.

### 4.3 Exact Set Normalization

Fix an event with candidates $j_1, \ldots, j_M$. An assignment $\mathbf{s} = (s_1, \ldots, s_M)$ encodes $S = \{(j_m, s_m) : s_m \neq 0\}$, with unary log weights $\ell_m(r) = \log \widehat{p}_{(h,j_m,v)}(r)$, including $r = 0$. Its total count is $k(\mathbf{s}) = \sum_m \mathbf{1}\{s_m \neq 0\}$ and its saturated role counts are $c_r(\mathbf{s}) = \min(2, \sum_m \mathbf{1}\{s_m = r\})$. For an admissible assignment, CoRISP defines

$$E(\mathbf{s}) = \sum_m \ell_m(s_m) + B(k(\mathbf{s})) + G(\mathbf{c}(\mathbf{s})), \tag{4}$$

$$p(\mathbf{s} \mid I, h, v) = \exp E(\mathbf{s})/Z, \qquad Z = \sum_{\mathbf{s}' \in \Omega_{h,v}} \exp E(\mathbf{s}'). \tag{5}$$

The Cardinality Potential $B$ scores total count from $\mathbf{e}_{h,v}$. The Role-Multiplicity Potential $G$ combines each contextual role vector with its count embedding and scores the pooled representation against the event state. Whereas $B$ distinguishes event sizes, $G$ also distinguishes role compositions

at the same size, such as two targets versus a target and an instrument. Both vanish at the empty set. Saturating each role count at $2+$ retains the distinction between absence, presence, and repetition while permitting compact count-based inference. At set inference, the fixed inventory $C(v, n_j, r)$ defines admissible positive states in $\Omega_{h,v}$.

A log-space dynamic program accumulates log unary mass $D_m(k, \mathbf{c})$ over the first $m$ candidates. Each transition leaves a candidate unselected or assigns one allowed role. The partition is

$$\log Z = \mathrm{LSE}_{k,\mathbf{c}}[D_M(k, \mathbf{c}) + B(k) + G(\mathbf{c})]. \tag{6}$$

Here LSE denotes log-sum-exp. Count-based inference follows established dynamic programming principles (Tarlow et al., 2012); exactness is within an event's retained candidates and support. Holding $B$ and $G$ fixed, role marginals are $\pi_m(r) = \partial \log Z / \partial \ell_m(r)$ and interaction confidence is $q_m^{\mathrm{set}} = 1 - \pi_m(0)$. Contextual unaries carry identity-specific evidence; $B$ and $G$ couple assignments through their composition.

Set prediction replaces log-sum-exp with maximization to obtain the highest-scoring assignment per count state, ranked by its normalized probability. Relation confidence and complete-event confidence follow from the same learned distribution.

**Set supervision.** Let $\mathcal{A}(y) \subseteq \Omega_{h,v}$ contain the distinct complete assignments compatible with annotation $y$, using each proposal at most once. The Localization-Marginalized Set Likelihood sums their probability:

$$\mathcal{L}_{\mathrm{set}}(y) = \log Z - \mathrm{LSE}_{\mathbf{s} \in \mathcal{A}(y)} E(\mathbf{s}). \tag{7}$$

Summing compatible assignments accommodates several proposals for one annotated participant without selecting an arbitrary localization target. Equivalent matching paths count once per assignment. Observed negatives target the empty set; unknown or unrepresentable positive events are omitted from this loss. Matching and focal modulation are specified in Appendix B.

# 5 Experiments

## 5.1 Experimental Setup

**Benchmarks.** HEIR uses 2,957 test images, 1,104 role classes, and 9,955 sets over 101 actions. V-COCO (Gupta & Malik, 2015) uses eligible trainval images and all 4,946 test images.

**Implementation details.** CoRISP uses frozen H-DETR/Swin-L proposals (Jia et al., 2023). A DINOv3 ViT-L/16 backbone followed by a DINO.txt vision head supplies text-aligned visual features (Siméoni et al., 2025; Jose et al., 2025); both components are frozen. CoRISP trains its feature projections, interaction field, context aggregation, and set potentials, totaling 9.9M trainable parameters. Training runs for 30 epochs, with HEIR checkpoints selected by validation Role mAP. HEIR uses a benchmark-wide semantic compatibility inventory of 10,072 action–noun–role combinations. The inventory encodes whether a combination is semantically admissible and is shared by dagger-marked baselines before their native top-100 selection and applicable NMS; other baselines retain their native support. UniHOI (Yang et al., 2026) and HOI-IDiff (Hui et al., 2025) are reported under their published V-COCO conventions in a separate table block.

**Evaluation setting.** HEIR reports Role and Set mAP (Section 3) and HOI mAP, which drops role labels and takes their maximum score. Pair-output baselines are converted into set hypotheses from ranked relation scores, while CoRISP predicts complete assignments from its learned set distribution. All methods use the same set matching protocol and the same 100-set image budget. The HEIR component-ablation protocol uses CoRISP's set prediction rule. Single/Multi/Repeat/Shared denote one-member events, multi-member events, repeated roles, and shared-entity images. V-COCO reports role AP ($\mathrm{AP}_{\mathrm{role}}$), excluding `point`, and our Set mAP under S1/S2; Dual averages three two-slot actions. Appendix C specifies matching, annotation scope, encoders, and conditional strata.

## 5.2 Comparison with Existing Methods

**HEIR.** Table 1 places relation detection alongside event recovery. CoRISP leads on Repeat and Shared Set mAP (8.99 and 23.80), exceeding the strongest baselines by 2.87 and 3.82 points with 9.9M trainable parameters. These strata test assigning several entities to one role and recovering

Table 1: **HEIR (left) and V-COCO (right)** (%). V-COCO pairs are S1/S2; re-evaluated role AP excludes `point`. Params: HEIR trainable millions. †: shared HEIR inventory; W/T/A/R: released weights, trained official code, adapted official code, or reimplementation. Bold/underline: best/second, excluding literature-only scores.

| Method | HEIR | | | | | | | | V-COCO (S1/S2) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Params (M)↓ | HOI ↑ | Role ↑ | Set mAP↑ | | | | | $AP_{role}$ ↑ | Set mAP↑ | |
| | | | | Full | Single | Multi | Repeat | Shared | | All | Dual |
| *Visual models* | | | | | | | | | | | |
| QPIC R50 (Tamura et al., 2021)† | 41.6 | 15.11 | 14.82 | 14.14 | 16.03 | 4.92 | 3.68 | 14.71 | 58.79/60.97$^{W}$ | 53.87/57.77 | 41.64/45.84 |
| QPIC R101 (Tamura et al., 2021)† | 60.5 | 15.57 | 15.26 | 14.84 | 16.93 | 3.86 | 3.01 | 15.10 | 58.16/60.65$^{W}$ | 53.03/57.11 | 39.38/44.28 |
| MUREN (Kim et al., 2023)† | 75.1 | 16.38 | 16.17 | 13.98 | 16.04 | 6.57 | 4.66 | 15.75 | 68.72/70.97$^{W}$ | 62.97/67.33 | 54.94/60.95 |
| SOV-STG-S (Chen et al., 2025)† | 54.1 | 15.55 | 15.21 | 13.64 | 15.67 | 2.62 | 1.70 | 13.00 | 62.99/64.56$^{T}$ | 56.62/59.53 | 47.79/50.02 |
| SOV-STG Swin-L (Chen et al., 2025)† | 240.4 | 23.69 | 23.35 | 18.93 | 21.28 | **9.87** | 5.85 | 19.57 | 69.87/71.61$^{T}$ | 63.86/67.15 | 55.20/59.36 |
| SOV-STG R101 (Chen et al., 2025)† | 87.4 | 17.11 | 16.86 | 14.58 | 16.89 | 5.33 | 1.57 | 16.43 | 66.63/68.21$^{W}$ | 60.52/63.63 | 49.59/53.91 |
| PViC R50 (Zhang et al., 2023) | 12.2 | 16.24 | 16.02 | 13.83 | 16.39 | 3.93 | 4.78 | 15.17 | 58.70/65.88$^{T}$ | 55.38/64.05 | 27.66/41.76 |
| PViC Swin-L (Zhang et al., 2023) | 12.1 | 19.22 | 18.99 | 14.57 | 17.21 | 5.11 | 6.12 | 16.62 | 60.85/68.72$^{T}$ | 57.01/66.89 | 31.26/49.91 |
| *Vision–language models* | | | | | | | | | | | |
| GEN-VLKT-S (Liao et al., 2022)† | 46.9 | 17.23 | 17.13 | 12.95 | 15.50 | 4.33 | 4.54 | 15.92 | 65.05/67.19$^{W}$ | 59.47/63.58 | 50.40/55.11 |
| GEN-VLKT-L (Liao et al., 2022)† | 75.4 | 19.23 | 18.92 | 15.85 | 19.34 | 4.79 | 3.25 | 17.59 | 66.34/68.81$^{W}$ | 60.34/65.11 | 49.50/55.43 |
| SOV-STG-VLA-S (Chen et al., 2025)† | 88.2 | 21.60 | 21.01 | 15.45 | 19.11 | 4.01 | 1.45 | 15.33 | 66.49/68.46$^{W}$ | 60.87/64.65 | 50.69/55.52 |
| RLIPv2 Swin-T (Yuan et al., 2023)† | 213.4 | 22.19 | 21.87 | 17.52 | 20.27 | 5.40 | 3.78 | 18.09 | 68.99/71.12$^{W}$ | 63.22/67.67 | 55.67/62.35 |
| RLIPv2 Swin-L (Yuan et al., 2023)† | 382.8 | **26.38** | **25.91** | **22.21** | **25.14** | 7.26 | 4.66 | 19.98 | 71.92/74.20$^{W}$ | 66.49/71.17 | 58.02/64.82 |
| GroupHOI-S (Hong et al., 2025) | 82.0 | 13.68 | 13.32 | 14.52 | 16.74 | 4.22 | 3.43 | 15.24 | 66.54/69.54$^{A}$ | 60.72/66.18 | 51.36/58.57 |
| InCoM-Net (Seo et al., 2026) | 42.2 | 15.07 | 14.79 | 14.94 | 17.63 | 4.58 | 5.18 | 17.75 | 70.11/73.80$^{R}$ | **67.66**/71.69 | 59.43/64.88 |
| SL-HOI (Sun et al., 2026) | 31.1 | 26.01 | 24.92 | 10.93 | 13.48 | 1.78 | 0.27 | 10.87 | 48.16/50.24$^{A}$ | 43.31/46.77 | 24.25/27.50 |
| *Generative models* | | | | | | | | | | | |
| HOI-IDiff (Hui et al., 2025) | | | | | – | | | | 73.4/76.1 | – | – |
| UniHOI (Yang et al., 2026) | | | | | – | | | | 72.91/77.45 | – | – |
| **CoRISP** | **9.9** | 21.21 | 20.41 | 18.04 | 20.57 | 8.94 | **8.99** | **23.80** | **73.72/76.23** | 67.11/**72.20** | **61.06/68.59** |

events in shared-participant images. RLIPv2 Swin-L leads on overall Set mAP (22.21), while SOV-STG Swin-L leads on Multi (9.87).

**V-COCO.** Under the common evaluation, CoRISP leads in $AP_{role}$ (73.72/76.23) and Dual Set mAP (61.06/68.59), improving Dual by 1.63/3.71 points over the strongest baseline. It also leads on overall S2 Set mAP (72.20); InCoM-Net leads on S1 (67.66). Dual tests joint recovery of both native slots, whereas HEIR additionally tests variable cardinality and repeated roles.

### 5.3 What HEIR Reveals

**Role ambiguity depends on the action–noun context.** On the 47 test-supported action–noun pairs observed with multiple roles, retaining role labels reduces pair-balanced AP by 4.78–12.66 points across 16 baselines. The absolute difference is at most 0.34 on the 1,010 single-role pairs (Figure 3a). Aggregate HOI–Role differences therefore mask the difficulty of assigning a participant's function when its action and noun admit several roles.

**Better relations need not yield better events.** Role mAP weights semantic classes, whereas Set mAP weights actions. After averaging Role AP within actions and then over the same 101 actions, 41 of 120 pairs among the 16 baselines still reverse order between the two metrics (Figure 3b). Ten of 55 pairs reverse within the shared-inventory baselines alone. SL-HOI exceeds PViC Swin-L by 7.49 points in action-balanced Role AP but trails it by 3.64 in Set mAP. Thus, event recovery measures more than a different weighting of relation accuracy.

**Completeness requires excluding plausible extras.** For five baselines, only 53.58–61.81% of edge-covered multi-participant events are recovered exactly (Appendix C). Each failure has an extra member in every prefix containing all true members. Within these covered events, the bottleneck is separating required members from interleaved false edges. On V-COCO, the same models recover 98.52–100% of fully covered two-slot events under S1. Native slots fix the role positions; HEIR additionally requires selecting a variable number of participants. Each rate conditions on that model's covered events.

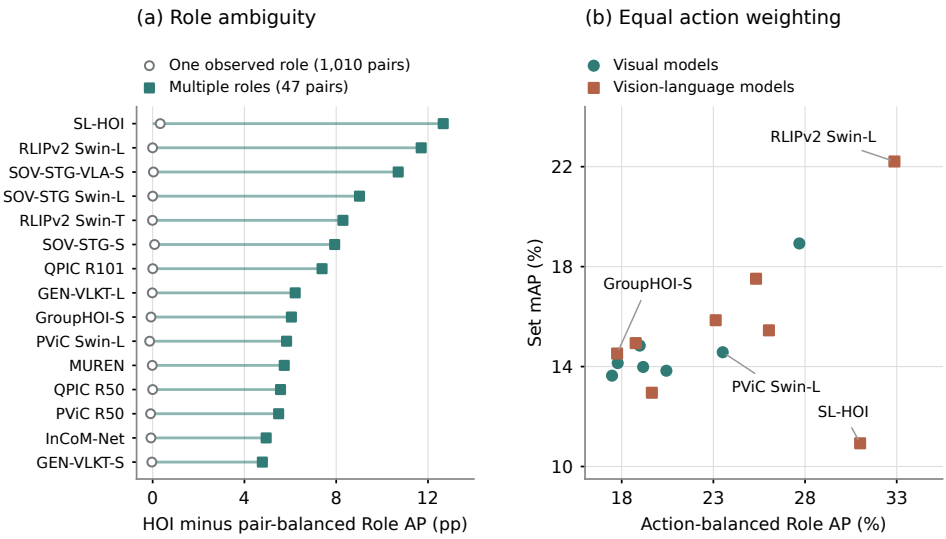



Figure 3: **Role ambiguity and event recovery across all 16 baselines.** (a) HOI–Role AP gaps with equal action–noun pair weights. (b) Role and Set AP with equal weights over the same 101 actions; sets use common-prefix decoding.

### 5.4 Evidence Availability and Selection

Following the distinction between localization and interaction errors (Zhu et al., 2025), we examine whether complete visual evidence is available and correctly ranked. The oracle supplies the true action and participant count $k^\star$ at a fixed budget of 100 relations/image. Coverage $C$ is the fraction of events with noun-compatible boxes for the actor and every participant; recovery $R$ is the fraction whose top-$k^\star$ edges form the correct set (Table 3; all models in Table 6).

CoRISP recovers 27.48% of Multi events and 33.41% of Repeat events, exceeding the strongest baselines by 3.26 and 6.52 points. Its Multi coverage is 48.92%, compared with SOV-STG Swin-L's 51.21%, yet its recovery is higher: 27.48% versus 24.22%. Thus, higher candidate coverage alone does not determine which system ranks complete sets correctly. CoRISP also provides the highest Repeat coverage (62.01%) and recovery. Coverage reflects the boxes and nouns present in the retained relations; recovery additionally requires correct member and role ranking. Together, these results identify complementary priorities: improve complete-participant coverage through localization, noun recognition, and candidate retention, while preserving CoRISP's stronger ranked-set recovery.

A complementary comparison holds the evaluated person–participant pairs fixed and requires both models to localize them with the correct noun. On 10,383 such pairs, including negative pairs, CoRISP attains 61.50 conditional Role mAP versus SL-HOI's 50.26 over the same 584 supported classes (Appendix C). Actions and roles remain predicted. This comparison separates classification on jointly recoverable pairs from the coverage losses measured above.

### 5.5 CoRISP Component Analysis

Removing $G$ retains cardinality scoring but removes the role-composition potential. Removing context messages disables event, pair, and shared-entity aggregation while retaining the recurrent field. Removing role feedback excludes the conditional role vector from recurrent messages while preserving confidence weighting. These removals reduce Multi Set mAP from 8.94 to 6.89, 6.70, and 6.97, respectively. Removing context messages causes the largest drop in overall Set mAP, from 18.04 to 16.42. Interestingly, each deletion slightly increases aggregate Role mAP while reducing Set mAP, reinforcing the distinction between scoring individual relations and recovering complete event composition. With the primary checkpoint fixed, retaining up to eight count-state winners improves Set mAP from 17.56 for the single best nonempty assignment to 18.04 (Table 5).

**Qualitative analysis.** Figure 4 contrasts finding a plausible partner with recovering all participants. In the four-person high-five, CoRISP ranks the complete three-target set first, while both baselines

Table 2: **HEIR component ablations** (%). The model and evaluation settings follow Table 1

| Variant | Role Full↑ | Set mAP↑ | | | | |
|---|---|---|---|---|---|---|
| | | Full | Single | Multi | Repeat | Shared |
| w/o $G$ | 20.67 | 17.09 | 20.54 | 6.89 | 5.39 | 19.96 |
| w/o context messages | 20.93 | 16.42 | 19.40 | 6.70 | 4.38 | 19.75 |
| w/o role feedback | **20.95** | 17.05 | 20.18 | 6.97 | 4.17 | 19.97 |
| CoRISP | 20.41 | **18.04** | **20.57** | **8.94** | **8.99** | **23.80** |

Table 3: **Action–cardinality oracle** (%). The true action and count are supplied. $C$: coverage; $R$: recovery. All use 100 relations/image; emphasis ranks shown models.

| Method | Multi | | Repeat | |
|---|---|---|---|---|
| | $C\uparrow$ | $R\uparrow$ | $C\uparrow$ | $R\uparrow$ |
| SOV-STG Swin-L (Chen et al., 2025) | **51.21** | 24.22 | 56.06 | 24.94 |
| PViC Swin-L (Zhang et al., 2023) | 47.47 | 22.55 | 60.41 | 26.89 |
| RLIPv2 Swin-L (Yuan et al., 2023) | 48.30 | 23.94 | 54.58 | 25.51 |
| CoRISP | 48.92 | **27.48** | **62.01** | **33.41** |

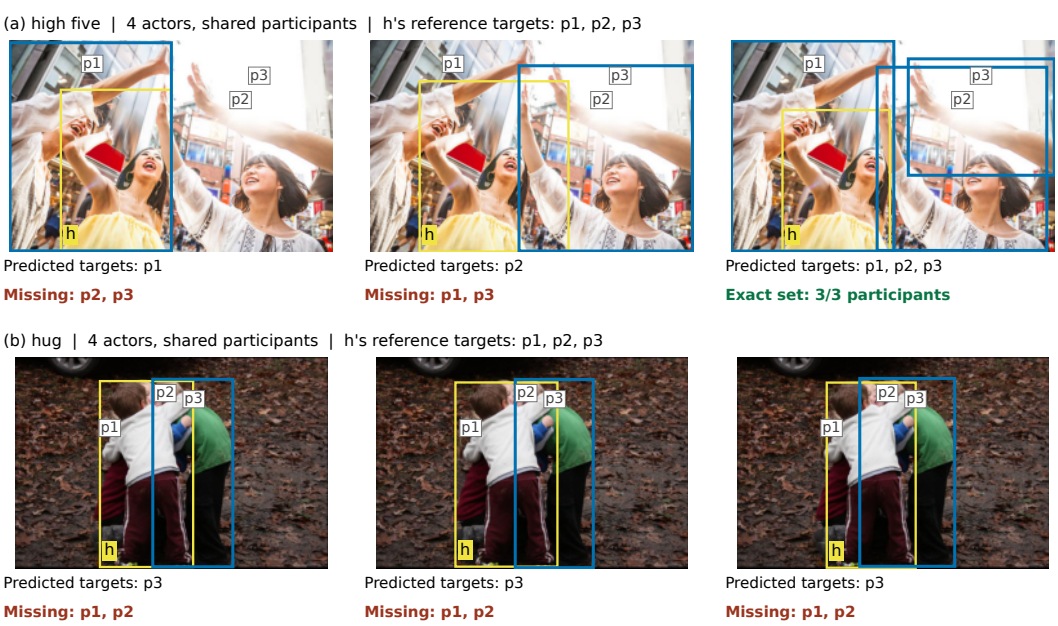



Figure 4: **Composition in four-person scenes.** Highest-scoring set for the indicated actor/action, on identical uncropped images. Gray IDs mark reference participants; solid yellow/blue boxes are predicted actors/targets. In (a), CoRISP recovers three targets sharing one role; in (b), all models omit two participants in an occluded group hug.

rank a singleton first. The distinction is between confidence in an individual relation and confidence that an event has no missing participants. In the group hug, all three models recover a valid target but omit two heavily occluded participants.

## 6 Discussion and Limitations

HEIR makes complete event composition measurable alongside relation detection. The ranking reversals and covered-but-incomplete events show why both views matter. CoRISP combines role-conditioned evidence with a normalized participant-set model, with its strongest system-level gains on repeated roles, shared-participant images, and V-COCO's two-slot actions. Candidate coverage remains a practical limit: an event requires suitable boxes and noun labels for every participant. The oracle results motivate improving this front end while preserving the model's stronger ranked-set recovery.

### AI USE STATEMENT

Generative AI was used to aid and polish manuscript writing, as well as to assist with retrieval and discovery during literature review. Additionally, multimodal models were used to propose initial dataset annotations prior to human correction and review, as described in Appendix A. The authors take full responsibility for the final text, claims, code, and data, including all AI-assisted contributions.

### ETHICS STATEMENT

HEIR reannotates images from existing public sources, including interpersonal and clinical scenes. Source provenance is listed in Appendix Table 4. Redistribution is subject to each source's license and privacy conditions; where image redistribution is not permitted, the release will provide identifiers and acquisition references. Annotation and review coverage are described in Appendix A.

### REPRODUCIBILITY STATEMENT

Appendix B specifies CoRISP's probabilities, potentials, dynamic program, and supervision. Appendix C defines matching, decoding, output budgets, model selection, and diagnostic protocols. We plan to release the code, checkpoints, annotations, and evaluation tools.

## REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Hymalai Bello, Lala Ray, Joanna Sorysz, Sungho Suh, and Paul Lukowicz. Openmarcie: Dataset for multimodal action recognition in industrial environments. *arXiv preprint arXiv:2603.02390*, 2026.

Alessandro Carfì, Francesco Foglino, Barbara Bruno, and Fulvio Mastrogiovanni. A multi-sensor dataset of human-human handover. *Data in brief*, 22:109–117, 2019.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Matthew Chang, Gunjan Chhablani, Alexander Clegg, Mikael Dallaire Cote, Ruta Desai, Michal Hlavac, Vladimir Karashchuk, Jacob Krantz, Roozbeh Mottaghi, Priyam Parashar, et al. Partnr: A benchmark for planning and reasoning in embodied multi-agent tasks. In *International Conference on Learning Representations*, volume 2025, pp. 65205–65268, 2025.

Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *IEEE Winter Conference on Applications of Computer Vision*, 2018. URL https://arxiv.org/abs/1702.05448.

Vivek Chavan, Yasmina Imgrund, Tung Lam Dao, Sanwantri Bai, Bosong Wang, Ze Lu, Oliver Heimann, and Jörg Krüger. Indego: a dataset of industrial scenarios and collaborative work for egocentric assistants. In *Wearable AI Workshop@ ECCV 2026*, 2026.

Junwen Chen, Yingcheng Wang, and Keiji Yanai. Focusing on what to decode and what to train: SOV decoding with specific target guided DeNoising and vision language advisor. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/Chen_Focusing_on_What_to_Decode_and_What_to_Train_SOV_WACV_2025_paper.html.

Kyunghyun Cho, Bart Van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1724–1734, 2014.

Mejdi Dallel, Vincent Havard, David Baudry, and Xavier Savatier. Inhard-industrial human action recognition dataset in the context of industrial collaborative robotics. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pp. 1–6. IEEE, 2020.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14987–14997. IEEE, 2025.

Shenghao Fu, Yukun Su, Fengyun Rao, Jing Lyu, Xiaohua Xie, and Wei-Shi Zheng. Wedetect: Fast open-vocabulary object detection as retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20377–20387, 2026.

Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *European conference on computer vision*, pp. 696–712. Springer, 2020.

Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *2018 ieee/cvf conference on computer vision and pattern recognition*, pp. 8359–8367. IEEE, 2018.

Google. Gemini 3.1 Flash-Lite, 2026a. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-lite. Gemini API documentation. Accessed: June 2026.

Google. Gemini 3.6 Flash, 2026b. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.6-flash. Gemini API documentation. Accessed: June 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6047–6056. IEEE, 2018.

Saurabh Gupta and Jitendra Malik. Visual semantic role labeling, 2015. URL https://arxiv.org/abs/1505.04474.

Ru Han, Xinyi Liang, Lei Shu, Xiaoyuan Jing, Fan Yang, and Renjie Tian. Teaweeding-action: a vision-based dataset for weeding behavior recognition in tea plantations. *Frontiers in Plant Science*, 16:1722007, 2025.

Umair Haroon, Amin Ullah, Tanveer Hussain, Waseem Ullah, Muhammad Sajjad, Khan Muhammad, Mi Young Lee, and Sung Wook Baik. A multi-stream sequence learning framework for human interaction recognition. *IEEE Transactions on Human-Machine Systems*, 52(3):435–444, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jiajun Hong, Jianan Wei, and Wenguan Wang. Learning human-object interaction as groups. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2510.18357.

Ming Hu, Lin Wang, Siyuan Yan, Don Ma, Qingli Ren, Peng Xia, Wei Feng, Peibo Duan, Lie Ju, and Zongyuan Ge. Nurvid: A large expert-level video database for nursing procedure activity understanding. *Advances in Neural Information Processing Systems*, 36:18146–18164, 2023.

Xiaofei Hui, Haoxuan Qu, Hossein Rahmani, and Jun Liu. An image-like diffusion method for human-object interaction detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2503.18134.

Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2207.13080.

Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25472–25483, 2026.

Cijo Jose et al. Dinov2 meets text: A unified framework for image- and pixel-level vision-language alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2412.16334.

Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 74–83. IEEE, 2021.

Sanghyun Kim, Deunsol Jung, and Minsu Cho. Relational context learning for human-object interaction detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Kim_Relational_Context_Learning_for_Human-Object_Interaction_Detection_CVPR_2023_paper.html.

Sanghyun Kim, Deunsol Jung, and Minsu Cho. Locality-aware zero-shot human-object interaction detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20190–20200. IEEE, 2025.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

Qinqian Lei, Bo Wang, and Robby T. Tan. CrossHOI-Bench: A unified benchmark for HOI evaluation across vision-language models and HOI-specific methods. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2508.18753.

Ting Lei, Shaofeng Yin, Qingchao Chen, Yuxin Peng, and Yang Liu. Open-vocabulary hoi detection with interaction-aware prompt and concept calibration. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23945–23957. IEEE, 2025.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pp. 80–93. PMLR, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023b.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liao_GEN-VLKT_Simplify_Association_and_Enhance_Interaction_Understanding_for_HOI_Detection_CVPR_2022_paper.html.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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.

Si Liu, Zitian Wang, Yulu Gao, Lejian Ren, Yue Liao, Guanghui Ren, Bo Li, and Shuicheng Yan. Human-centric relation segmentation: Dataset and solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4987–5001, 2021a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021b.

Yuanhao Luo, Di Wen, Kunyu Peng, Ruiping Liu, Junwei Zheng, Yufan Chen, Jiale Wei, and Rainer Stiefelhage. Rethinking video human-object interaction: Set prediction over time for unified detection and anticipation. *arXiv preprint arXiv:2604.10397*, 2026.

Friedrich Niemann, Fernando Moya Rueda, Moh'd Khier Al Kfari, Nilah Ravi Nair, Dustin Schauten, Veronika Kretschmer, Stefan Lüdtke, and Alice Kirchheim. Dara dataset: Combining wearable sensors, location tracking, and process knowledge for enhanced human activity and human context recognition in warehousing. *Sensors*, 26(2):739, 2026.

OpenAI. Introducing GPT-5.4, 2026. URL https://openai.com/index/introducing-gpt-5-4/. Accessed: June 2026.

Openverse. Openverse, 2026. URL https://openverse.org/. Accessed: June 2026.

Ege Özsoy, Chantal Pellegrini, Tobias Czempiel, Felix Tristram, Kun Yuan, David Bani-Harouni, Ulrich Eck, Benjamin Busam, Matthias Keicher, and Nassir Navab. Mm-or: A large multimodal operating room dataset for semantic understanding of high-intensity surgical environments. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19378–19389. IEEE, 2025.

Ege Özsoy, Arda Mamur, Felix Tristram, Chantal Pellegrini, Magdalena Wysocki, Benjamin Busam, and Nassir Navab. Egoexor: An ego-exo-centric operating room dataset for surgical activity understanding. *Advances in Neural Information Processing Systems*, 38, 2026.

Konstantinos Papoutsakis, Nikolaos Bakalos, Konstantinos Fragkoulis, Athena Zacharia, Georgia Kapetadimitri, and Maria Pateraki. A vision-based framework for human behavior understanding in industrial assembly lines. In *European Conference on Computer Vision*, pp. 324–342. Springer, 2024.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *European Conference on Computer Vision*, 2020. URL https://arxiv.org/abs/2003.12058.

Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *European conference on computer vision*, pp. 407–423. Springer, 2018.

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.

Francesco Ragusa, Rosario Leonardi, Michele Mazzamuto, Daniele Di Mauro, Camillo Quattrocchi, Alessandro Passanisi, Irene D'Ambra, Antonino Furnari, and Giovanni Maria Farinella. Enigma-360: An ego-exo dataset for human behavior understanding in industrial scenarios. *arXiv preprint arXiv:2603.09741*, 2026.

S Hamid Rezatofighi, Vijay Kumar Bg, Anton Milan, Ehsan Abbasnejad, Anthony Dick, and Ian Reid. Deepsetnet: Predicting sets with deep neural networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5257–5266. IEEE, 2017.

Isaac Robinson, Peter Robicheaux, Matvei Popov, Deva Ramanan, and Neehar Peri. Rf-detr: Neural architecture search for real-time detection transformers. In *International Conference on Learning Representations*, volume 2026, pp. 56087–56100, 2026.

Michael S Ryoo and Jake K Aggarwal. Ut-interaction dataset, icpr contest on semantic description of human activities (sdha). In *IEEE International Conference on Pattern Recognition Workshops*, volume 2, 2010.

Soo Won Seo, KyungChae Lee, Hyungchan Cho, Taein Son, Nam Ik Cho, and Jun Won Choi. Mining instance-centric vision-language contexts for human-object interaction detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2604.02071.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Vinkle Srivastav, Thibaut Issenhuth, Abdolrahim Kadkhodamohammadi, Michel de Mathelin, Afshin Gangi, and Nicolas Padoy. Mvor: A multi-view rgb-d operating room dataset for 2d and 3d human pose estimation. *arXiv preprint arXiv:1808.08180*, 2018.

Chang Sun, Dongliang Liao, and Changxing Ding. Streamlined open-vocabulary human-object interaction detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Sun_Streamlined_Open-Vocabulary_Human-Object_Interaction_Detection_CVPR_2026_paper.html.

Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Tamura_QPIC_Query-Based_Pairwise_Human-Object_Interaction_Detection_With_Image-Wide_Contextual_Information_CVPR_2021_paper.html.

Gokhan Tanisik, Cemil Zalluhoglu, and Nazli Ikizler-Cinbis. Facial descriptors for human interaction recognition in still images. *Pattern Recognition Letters*, 73:44–51, 2016.

Daniel Tarlow, Kevin Swersky, Richard S Zemel, Ryan Prescott Adams, and Brendan J Frey. Fast exact inference for recursive cardinality models. *arXiv preprint arXiv:1210.4899*, 2012.

Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13614–13623. IEEE, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Shihao Wang, Shilong Liu, Yuanguo Kuang, Xinyu Wei, Yangzhou Liu, Zhiqi Li, Yunze Man, Guo Chen, Andrew Tao, Guilin Liu, et al. Locateanything: Fast and high-quality vision-language grounding with parallel box decoding. In *European Conference on Computer Vision*, pp. 336–357. Springer, 2026.

Suchen Wang, Kim-Hui Yap, Henghui Ding, Jiyan Wu, Junsong Yuan, and Yap-Peng Tan. Discovering human interactions with large-vocabulary objects via query and multi-scale detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13455–13464. IEEE, 2021.

Jianan Wei, Tianfei Zhou, Yi Yang, and Wenguan Wang. Nonverbal interaction detection. In *European Conference on Computer Vision*, pp. 277–295. Springer, 2024.

Di Wen, Kunyu Peng, Kailun Yang, Yufan Chen, Ruiping Liu, Junwei Zheng, Alina Roitberg, Danda Pani Paudel, Luc Van Gool, and Rainer Stiefelhagen. Rohoi: Robustness benchmark for human-object interaction detection. *arXiv preprint arXiv:2507.09111*, 2025.

Di Wen, Zeyun Zhong, David Schneider, Manuel Zaremski, Linus Kunzmann, Yitian Shi, Ruiping Liu, Yufan Chen, Junwei Zheng, Jiahang Li, et al. Impact: A dataset for multi-granularity human procedural action understanding in industrial assembly. *arXiv preprint arXiv:2604.10409*, 2026.

Wikimedia Foundation. Wikimedia Commons, 2026. URL https://commons.wikimedia.org/. Accessed: June 2026.

Yicheng Xu, Yue Wu, Jiashuo Yu, Ziang Yan, Tianxiang Jiang, Yinan He, Qingsong Zhao, Kai Chen, Yu Qiao, Limin Wang, et al. Expvid: a benchmark for experiment video understanding & reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 83310–83348, 2026.

Meng Yang, Chengke Wu, Yuanjun Guo, Rui Jiang, Feixiang Zhou, Jianlin Zhang, and Zhile Yang. Transformer-based deep learning model and video dataset for unsafe action identification in construction projects. *Automation in Construction*, 146:104703, 2023.

Panqi Yang, Haodong Jing, Nanning Zheng, and Yongqiang Ma. UniHOI: Unified human-object interaction understanding via unified token space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. URL https://ojs.aaai.org/index.php/AAAI/article/view/38148.

Hangjie Yuan, Shiwei Zhang, Xiang Wang, Samuel Albanie, Yining Pan, Tao Feng, Jianwen Jiang, Dong Ni, Yingya Zhang, and Deli Zhao. RLIPv2: Fast scaling of relational language-image pre-training. In *IEEE/CVF International Conference on Computer Vision*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Yuan_RLIPv2_Fast_Scaling_of_Relational_Language-Image_Pre-Training_ICCV_2023_paper.pdf.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024.

Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. *Advances in neural information processing systems*, 34:17209–17220, 2021.

Frederic Z. Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring predicate visual context in detecting of human-object interactions. In *IEEE/CVF International Conference on Computer Vision*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhang_Exploring_Predicate_Visual_Context_in_Detecting_of_Human-Object_Interactions_ICCV_2023_paper.html.

Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. HOI-M3: Capture multiple humans and objects interaction within contextual environment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2404.00299.

Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19548–19557, 2022.

Hao Zheng, Regina Lee, and Yuqian Lu. Ha-vid: A human assembly video dataset for comprehensive assembly knowledge understanding. *Advances in Neural Information Processing Systems*, 36:67069–67081, 2023.

Fangrui Zhu, Yiming Xie, Weidi Xie, and Huaizu Jiang. Diagnosing human-object interaction detectors. *International Journal of Computer Vision*, 2025. doi: 10.1007/s11263-025-02369-8. URL https://arxiv.org/abs/2308.08529.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

Table 4: **Source datasets and collections.** Released images per source and the setting each source contributes. HICO-DET, V-COCO, NVI, and SWiG enter with native interaction labels; all other sources are annotated from model proposals.

| Setting | Source | Released | What it contributes |
|---|---|---|---|
| Everyday | HICO-DET (Chao et al., 2018) | 6,347 | Web photographs, 117 actions, COCO nouns |
| | V-COCO (Gupta & Malik, 2015) | 1,848 | COCO images with action-specific roles |
| | SWiG (Pratt et al., 2020) | 1,828 | Situations with grounded semantic roles |
| | Visual Genome (Krishna et al., 2017) | 216 | Dense scene-graph relations |
| | Open Images (Kuznetsova et al., 2020) | 191 | Visual relationship annotations |
| | Wikimedia Commons (Wikimedia Foundation, 2026) | 157 | Openly licensed photographs |
| | Openverse (Openverse, 2026) | 103 | Openly licensed photographs |
| Human–human | HIIv2 (Haroon et al., 2022) | 2,529 | Two-person and group interactions |
| | HII (Tanisik et al., 2016) | 1,471 | Two-person interactions |
| | PIC2.0 (Liu et al., 2021a) | 1,108 | Nonverbal interaction in crowds |
| | AVA (Gu et al., 2018) | 239 | Film frames, person-to-person actions |
| | NVI (Wei et al., 2024) | 113 | Nonverbal interaction detection |
| | Handover (Carfì et al., 2019) | 66 | Object hand-over between people |
| | UT-Interaction (Ryoo & Aggarwal, 2010) | 45 | Surveillance-view interactions |
| Procedural | NurViD (Hu et al., 2023) | 401 | Nursing procedures |
| | MVOR (Srivastav et al., 2018) | 311 | Multi-view operating room |
| | ExpVid (Xu et al., 2026) | 161 | Laboratory experiments |
| | DaRA (Niemann et al., 2026) | 100 | Warehouse and logistics activities |
| | MM-OR (Özsoy et al., 2025) | 52 | Robotic surgery operating room |
| | EgoExOR (Özsoy et al., 2026) | 38 | Exocentric operating-room views |
| Industrial and agricultural | TeaWeed-Action (Han et al., 2025) | 590 | Field work with hand tools |
| | HA-ViD (Zheng et al., 2023) | 384 | Assembly with tools and parts |
| | ENIGMA-360 (Ragusa et al., 2026) | 229 | Industrial procedures |
| | Construction-CMA (Yang et al., 2023) | 81 | Construction site actions |
| | IMPACT (Wen et al., 2026) | 57 | Industrial manual tasks |
| | InHARD (Dallel et al., 2020) | 28 | Human–robot assembly |
| | CarDA (Papoutsakis et al., 2024) | 18 | Car-part assembly |
| | OpenMarcie (Bello et al., 2026) | 10 | Industrial maintenance |
| | IndEgo (Chavan et al., 2026) | 9 | Industrial exocentric views |

# A Annotation Provenance and Quality

**Sources.** Table 4 lists the 29 sources. Sources with native interactions contribute their boxes and labels as annotation evidence; other sources enter through model proposals. Each image, including a sampled video frame, is annotated independently with image-local entity identities.

**Split isolation.** Frames from the same source video or capture sequence, together with near-duplicate images, are assigned to a single split.

**Model assistance.** Annotation proposals combine multimodal labeling, entity localization, and interaction detection. On sources without native interactions, GPT-5.4 (OpenAI, 2026) provides initial labels; RF-DETR, LocateAnything, LLMDet, WeDetect, and Rex-Omni provide entity candidates (Robinson et al., 2026; Wang et al., 2026; Fu et al., 2025; 2026; Jiang et al., 2026). SAM 3 (Carion et al., 2026) refines boxes, SL-HOI (Sun et al., 2026) proposes interaction pairs, and Gemini 3.6 Flash (Google, 2026b) labels merged candidates. For native-interaction sources, Gemini 3.1 Flash-Lite (Google, 2026a) proposes roles while preserving source boxes, nouns, and actions before human correction. Geometric duplicates merge at IoU $\geq 0.95$; conflicting nouns remain for human adjudication.

**Human correction.** Thirteen annotators correct entity identities, boxes, nouns, directed action links, and roles, and ten reviewers cross-check the annotations. All 18,730 released images have complete human annotations of interactions and receive a second review by someone other than the initial annotator. Structural validation checks vocabulary membership, person actors, one role per actor–action–participant triple, and duplicates.

**Role definitions.** `target`: the entity directly acted upon, attended to, indicated, displayed, read, moved, or changed in this action; a racket being held and a box being opened are both targets. `instrument`: the tool, equipment, controller, or external implement the person uses to perform the action. `support`: the entity bearing body or object weight or supporting posture, standing, or riding; distinct from the action `support` (assisting a person), whose assisted person is the target. `source`: the origin that a person, entity, or content leaves, including removal, extraction, pouring

out, departure, and the vessel drunk, eaten, or scooped from. `destination`: the endpoint that a person, entity, or content reaches, including placement, insertion, pouring into, hand-over, and entry. A container can therefore be a source or destination, depending on the action. `constraint`: the entity limiting, obstructing, fixing, or restricting movement or state. A role describes the entity's function in this action, not its noun category. Fixed-role actions take one role by definition (`hold`, `grasp`, `touch`, `wear`: `target`; `sit`, `lie`, `stand`, `ride`: `support`); other actions distribute their participants over roles as the event requires.

**Participant identity.** People, objects, and body regions have image-local entity identities. A person participating in another actor's action retains the same identity across events. In the benchmark definition, an actor's affected body region is localized separately. Merely wearing or holding an object does not make the whole actor a self-directed participant.

**Class support.** Of 2,972 observed action–noun–role classes, 504 have at least ten training examples, 2,323 have one to nine, and 145 have no training examples. The corresponding test-supported counts are 424, 547, and 133, with 10,374, 1,201, and 208 test relations. The six roles have 51,142 target, 3,311 instrument, 12,036 support, 761 source, 1,081 destination, and 472 constraint relations across the release.

## B CoRISP Parameterization and Training

**Visual inputs.** Frozen H-DETR (Jia et al., 2023) with a Swin-L backbone (Liu et al., 2021b) supplies proposals. DINOv3 ViT-L/16 features pass through a pretrained DINO.txt vision head for text alignment (Siméoni et al., 2025; Jose et al., 2025); both components are frozen. A trainable projection maps person-proposal features into queries processed with image tokens by this vision head. Gradients pass through the head to the projection. Person, entity, union-region, geometry, and global features initialize 384-dimensional interaction states.

**Local distribution.** For a visible candidate $i = (h, j, v)$, semantic action logit $\zeta_i$, interaction residual $\eta_i$, role residual $d_{i,r}$, and learned scale $\lambda$, the logits are

$$z_{i,0} = 0, \qquad z_{i,r} = \zeta_i - \log|\mathcal{R}_v| + \lambda(\eta_i + d_{i,r} - \bar{d}_i), \quad \bar{d}_i = |\mathcal{R}_v|^{-1} \sum_{r \in \mathcal{R}_v} d_{i,r}. \tag{8}$$

The categorical distribution is $p_i = \mathrm{softmax}(z_i)$. At zero residual, $q_i = \mathrm{sigmoid}(\zeta_i)$ independently of role count. The conditional role mean uses $\epsilon = 10^{-30}$. Event messages are averaged with $q_i$; entity messages first average over actions and then across incoming pairs with weights $1 - \prod_v (1 - q_{hjv})$. Entity states with zero incoming mass are retained. Positive-role probabilities are multiplied by detector-confidence and native visibility factors, yielding $\widetilde{p}_i(r)$ and $\widetilde{p}_i(0) = 1 - \sum_r \widetilde{p}_i(r)$.

**Role pooling.** The neighborhoods are $\mathcal{N}_{\mathrm{event}}(i) = \{(h, j', v)\}$, $\mathcal{N}_{\mathrm{pair}}(i) = \{(h, j, v')\}$, and $\mathcal{N}_{\mathrm{entity}}(i) = \{(h', j, v')\}$ over valid indices. Let $w_{i,r} = \mathrm{stopgrad}(\widetilde{p}_i(r))$. For query role $r$ and source role $r'$, the summary is

$$\boldsymbol{\mu}^{g}_{i,r'|r} = \frac{\sum_{i' \in \mathcal{N}_g(i)} w_{i',r'} \mathbf{v}_{i',r'} - \mathbf{1}\{r' = r\} w_{i,r} \mathbf{v}_{i,r}}{\sum_{i' \in \mathcal{N}_g(i)} w_{i',r'} - \mathbf{1}\{r' = r\} w_{i,r}}. \tag{9}$$

The neighborhood includes $i$ before subtraction. Zero-mass summaries are masked; other denominators are floored numerically. Scaled dot-product attention (Vaswani et al., 2017) uses local features to query summaries augmented with role and relation embeddings. A second attention fuses neighborhood messages. Refinement uses $\delta_{i,r} = 4 \tanh\beta \, \tanh\Delta_{i,r}$, where $\Delta_{i,r}$ is predicted from the bound state and $\beta$ is initialized to zero. The same detached weights pool bound states within each event and role; adding event and action–role features followed by layer normalization yields $\mathbf{U}^{\mathrm{role}}_{h,v}$.

**Count potentials.** A linear head on the normalized event state predicts $B(k) = f_B(\mathbf{e}_{h,v})_k - f_B(\mathbf{e}_{h,v})_0$. For role count embeddings $\boldsymbol{\kappa}_t$, $t \in \{0, 1, 2\}$, and learned projections of width $d_G$, let

$R = |\mathcal{R}_v|$ and $\mathbf{U}_r$ denote one role-context row:

$$\mathbf{f}_r(c_r) = \mathrm{LN}\big(\phi_U(\mathbf{U}_r) + \boldsymbol{\kappa}_{c_r} + \phi_U(\mathbf{U}_r) \odot \boldsymbol{\kappa}_{c_r}\big), \tag{10}$$

$$g(\mathbf{c}) = d_G^{-1/2} \phi_E(\mathbf{e}_{h,v})^\top \phi_S\left(R^{-1/2} \sum_r \mathbf{f}_r(c_r)\right), \tag{11}$$

$$G(\mathbf{c}) = 4 \tanh \beta_G \, \tanh\big(g(\mathbf{c}) - g(\mathbf{0})\big). \tag{12}$$

The scale $\beta_G$ is initialized to zero. Assignment-dependent inputs to $B$ and $G$ are counts; candidate-specific compatibility enters through the contextual unaries.

**Dynamic program.** Set disallowed positive unary weights to $-\infty$ without changing $\ell_m(0)$. Initialize $D_0(0, \mathbf{0}) = 0$ and all other states to $-\infty$. Let $T_r$ increment the role-$r$ count, saturating at two. Then

$$A_{m+1}(k, \mathbf{c}) = \mathrm{LSE}_{\substack{r \in \mathcal{R}_v,\ \mathbf{c}': \\ T_r(\mathbf{c}') = \mathbf{c}}} [D_m(k-1, \mathbf{c}') + \ell_{m+1}(r)], \tag{13}$$

$$D_{m+1}(k, \mathbf{c}) = \mathrm{LSE}\big(D_m(k, \mathbf{c}) + \ell_{m+1}(0),\ A_{m+1}(k, \mathbf{c})\big). \tag{14}$$

The predecessor sum includes only valid states; both counts 1 and 2+ can precede a saturated count. Applying the terminal potentials gives Equation (6). The dense recurrence takes $O(M^2 R 3^R)$ time and $O(M 3^R)$ rolling-table storage, excluding backward intermediates. Removing unreachable states is exact under the same support. For V-COCO, role-specific null candidates represent missing fillers and enter the same event-level inference.

**Matching and loss.** HEIR matches people one-to-one at IoU $\geq 0.5$; participant proposals require the annotated noun and IoU $\geq 0.5$. Compatible complete assignments use each proposal at most once and deduplicate matching paths. Ambiguous person targets and positives without a compatible proposal assignment are ignored during training, not removed from evaluation. Unobserved actions are treated as negatives only where the source annotation defines the corresponding supervision as complete. Cropping away a participant makes the positive event unknown rather than a smaller complete event.

We apply focal modulation (Lin et al., 2017) to the set likelihood. For $L_y = \mathcal{L}_{\mathrm{set}}(y)$, supervised events $\mathcal{O}$ in a local microbatch, $W$ distributed ranks, and positive participant-group counts $N_w^+$, the loss is

$$\mathcal{L} = \frac{\sum_{y \in \mathcal{O}} 0.5(1 - e^{-L_y})^{0.1} L_y}{\max(1, W^{-1} \sum_{w=1}^{W} N_w^+)}. \tag{15}$$

Gradients are averaged across ranks; normalization is per microbatch. Visual computations use BF16 autocast, and log-space DP uses FP32. HEIR also evaluates structured-potential projections in FP32; V-COCO retains autocast for those projections. Frozen modules are excluded from trainable-parameter counts but remain part of inference.

## C Evaluation and Supporting Results

**HEIR relations.** Role mAP averages AP over 1,104 test-supported action–noun–role classes. HOI mAP instead uses 1,057 action–noun classes, retaining the maximum role score for each person–participant–action tuple. This common projection is applied to every model. For CoRISP, the benchmark score is $\max_r \pi_m(r)$, while the model's interaction probability is $q_m^{\mathrm{set}} = \sum_r \pi_m(r)$. Rare, non-rare, and unseen classes have 1–9, at least 10, and zero training instances, respectively.

**V-COCO.** V-COCO (Gupta & Malik, 2015) annotates COCO images (Lin et al., 2014), with 2,533 training, 2,867 validation, and 4,946 test images. We train on trainval and evaluate all test images. Role AP under Scenarios 1 and 2 is denoted by $\mathrm{AP}^{\mathrm{S1}}_{\mathrm{role}}$ and $\mathrm{AP}^{\mathrm{S2}}_{\mathrm{role}}$, respectively. Our re-evaluations average the 24 action–role classes remaining after excluding `point`. Both scenarios use the authors' matching rules[1]: a correct detection requires the action and role labels and IoU $\geq 0.5$ for the person and each evaluated visible filler. For an absent ground-truth filler, S1 requires a null prediction, whereas S2 ignores its object localization.

[1] https://github.com/s-gupta/v-coco

**HEIR baseline sets.** Same-noun prediction boxes are clustered at IoU $\geq 0.7$ against cluster representatives, and duplicate edges retain their maximum score. For each predicted person identity and action, edges are sorted by confidence. Every nonempty prefix forms a set hypothesis, scored by its minimum member confidence. Baselines use this decoder, retaining at most 100 sets per image without an additional relation cap or a member-score threshold. Both baseline and CoRISP sets are evaluated with one shared, noun-compatible, one-to-one entity correspondence per image at IoU $\geq 0.5$. The test set contains 9,955 annotated sets across 101 actions; 85.5% have one member.

**Shared entity correspondence.** Each predicted entity appearing in a submitted set receives the maximum score of any set containing it as actor or participant. Entities are processed in descending order of this priority, then descending entity confidence, then ascending entity ID. Each is assigned to the unused ground-truth entity with the same noun and greatest IoU, provided IoU $\geq 0.5$; equal IoUs are resolved by ascending ground-truth ID. The resulting correspondence is reused across all events in the image. Action and role agreement are checked when matching complete sets.

**Set AP.** Within each action, hypotheses are ordered by decreasing set score, with image and hypothesis IDs resolving ties for matching. A true positive must have every endpoint matched and exactly equal an unused annotated event's entity–role set. Each annotated event can be used once; unmatched hypotheses and subsequent duplicates count as false positives. Equal-score outcomes are grouped before computing precision and recall. AP integrates the all-point interpolated precision envelope, and Set mAP averages AP over the 101 test-supported actions.

**Structure-specific Set AP.** Multi restricts both ground-truth sets and submitted hypotheses to at least two entity–role members; Repeat requires at least two distinct entities with the same role. These strata contain 1,441 and 874 ground-truth sets over 62 and 34 actions, respectively. The 100-set budget is applied before stratum filtering, and correspondence is recomputed within each stratum. These scores measure exact recovery conditional on the specified member structure. Shared evaluates all 3,217 person–action sets over 64 actions in the 534 images containing cross-actor participant sharing. Full-test Set mAP and these structural breakdowns together describe overall and composition-specific performance.

**CoRISP set prediction.** CoRISP submits sets directly from its event potentials. A max-plus dynamic program finds the highest-scoring assignment for each reachable total-count and saturated-role-count state. Each state contributes one winner. We rank these winners by their exact probability under the full distribution, retain at most $K = 8$ per person–action event, and apply the same 100-set image budget. The resulting alternatives represent different count states. This readout is fixed across CoRISP evaluations and component ablations.

**V-COCO sets.** Our complete-set metric extends V-COCO evaluation to joint slot recovery over 21 role-bearing actions, excluding `point`. Predicted person boxes are clustered at IoU $\geq 0.7$ against fixed representatives, in descending order of their maximum role score. Each cluster–action pair forms one hypothesis using the highest-scoring prediction for each native role slot, scored by the minimum slot confidence. Every slot requires an explicit prediction, including an explicit null for a missing filler. No confidence threshold or additional output cap is applied. Matching follows the native person-matching convention and S1/S2 missing-filler rules, requiring all role slots to be correct. Set mAP averages interpolated AP over actions, grouping tied scores. Dual averages the three two-slot actions (`hit`, `eat`, and `cut`), including their absent-filler cases under S1/S2.

**Training.** CoRISP keeps the proposal detector and visual–semantic encoders frozen while training its feature projections, recurrent interaction field, context aggregation, and set potentials. The complete model has 9.92M trainable parameters on HEIR and 9.91M on V-COCO, excluding frozen components. The same architecture uses six functional roles on HEIR and native `obj`/`instr` slots on V-COCO. Both datasets use a 30-epoch training schedule. HEIR checkpoints are selected by validation Role mAP. V-COCO uses 5,267 eligible trainval images with visible-participant or missing-filler supervision and evaluates all 4,946 test images.

**Optimization.** The V-COCO model and HEIR component runs use AdamW with learning rate $10^{-4}$, weight decay $10^{-4}$, and gradient-norm clipping at 0.1. The learning rate is multiplied by 0.2 after epoch 20. Two accumulated microbatches give an effective batch size of 16; the random seed is 42. V-COCO uses the final epoch, and the HEIR component runs use validation Role mAP for checkpoint selection.

Table 5: **Effect of the set hypothesis budget** (Set mAP, %). The CoRISP checkpoint is fixed. $K$ counts state winners; both settings retain at most 100 sets/image.

| Readout | Full↑ |
|---|---|
| Single best nonempty assignment | 17.56 |
| State-wise MAP, $K = 8$ | **18.04** |

**Component protocol.** Each ablation is trained independently with the same optimization schedule, checkpoint criterion, set prediction rule, and output budget as the full model.

**Model configurations.** SOV-STG Swin-L uses resolution 384 on HEIR. Its V-COCO model and SOV-STG R101 on HEIR are trained with official code (T); SOV-STG R101 on V-COCO uses the released SOV-STG-L weights (W). MUREN and SOV-STG-S use ResNet-50 (He et al., 2016), while GEN-VLKT-S/L use ResNet-50/101. GroupHOI-S uses ResNet-50 with CLIP-B/16 (Radford et al., 2021), InCoM-Net uses ResNet-50 with CLIP-L/14, and SL-HOI uses DINOv3-L/16. SOV-STG-VLA-S uses ResNet-50 with BLIP-2 (Li et al., 2023b) on HEIR and CLIP-B/32 on V-COCO; its HEIR parameter count does not describe the V-COCO configuration. The published HOI-IDiff result lists Deformable DETR (Zhu et al., 2020); UniHOI uses VQGAN (Esser et al., 2021) and Llama3-8B (Grattafiori et al., 2024) with 550K additional image–text pairs. Score superscripts W/T/A/R indicate released weights, our training with official code, adapted official code, and our reimplementation, respectively.

UniHOI's released materials do not provide its task checkpoint and complete HOI inference pipeline; HOI-IDiff does not specify the exact task detector checkpoint and the full diffusion-training and role-output configuration. We retain their published V-COCO results under the cited protocols.

**Comparison scope.** Baselines retain the encoders and pretraining indicated in the main tables. Dagger-marked HEIR baselines use the shared inventory within their standard inference pipelines; other baselines retain their native support. All set outputs share the same 100-set budget. Their set outputs share the 100-set budget. Frozen upstream components are included in the system, but not in trainable-parameter counts. Native V-COCO role averaging and complete-set evaluation remain distinct from HEIR's noun-qualified, image-level identity criterion.

**Number of set hypotheses.** Table 5 holds the primary HEIR checkpoint fixed. Both settings score complete assignments using the event potentials: one returns the best nonempty assignment, while the other retains up to eight winners from different count states.

**Action–cardinality oracle.** For each annotated person–action event $e$, we reveal its true action and member count $k_e^\star = |S_e^\star|$. Predicted boxes are first clustered with same-noun IoU $\geq 0.7$. People are matched greedily by IoU, and each ground-truth event then receives a noun-compatible, one-use participant matching at IoU $\geq 0.5$. Let $\sigma_e$ rank that event's predicted entity–role edges by confidence. The oracle recovery rate is

$$R_{k^\star} = \frac{1}{|\mathcal{D}|} \sum_{e \in \mathcal{D}} \mathbf{1}\left[\mathrm{Top}_{k_e^\star}(\sigma_e) = S_e^\star\right]. \tag{16}$$

Unmatched events count as failures. Coverage $C$ is the fraction whose subject and every participant admit the required box-and-noun matching. Both $C$ and $R_{k^\star}$ are event-averaged recalls with the true action and set size supplied. Matching is performed within each event for this diagnostic, following the procedure above.

**Benchmark diagnostics.** Figure 3 compares all 16 HEIR baselines in Table 1. Action-balanced Role AP first averages supported noun–role classes within each action, then averages over the same 101 actions as Set mAP. Under equal action weights, 41 of 120 model pairs reverse order; ten of 55 reverse among shared-inventory baselines. Multi-role-pair Role AP weights each of the 47 test-supported action–noun pairs with multiple roles equally, averaging its role AP values first. Both analyses use the full test predictions.

For the pair-balanced comparison in Figure 3(a), HOI AP averages one AP per action–noun pair, while Role AP first averages that pair's supported role classes and then averages pairs. The multi-role group contains 47 pairs, 94 role classes, and 630 relations; the single-role group contains 1,010 pairs and 11,153 relations. Their AP gap measures sensitivity to role labels under equal pair weights.

Table 6: **Action–cardinality oracle at 100 relations per image** (%). $C$: box-and-noun coverage; $R$: exact top-$k^\star$ recovery with the true action and count supplied. Denominators include all 1,441 Multi and 874 Repeat events. Denominators include all 1,441 Multi and 874 Repeat events.

| **Method** | **Multi** | | **Repeat** | |
|---|---|---|---|---|
| | $C\uparrow$ | $R\uparrow$ | $C\uparrow$ | $R\uparrow$ |
| *Visual models* | | | | |
| QPIC R50 (Tamura et al., 2021) | 28.94 | 14.99 | 31.69 | 16.25 |
| QPIC R101 (Tamura et al., 2021) | 31.58 | 15.75 | 33.98 | 16.48 |
| MUREN (Kim et al., 2023) | 34.91 | 16.79 | 40.16 | 17.85 |
| SOV-STG-S (Chen et al., 2025) | 42.19 | 17.49 | 49.08 | 18.88 |
| SOV-STG Swin-L (Chen et al., 2025) | **51.21** | <u>24.22</u> | 56.06 | 24.94 |
| SOV-STG R101 (Chen et al., 2025) | 43.44 | 18.46 | 49.31 | 18.76 |
| PViC R50 (Zhang et al., 2023) | 37.82 | 17.77 | 46.11 | 21.85 |
| PViC Swin-L (Zhang et al., 2023) | 47.47 | 22.55 | <u>60.41</u> | <u>26.89</u> |
| *Vision–language models* | | | | |
| GEN-VLKT-S (Liao et al., 2022) | 44.97 | 16.10 | 49.77 | 17.96 |
| GEN-VLKT-L (Liao et al., 2022) | 48.09 | 19.08 | 52.63 | 20.14 |
| SOV-STG-VLA-S (Chen et al., 2025) | <u>49.13</u> | 15.61 | 53.32 | 16.93 |
| RLIPv2 Swin-T (Yuan et al., 2023) | 41.71 | 21.58 | 49.31 | 25.06 |
| RLIPv2 Swin-L (Yuan et al., 2023) | 48.30 | 23.94 | 54.58 | 25.51 |
| GroupHOI-S (Hong et al., 2025) | 40.60 | 12.28 | 47.94 | 13.39 |
| InCoM-Net (Seo et al., 2026) | 31.51 | 17.14 | 41.42 | 21.62 |
| SL-HOI (Sun et al., 2026) | 25.75 | 6.87 | 30.09 | 6.29 |
| CoRISP | 48.92 | **27.48** | **62.01** | **33.41** |

Edge coverage requires every ground-truth member to occur in the union of retained hypotheses; exact recovery requires one hypothesis to equal the entire set. Under the same 100-set budget, the latter recovers 54.86%, 56.49%, 53.85%, 61.81%, and 53.58% of edge-covered multi-participant events for PViC R50, PViC Swin-L, GroupHOI-S, InCoM-Net, and SL-HOI. Every covered but unrecovered event has an extra member in each covering prefix. Each recall is conditioned on that model's own covered events.

**Classification with correct grounding.** We compare CoRISP with SL-HOI, the strongest relation-level baseline under this common-grounding analysis, on a shared set of person–participant pairs. Each model's predicted pairs are matched to annotated pairs once, first maximizing the number with person and participant IoU $\geq 0.5$ and the correct foreground noun, then maximizing the sum of their minimum endpoint IoUs. Matching uses geometry and nouns only. The intersection contains 10,383 pairs from 2,495 images, including 4,568 pairs without a positive annotated relation. Both models retain their predicted scores; actions and roles are predicted over identical pairs and supervision scopes. AP averages classes with at least one positive in this shared population. CoRISP and SL-HOI obtain 63.05/50.97 conditional HOI mAP over 562 action–noun classes and 61.50/50.26 conditional Role mAP over 584 action–noun–role classes. Among the 214 supported rare role classes, the scores are 58.68/47.28. These values measure classification on common correctly grounded candidates; the features and detection confidences remain those of each model.

**Scoring convention.** Identical GEN-VLKT-S predictions score 62.48/64.53 over 25 action–role classes and 65.05/67.19 over the 24 classes excluding `point`. The change follows solely from the averaging convention. All re-evaluated $\mathrm{AP}_{\mathrm{role}}$ scores in Table 1 use the latter convention; literature-only rows retain source averaging conventions and are displayed separately.

Table 7: **Published V-COCO role AP** (%, S1/S2). Source averaging conventions differ; scores are therefore unranked. R/C/D: ResNet/CLIP/DINOv3; DDETR: Deformable DETR.

| **Method** | **Encoder(s)** | **Reported role AP↑** |
|---|---|---|
| *Visual models* | | |
| QPIC (Tamura et al., 2021) | R50 | 58.8/61.0 |
| QPIC (Tamura et al., 2021) | R101 | 58.3/60.7 |
| MUREN (Kim et al., 2023) | R50 | 68.8/71.0 |
| SOV-STG-S (Chen et al., 2025) | R50 | – |
| SOV-STG R101 (Chen et al., 2025) | R101 | 63.9/65.4 |
| PViC (Zhang et al., 2023) | R50 | – |
| PViC (Zhang et al., 2023) | Swin-L | 64.1/70.2 |
| *Vision–language models* | | |
| GEN-VLKT-S (Liao et al., 2022) | R50 | 62.41/64.46 |
| GEN-VLKT-L (Liao et al., 2022) | R101 | 63.58/65.93 |
| SOV-STG-VLA-S (Chen et al., 2025) | R50 + C-B/32 | 63.8/65.7 |
| RLIPv2 (Yuan et al., 2023) | Swin-T | 68.8/70.8 |
| RLIPv2 (Yuan et al., 2023) | Swin-L | 72.1/74.1 |
| GroupHOI-S (Hong et al., 2025) | R50 + C-B/16 | 65.0/66.0 |
| InCoM-Net (Seo et al., 2026) | R50 + C-L/14 | 73.6/75.4 |
| SL-HOI (Sun et al., 2026) | D-L/16 | – |