FINE-GRAINED COMPOSITIONAL EVALUATION OF HUMAN-OBJECT INTERACTION UNDERSTANDING IN MULTI-PERSON SCENES

Anonymous authors
Paper under double-blind review

ABSTRACT

Human–object interaction (HOI) detection requires localizing people and objects and recognizing their interactions. Existing evaluations typically report aggregate metrics, offering limited insight into which scene structures challenge a model. Recent diagnostic studies examine individual interaction patterns, such as whether people perform the same or different actions on shared or separate objects. Yet real-world scenes often combine multiple patterns, raising an underexplored question: how well do HOI models handle such combinations? Answering this question systematically is difficult because existing benchmarks provide sparse and uneven coverage of these combinations. To address this gap, we introduce HICO-DET-SYN, a controlled synthetic benchmark that expands HICO-DET's coverage of coexisting interaction patterns. Our multi-agent image-editing pipeline constructs underrepresented combinations by adding people, objects, and interactions to real images while preserving the original scene context and existing interactions. The synthetic images complement the original data to provide more balanced coverage of diverse combinations, enabling systematic evaluation of specialized HOI detectors and general-purpose vision–language models on scenes containing individual or coexisting patterns. Our evaluation shows that specialized HOI models generally outperform general-purpose vision-language models, while compositional difficulty follows a non-monotonic trend rather than increasing steadily with scene complexity. These findings suggest that the structure of interaction composition, rather than complexity alone, is a key factor in HOI understanding. By making underrepresented interaction combinations testable, HICO-DET-SYN enables targeted diagnosis of compositional limitations and lays the groundwork for more reliable HOI understanding in real-world scenes.

1 INTRODUCTION

Human–object interaction (HOI) detection aims to localize human–object pairs and recognize their interactions in visual scenes (Gkioxari et al., 2018; Tamura et al., 2021). In realistic scenes, multiple people and objects often coexist, requiring models to understand multiple human–object interactions and correctly associate each interaction with the corresponding people and objects (Wang et al., 2024; Geng et al., 2025; Seo et al., 2026). Existing methods are typically evaluated using mean Average Precision (mAP) over the entire benchmark (Li et al., 2023b; Zhang et al., 2021), such as HICO-DET (Chao et al., 2018), V-COCO (Gupta & Malik, 2015), and SWiG-HOI (Wang et al., 2021). While such aggregate metrics provide a convenient measure of overall performance, they offer limited insight into the types of scenes in which a model succeeds or fails. As illustrated in Fig. 1(c), models with similar overall mAP can exhibit different behavior across multi-person interaction patterns. For example, HOLa and CMMP achieve comparable overall mAP, but their relative performance reverses across different scenes.

Recent diagnostic evaluation characterizes multi-person HOI scenes using interpretable interaction patterns, such as whether people interact with the same or different object instances and whether they perform the same or different interactions (Wang et al., 2026). However, such analyses typically describe a scene using a single interaction pattern. Real-world multi-person scenes can contain multiple patterns simultaneously. For example, as illustrated in Fig. 1, several people are riding the

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.

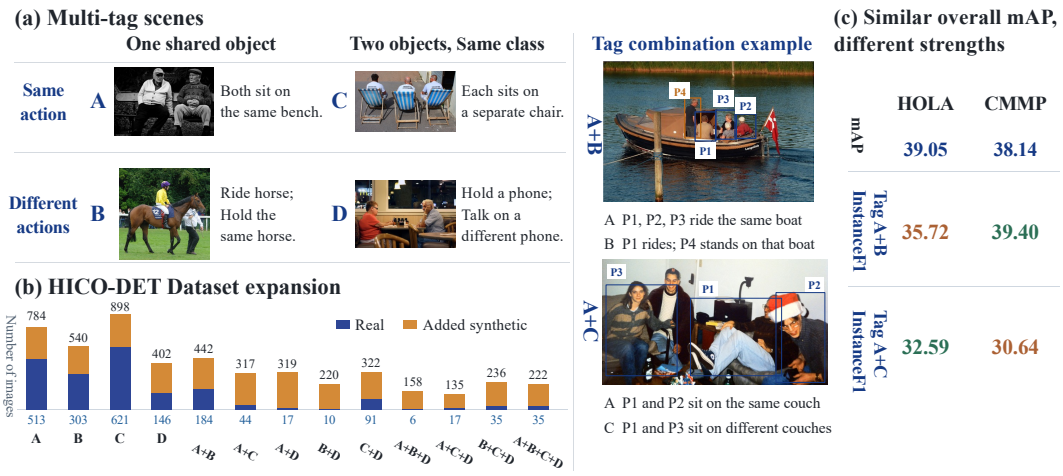


Figure 1: (a) Examples illustrate individual tags and their combination within a scene. (b) Synthetic data expands coverage of Tag A–D and their combinations. (c) Despite similar overall mAP, HOLa (Lei et al., 2025) and CMMP (Lei et al., 2024) exhibit opposite rankings in instance F1 on Tag A+B and A+C, displaying different advantages.

same boat, while another person is standing on that boat. We refer to such cases as **multi-tag scenes**, where multiple interaction patterns coexist within a single image.

This motivates a natural question: how does HOI understanding change when multiple interaction patterns coexist within the same scene? Systematically answering this question using existing HOI benchmarks is difficult, as naturally collected test sets provide highly uneven coverage of multi-tag scenes. While individual tags are reasonably represented, many multi-tag combinations contain only a small number of images, and some are entirely absent. As shown in Fig. 1(b), the available multi-tag scenes are therefore insufficient for reliable and systematic analysis across different combinations.

To address this limitation, we introduce **HICO-DET-SYN**, a controlled synthetic benchmark built from real HICO-DET images to provide systematic coverage of multi-tag scenes. We construct HICO-DET-SYN through a controlled image-editing pipeline that transforms real HICO-DET images into specified multi-tag scenes while preserving the original scene context, camera viewpoint, visual appearance, and existing interactions. Because the target tag combination and HOI classes are explicitly specified during generation, the corresponding annotations are obtained by construction. We further apply automated consistency checks and human validation to verify that the intended interactions are accurate and that the edited images remain visually coherent. HICO-DET-SYN therefore complements the natural HICO-DET test set by enabling controlled evaluation across multi-tag combinations that cannot be reliably analyzed from existing HICO-DET benchmark alone.

Using the original HICO-DET and our synthetic HICO-DET-SYN, we systematically analyze representative HOI detectors and general-purpose vision-language models (VLMs) across individual tags and multi-tag scenes. This allows us to examine whether model behavior observed under simpler interaction patterns persists when multiple patterns coexist within the same image. Our analysis reveals several notable trends. First, specialized HOI models generally outperform general-purpose LVLMs, suggesting that task-specific supervision remains important for precise human-object interaction recognition and attribution. Second, difficulty does not increase monotonically with the number of coexisting interaction patterns. Instead, performance follows a roughly V-shaped trend, decreasing from single-tag to two-tag scenes before partially recovering for three- and four-tag compositions. Finally, certain two-tag configurations, particularly B+D and C+D, are consistently more challenging than even higher-order compositions. These results indicate that the structure of the interaction composition, rather than tag cardinality alone, is a key factor governing HOI difficulty.

Our main contributions are:

- We highlight an overlooked problem in current HOI evaluation: real-world multi-person scenes often contain multiple interaction patterns simultaneously, but existing benchmarks and analyses provide limited coverage and little systematic diagnosis of model behavior in such multi-tag scenes.
- We introduce **HICO-DET-SYN**, a controlled synthetic benchmark built from real HICO-DET images to provide systematic coverage of multi-tag scenes. A controlled image-editing pipeline constructs specified tag combinations while preserving realistic scene context, with verified annotations.
- We evaluate representative specialized HOI detectors and general-purpose VLMs on real and synthetic single- and multi-tag subsets under a common protocol, and show that compositional difficulty is not determined solely by the number of tags: performance drops from single- to two-tag compositions but can partially recover for three- and four-tag combinations, with substantial variation across specific tag combinations.

2 Related Work

Human–Object Interaction Benchmarks. HOI benchmarks evaluate interaction understanding from different perspectives. HICO-DET and V-COCO focus on instance-level HOI detection and person-centric action roles, respectively (Chao et al., 2018; Gupta & Malik, 2015). Later benchmarks broaden evaluation, with SWiG-HOI scaling up interaction annotations (Wang et al., 2021) and Bongard-HOI targeting few-shot visual reasoning (Jiang et al., 2022). Recent work further revisits evaluation: CrossHOI-Bench compares VLMs with HOI-specific models (Lei et al., 2026), and SHOE introduces semantic evaluation for open-vocabulary HOI (Noack et al., 2026). However, these benchmarks do not provide fine-grained diagnostic analysis of model behavior across different interaction patterns. DiagHOI takes a step in this direction by organizing scenes according to interpretable interaction patterns (Wang et al., 2026), but analyzes each pattern separately. In contrast, we focus on multi-tag scenes, where multiple interaction patterns coexist within one image.

Vision–Language Models for HOI. Large vision–language models (VLMs) have shown strong general-purpose visual understanding and open-vocabulary recognition across a wide range of tasks (Radford et al., 2021; Li et al., 2023a; Liu et al., 2024a; Chen et al., 2024; Team et al., 2024; Bai et al., 2025; OpenAI, 2026). Their language-aligned representations have also been adopted by HOI-specific methods to improve semantic generalization and recognition of unseen interactions (Luo et al., 2026; Mao et al., 2023; Hu et al., 2025). More recently, CrossHOI-Bench directly evaluates standalone VLMs for HOI understanding and reports challenges in scenes involving multiple people, concurrent actions, and interaction attribution (Lei et al., 2026). However, existing evaluations largely characterize HOI capability at the aggregate level or through broad scene categories, without systematically disentangling how different forms of interaction composition affect model behavior, which is the focus of this work.

Synthetic Benchmarking Datasets. Synthetic and procedurally generated data provide a useful way to construct controlled evaluations in which individual factors of variation can be manipulated independently. Such datasets have been widely used to study compositional visual reasoning, temporal reasoning, physical understanding, and robustness, including CLEVR (Johnson et al., 2017), CATER (Girdhar & Ramanan, 2019), Physion (Bear et al., 2021), Kubric (Greff et al., 2022), CVR (Zerroug et al., 2022), and Super-CLEVR (Li et al., 2023c). Yang et al. (2024) also introduces synthetic data for HOI detection, using generative models to alleviate data scarcity for rare classes in training. In contrast, we use controlled image editing on real HICO-DET images to construct multi-tag scenes for controlled evaluation rather than training augmentation.

3 Methodology

Preliminaries: Multi-Person HOI Interaction Tags. To establish a fine-grained compositional evaluation of human-object interaction (HOI) models in multi-person scenes, we build upon the interaction taxonomy introduced by Wang et al. (2026). Under this formulation, multi-person interactions are classified along two key factors: (1) object instance relationships (sharing the same instance vs. distinct instances) and (2) action equivalence (identical vs. distinct actions), yielding atomic interaction tags (Tags A–D in Figure 1(a)). However, real-world scenes with three or more

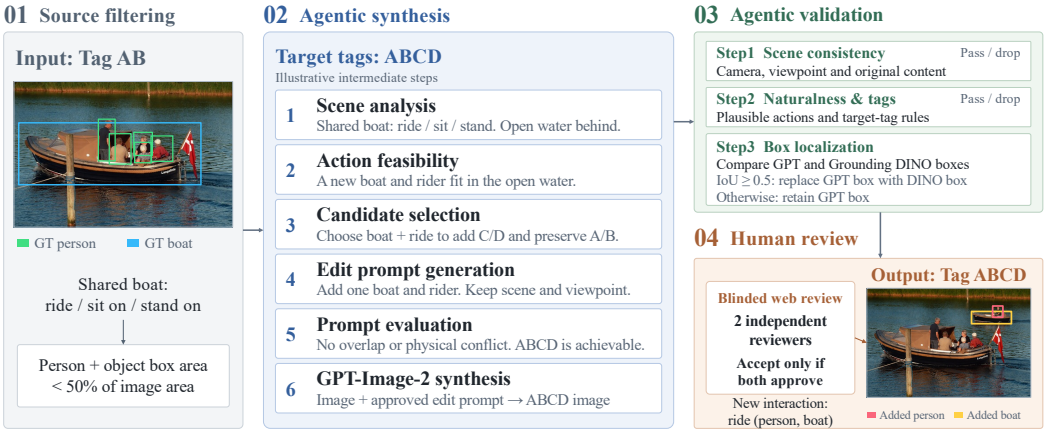


Figure 2: Overall framework of our proposed multi-agent image synthesis pipeline. The boat example illustrates an Tag AB-to-ABCD transformation.

individuals exhibit multiple interaction patterns simultaneously. To evaluate compositional understanding under such complex dynamics, we define these configurations as **Multi-Tags** (e.g., $A+B$, $A+C$, $B+D$, $B+C+D$, and $A+B+C+D$). As illustrated in Figure 1, an $A+B$ scene requires parsing both shared and distinct actions on an identical object instance, whereas an $A+C$ scene demands distinguishing identical actions across both shared and separate instances. Multi-Tags thereby enable fine-grained, diagnostic evaluation beyond conventional aggregated metrics by decomposing model performance across specific compositional structures.

Image Synthesis Pipeline. The synthesis pipeline operates through four sequential steps (Figure 2): (1) source image pre-filtering, (2) Agentic synthesis to produce realistic interactions using generative models, (3) automated agentic validation to ensure visual consistency and correct semantics, and (4) final human review where two independent annotators verify overall visual quality and annotation fidelity.

STEP 1: SOURCE IMAGE PRE-FILTERING. To prevent synthesis artifacts, we filter out excessively crowded scenes lacking sufficient space for natural object placement (details in Appendix B).

STEP 2: MULTI-AGENT SYNTHESIS STAGE. This stage coordinates six specialized agents to sequentially draft, evaluate, and execute image edits. First, the scene analysis agent examines the source image and ground-truth annotations to extract environmental geometry, lighting, and camera viewpoint. Next, the feasibility generation agent proposes candidate actions grounded in real-world HICO-DET co-occurrence statistics, from which the candidate selection agent chooses valid object–action pairs that satisfy the target multi-tag rules without altering pre-existing subjects. An edit prompt generation agent then converts these selected pairs into explicit editing instructions, which are strictly vetted by a prompt evaluation agent to filter out physical inconsistencies. Finally, the approved prompt and reference image are passed to `GPT-Image-2` to generate the final synthesized image.

STEP 3: MULTI-AGENT VALIDATION AND BOUNDING BOX REFINEMENT. The synthesized image is sequentially evaluated by three automated validation agents. A consistency agent first verifies that the camera angle, background, and original subjects remain unchanged, discarding failed cases immediately. A naturalness and tag verification agent then confirms visual nationality and validates that the added interaction satisfies the target multi-tag rules. Finally, a localization agent estimates 2D bounding boxes for the newly introduced person and object. To compensate for coarse coordinate predictions from generative models, we refine these boxes using Grounding DINO Liu et al. (2024b): if a detected box has an IoU ≥ 0.5 with the predicted box, it replaces the candidate annotation (accounting for 82.5% of added objects, while 17.5% retain their initial predictions).

STEP 4: FINAL REVIEW BY HUMAN ANNOTATORS. Images that passed all automated validation steps were manually reviewed by human annotators via a web-based interface. Annotators checked both the original and synthesized images, with and without bounding box overlays, to detect an-

notation errors and visual artifacts. Images were included in the final dataset only after receiving approval from at least two independent annotators. The main reason by Human reviewer was due to physical or semantic inconsistencies, as well as annotation corruption. Common failure cases included physically implausible interactions (e.g., two distant individuals simultaneously flying a single kite, or an inserted person holding skis mid-air while another person was jumping) and unintentional modifications to existing subjects or objects that invalidated the original HICO-DET ground-truth annotations.

4 EXPERIMENTS

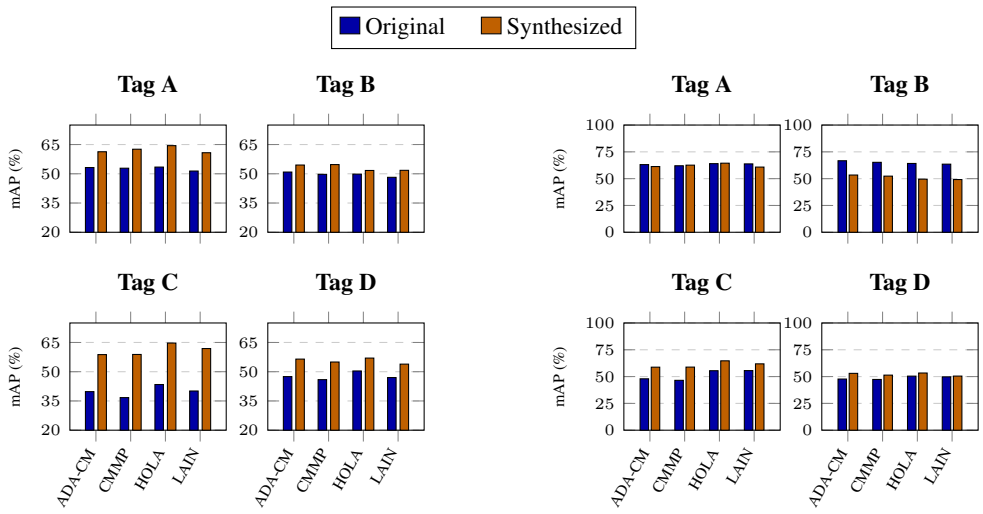
Dataset and Evaluation Setting. We build our benchmark upon HICO-DET (Chao et al., 2018), which contains 38,118 training images and 9,658 test images spanning 600 HOI categories over 80 object classes. Its standard evaluation reports mAP over Full, Rare, and Non-Rare category sets, providing an established measure of HOI detection performance. These HOI-class-level summaries, however, do not explicitly distinguish different interaction patterns examined in our benchmark. Applying our tag taxonomy to real images from HICO-DET reveals uneven coverage across interaction patterns (Fig. 1). Some multi-tag combinations contain few examples, limiting reliable comparisons of model behavior within these scenarios. Moreover, aggregated metrics such as mAP can obscure differences in how models handle shared objects, distinct object instances, and co-occurring actions. We therefore complement real-image evaluation with targeted synthetic examples and report performance separately for each interaction pattern and their combinations.

Analysis of the Synthetic Dataset. Through over 100 synthesis batches, 8,897 candidate images passed the automated multi-agent validation stage. After human review, we retained 3,308 images, yielding approximately 200 instances per target interaction tag. Although these images passed human review for visual naturalness and annotation accuracy, it remains questionable whether HOI models treat them equivalently to real images. Following previous work (Wang et al., 2026), we evaluated their mAP results to address this. However, a direct mAP comparison (Figure 3a) presents a limitation. Because each dataset has a different HOI distribution, containing different proportions of easy and hard interaction categories. Comparing performance without accounting for these differences can lead to misleading conclusions. Specifically, we designed our prompts using the frequent HOI categories during image generation. As a result, interactions that appear only rarely in the original dataset were not included in our synthetic samples, making a direct overall comparison unfair. Therefore, to ensure a fair comparison, we restricted our evaluation to the shared set of HOIs. For example, we selected the top 15 most frequent HOIs from the synthesized images for Tag A, and then evaluated the original images using this exact same set of 15 interactions. We applied the same procedure to Tags B, C, and D. Although the images themselves differ and cannot be identical, the models showed similar performance patterns across both datasets (Figure 3b). We maintained this controlled setting using shared HOIs in subsequent Performance analysis as well. To ensure rigorous comparisons despite compositional sparsity across combinations, we established unified interaction subsets based on occurrence thresholds (detailed in Appendix F):

- **Two-Tag Combinations Multi-Tags (19 Shared HOIs):** Filtered with an occurrence threshold of ≥ 5 across the 5 Two-Tag Combinations Multi-Tags, resulting in 19 common HOI classes.
- **Multi-Tag Combinations (17 Shared HOIs, 2- to 4-Tag):** Due to the increased compositional sparsity across all 9 combinations (2-, 3-, and 4-way tags), the threshold is set to ≥ 2, establishing a unified subset of 17 common HOIs.

User Study on Visual Realism of Synthesized Images. While quantitative mAP evaluations demonstrate metric parity, verifying whether synthesized images exhibit perceptible artifacts to the human eye is equally critical. To this end, we conducted a double-blind human perceptual study.

As illustrated in Fig. A.9, annotators were presented with image pairs and instructed to identify the synthesized sample(s) among four options: Image 1 Only, Image 2 Only, Both Synthesized, or Both Real. To prevent positional bias, synthesized instances were randomly allocated within each target interaction tag. Across 98 balanced pairs evaluated by two independent annotators, the classification accuracy remained near chance level (47.96% for one annotator and 54.59% on



(a) Direct mAP comparison

(b) Controlled comparison on shared HOIs

Figure 3: Direct vs. Controlled comparison of HOI detection performance across models.

average). This indicates that humans struggle to reliably distinguish our synthesized images from real ones, confirming that the generated samples achieve satisfactory visual fidelity.

Single-Tag Performance. We first examine model performance under individual tags A–D, as shown in Fig. 4. Overall, specialized HOI models tend to outperform LVLMs across the single-tag settings, suggesting that task-specific HOI models retain an advantage in fine-grained human–object interaction recognition and localization. However, the performance differences across tags A–D are relatively modest and are not fully consistent across models. Some models perform better on particular tags than others, and no single tag is uniformly the most challenging across all methods. This suggests that, in the single-tag setting, model-specific characteristics may play a larger role than the tag type itself.

Two-Tag Performance. We next consider two-tag compositions, shown in Fig. 5 and detailed quantitative results reported in Table 1. Moving from single-tag to two-tag compositions generally leads to a noticeable drop in performance, indicating that combining multiple interaction configurations introduces additional difficulty. However, the difficulty varies substantially across different tag combinations rather than being uniform across all two-tag settings. In particular, B+D and C+D tend to be among the most challenging combinations across models, while A+B is generally easier. This suggests that the type of compositional ambiguity matters more than simply having two tags. The relatively poor performance on D-related combinations may reflect the added difficulty of correctly assigning different interactions to different object instances, especially when multiple similar objects or people are present.

Multi-Tag Performance. Finally, we extend the comparison to all two-way, three-way, and four-way tag compositions, as summarized in Fig. 6 and detailed quantitative results reported in Table 2. We observe a clear non-monotonic trend in model performance. Although two-tag compositions are generally more difficult than single-tag cases, adding more tags does not necessarily lead to further performance degradation. In fact, several three-tag and four-tag combinations achieve higher performance than the most difficult two-tag settings, particularly B+D. This suggests that compositional difficulty is not determined solely by the number of tags. Instead, the specific structure of the composition and the type of ambiguity introduced by the tags appear to play a more important role.

Interestingly, the results exhibit a roughly V-shaped trend across tag cardinalities: performance decreases from single-tag to two-tag compositions, but partially recovers for three-tag and four-tag cases. One possible explanation is that additional tags may introduce not only greater compositional complexity, but also more contextual or redundant cues that help the model interpret the scene. In

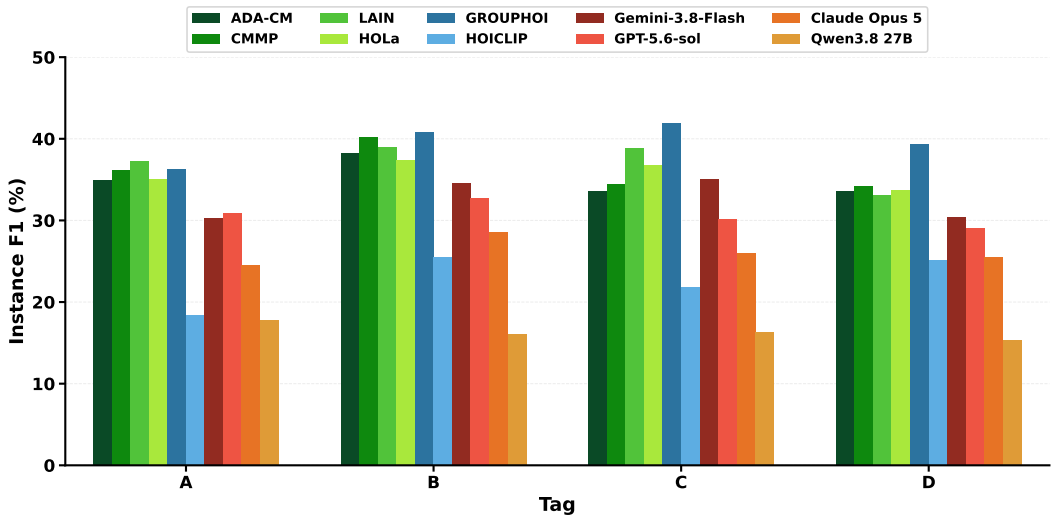


Figure 4: CrossHOI evaluation results on the Original subset across Tags A–D, measured by Instance F1 (%).

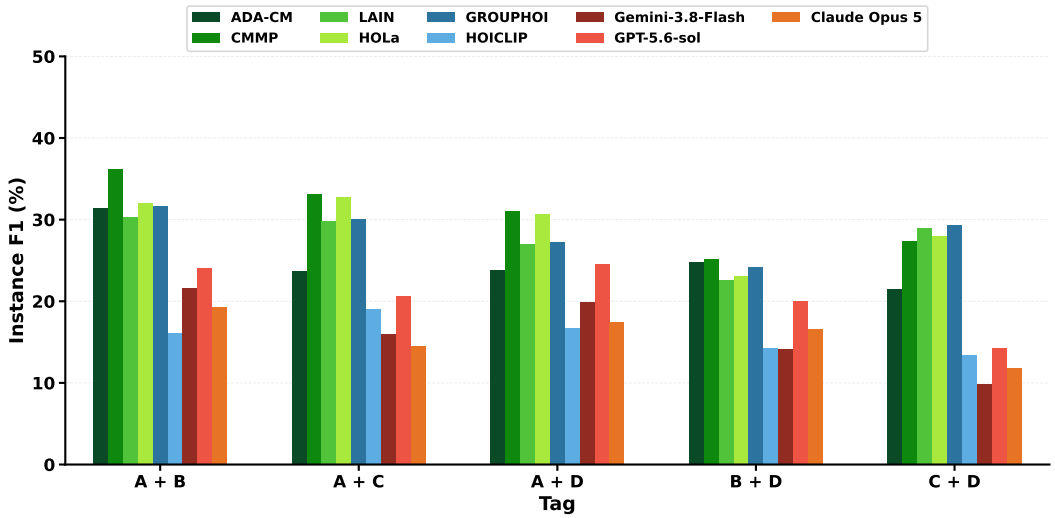


Figure 5: CrossHOI evaluation results on 19 shared HOIs across two-way tag combinations in synthesized images. Performance is reported using Instance F1 (%) for specialized HOI models and LVLMs across the five tag combinations: A+B, A+C, A+D, B+D, and C+D. The detailed list of the 19 shared HOIs is provided in the Appendix.

contrast, some two-tag combinations may create particularly strong ambiguity in interaction attribution or object-instance assignment. Overall, these results suggest that increasing compositional richness does not necessarily make HOI understanding harder; rather, difficulty depends on how different interaction and instance-level ambiguities are combined.

Comparison across two-tag combinations. Table 1 compares models on 19 shared HOI classes across five two-tag combinations. CMMP achieves the highest Instance F1 on four combinations, while GroupHOI performs best on C+D (29.38%). The relative difficulty of the combinations differs across model families: B+D yields the lowest score for CMMP, HOLA, LAIN, and GroupHOI, whereas C+D is the most challenging combination for all three LVLMs. GPT-5.6-Sol performs best among the evaluated LVLMs on every combination, but remains below the strongest specialized detector in each case. The gap is particularly pronounced on C+D, where GroupHOI exceeds GPT-

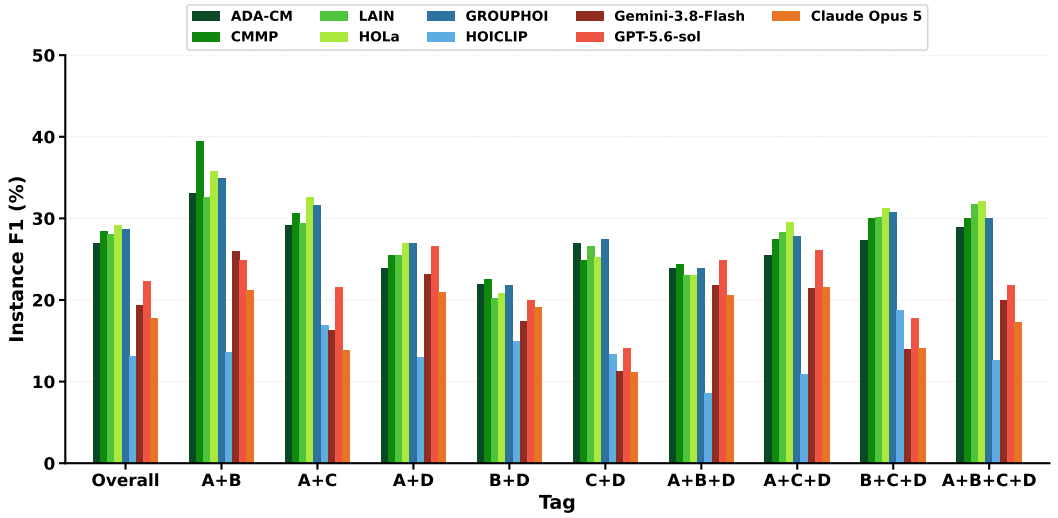


Figure 6: CrossHOI evaluation results on 17 shared HOIs across all multi-tag combinations in synthesized images. Performance is reported using Instance F1 (%) for specialized HOI models and LVLMs across two-way, three-way, and four-way tag compositions. The detailed list of the 17 shared HOIs is provided in the Appendix.

Table 1: CrossHOI evaluation results: **Instance F1 (%)** on 19 shared HOIs across two-way tag combinations in synthesized images.

Model	**A + B**	**A + C**	**A + D**	**B + D**	**C + D**
Specialized HOI Models					
ADA-CM	31.44	23.74	23.86	24.84	21.47
CMMP	36.16	33.19	31.08	25.17	27.35
HOLA	32.00	32.80	30.74	23.04	27.99
LAIN	30.38	29.82	27.07	22.55	29.03
HOICLIP	16.08	19.00	16.72	14.24	13.40
GroupHOI	31.68	30.13	27.28	24.20	29.38
Vision-Language Foundation Models (LVLMs)					
GPT-5.6-Sol	24.07	20.60	24.57	20.00	14.30
Claude-Opus-5	19.34	14.53	17.49	16.59	11.77
Gemini-3.8-Flash	21.68	15.94	19.91	14.18	11.43

Table 2: CrossHOI evaluation results: **Instance F1 (%)** on 17 shared HOIs across all multi-tag combinations in synthesized images.

Model	**Overall**	**2-Tag Combinations**					**3-Tag Combinations**			**4-Tag**
		A+B	**A+C**	**A+D**	**B+D**	**C+D**	**A+B+D**	**A+C+D**	**B+C+D**	**A+B+C+D**
Specialized HOI Models										
ADA-CM	26.99	33.06	29.17	23.86	21.87	26.97	23.94	25.47	27.34	28.95
CMMP	28.45	39.40	30.64	25.51	22.55	24.84	24.34	27.40	30.02	30.06
HOLA	29.10	35.72	32.59	26.92	20.88	25.24	23.06	29.51	31.23	32.12
LAIN	28.01	32.53	29.46	25.51	20.23	26.60	23.04	28.29	30.08	31.74
HOICLIP	13.14	13.60	16.85	12.97	14.88	13.39	8.56	10.89	18.77	12.57
GroupHOI	28.62	34.96	31.55	26.89	21.84	27.39	23.88	27.78	30.69	29.99
Vision-Language Foundation Models (LVLMs)										
GPT-5.6-Sol	22.34	24.86	21.61	26.56	19.93	14.08	24.90	26.04	17.76	21.77
Claude-Opus-5	17.78	21.13	13.78	20.89	19.14	11.09	20.60	21.61	14.05	17.27
Gemini-3.8-Flash	19.39	25.93	16.30	23.13	17.36	11.28	21.82	21.38	13.99	19.98

5.6-Sol by 15.08 percentage points. These results highlight differences in how the evaluated models handle particular interaction compositions, even when the HOI class vocabulary is shared.

Comparison across two-, three-, and four-tag combinations. Table 2 extends the comparison to nine combinations using 17 shared HOI classes. HOLA obtains the highest reported overall Instance F1 (29.10%), while GPT-5.6-Sol leads the LVLMs overall (22.34%). However, model rankings vary across combinations: GPT-5.6-Sol achieves the highest score on A+B+D (24.90%), slightly exceeding CMMP (24.34%). Increasing the number of tags does not produce a monotonic decrease in performance. For example, HOLA scores 20.88% on B+D, 31.23% on B+C+D, and 32.12% on A+B+C+D. Similarly, all three LVLMs score higher on A+C+D than on C+D. Thus, tag count alone does not consistently predict difficulty in these evaluated subsets. Although the shared-class protocol controls the HOI vocabulary, class frequencies and other image characteristics may still differ, so these comparisons do not isolate the causal effect of adding a tag.

5 Limitations and Future Work

Our evaluation uses shared HOI classes to improve comparability across tag combinations. Requiring sufficient support in every combination narrows the evaluation vocabulary relative to HICO-DET's 600 classes, limiting coverage of rare interactions and classes unique to particular configurations. Matching class identities also does not ensure matched frequencies or visual characteristics. Observed performance differences therefore reflect associations with tag composition rather than an isolated causal effect of adding a tag. Future work could expand coverage through targeted synthesis and assess sensitivity to class-support thresholds and distribution matching.

Images with the same tags can differ in human and object scale, spatial arrangement, occlusion, and background clutter, all of which may affect recognition. Tag count therefore should not be interpreted as a universal measure of difficulty, and the observed non-monotonic trends may not generalize to other distributions. Relative object size alone is also insufficient to establish occlusion. Detailed visibility annotations and comparisons between matched scenes would help distinguish interaction structure from visual complexity. Larger evaluations of three- and four-tag configurations would strengthen conclusions about higher-order composition.

Synthetic generation and model-assisted annotation may introduce artifacts, selection effects, and model-family biases. Retaining source annotations, incorporating Grounding DINO, and conducting human review provide complementary checks, but cannot guarantee that all errors or biases are removed. Evaluating models from a family also used during dataset construction therefore warrants caution. Human verification limits dataset scale, while externally hosted models may change over time. Future work could combine multiple generation and annotation model families, conduct independent annotation audits, and improve human verification efficiency to increase diversity, reliability, and scalability.

6 Conclusion

In this work, we study fine-grained compositional evaluation for human–object interaction understanding in multi-person scenes. We introduce HICO-DET-SYN, a controlled synthetic benchmark that expands the coverage of underrepresented multi-tag interaction configurations while preserving the context and existing interactions of real HICO-DET images. Using this benchmark, we systematically compare specialized HOI detectors and general-purpose vision-language models across single- and multi-tag scenes. Our analysis shows that specialized HOI models generally retain an advantage over LVLMs, while compositional difficulty does not increase monotonically with the number of coexisting interaction patterns. Instead, performance exhibits a non-monotonic trend, indicating that the specific structure of interaction composition plays a more important role than tag cardinality alone. We hope HICO-DET-SYN provides a useful foundation for developing HOI systems that are not only accurate on average, but also robust to the diverse compositional structures encountered in real-world multi-person scenes.

AI USE STATEMENT

We used generative AI tools to generate candidate HOI images through controlled editing of source images, assist with annotation, and perform automated validation and filtering using vision-language models. We also used AI assistants for code development and debugging, experimental workflow support, research planning and methodological feedback, data analysis and interpretation support, literature discovery, and manuscript drafting and editing. The authors reviewed AI-assisted outputs and retained responsibility for dataset selection, human validation, experimental design, and interpretation of results. We take full responsibility for the final manuscript, code, data, and scientific claims.

ETHICS STATEMENT

Our dataset is derived from HICO-DET images through controlled image editing and is intended for research on human-object interaction understanding. Automated checks and human review assess image plausibility and annotation quality, but do not eliminate potential biases inherited from the source dataset or the models used in generation and annotation. For human validation and the perceptual user study, participants were informed of the study's purpose and consented to the use of their annotations for research. Dataset use and release are subject to the applicable source-data licenses and model-provider terms.

REPRODUCIBILITY STATEMENT

Section 3 describes our dataset construction pipeline, including prompt generation, automated validation, and human verification. Our evaluation protocols specify the shared HOI class sets, minimum per-class support criteria, model settings, and metrics used for each comparison, including mAP, Instance F1, Macro F1, and Exact Match. Upon acceptance, we will release the HICO-DET-SYN dataset and annotations, subject to applicable redistribution terms, together with evaluation scripts and configuration files. Because the pipeline uses externally hosted generative models, exact regeneration may vary with sampling and service updates; the released dataset will provide a fixed basis for evaluation.

REFERENCES

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.

Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Pei Geng, Jian Yang, and Shanshan Zhang. Horp: Human-object relation priors guided hoi detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25325–25335. IEEE, 2025.

Rohit Girdhar and Deva Ramanan. Cater: A diagnostic dataset for compositional actions and temporal reasoning. *arXiv preprint arXiv:1910.04744*, 2019.

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.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3749–3761, 2022.

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

Yupeng Hu, Changxing Ding, Chang Sun, Shaoli Huang, and Xiangmin Xu. Bilateral collaboration with large vision-language models for open vocabulary human-object interaction detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20126–20136. IEEE, 2025.

Huaizu Jiang, Xiaojian Ma, Weili Nie, Zhiding Yu, Yuke Zhu, and Anima Anandkumar. Bongard-hoi: Benchmarking few-shot visual reasoning for human-object interactions. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19034–19043. IEEE, 2022.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Qinqian Lei, Bo Wang, and Robby T. Tan. Hola: Zero-shot HOI detection with low-rank decomposed VLM feature adaptation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 1825–1835. IEEE, 2025. doi: 10.1109/ICCV51701.2025.00178. URL https://doi.org/10.1109/ICCV51701.2025.00178.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38520–38531, 2026.

Ting Lei, Shaofeng Yin, Yuxin Peng, and Yang Liu. Exploring conditional multi-modal prompts for zero-shot hoi detection. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2024.

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, 2023a.

Liulei Li, Jianan Wei, Wenguan Wang, and Yi Yang. Neural-logic human-object interaction detection. *Advances in Neural Information Processing Systems*, 36:21158–21171, 2023b.

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14963–14973, 2023c.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024b.

Jinguo Luo, Weihong Ren, Quanlong Zheng, Yanhao Zhang, Zhenlong Yuan, Zhiyong Wang, Haonan Lu, and Honghai Liu. Instructhoi: Context-aware instruction for multi-modal reasoning in human-object interaction detection. *Advances in Neural Information Processing Systems*, 38: 142558–142577, 2026.

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.

Maja Noack, Qinqian Lei, Taipeng Tian, Bihan Dong, Robby T Tan, Yixin Chen, John Young, Saijun Zhang, and Bo Wang. Shoe: Semantic hoi open-vocabulary evaluation metric. *arXiv preprint arXiv:2604.01586*, 2026.

OpenAI. Gpt-5.6: Frontier intelligence that scales with your ambition. `https://openai.com/index/gpt-5-6/`, July 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.

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. *arXiv preprint arXiv:2604.02071*, 2026.

Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10405–10414. IEEE, 2021.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Guangzhi Wang, Yangyang Guo, Ziwei Xu, and Mohan Kankanhalli. Bilateral adaptation for human-object interaction detection with occlusion-robustness. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27970–27980. IEEE, 2024.

Lemeng Wang, Qinqian Lei, Vidhi Bakshi, Daniel Yi, Yifan Liu, Jiacheng Hou, Asher Seng Hao, Zheda Mai, Wei-Lun Chao, Robby T Tan, et al. A study of failure modes in two-stage human-object interaction detection. *arXiv preprint arXiv:2604.13448*, 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.

Jie Yang, Bingliang Li, Ailing Zeng, Lei Zhang, and Ruimao Zhang. Open-world human-object interaction detection via multi-modal prompts. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16954–16964. IEEE, 2024.

Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Musslick, and Thomas Serre. A benchmark for compositional visual reasoning. *Advances in neural information processing systems*, 35: 29776–29788, 2022.

Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human–object interactions. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13299–13307. IEEE, 2021.

Appendix

A Comprehensive Dataset Statistics across Various Tags

Table A.1: Complete dataset statistics across single tags (Tags A–F) and multi-tag combinations, with counts of original real images, synthesized images.

Configuration	Real Images	Synthesized Images	Total Images
Single Tags			
Tag A	513	271	784
Tag B	303	237	540
Tag C	621	277	898
Tag D	146	256	402
Tag E	1	13	14
Tag F	9	34	43
2-Tag Combinations			
Tag A + B	184	258	442
Tag A + C	44	273	317
Tag A + D	17	302	319
Tag A + E	1	0	1
Tag A + F	6	0	6
Tag B + D	10	210	220
Tag B + F	2	0	2
Tag C + D	91	231	322
Tag C + E	1	0	1
Tag C + F	6	0	6
3-Tag Combinations			
Tag A + B + D	6	152	158
Tag A + B + F	3	0	3
Tag A + C + D	17	118	135
Tag A + C + F	8	0	8
Tag B + C + D	35	201	236
Tag C + D + F	1	0	1
Tag C + E + F	1	0	1
4-Tag Combinations			
Tag A + B + C + D	35	187	222
Tag A + B + C + F	1	0	1
Tag A + C + D + F	1	0	1
Tag B + C + D + F	1	0	1
Tag B + C + E + F	1	0	1
Tag C + D + E + F	1	0	1
5-Tag Combinations			
Tag A + B + C + D + F	1	0	1
Total	**2,075**	**3,019**	**5,094**

B Source Image Pre-filtering Details and Artifact Examples

When a source image has limited empty space, inserting new persons or objects forces the generative model to place them unnaturally over existing subjects or in physically impossible locations. Figure A.1 shows an example of such artifacts when crowded images are edited without pre-filtering. Enforcing the 50% bounding-box area constraint effectively prevents these structural failures before passing images to the synthesis pipeline.



(a) Original Image (b) Synthesis Artifact (Without Filtering)

Figure A.1: Example of synthesis failure caused by dense foreground objects. Crowded scenes without sufficient empty space lead to unnatural object overlaps.

C Prompt Templates for Multi-Agent Synthesis and Validation Pipeline

This section details the system prompt templates and execution schemes employed by the synthesis and validation agents.

C.1 Image Synthesis Agents

Synthesis Agent 1 (Scene Analysis): Parses existing GT HOI pairs and extracts scene geometry, background context, lighting, and camera viewpoint.

Scene Analysis Prompt Template

The image file is '{`selected_img_name`}' with Current HOI Tag '{`current_tag`}'. Ground Truth HOI object/action pairs: {`gt_pairs_str`}. Analyze the image and describe the scene in detail.

Synthesis Agent 2 (Feasibility Generation): Based on current image description and target HOI list, generates plausible text prompts.

Feasibility Generation Prompt Template

Analyze the input image, image description, current HOI interactions, and candidate (`added_object, dataset_action_combinations`) pairs grouped by target tag.
Refer to the real-world HOI action combination statistics derived from the HICO-DET dataset (`action_combinations_analysis.json`) provided in the user prompt.
Generate up to 5 plausible, contextually realistic, and physically achievable action lists (`action_list: ["act1", "act2", ...]`) for each candidate object and target tag, considering the scene environment, background, lighting, and camera viewpoint.

Synthesis Agent 3 (Candidate Selection): Evaluates the scene context to select candidate target object–action pairs produced by Agent 2.

Candidate Selection Prompt Template

Given the input image '{selected_img_name}', scene description, ground-truth HOI annotations, and feasible action lists proposed by Agent 2:

1. Maintain camera angle, viewpoint, and all existing human/object positions strictly unchanged.
2. Select all viable (added_object, added_action_list) pairs that strictly satisfy the target Tag rules.

Synthesis Agent 4 (Edit Prompt Generation): Converts each selected candidate pair into an explicit instruction prompt for image synthesis.

Edit Prompt Generation Template

The input image is '{selected_img_name}' with original primary HOI ('{original_object}', '{original_action}'). Below are candidate pairs for Multi-Tag synthesis: {flattened_candidates}. Generate a precise image edit instruction for each candidate pair under the specified constraints.

Synthesis Agent 5 (Prompt Evaluation & Selection): Evaluates all candidate prompts to filter out physical contradictions and retain only high-quality prompts for target tag formation.

Prompt Evaluation Prompt Template

Evaluate the generated candidate prompts: {generated_case_prompts}. Select only the optimal prompts that satisfy all criteria.
Return the evaluated summary and the list of approved prompts.

Synthesis Agent 6 (Image Generation Execution): Feeds the original image reference and approved edit prompt into the generative model to produce the final edited image.

Image Generation Execution

Input: Original image reference (data_uri) + Approved edit prompt (edit_prompt)
Model: GPT-Image-2
Output: Synthesized image with newly added person-object interaction completing the target multi-tag.

C.2 Multi-Agent Validation Stage

Validation Agent 1 (Camera Angle & Scene Consistency): Compares the synthesized image with the original image to verify that viewpoint, perspective, and background elements remain unchanged.

Stage 1 Validation Prompt Template

Compare the original image and the synthesized edited image side-by-side:

1. **Camera Viewpoint**: Verify that camera angle (pitch, yaw, roll), framing, and perspective have not shifted.
2. **Original Scene Preservation**: Confirm that all pre-existing persons, objects, poses, and backgrounds are faithfully preserved in their exact locations.

Output: is_angle_and_view_preserved (bool), is_original_image_preserved (bool). If failed, the image is immediately dropped.

Validation Agent 2 (Naturalness & Tag Verification): Evaluates visual naturalness and confirms whether the newly added interaction fulfills the intended target tag.

Stage 2 Validation Prompt Template

Evaluate the synthesized image for:

1. **Visual Naturalness**: Reject severe physical distortions, floating limbs, or impossible kinematics (e.g., standing on water, dual-riding single-person boards).
2. **Action Execution**: Verify that the newly added person successfully performs the specified `added_action_list`.
3. **Tag Satisfaction**: Confirm that the full scene satisfies target tag criteria.

Output: `is_natural` (bool), `naturalness_score` (1–10), `is_action_list_valid` (bool), `is_tag_valid` (bool).

Validation Agent 3 (Bounding Box Localization): Detects normalized 2D bounding boxes for newly synthesized instances, refined using Grounding DINO to construct ground-truth annotations.

Stage 3 Localization Prompt Template

Identify and localize the newly added entities in the synthesized image:

- **Added Person Box**: Detect 2D coordinates $[y_{\min}, x_{\min}, y_{\max}, x_{\max}]$ on a 0–1000 scale for the added actor.
- **Added Object Box**: Detect 2D coordinates for the added object instance (if applicable).

D Detection Score Threshold Calibration for 1-Stage and 2-Stage Models

To ensure fair evaluation across different architectures, we analyze the sensitivity of evaluation performance to detection score thresholds ($S_{\text{thresh}} \in [0.1, 0.9]$). As summarized in Table A.2, the optimal threshold diverges substantially depending on the model paradigm:

- **1-Stage Models (GroupHOI, HOICLIP):** In these architectures, both the object detection confidence (s_{obj}) and the interaction prediction confidence (s_{hoi}) are individually normalized to the range $[0, 1]$. The final interaction score is computed using their geometric mean ($s = \sqrt{s_{\text{obj}} \cdot s_{\text{hoi}}}$) or joint probability product, resulting in a well-calibrated score strictly bounded within $[0, 1]$. Because 1-stage models generate dense candidate triplets across the image, low thresholds (0.1–0.3) lead to a sharp drop in precision due to false positive associations. As shown in the stage-level average, performance steadily climbs and peaks at $S_{\text{thresh}} = 0.7$ (0.0881), where GroupHOI achieves its peak performance (0.1659) and HOICLIP reaches near-optimal performance (0.0103, slightly peaking at 0.8 with 0.0117). We therefore set $S_{\text{thresh}} = 0.7$ as the unified evaluation threshold for 1-stage models.
- **2-Stage Models (ADA-CM, CMMP, HOLA, LAIN):** These pipelines seperate human-object proposal detection from relation classification. Due to cascaded thresholding and proposal filtering in the primary stage, downstream verb confidence scores are inherently conservative. Consequently, higher thresholds severely penalize recall, leading to monotonic degradation as threshold increases (average score drops from 0.1419 at 0.1 to 0.0076 at 0.9)[cite: 1]. All four 2-stage models peak at $S_{\text{thresh}} = 0.1$ (ADA-CM: 0.1023, CMMP: 0.1576, HOLA: 0.1619, LAIN: 0.1456)[cite: 1]. Thus, $S_{\text{thresh}} = 0.1$ is adopted as the canonical operational point.

Applying a unified threshold across paradigms would artificially degrade either the precision of 1-stage models or the recall of 2-stage models. Selecting paradigm-specific optimal thresholds (0.7 and 0.1) reflects each architecture's best operational capability under the CrossHOI protocol.

E Sample Cases for Human Person Validation

During the final human validation stage, independent annotators inspect the consistency of the original subjects and verify the precision of the newly added person and interaction annotations. Fig-

Table A.2: Score threshold sensitivity across 1-Stage and 2-Stage models. Optimal values are highlighted in **bold**.

Stage	Model	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1-Stage	GroupHOI	0.0046	0.0060	0.0154	0.0501	0.1042	0.1435	**0.1659**	0.1049	0.0366
	HOICLIP	0.0041	0.0041	0.0041	0.0050	0.0058	0.0063	0.0103	**0.0117**	0.0075
	Average	0.0044	0.0051	0.0098	0.0276	0.0550	0.0749	**0.0881**	0.0583	0.0221
2-Stage	ADA-CM	**0.1023**	0.0893	0.0633	0.0802	0.0686	0.0593	0.0495	0.0212	0.0127
	CMMP	**0.1576**	0.1461	0.1244	0.1003	0.0768	0.0548	0.0344	0.0184	0.0064
	HOLA	**0.1619**	0.1539	0.1382	0.1187	0.0905	0.0645	0.0405	0.0203	0.0064
	LAIN	**0.1456**	0.1424	0.1247	0.1067	0.0787	0.0527	0.0312	0.0129	0.0050
	Average	**0.1419**	0.1329	0.1127	0.1015	0.0787	0.0578	0.0389	0.0182	0.0076

ure A.2 shows qualitative verification examples presented to annotators, comparing the original scenes with the synthesized pairs and their corresponding localized bounding boxes.



(a) Validation Sample 28 (b) Validation Sample 306

Figure A.2: Sample inspection cases presented during the final human validation stage. Annotators evaluate whether pre-existing subjects remain intact and ensure that newly added persons and bounding-box coordinates accurately reflect the target interaction.

F SHARED HOI DETAILS

To ensure consistent and fair comparisons across varying multi-tag settings, we define unified subsets of shared HOI classes based on occurrence frequency thresholds:

- **Two-Tag Combinations (19 Shared HOIs):** Filtered with an occurrence threshold of ≥ 5 across the 5 pairwise tag combinations (A+B, A+C, A+D, B+D, and C+D). The 19 retained interaction classes are listed in Table A.3.
- **Multi-Tag Combinations (17 Shared HOIs):** Due to sparsity across all 9 configurations spanning 2-, 3-, and 4-tag compositions, the occurrence threshold is set to ≥ 2. The resulting 17 unified interaction classes are listed in Table A.4.

Table A.3: List of the 19 shared HOIs evaluated across 5 two-way multi-tag combinations (Threshold ≥ 5).

HOI ID	HOI Name	Object	Action
HOI_012	inspect bicycle	bicycle	inspect
HOI_017	repair bicycle	bicycle	repair
HOI_019	sit_on bicycle	bicycle	sit_on
HOI_020	straddle bicycle	bicycle	straddle
HOI_021	walk bicycle	bicycle	walk
HOI_022	wash bicycle	bicycle	wash
HOI_039	row boat	boat	row
HOI_132	hug horse	horse	hug
HOI_140	run horse	horse	run
HOI_141	straddle horse	horse	straddle
HOI_142	train horse	horse	train
HOI_147	inspect motorcycle	motorcycle	inspect
HOI_154	sit_on motorcycle	motorcycle	sit_on
HOI_155	straddle motorcycle	motorcycle	straddle
HOI_156	turn motorcycle	motorcycle	turn
HOI_371	hold kite	kite	hold
HOI_372	inspect kite	kite	inspect
HOI_517	drag surfboard	surfboard	drag
HOI_519	inspect surfboard	surfboard	inspect

Table A.4: List of the 17 shared HOIs evaluated across 9 multi-tag combinations (Threshold ≥ 2).

HOI ID	HOI Name	Object	Action
HOI_012	inspect bicycle	bicycle	inspect
HOI_017	repair bicycle	bicycle	repair
HOI_019	sit_on bicycle	bicycle	sit_on
HOI_020	straddle bicycle	bicycle	straddle
HOI_021	walk bicycle	bicycle	walk
HOI_022	wash bicycle	bicycle	wash
HOI_039	row boat	boat	row
HOI_042	stand_on boat	boat	stand_on
HOI_043	tie boat	boat	tie
HOI_147	inspect motorcycle	motorcycle	inspect
HOI_154	sit_on motorcycle	motorcycle	sit_on
HOI_155	straddle motorcycle	motorcycle	straddle
HOI_156	turn motorcycle	motorcycle	turn
HOI_339	spin frisbee	frisbee	spin
HOI_370	fly kite	kite	fly
HOI_372	inspect kite	kite	inspect
HOI_373	launch kite	kite	launch

G ADDITIONAL CROSSHOI EVALUATION RESULTS.

We provide the complete CrossHOI evaluation results under three metrics—Instance F1, Macro F1, and Exact Match—across three evaluation settings.

For the **single-tag setting**, Figures 4, A.4, and A.3 visualize the Instance F1, Macro F1, and Exact Match results, respectively, on the Original subset across Tags A–D. The corresponding numerical results are reported in Tables A.5, A.6, and A.7.

For the **two-tag setting**, Figures 5, A.6, and A.5 show the corresponding results across the five two-way tag combinations in synthesized images. The exact numerical values are provided in Tables 1, A.8, and A.9.

Finally, for the **full multi-tag setting**, Figures 6, A.8, and A.7 present results across all two-way, three-way, and four-way tag compositions. Their corresponding numerical results are reported in Tables 2, A.11, and A.12.

Overall, the Macro F1 and Exact Match results exhibit trends broadly consistent with those observed under Instance F1, with performance varying substantially across different tag compositions and model families.

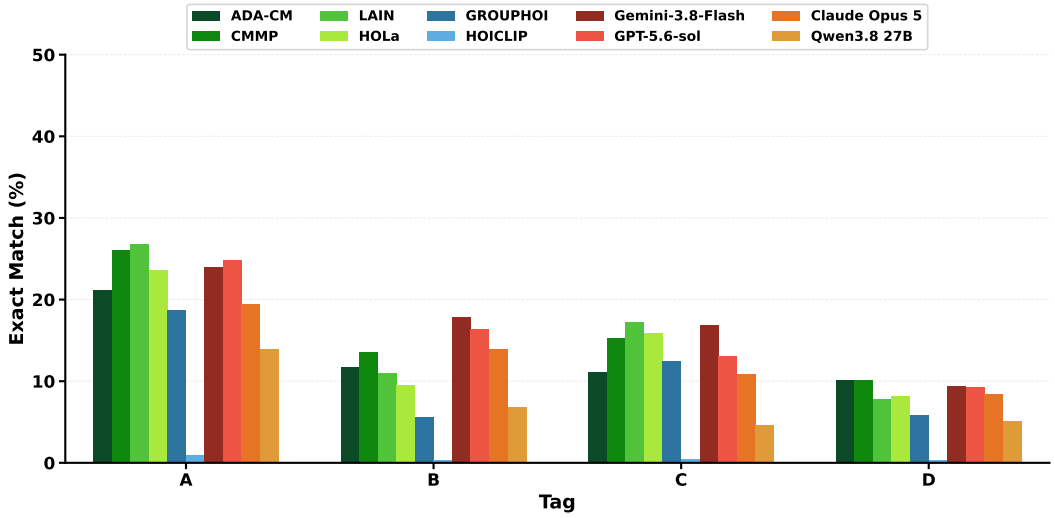


Figure A.3: CrossHOI evaluation results on the Original subset across individual Tags A–D, measured by Exact Match (%). Performance is reported for specialized HOI models and LVLMs.

Table A.5: CrossHOI evaluation results: **Instance F1 (%)** on the Original subset across Tags A–D.

Model	A	B	C	D
Specialized HOI Models				
ADA-CM	34.91	38.21	33.56	33.64
CMMP	36.17	40.25	34.46	34.17
LAIN	37.29	39.02	38.79	33.12
HOLa	35.06	37.42	36.79	33.71
GROUPHOI	36.33	40.81	41.90	39.38
HOICLIP	18.40	25.44	21.86	25.11
Vision-Language Foundation Models (LVLMs)				
Gemini-3.8-Flash	30.22	34.57	35.02	30.45
GPT-5.6-sol	30.90	32.67	30.21	29.03
Claude Opus 5	24.48	28.60	26.00	25.46
Qwen3.8_27B	17.77	16.07	16.27	15.38

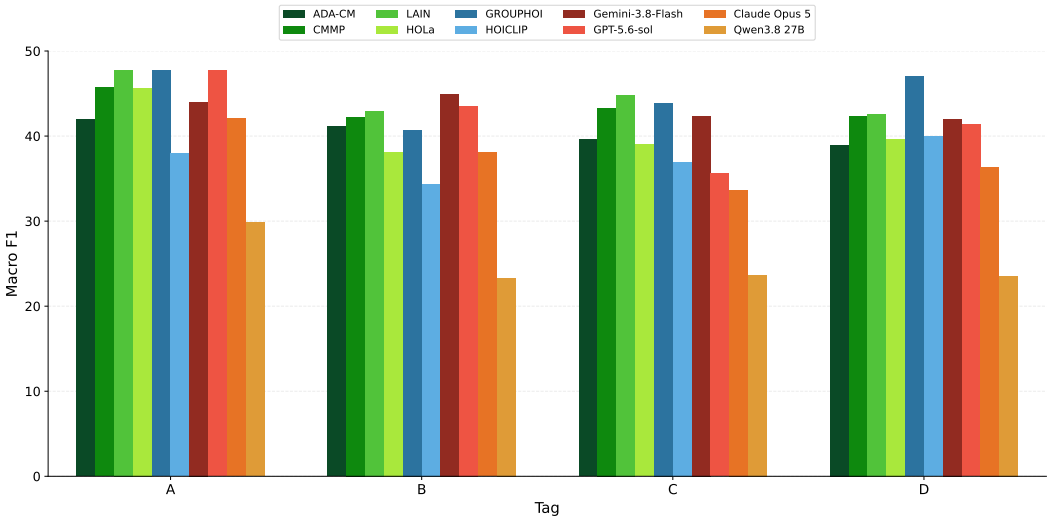


Figure A.4: CrossHOI evaluation results on the Original subset across individual Tags A–D, measured by Macro F1 (%). Performance is reported for specialized HOI models and LVLMs.

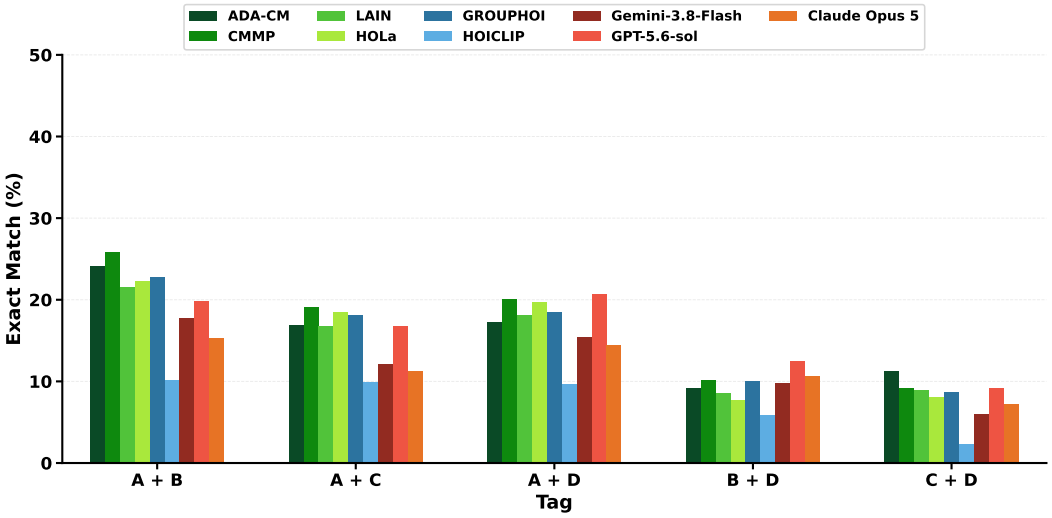


Figure A.5: CrossHOI evaluation results on 19 shared HOIs across two-way tag combinations in synthesized images, measured by Exact Match (%). Results are reported for A+B, A+C, A+D, B+D, and C+D.

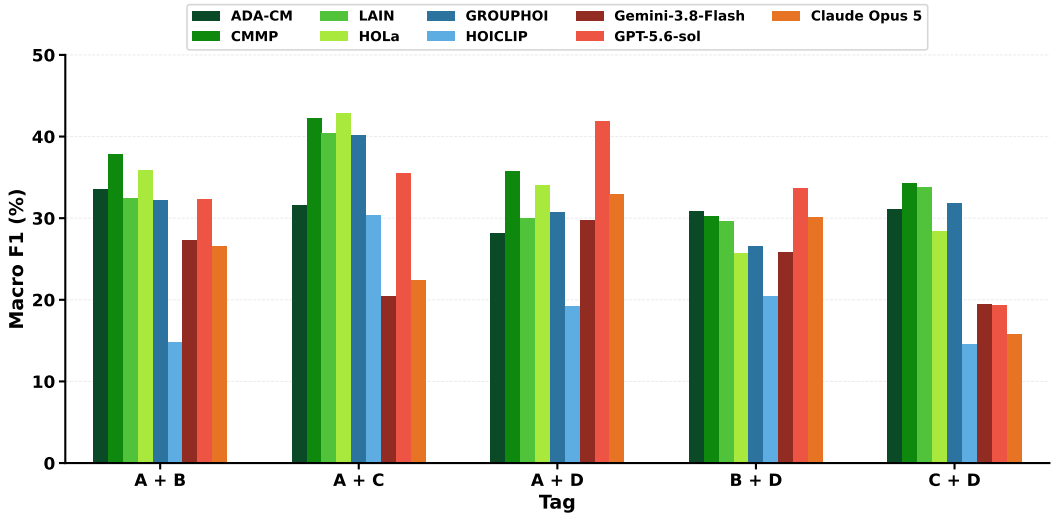


Figure A.6: CrossHOI evaluation results on 19 shared HOIs across two-way tag combinations in synthesized images, measured by Macro F1 (%). Results are reported for A+B, A+C, A+D, B+D, and C+D.

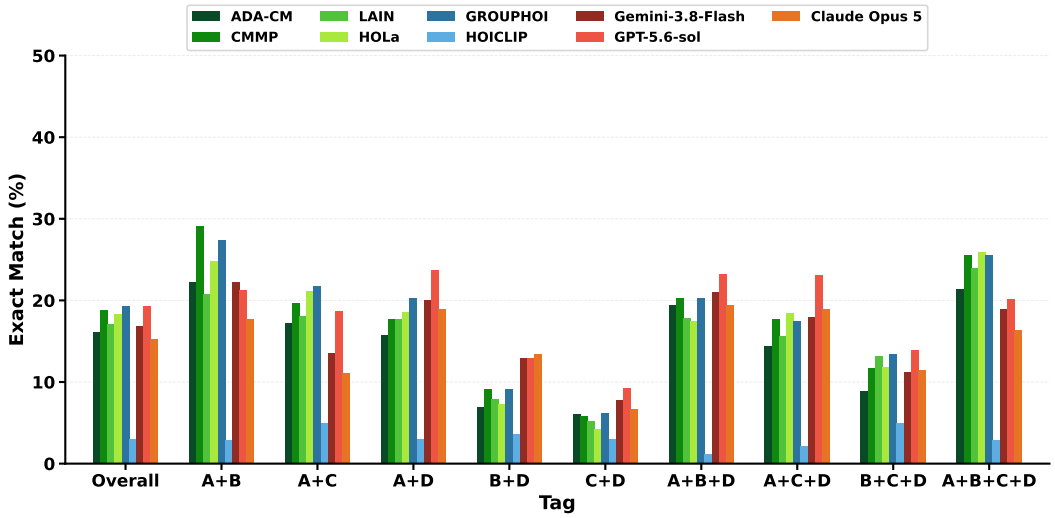


Figure A.7: CrossHOI evaluation results on 17 shared HOIs across all multi-tag combinations in synthesized images, measured by Exact Match (%). Results include two-way, three-way, and four-way tag compositions.

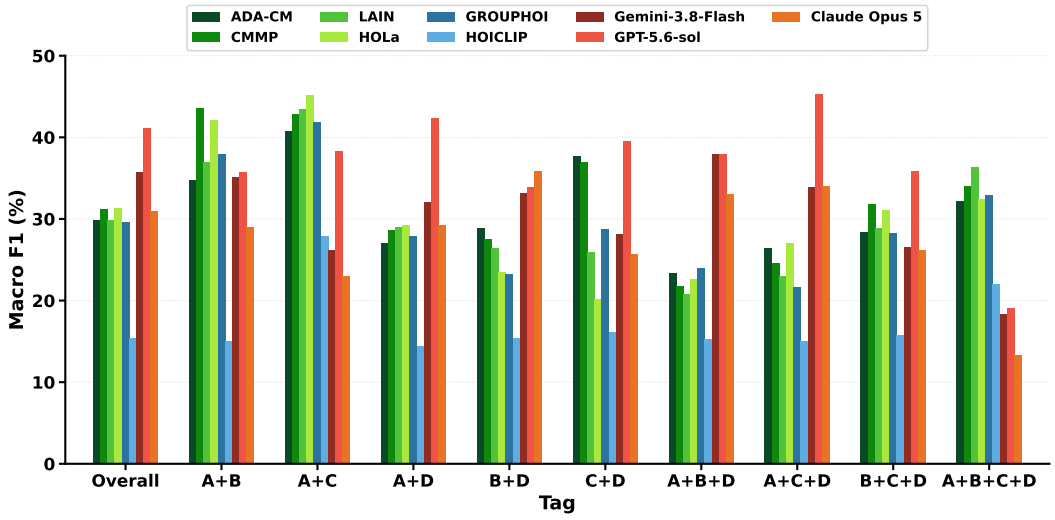


Figure A.8: CrossHOI evaluation results on 17 shared HOIs across all multi-tag combinations in synthesized images, measured by Macro F1 (%). Results include two-way, three-way, and four-way tag compositions.

Table A.6: CrossHOI evaluation results: **Macro F1 (%)** on the Original subset across Tags A–D.

Model	**A**	**B**	**C**	**D**
Specialized HOI Models				
ADA-CM	41.95	41.14	39.61	38.93
CMMP	45.69	42.23	43.28	42.33
LAIN	47.75	42.92	44.83	42.58
HOLa	45.67	38.12	39.02	39.63
GROUPHOI	47.70	40.69	43.88	47.02
HOICLIP	37.93	34.35	36.93	39.97
Vision-Language Foundation Models (LVLMs)				
Gemini-3.8-Flash	44.00	44.98	42.35	41.99
GPT-5.6-sol	47.72	43.51	35.62	41.38
Claude Opus 5	42.08	38.08	33.68	36.37
Qwen3.8_27B	29.83	23.30	23.70	23.56

Table A.7: CrossHOI evaluation results: **Exact Match (%)** on the Original subset across Tags A–D.

Model	**A**	**B**	**C**	**D**
Specialized HOI Models				
ADA-CM	21.15	11.65	11.09	10.06
CMMP	26.02	13.51	15.30	10.06
LAIN	26.80	11.02	17.17	7.79
HOLa	23.59	9.47	15.91	8.12
GROUPHOI	18.72	5.59	12.48	5.84
HOICLIP	0.89	0.31	0.48	0.32
Vision-Language Foundation Models (LVLMs)				
Gemini-3.8-Flash	23.92	17.77	16.80	9.43
GPT-5.6-sol	24.82	16.31	13.00	9.30
Claude Opus 5	19.45	13.85	10.79	8.44
Qwen3.8_27B	13.87	6.86	4.58	5.14

Table A.8: CrossHOI evaluation results: **Macro F1 (%)** on 19 shared HOIs across two-way tag combinations in synthesized images.

Model	A + B	A + C	A + D	B + D	C + D
Specialized HOI Models					
ADA-CM	33.56	31.60	28.19	30.83	31.05
CMMP	37.75	42.18	35.78	30.17	34.23
HOLA	35.90	42.79	34.07	25.71	28.39
LAIN	32.44	40.33	29.98	29.61	33.80
HOICLIP	14.78	30.30	19.24	20.42	14.57
GroupHOI	32.22	40.15	30.69	26.56	31.79
Vision-Language Foundation Models (LVLMs)					
GPT-5.6-Sol	32.25	35.49	41.88	33.68	19.30
Claude-Opus-5	26.59	22.35	32.86	30.09	15.79
Gemini-3.8-Flash	27.29	20.42	29.74	25.77	16.51

Table A.9: CrossHOI evaluation results: **Exact Match (%)** on 19 shared HOIs across two-way tag combinations in synthesized images.

Model	A + B	A + C	A + D	B + D	C + D
Specialized HOI Models					
ADA-CM	24.03	16.89	17.20	9.19	11.16
CMMP	25.84	19.01	20.07	10.07	9.09
HOLA	22.18	18.46	19.71	7.72	7.99
LAIN	21.53	16.70	18.04	8.49	8.85
HOICLIP	10.13	9.82	9.60	5.83	2.22
GroupHOI	22.75	18.12	18.48	10.04	8.66
Vision-Language Foundation Models (LVLMs)					
GPT-5.6-Sol	19.78	16.75	20.65	12.38	9.18
Claude-Opus-5	15.22	11.19	14.38	10.63	7.14
Gemini-3.8-Flash	17.75	12.05	15.44	9.79	7.63

Table A.10: Overall performance summary on 19 shared HOIs (Overall 5 Two-Way Tags).

Model	Stage	Threshold	Instance F1 (%)	Macro F1 (%)	EM (%)
Specialized HOI Models					
ADA-CM	2-Stage	0.1	24.98	30.66	16.35
CMMP	2-Stage	0.1	31.09	34.53	17.51
HOLA	2-Stage	0.1	30.04	33.31	16.21
LAIN	2-Stage	0.1	28.27	30.46	15.58
HOICLIP	1-Stage	0.7	16.34	17.74	8.22
GroupHOI	1-Stage	0.7	28.89	31.38	16.36
Vision-Language Foundation Models (LVLMs)					
GPT-5.6-Sol	Prompt	–	20.83	35.45	16.19
Claude-Opus-5	Prompt	–	15.83	27.99	11.83
Gemini-3.8-Flash	Prompt	–	16.92	28.09	12.84

Table A.11: CrossHOI evaluation results: **Macro F1 (%)** on 17 shared HOIs across multi-tag combinations in synthesized images.

Model	Overall	2-Tag Combinations					3-Tag Combinations			4-Tag
		A+B	A+C	A+D	B+D	C+D	A+B+D	A+C+D	B+C+D	A+B+C+D
Specialized HOI Models										
ADA-CM	29.88	34.74	40.71	27.01	28.87	37.66	23.31	26.38	28.38	32.11
CMMP	31.19	43.61	42.80	28.59	27.45	36.98	21.77	24.51	31.81	34.06
HOLA	31.28	42.09	45.13	29.24	23.43	20.13	22.55	26.97	31.06	32.43
LAIN	29.82	36.95	43.40	29.00	26.44	25.91	20.78	22.93	28.88	36.29
HOICLIP	15.38	15.00	27.85	14.43	15.36	16.17	15.20	15.07	15.70	22.01
GroupHOI	29.61	37.88	41.84	27.92	23.21	28.68	23.90	21.57	28.22	32.96
Vision-Language Foundation Models (LVLMs)										
GPT-5.6-Sol	41.08	35.76	38.28	42.33	33.86	39.57	37.89	45.26	35.89	19.00
Claude-Opus-5	30.99	28.93	23.00	29.21	35.88	25.66	32.98	34.00	26.19	13.30
Gemini-3.8-Flash	35.74	35.14	26.22	32.09	33.17	28.15	37.97	33.82	26.50	18.26

Table A.12: CrossHOI evaluation results: **Exact Match (%)** on 17 shared HOIs across multi-tag combinations in synthesized images.

Model	Overall	2-Tag Combinations					3-Tag Combinations			4-Tag
		A+B	A+C	A+D	B+D	C+D	A+B+D	A+C+D	B+C+D	A+B+C+D
Specialized HOI Models										
ADA-CM	16.11	22.20	17.25	15.73	6.96	6.12	19.44	14.41	8.94	21.43
CMMP	18.81	29.07	19.71	17.65	9.18	5.85	20.32	17.71	11.73	25.54
HOLA	18.35	24.78	21.16	18.60	7.27	4.27	17.52	18.50	11.83	25.89
LAIN	17.08	20.80	18.13	17.67	7.85	5.24	17.82	15.66	13.17	23.96
HOICLIP	3.04	2.87	4.98	3.04	3.57	2.96	1.20	2.14	4.92	2.89
GroupHOI	19.30	27.40	21.73	20.23	9.12	6.16	20.23	17.47	13.45	25.56
Vision-Language Foundation Models (LVLMs)										
GPT-5.6-Sol	19.26	21.24	18.66	23.70	12.94	9.30	23.20	23.06	13.95	20.18
Claude-Opus-5	15.29	17.65	11.11	18.89	13.37	6.69	19.46	18.90	11.44	16.40
Gemini-3.8-Flash	16.79	22.19	13.53	20.04	12.95	7.83	21.01	17.94	11.26	18.89

Table A.13: Overall performance summary on 17 shared HOIs across all 9 multi-tags in synthesized images.

Model	Stage	Threshold	Pairs	Instance F1 (%)	Macro F1 (%)	EM (%)
Specialized HOI Models						
ADA-CM	2-Stage	0.1	4,921	26.99	29.88	16.11
CMMP	2-Stage	0.1	4,710	28.45	31.19	18.81
HOLA	2-Stage	0.1	4,742	29.10	31.28	18.35
LAIN	2-Stage	0.1	4,855	28.01	29.82	17.08
HOICLIP	1-Stage	0.7	5,370	13.14	15.38	3.04
GroupHOI	1-Stage	0.7	4,550	28.62	29.61	19.30
Vision-Language Foundation Models (LVLMs)						
GPT-5.6-Sol	Prompt	–	3,599	22.34	41.08	19.26
Claude-Opus-5	Prompt	–	3,598	17.78	30.99	15.29
Gemini-3.8-Flash	Prompt	–	3,436	19.39	35.74	16.79

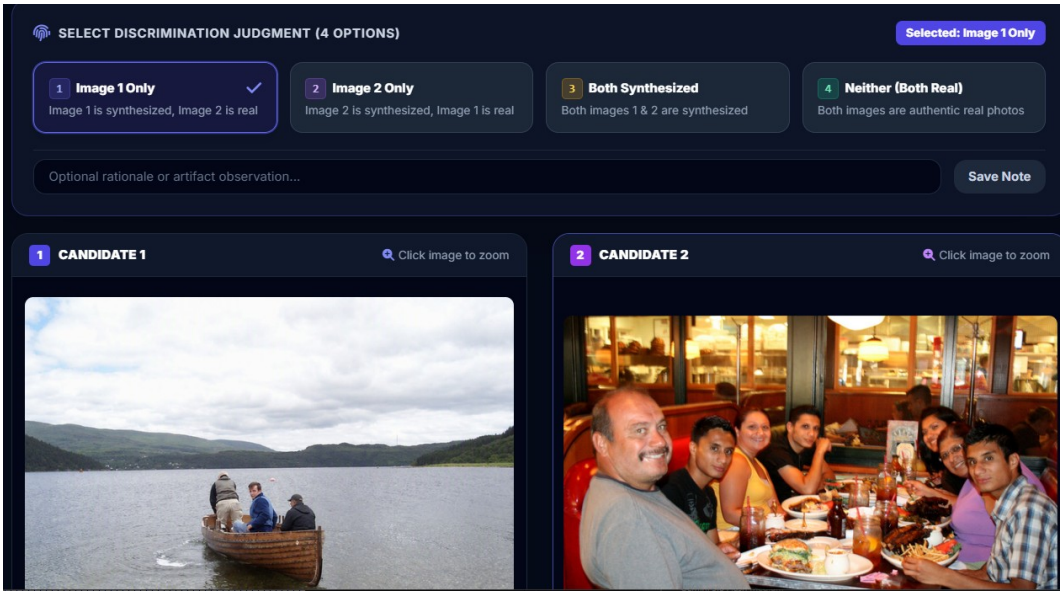


Figure A.9: **Web-based user study interface for visual realism evaluation.** In each trial, annotators are presented with two images from the same interaction tag and tasked with identifying which image is synthesized. Each pair randomly contains 0, 1, or 2 synthesized images across four choices to prevent forced-choice bias. Filenames and annotations are fully masked to maintain a strict double-blind setup.