Supplementary Material for CrossHOI-Bench: A Unified Benchmark for HOI Evaluation across Vision-Language Models and HOI-Specific Methods

Qinqian Lei[1] Bo Wang[2] Robby T. Tan[1,3]
[1]National University of Singapore [2]University of Mississippi
[3]ASUS Intelligent Cloud Services (AICS)
qinqian.lei@u.nus.edu hawk.rsrch@gmail.com robby.tan@nus.edu.sg
https://github.com/ChelsieLei/CrossHOI-Bench

1. Benchmarking Dataset Discussion

1.1. Limitations of the HICO-DET Benchmark

Incomplete Annotation To understand why existing HOI benchmarks fail to provide reliable evaluation for both VLMs and HOI-specific models, we manually check the annotation quality of HICO-DET. Through this inspection, we identify two major issues: *information incompleteness* and *annotation sparsity*. These issues significantly impact evaluation correctness, especially for VLMs whose predictions may be valid but absent from the ground-truth annotations.

The first type, information incompleteness, appears when the visual or temporal evidence in a static image is insufficient to determine the correct interaction, leading to two issues in existing HOI benchmarks' annotation. One issue is that the available visual evidence suggests a reasonable interaction, but does not provide enough visual information to confirm it. Fig. 2(a) is an example, where inspecting a hair dryer is a plausible interaction, but the static image alone is not enough to verify this. Because such interactions are plausible but visually unconfirmed, they are often considered as negatives in HICO-DET benchmark, although they are potentially valid interpretations. The other issue is the inherently ambiguous temporal states (e.g., *catch* vs. *throw frisbee* in Fig. 2(c)). These cases arise when multiple actions are visually indistinguishable in a single frame and only one of the plausible interactions is annotated (e.g., HICO-DET). In our manual review of 200 randomly sampled images from the HICO-DET test set, 29 exhibited insufficient visual evidence and 12 exhibited ambiguous temporal states (combined accounts for 20.5% of the sampled images). These examples illustrate that such ambiguities occur non-trivially in the existing HOI benchmark dataset, thus, affecting the evaluation when these images and their ground truths are used.

The second type, annotation sparsity, refers to missing or partial labels in multi-person scenes or within individual human–object pairs. In multi-person or multi-object cases, annotations are often provided for only a subset of individuals or objects, leaving other valid interactions unlabeled. As shown in Fig. 2(d), the female sitting on the couch is annotated, while the male on the adjacent couch is not. In Fig. 2(e), *cut cake* is labeled but *cut with knife* is missing. Sparsity also occurs within a single human–object pair when multiple concurrent or semantically related actions are possible but only a subset is annotated. For example, in Fig. 2(f), *flip skateboard* is labeled, while other plausible actions such as *jump skateboard* are not labeled. In our 200-image check, 58 images (29%) contained unannotated persons or objects, and 49 images (24.5%) contained incomplete action labels within annotated human–object pairs. These observations highlight the prevalence of sparse an-

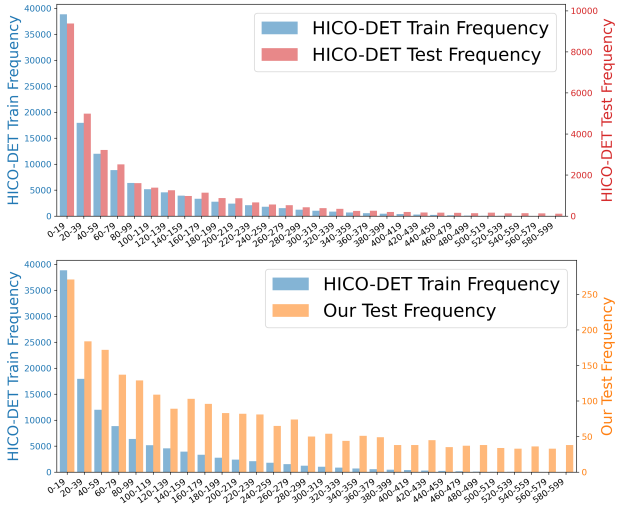


Figure 1. Comparison of HOI class frequency distributions between training and testing sets.

notations that treat valid interactions as negatives, leading to false negatives and incorrect evaluation.

Overall, these two categories reveal that HICO-DET benchmark suffers from two fundamental sources of annotation noise: information incompleteness and annotation sparsity. Together, they introduce evaluation bias and hinder fair comparison across methods, highlighting the need for a benchmarking dataset that enables more reliable HOI evaluation.

Similar Train and Test Splits Distribution Another issue with HICO-DET is the nearly identical distribution between its training and test splits (Fig. 1(a)). The KL divergence between the two splits is only 0.088, indicating that the test set closely mirrors the training distribution. As a result, models trained on this distribution may exhibit artificially high performance. In contrast, the KL divergence between the training split and our redistributed test set increases to 0.629, introducing meaningful variation without altering the inherent long-tail nature of HOI classes in the real world. Our goal is not to impose an artificial or unrealistic class distribution, but to avoid a test set that is effectively a resampled copy of the training distribution, thereby enabling a more reliable evaluation.

1.2. Comparison with Existing HOI Benchmarks

In Table 1, we provide a systematic comparison of our benchmark with existing HOI datasets across multiple dimensions. In HICO-DET [4], the majority of cases (60.2%) involve a single person interacting with a single object, which often results in relatively easy recognition. Our main benchmark, constructed from HICO-DET, reduces this proportion to 33.1%, thereby shifting the focus toward more challenging multi-person scenarios. In addition, a key strength of our benchmark is the inclusion of multi-person images with different interactions. While only 7.5% of HICO-DET and 22.5% of V-COCO [24] contain such cases, our dataset increases this proportion to 31.2%, providing a richer evaluation of compositional reasoning across individuals. SWiG-HOI [34] and Bongard-HOI [14] contain no such cases, because they only provide one annotated person for each image, in contrast to our benchmark's focus on multi-person HOI scenarios. Beyond the main benchmark, we create two sub-benchmarks focusing on multi-person and human-human interaction scenes. On the combined main and two sub-benchmarks, the proportion of single-person single-object cases drops to 11.2%, while multi-person different-HOI cases rise to 67.5%, which expands the diversity and difficulty of the evaluation.

Moreover, similar to existing HOI datasets, our benchmark remains compatible with HOI-specific methods, enabling evaluation under a unified protocol. While HOI models output a confidence score for every pre-defined HOI class, practical applications always require a selection step for final prediction, either via a threshold [19, 40] or by choosing the top-ranked predictions [35], a common step in HOI-specific inference pipelines. Although this selection is not part of the standard HOI evaluation protocol (e.g., mAP), where all predictions are used directly for evaluation, such a step is routinely applied in real-world HOI applications. In our evaluation, we adopt a top-K selection strategy, which evaluates HOI-specific models in a practical way. In addition, we also compare against threshold-based filtering and find that the top-5 selection outperforms other choices in Table 5 of the main paper. Importantly, this conversion only selects the predicted classes before matching them to each question choices, and does not alter model behavior, as it simply aligns the output format for comparison.

Apart from supporting HOI-specific methods, our benchmark is explicitly designed to support general-purpose VLMs, unlike HICO-DET, V-COCO, and SWiG-HOI. This is achieved by framing the task as multiple-choice question answering, naturally aligning with the input–output format of modern VLMs. Bongard-HOI, although relevant for HOI recognition, is limited to binary classification tasks (i.e., "is this interaction present or not?"). Our benchmark instead requires multi-class, multi-label prediction, reflecting the true complexity of HOI understanding in realistic images. Although there are some recent vision-language approaches (e.g., DAM [31], COCONut-PanCap [10]) related to interaction understanding, they do not provide a unified evaluation protocol that can fairly compare VLMs with HOI-specific detectors.

Taken together, our benchmarking dataset uniquely combines the strengths of previous HOI datasets while addressing their shortcomings. First, our datasets reduces oversimplified single-person cases, emphasizes multi-person cases with different interactions, remains compatible with HOI-specific methods, introduces explicit support for VLM evaluation, and requires full multi-class HOI prediction. This makes it the first benchmark to emphasize challenging cases and enable comparison across both specialized HOI models and general-purpose VLMs.

MCQA vs. Open-Set HOI Generation Our CrossHOI-Bench adopts MCQA to make evaluation well-defined across VLMs and HOI detectors. VLMs often generate free-form text. Mapping free-form text to fixed HOI labels is noisy and requires ad hoc rules. MCQA avoids this by fixing the output space, while still testing visual grounding among plausible interaction alternatives. We acknowledge that MCQA is easier than fully open-set HOI generation, so we treat it as a diagnostic benchmark, not a replacement for open-set evaluation. In this sense, MCQA provides a conservative test of interaction understanding: while strong performance does not necessarily imply full open-set HOI capability, poor MCQA performance is strong evidence that the model struggles with the unconstrained setting.

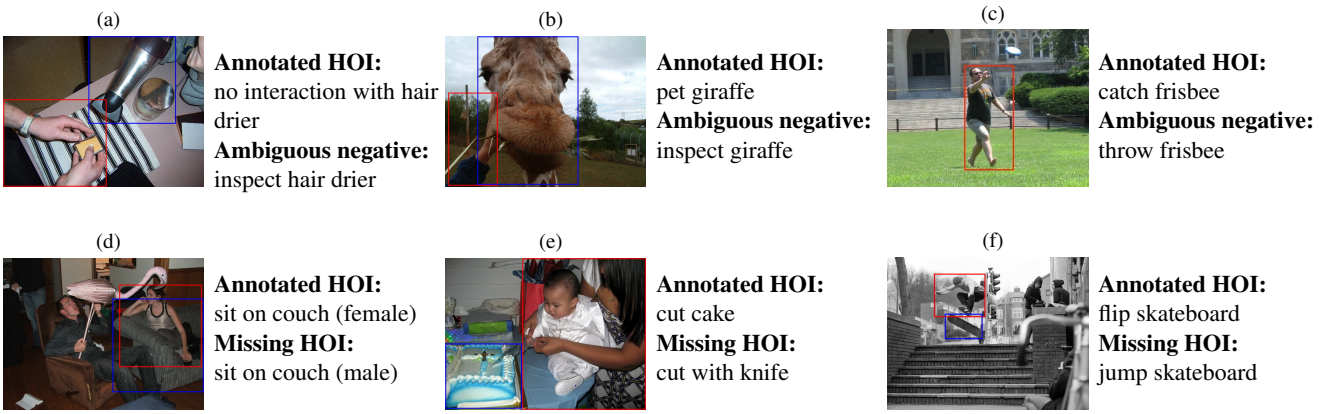


Figure 2. Illustration of HOI dataset annotation problems.

Dataset	Single-Person Single Obj ↓	Multi-Person Diff. HOI ↑	Applicable to HOI methods	Applicable to VLMs	Multi-Class HOI Prediction
HICO-DET [4]	60.2%	7.5%	✓	✗	✓
V-COCO [24]	51.2%	22.5%	✓	✗	✓
SWiG-HOI [34]	62.2%	0.0%	✓	✗	✓
Bongard-HOI [14]	100.0%	0.0%	✓	✓	✗
Ours (main benchmark)	33.1%	31.2%	✓	✓	✓
Ours (main+ sub- benchmarks)	**11.2%**	**67.5%**	✓	✓	✓

Table 1. Comparison between existing HOI benchmarks and ours.

1.3. Dataset Construction Details

We construct our benchmarking dataset using a two-stage approach: coarse screening and manual refinement. We refer readers to Sec. 3.1 of the main paper for the full process for coarse screening. The prompts used in the coarse screening stage are provided later in Sec. 2.3. During the manual refinement stage, we first remove overly simple scenes that offer limited diagnostic value. As illustrated in Fig. 3, cases such as multiple people riding bicycles unambiguously or a person washing a toothbrush in a simple background are straightforward. Therefore, these images are excluded from our benchmark.

For the remaining questions, we then design more challenging choices. To create hard positives, we select the target person to perform interactions that differ from those of surrounding individuals (e.g., "launch boat" in Fig. 4(a)). In ambiguous temporal scenes, we allow multiple plausible actions to be simultaneously correct (e.g., both *boarding* and *exiting* boat in Fig. 4(b)). We also include images with people performing multiple actions that models might miss some of these interactions, such as "load truck" and "sit on truck" in Fig. 4(c). To create hard negatives, we introduce interactions performed by nearby people that could be mistakenly attributed to the target person (e.g., *hold banana* in Fig. 4(d)), and we incorporate fine-grained distinctions between visually similar actions (e.g., *repair* vs. *type on* a laptop in Fig. 4(e)).

In total, the manual refinement stage updated 1,956 out of 5,096 total choices in our main benchmark, accounting for 38.39% of all choices and indicating a substantial level of human correction. Among them, 460 updates were made to positive choices (19.74% of all 2,330 positives). The remaining 1,496 updates correspond to negative choices (54.10% of all 2,765 negatives). Four HOI-focused annotators involve in the manual refinement and final decisions are made by majority vote. The annotators follow a shared guideline to independently verify coarse-screened negatives and propose hard positives and hard negatives. Inter-annotator agreement is high, with 95.22% mean pairwise agreement and 91.00% unanimous agreement across all four annotators.

For the V-COCO-based sub-benchmark, we focus specifically on multi-person scenarios, so we retain only images containing multiple people before applying the same coarse screening and manual refinement. During refinement, we add positive choices according to the predefined HOI classes in HICO-DET, since the original V-COCO benchmark includes only 24 action classes, which is insufficient for creating challenging choices. For the SWiG-HOI sub-benchmark, we target human–human interaction scenarios, therefore we keep only images annotated with interactions between pairs of humans. We then perform the same coarse screening and manual refinement.

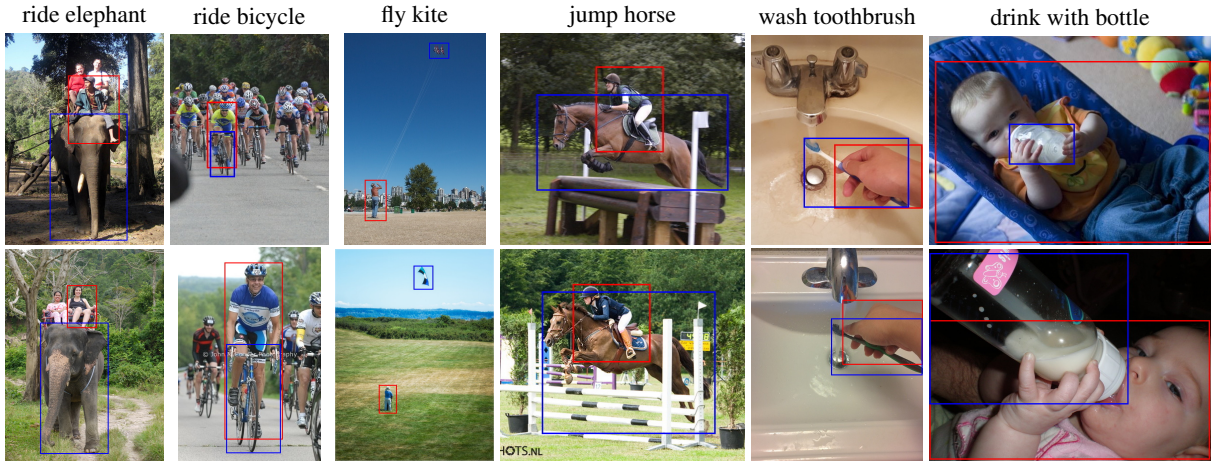


Figure 3. Removed simple scenes from HICO-DET test set during manual refinement stage.

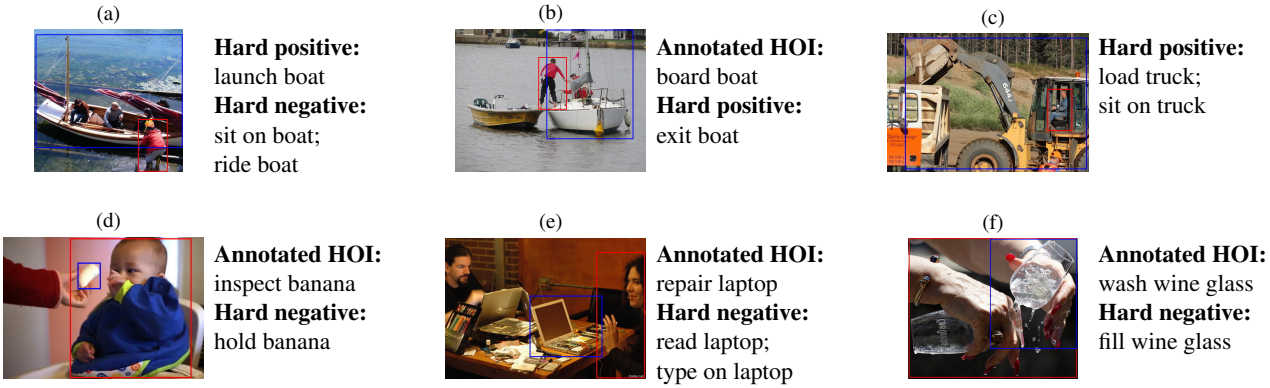


Figure 4. Hard choices modified during manual refinement.

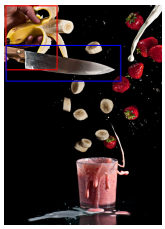
Mitigated Potential Artifacts or Bias During our dataset construction, VLMs are only used to generate and coarsely screen candidate negative options. They do not decide the final Multiple-Choice Question Answering (MCQA) options in the benchmark. Every negative option that is kept is checked by humans and must be judged invalid for the image. This addresses the artifact concern that a negative option is actually a valid interaction. As for the potential bias, where VLM-based coarse screening could change which types of negatives enter the pool and thus affect task difficulty or model rankings, Table 4 in the main paper provides evidence that the kept negatives are not just "VLM opinions." HOI-specific methods agree with VLM coarse screened negatives at about 99% agreement, while VLMs are at 95–97%. This suggests the negatives are generally wrong, not just wrong according to the VLM.

1.4. Our Dataset Examples

Examples in Fig. 5 illustrate the main challenges our benchmark emphasizes. In multi-person scenarios (e.g., the surfboard, frisbee, and cell phone related examples), different individuals perform distinct interactions, which is potentially confusing and leads to misattributing actions across people. At the same time, certain single-person cases are difficult due to either contactless interactions (e.g., peel apple) or visually similar categories (e.g., hold person vs. hug person). As a result, our benchmark provides a challenging evaluation of HOI understanding.

1.5. Our HOI Training Dataset

We primarily position our benchmark as an evaluation resource, with a central focus on zero-shot VLM performance and unified comparison against HOI-specific methods. However, for completeness and to support future work that adapts VLMs to HOI tasks, we also provide a standardized training split constructed from the HICO-DET training data. This split includes all three question types required by our benchmark: (i) a *Setting 1* question covering interactions of all annotated people, (ii) a *Setting 2* question focusing on the target person, and (iii) an additional object detec-



Setting 1:
First detect all persons, then Select interactions for the target person from choices [GT: A,B,C (hum.) | A,B (hum.-obj.)]
Choices: (A) wield knife; (B) cut with knife; (C) cut banana; (D) peel banana

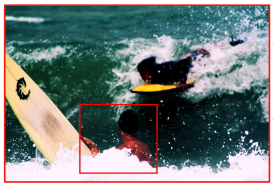
Setting 2:
Select interactions for the target person (box given) from choices [GT: A,B,C (hum.) | A,B (hum.-obj.)]
Choices: (A) wield knife; (B) cut with knife; (C) cut banana; (D) peel banana

Setting 3:Select interactions for all persons in the image from choices [GT: A,B,C]
Choices: (A) wield knife; (B) cut with knife; (C) cut banana; (D) peel banana

Setting 1:
First detect all persons, then Select interactions for the target person from choices [GT: A,C (hum.) | A,C (hum.-obj.)]
Choices: (A) catch frisbee; (B) spin frisbee; (C) throw frisbee; (D) hold frisbee

Setting 2:
Select interactions for the target person (box given) from choices [GT: A,C (hum.) | A,C (hum.-obj.)]
Choices: (A) catch frisbee; (B) spin frisbee; (C) throw frisbee; (D) hold frisbee

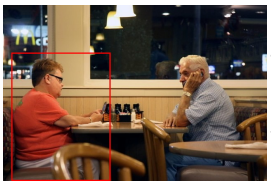
Setting 3:
Select interactions for all persons in the image from choices [GT: A,C]
Choices: (A) catch frisbee; (B) spin frisbee; (C) throw frisbee; (D) hold frisbee



Setting 1:
First detect all persons, then Select interactions for the target person from choices [GT: A (hum.) | A (hum.-obj.)]
Choices: (A) hold surfboard; (B) lie on surfboard; (C) stand on surfboard; (D) drag surfboard

Setting 2:
Select interactions for the target person (box given) from choices [GT: A (hum.) | A (hum.-obj.)]]
Choices: (A) hold surfboard; (B) lie on surfboard; (C) stand on surfboard; (D) drag surfboard

Setting 3:
Select interactions for all persons in the image from choices [GT: A,B]
Choices: (A) hold surfboard; (B) lie on surfboard; (C) stand on surfboard; (D) drag surfboard



Setting 1:
First detect all persons, then Select interactions for the target person from choices [GT: A,B (hum.) | A,B (hum.-obj.)]
Choices: (A) read cell phone; (B) hold cell phone; (C) repair cell phone; (D) talk on cell phone

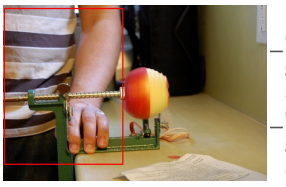
Setting 2:
Select interactions for the target person (box given) from choices [GT: A,B (hum.) | A,B (hum.-obj.)]
Choices: (A) read cell phone; (B) hold cell phone; (C) repair cell phone; (D) talk on cell phone

Setting 3:
Select interactions for all persons in the image from choices [GT: A,B,D]
Choices: (A) read cell phone; (B) hold cell phone; (C) repair cell phone; (D) talk on cell phone

Setting 1:
First detect all persons, then Select interactions for the target person from choices [GT: A,B (hum.) | A,B (hum.-obj.)]
Choices: (A) hold person; (B) carry person; (C) hug person; (D) kiss person

Setting 2:
Select interactions for the target person (box given) from choices [GT: A,B (hum.) | A,B (hum.-obj.)]
Choices: (A) hold person; (B) carry person; (C) hug person; (D) kiss person

Setting 3:
Select interactions for all persons in the image from choices [GT: A,B]
Choices: (A) hold person; (B) carry person; (C) hug person; (D) kiss person



Setting 1:
First detect all persons, then Select interactions for the target person from choices [GT: C (hum.) | C (hum.-obj.)]
Choices: (A) hold apple; (B) cut apple; (C) peel apple; (D) no interaction with apple

Setting 2:
Select interactions for the target person (box given) from choices [GT: C (hum.) | C (hum.-obj.)]
Choices: (A) hold apple; (B) cut apple; (C) peel apple; (D) no interaction with apple

Setting 3:
Select interactions for all persons in the image from choices [GT: C]
Choices: (A) hold apple; (B) cut apple; (C) peel apple; (D) eat apple

Figure 5. Example questions in our benchmark under the three evaluation settings.

tion question for *Setting 3*. This results in 111,459 training questions, offering a protocol for fine-tuning models.

1.6. Dataset Licenses and Release

Licenses We use the HICO-DET dataset [4], which is publicly released under a CC0: Public Domain license. Our sub-benchmarks also use V-COCO [24] (released under a CC-BY license) and SWiG-HOI [34] (released under its academic research license).

Data Release and Ethical Considerations We do not release any images or raw annotations from HICO-DET, V-COCO, or SWiG-HOI. Our benchmark only provides derived multiple-choice questions built on top of these datasets. Each question references the original image in-

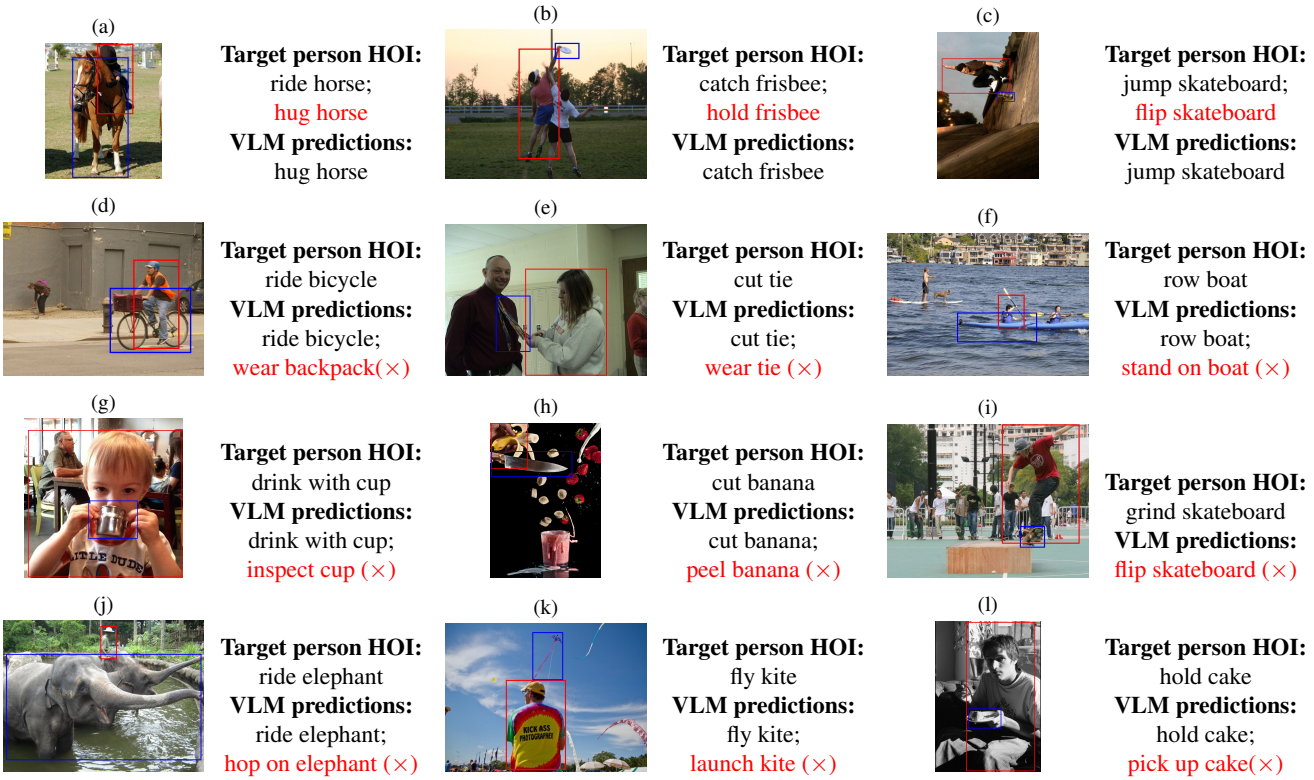


Figure 6. Illustration of VLM (Qwen2.5-VL-32B) failure cases in *Setting 1*, where HOI-specific model (ADA-CM) predicts correctly.

dex and is derived from the HOI annotations provided in the respective dataset, supplemented with minor modifications when needed to resolve annotation noise or include additional HOI classes.

No images, bounding boxes, or HOI annotations are redistributed; users must obtain the original datasets separately under their respective licenses. Since our release contains only question–answer pairs and index mappings, the risk of exposing personally identifiable information or offensive content is minimal. All consent and licensing considerations follow those of the original datasets, and no additional consent verification is conducted beyond their official releases.

2. Experiments

Baseline Details We evaluate two groups of baselines on our benchmark: general-purpose VLMs and HOI-specific methods. Recent large VLMs represent the frontier of general-purpose image understanding. Qwen2-VL, Qwen2.5-VL (7B / 32B) and Qwen3-VL-instruct [2] are selected as they excel in fine-grained spatial localization and visual reasoning, making them suitable for HOI tasks. InternVL2.5 and InternVL3 (8B / 38B) [7, 36, 41] are included because they achieve leading performance across diverse multimodal benchmarks and emphasize high-resolution perception, which is relevant for recognizing human–object interactions. LLaVA-OV-7B [23] is an instruction-tuned VLM designed for open-vocabulary understanding demonstrating versatility and strong performance across multiple vision–language tasks, making it a relevant baseline for HOI evaluation.

For completeness, we additionally report supervised fine-tuning results for Qwen2.5-VL-7B, providing a baseline under a finetuned setup using our HOI training dataset (see Sec. 1.5). We finetune the VLM using Low-Rank Adaptation (LoRA) [13] applied to the text-decoder attention projection layers (query, key, value) and the final output projection. Training is conducted for 5 epochs with a batch size of 32.

Beyond VLM baselines, we also evaluate recent HOI detection methods. ADA-CM [19], CMMP [22], LAIN [16] and HOLa [18] demonstrate competitive performance on the existing HICO-DET benchmark [4]. In addition, CMD-SE [21] is a recent open-vocabulary HOI detection method emphasizing generalization ability, achieving competitive performance on SWiG-HOI [34] and HICO-DET benchmarks. We use best-performing pre-trained checkpoints when available, and otherwise reproduce results with the authors' code under the closest available configurations. Specifically, ADA-CM, CMMP, and HOLa are evaluated

with the ViT-L vision backbone, while CMD-SE and LAIN are based on ViT-B.

2.1. Additional Findings

Additional Finding 1: Analysis of VLM Failure Cases To further understand the performance of VLMs and HOI-specific models, we analyzed 200 failure cases of Qwen2.5-VL-32B [2] in *Setting 1*, where the HOI model (ADA-CM [22]) successfully produced the correct prediction but the VLM failed. Here, we exclude the detection failure cases, which is one of the limitations of VLMs shown in Table 1 (main paper) and instead focus on the remaining failure cases. We discuss the detection limitations of VLMs in detail in paragraph *Detection Limitations of VLMs*. These non-detection failure cases are grouped into four main categories.

1. Incomplete multi-action recognition (54% out of the 200 failure cases). The most frequent failure occurs when the VLM predicts only a subset of multiple valid actions for a human–object pair. For example, in Fig. 6(a), in an image where a person is riding and hugging a horse, the model recognizes only riding but omits hugging. This suggests that the VLM tends to capture a certain action while ignoring concurrent interactions. In contrast, HOI-specific models are less affected by this issue, largely because their training data include co-occurring action annotations. Through supervised learning on such examples, HOI models implicitly learn that multiple actions can jointly occur on the same human–object pair.

2. Cross-person HOI misattribution (16% out of the 200 failure cases). The VLM often misattributes the interaction of a surrounding person to the target person. When multiple people appear in close proximity performing different actions, the global attention mechanism can mix the features of nearby persons and assign the wrong action to the target. As shown in Fig. 6(e), the model incorrectly predicts wearing a tie for the woman who is cutting the tie. In contrast, HOI-specific models explicitly preserve spatial topology through region cropping or query anchoring, which alleviates such confusion.

3. HOI similarity confusion (13% out of the 200 failure cases). Another source of failure arises from visually similar HOI classes. In many cases, two HOIs differ only by subtle local cues that occupy a very small region of the image (e.g., just a few dozen pixels in a 400×600 image). For instance, in Fig. 6(g), when a person drinks from the cup with their gaze directed toward the camera, the VLM predicts *inspect cup* instead. These errors suggest that current VLMs struggle with fine-grained visual discrimination, which aligns with prior analyses showing VLMs have difficulty distinguishing categories that differ by subtle visual or semantic differences [12, 38].

4. Hallucinated HOI inference (8% out of the 200 failure cases). Some failures occur when the image itself does not provide sufficient visual evidence for the action. For instance, in Fig. 6(j), when the image shows a person already riding the elephant, rather than mounting it, the VLM predicts *hop on elephant*. Despite the absence of visual cues, the VLM occasionally predicts hallucinated actions, likely driven by language priors or common object–action co-occurrence patterns. This behavior resembles the object hallucination phenomenon reported in previous studies [6, 9], but here it reflects at the action level, where the model infers plausible yet visually ungrounded interactions. The HOI-specific model, relying on explicit region-level visual features, is less prone to such hallucinated inferences.

Additional Finding 2: Problems Persist in Both VLMs and HOI-specific Models Although the previous failure analysis showed that HOI-specific models outperform VLMs in four categories, we find that for two of them, cross-person HOI misattribution and HOI similarity confusion, HOI-specific models only partially alleviate the problem. In the following, we discuss why HOI models improve these cases compared to VLMs, and why the problems still persist.

Both types of errors originate from how VLMs represent spatial structure and instance-level information. VLMs [2, 28, 41] encode an image as a flattened sequence of patch embeddings, where global self-attention treats all tokens uniformly. Although two-dimensional positional encodings are included, instance-level associations, such as determining which hand belongs to which person, are only learned implicitly rather than structurally enforced. As a result, tokens from nearby persons or objects can interfere, causing cross-person misattribution. Meanwhile, distinguishing similar HOIs often depends on extremely sparse cues (e.g., hands or gaze direction in Fig. 6(e)-(f)), which account for only a few patches. Because these cues are not explicitly emphasized, the model relies purely on data-driven learning to capture them, which can be unreliable in practice.

HOI-specific methods mitigate these issues by explicitly preserving spatial topology and emphasizing local features. Two-stage approaches [16, 17, 19] extract a two-dimensional feature map after a Transformer encoder and apply RoIAlign to crop human and object regions, isolating instance-specific features and narrowing the search space for fine-grained cues. One-stage transformer methods [15, 20, 26, 30] achieve a similar effect through object queries that serve as implicit region anchors focusing on distinct human–object pairs. However, these strategies only partially solve the problem, because cropped regions may still contain unrelated body parts, while query attention can overlap when people are close. Moreover, the critical visual cues for distinguishing actions may lie in extremely small areas, such as a few pixels around the hand or gaze, making

these cues too localized to be consistently captured.

In our analysis of 200 multi-person cases where surrounding individuals perform different actions, the VLM (Qwen2.5-VL-32B [2]) resulted in 44 misattributed failure cases (22%), while the HOI model (CMMP [22]) had 30 (15%). We further analyze 400 questions in the test set focusing on situations where the correct class is easily confused with a semantically related one (e.g., "cut banana" vs. "peel banana" in Fig. 6(h)). Among these questions, we identify 110 HOI similarity confusions in total. Of these, 65 (59%) are made by the VLM and 57 (52%) by the HOI model, with overlap where both models make the same confusion. Both experiment results consistently indicate that HOI-specific models alleviate, but do not eliminate errors related to cross-person misattribution and fine-grained similar HOI confusion.

Discussion: Why VLMs Outperform HOI Models in Interaction Understanding Although HOI-specific models are designed for HOI detection, based on our experiment results across different settings and benchmarks (main and two complementary), we observe that VLMs often exhibit stronger interaction recognition and contextual understanding abilities. We hypothesize three potential factors behind this advantage.

First, VLMs benefit from vastly broader and more diverse training data [11], including internet-scale image–text pairs and instruction-tuned datasets covering a wide spectrum of visual reasoning and commonsense knowledge [8, 37]. In contrast, HOI datasets such as HICO-DET and V-COCO contain (e.g., HICO-DET train set includes 38,118 images, V-COCO train set includes 5,400 images), which are smaller than datasets VLMs trained on. Second, VLMs possess far larger model capacities, often scaling to tens of billions of parameters, which enhance their representational power and enable stronger cross-modal generalization [5, 29]. Third, the generative nature of VLMs may inherently enhance image understanding and reasoning ability. Unlike HOI-specific discriminative models that map vision features to fixed categories, the generative modeling principles in VLMs have been shown to drive scalability, in-context learning, and compositional generalization [1, 27], which likely contribute to their stronger contextual understanding [25]. However, each factor remains an open question and requires future work to understand their individual contributions.

Detection Limitations of VLMs The performance drop from *Setting 2* to *Setting 1* highlights that VLMs are inferior than at detection (e.g., Qwen2.5-VL-32B drops 16.6% Instance-F1). When ground-truth boxes are provided in *Setting 2*, VLMs can recognize interactions effectively, but their performance drops significantly in *Setting 1*, where detection must be performed by the VLMs themselves (Table 1, main paper). HOI-specific methods, in contrast, integrate detection within their pipelines and therefore cannot directly exploit ground-truth boxes, but under *Setting 1* they often perform competitively, sometimes surpassing even large VLMs (e.g., ADA-CM achieves 13.6% higher Micro-F1 than InternVL3-38B and 0.3% higher than Qwen2.5-VL-32B). This indicates that detection remains a major bottleneck for VLMs, while HOI-specific models strike a more balanced trade-off between detection and recognition.

We conduct error analysis using Qwen2.5-VL-32B, which achieves the highest performance in *Setting 1* among VLM baselines. As shown in Fig. 7, failures mainly occur in multi-person scenes, where the VLM struggles to correctly localize individuals in crowded settings, and in occluded scenes, where heavy occlusion of the person leads to missed or inaccurate bounding boxes. These cases show that complex layouts and occlusions remain key challenges for VLM detection.

MCQA Option Bias Analysis A potential concern with multiple-choice evaluation is that VLMs may exploit option-level shortcuts rather than genuine interaction understanding. To verify that CrossHOI-Bench does not introduce such biases, we analyze the distribution of answer positions and model predictions. As shown in Table 2, the correct option position is randomized and approximately uniform. Importantly, the prediction distributions of strong VLM baselines do not align with the ground-truth distribution, indicating that the models are not exploiting positional biases. We also use a fixed template for all options and construct negatives with the same or closely related objects, so models cannot win by option wording or object-only cues.

Statistical Significance Analysis For statistical significance testing, we report 95% bootstrap confidence intervals in Table 3, using 1,000 question-level resampling iterations in *Setting 1*. This shows our results are statistically stable despite the moderate test size.

2.2. Additional Results

***Setting 2* Discussion** *Setting 2* is a diagnostic extension of *Setting 1*. It tests whether VLM underperformance in *Setting 1* is mainly due to localization. We provide ground-truth boxes, so detection failures are removed. This lets us evaluate recognition in isolation. Most HOI methods cannot be evaluated in *Setting 2* with GT boxes: one-stage (end-to-end) methods require joint detection, and among two-stage methods, those relying on detector embeddings cannot operate with GT-box input. However, *Setting 2* is not used to assess paradigm gaps, which are evaluated in *Setting 1*. For completeness, we include two HOI methods (ADA-CM and CMMP) that can accept GT boxes in Table 4. They improve

Figure 7. Failure detection cases of the Qwen2.5-VL-32B model in *Setting 3*. The red box marks the target person specified in the question. The first row shows failures in multi-person scenarios, while the second row shows failures under occlusion.

Choices	A	B	C	D		A	B	C	D
Qwen2.5-Vl-32B	0.20	0.15	0.27	0.38	GT	0.24	0.25	0.25	0.26
Internvl3-38B	0.27	0.25	0.26	0.22	GT	0.25	0.24	0.26	0.25

Table 2. Analysis of MCQA option bias. We report the distribution of predicted answer positions for two VLM baselines and compare it with the ground-truth distribution. The approximately uniform ground-truth distribution and the mismatch with model predictions indicate that the benchmark does not introduce positional bias.

over their *Setting 1* results, but remain below the best large VLMs.

VLM Zero-shot Evaluation with Off-the-shelf Object Detector Since VLMs often struggle with reliable person detection, we follow the two-stage HOI detection paradigm [19, 32, 40] and leverage a widely used off-the-shelf object detector, DETR [3] pre-trained on HICO-DET. Table 5 shows that incorporating DETR helps improve performance in the *Setting 1* evaluation, though remains lower than in *Setting 2* due to detection errors. Among small VLMs, Qwen2.5-VL-7B achieves the best overall performance among most evaluation metrics, competitive to SOTA HOI-specific methods. For large models, Qwen2.5-VL-32B outperforms InternVL3-38B most of the time and other VLMs included in the comparison. By comparing Table 5 and Table 1 (main paper), we observe a clear performance drop when VLMs perform detection on their own, as opposed to relying on off-the-shelf detectors (i.e., +6.6% Macro-F1 for Qwen2.5-VL-7B, +4.34% for Qwen2.5-VL-32B, +19.92% for InternVL-8B and +8.35% for InternVL-38B). This highlights that the detection capability of current VLMs still lags behind that of specialized object detectors.

Prediction-Selection Strategies for HOI Methods Table 6 compares two common prediction-selection strategies for HOI detectors: selecting the Top-K predictions or applying a confidence threshold. We observe that Top-5 selection consistently outperforms threshold-based filtering. Applying confidence thresholds leads to substantial drops, as stricter filtering removes many correct predictions and severely hurts recall and F1. Table 7 further studies the effect of different K values. Increasing K from 3 to 5 significantly improves all F1 metrics, indicating that HOI models often predict multiple valid interactions for a person–object pair. However, increasing K to 10 introduces more false positives: recall-related metrics such as Instance-F1 and Micro-F1 continue to increase slightly, while EM begins to decline. Overall, Top-5 provides the best balance between recall and precision, and is therefore adopted in our main evaluation.

Potential Circular Bias Analysis In the original dataset construction, Qwen2.5-VL-32B was used in the Coarse Screening stage to filter obviously invalid negative options. Since the same model is also evaluated as a baseline, this raises a potential circular bias: the benchmark may favor Qwen models if the negative options were filtered using the same model family during dataset construction. To exam-

	Qwen2.5-VL-32B	InternVL3-38B	ADA-CM	HOLa
Instance-F1	52.94 ± 2.12	38.68 ± 2.24	47.76 ± 2.32	47.12 ± 2.22
Macro-F1	50.71± 2.23	38.04 ± 1.91	43.02 ± 2.07	43.06± 2.15

Table 3. Statistical significance analysis. We report 95% bootstrap confidence intervals computed using 1,000 question-level resampling iterations in *Setting 1*.

Method	Macro-F1 (%)	Instance-F1 (%)	Micro-F1 (%)	EM (%)	Avg. Prec. (%)	Avg. Rec. (%)
HOI-specific methods						
ADA-CM [19]	48.89	**56.25**	**67.19**	**21.74**	**73.11**	**62.16**
CMMP [22]	**49.04**	55.10	65.85	19.78	70.90	61.47
VLM zero-shot evaluation						
InternVL2.5-38B [7]	48.43	46.43	51.56	20.64	77.52	38.63
InternVL3-38B [41]	58.94	67.41	67.81	**35.64**	**81.90**	57.85
Qwen3-VL-30B-Instruct [2]	56.43	64.00	64.59	32.57	79.89	54.21
Qwen2.5-VL-32B [2]	**62.90**	**69.52**	**70.69**	35.01	75.30	**66.61**
LLaVA-OV-7B [23]	47.76	56.53	54.80	25.12	**77.43**	42.40
InternVL3-8B [41]	**49.88**	52.35	55.54	23.86	74.41	44.31
Qwen2-VL-7B [33]	46.90	53.93	53.61	23.23	76.84	41.16
Qwen2.5-VL-7B [2]	48.93	**57.25**	**57.53**	**25.98**	74.49	**46.87**

Table 4. *Setting 2* experiment results comparison. Best performance within each group is highlighted in **bold**. "Avg. Prec." and "Avg. Rec." denote the precision and recall averaged across the test set, respectively.

ine this effect, we construct a 12% random subset where Qwen2.5-VL-32B is excluded from the coarse screening step. Qwen2.5-VL-32B still achieves similar Setting-1 performance on this subset (Table 8). We also re-evaluate the main baselines on the same subset and observe similar ranking trends. This indicates that Qwen's performance and the overall comparisons are not driven by a "Qwen-on-Qwen" screening loop.

Evaluation for Target Human-Object Pairs Since we require the model to recognize localized interactions for target person by default in *Setting 1* and *Setting 2*, here we provide the experiment results where models are required to identify interactions for target human-object pairs. We name these two settings as **Extended *Setting 1*** and **Extended *Setting 2***.

In Table 9, each model first detects all human and object instances before predicting HOI classes. The *Extended Setting 1*, therefore, requires comprehensive localization of multiple objects and people in diverse scenes. The performance drop observed across all VLMs highlights their limited ability to perform reliable object detection. Even though the object categories are predefined (the 80 object classes from HICO-DET), large VLMs still struggle to consistently identify all relevant objects. In contrast, HOI-specific methods, which are trained jointly for detection and interaction recognition, achieve improved performance in this setting. These results indicate that while large VLMs exhibit promising zero-shot learning ability, their object detection remains a critical bottleneck for HOI detection.

We further leverage an off-the-shelf DETR model for object detection, which is also commonly adopted by HOI-specific methods [19, 39, 40]. With this detector, VLMs perform substantially better than when detecting objects by themselves, as shown in Table 9. However, smaller VLMs still lag behind HOI-specific methods, and even large models such as InternVL3-38B and Qwen2.5-VL-32B only slightly outperform the best HOI-specific method in a few metrics (e.g., Instance-F1, Macro-F1) while falling behind in others.

Extended *Setting 2* isolates the detection ability and focuses on localized human-object pair interaction recognition. However, compared with default *Setting 2*, the overall performance of VLMs drops slightly when the target object is specified, as shown in Table 10. This is because VLMs tend to predict interactions involving non-target objects. For example, when a person is sitting on a bench while holding a book, and the target object is bench, the model often predicts both "sit on bench" and "hold a book". Quantitatively, such wrong-object predictions account for 11.60% of all errors in InternVL3-38B and 9.51% in Qwen2.5-VL-32B. This suggests that even with explicit object localization cues, current VLMs struggle to restrict their predictions to the intended target object.

Sub-benchmarks Evaluation The sub-benchmark experiments extend our main evaluation by evaluating model performance on the V-COCO-based and SWiG-HOI-based

Method	Macro-F1 (%)	Instance-F1 (%)	Micro-F1 (%)	EM (%)	Avg. Prec. (%)	Avg. Rec. (%)
HOI-specific methods						
ADA-CM	43.02	**47.76**	**61.69**	19.15	76.25	51.80
CMMP	43.06	46.62	60.85	18.84	75.06	51.16
LAIN	41.28	45.64	59.09	19.31	73.42	49.44
HOLa	43.61	47.12	61.29	19.78	74.31	**52.15**
CMD-SE	**47.49**	44.66	58.71	**20.33**	**78.33**	46.96
VLM zero-shot evaluation (off-the-shelf object detector, DETR [3])						
InternVL2.5-38B	48.12	46.69	52.26	21.04	78.25	39.23
InternVL3-38B	59.33	66.13	67.38	**33.20**	**82.20**	57.08
Qwen3-VL-30B-Instruct	57.05	63.55	64.91	31.95	80.32	54.46
Qwen2.5-VL-32B	**61.59**	**67.29**	**69.48**	31.79	74.65	**64.98**
LLaVA-OV-7B	47.03	55.55	54.68	24.18	**78.43**	41.97
InternVL3-8B	**51.81**	53.47	57.05	24.25	76.77	45.41
Qwen2-VL-7B	46.06	51.85	52.71	21.51	77.28	40.00
Qwen2.5-VL-7B	49.56	**56.91**	**57.98**	**25.35**	75.22	**47.17**

Table 5. *Setting 1* experiment results comparison. VLM zero-shot evaluation is based on off-the-shelf detector (DETR) for object detection [3]. Results are grouped by VLM size (small vs. large). Best performance within each group is highlighted in **bold**. "Avg. Prec." means the precision averaged across test set and "Avg. Rec." means the recall averaged across test set.

Method	**Top-5**	**Score $\geq$ 0.3**	**Score $\geq$ 0.5**	**Score $\geq$ 0.7**
ADA-CM	47.76	33.84	24.31	14.26
CMMP	46.62	32.84	23.33	13.74
LAIN	45.64	30.96	21.10	10.44
HOLa	47.12	34.74	25.47	13.13
CMD-SE	44.66	39.95	35.00	29.05

Table 6. Comparison between Top-K selection and confidence-threshold filtering for HOI-specific models under Setting 1 (Instance-F1).

Method	Top-K	Macro-F1	Instance-F1	Micro-F1	EM
ADA-CM	Top3	35.17	40.86	54.77	17.58
	Top5	41.53	47.74	61.68	**19.07**
	Top10	**46.30**	**50.58**	**63.50**	15.54
CMD-SE	Top3	31.35	37.67	51.19	15.86
	Top5	**47.49**	44.66	58.71	**20.33**
	Top10	46.35	**48.14**	**61.93**	17.26

Table 7. Effect of different Top-K values for HOI prediction selection under Setting 1.

Coarse Screening	Macro-F1	Instance-F1	Micro-F1	EM
w/o Qwen2.5-VL-32B	49.00	57.51	65.42	28.00
w/ Qwen2.5-VL-32B	46.77	56.33	63.87	29.33

Table 8. Ablation study on the effect of using Qwen2.5-VL-32B in the coarse screening during dataset construction. The experiment is conducted in a 12% random subset. "w/o Qwen2.5-VL-32B" means that Qwen is not used for screening; "w/ Qwen2.5-VL-32B' means that Qwen is used for screening.

subsets. These subsets emphasize distinct challenges: V-COCO-based sub-benchmark focuses on multi-person scenes, whereas SWiG-HOI-based one highlights human-human interaction understanding. Table 11-15 display sub-benchmark evaluations across baselines.

For HOI-specific models, sub-benchmarks serve as the out-of-distribution evaluation, as we require one model to be tested across our three benchmarks (one main benchmark based on HICO-DET and two sub-benchmarks based on V-COCO and SWiG-HOI). All HOI-specific methods evaluated on sub-benchmarks are pre-trained on HICO-DET training dataset. Tables 11-15 show that HOI-specific methods experience clear performance drop when evaluated on the V-COCO– and SWiG-HOI–based subsets. The V-COCO-based sub-benchmark has a larger label space after we add additional HOI annotations based on HICO-DET pre-defined classes (Original V-COCO only include 24 action labels). For example, in *Setting 1*, CMMP decreases from 46.62 Instance-F1 on the HICO-based main benchmark (Table 1 of main paper) to 28.39 on the V-COCO-based sub-benchmark, an absolute drop of 18.23 points. In contrast, CMD-SE decreases from 44.66 Instance-F1 (main benchmark) to 7.37 on the SWiG-HOI sub-benchmark, a much larger drop of 37.29 points. This pattern is consistent across other settings. These numbers indicate that HOI-specific models struggle with cross-dataset generalization, suffering much larger performance drops on SWiG-HOI-based sub-benchmarks, whose HOI class distributions and interaction patterns differ more from those in HICO-DET-based main benchmark.

VLMs, under zero-shot evaluation, demonstrate consistent generalization across datasets, but they nevertheless experience the largest performance drop on the V-COCO-based sub-benchmark, posing the greatest challenge among

Method	Macro-F1 (%)	Instance-F1 (%)	Micro-F1 (%)	EM (%)	Avg. Prec. (%)	Avg. Rec. (%)
HOI-specific methods						
ADA-CM	48.96	**42.10**	50.57	23.00	77.79	46.97
CMMP	49.01	41.03	57.66	22.92	77.20	46.01
LAIN	47.13	40.77	56.65	22.21	75.53	45.33
HOLa	48.96	41.68	**58.16**	**23.23**	76.01	**47.10**
CMD-SE	**51.98**	32.35	47.64	16.72	**78.96**	34.11
VLM zero-shot evaluation						
InternVL2.5-38B	7.80	4.48	6.97	2.35	77.67	3.65
InternVL3-38B	**21.35**	**16.50**	**24.69**	**8.79**	**78.87**	**14.64**
Qwen3-VL-30B-Instruct	-	-	-	-	-	-
Qwen2.5-VL-32B	17.13	11.48	19.16	5.10	76.92	10.94
LLaVA-OV-7B	-	-	-	-	-	-
InternVL3-8B	3.07	0.55	0.45	0.08	**83.33**	0.23
Qwen2-VL-7B	-	-	-	-	-	-
Qwen2.5-VL-7B	**13.49**	**10.21**	**14.88**	**4.71**	68.54	**8.34**
VLM zero-shot evaluation (off-the-shelf object detector, DETR [3])						
InternVL2.5-38B	46.64	34.59	42.52	16.95	**78.34**	29.18
InternVL3-38B	52.72	44.47	52.77	**22.37**	76.70	40.22
Qwen3-VL-30B-Instruct	49.17	42.55	49.95	20.17	73.80	37.76
Qwen2.5-VL-32B	**53.03**	**45.56**	**54.89**	21.11	71.61	**44.51**
LLaVA-OV-7B	**45.77**	38.15	42.30	**18.29**	**74.40**	29.55
InternVL3-8B	45.50	36.31	**44.55**	15.78	71.17	32.42
Qwen2-VL-7B	43.89	37.28	41.60	17.82	72.70	29.14
Qwen2.5-VL-7B	43.56	**38.18**	44.16	16.25	68.59	**32.56**

Table 9. Extended *Setting 1* experiment results comparison, requiring detection for both humans and objects, and then predicting HOI classes based on detected human and object boxes. Results are reported for VLMs and HOI-specific methods. Best performance within each group is highlighted in **bold**. "Avg. Prec." means the precision averaged across test set and "Avg. Rec." means the recall averaged across test set.

Method	Macro-F1 (%)	Instance-F1 (%)	Micro-F1 (%)	EM (%)	Avg. Prec. (%)	Avg. Rec. (%)
VLM zero-shot evaluation						
InternVL2.5-38B	51.09	48.74	52.95	26.22	76.49	40.49
InternVL3-38B	58.02	65.20	65.92	**34.30**	**76.58**	57.87
Qwen3-VL-30B-Instruct	55.00	62.45	62.66	30.46	73.79	54.45
Qwen2.5-VL-32B	**60.00**	**65.79**	**66.86**	31.08	70.34	**63.70**
LLaVA-OV-7B	49.82	**56.88**	55.38	**28.49**	**75.35**	43.78
InternVL3-8B	**53.27**	53.62	**56.46**	24.96	71.28	46.74
Qwen2-VL-7B	48.24	54.23	53.39	26.06	72.33	42.32
Qwen2.5-VL-7B	48.86	55.98	56.32	26.37	70.43	**46.92**

Table 10. Extended *Setting 2* experiment results comparison, with both human and object boxes provided from each question. Results are reported for VLMs and HOI-specific methods. Best performance within each group is highlighted in **bold**. "Avg. Prec." means the precision averaged across test set and "Avg. Rec." means the recall averaged across test set.

the three benchmarks. Detection-required settings amplify this difficulty. In *Setting 1*, Qwen2.5-VL-7B drops from 30.53 Instance-F1 on the main benchmark (Table 1, main paper) to 12.66 on V-COCO (-17.87), which is substantially lower than its performance on SWiG-HOI (47.95). This pattern becomes more noticeable in Extended *Setting 1*, where the requirement to detect both humans and objects causes even large VLMs drop drastically. InternVL3-38B falls from 16.50 Instance F1 on main benchmark to 10.07 on V-COCO (Table 12). These results highlight that scenarios on V-COCO with multiple people and multiple objects, significantly increase the difficulty of reliable person/object detection for VLMs.

When detection is not required, most VLMs show a clear

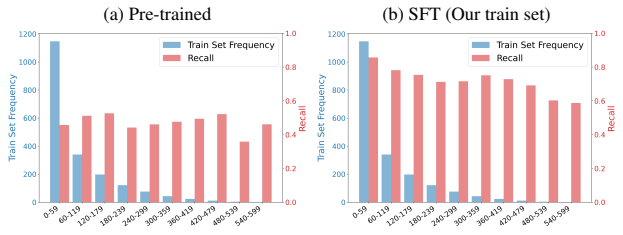


Figure 8. Comparison among pre-trained Qwen2.5-VL-7B [2] and SFT finetuned Qwen2.5-VL-7B on our train set. The blue histograms indicate the binned class frequency in our training dataset, while the red histograms present the recall rate. Head and tail classes are defined by HOI class frequency in our training set, and all HOI classes are ordered accordingly.

Macro-F1 drop when moving from the main benchmark to either sub-benchmark. For example, Qwen2.5-VL-32B decreases from 62.90 Macro-F1 on the main benchmark (Table 2, main paper) to 55.61 on V-COCO and 52.74 on SWiG-HOI (Table 13). A similar pattern appears in Extended *Setting 2*, where its Macro-F1 falls from 60.00 (Table 10) to 44.76 on V-COCO and 55.48 on SWiG-HOI (Table 14). These consistent Macro-F1 reductions across models indicate that both sub-benchmarks exhibit a distribution shift in HOI classes, under multi-person and human-human interaction scenes. In *Setting 3*, the EM accuracy drops a lot across VLMs, when comparing main benchmark and V-COCO sub-benchmarks. This indicates that it is challenging to recognize comprehensive interactions across the image under multiple people and multiple object scenarios.

Fine-tuned VLM Evaluation Table 16 provides results of the baseline under a finetuned setup, Qwen2.5-VL-7B fine-tuned on our training dataset. Before fine-tuning, VLMs typically exhibit much higher average precision than recall, reflecting a conservative prediction style that favors fewer outputs, shown in Table 1,2,3 (main paper). After training on our dataset, however, the model learns to adjust its strategy: recall surpasses precision across settings. This shift indicates that the model adapts to predict multiple interactions for a given question. Predicting more for each question increases the chance of covering all correct answers, although this comes at the cost of lower precision.

Moreover, in Fig. 8(b), recall (red bins) decreases steadily from head to tail classes, revealing a clear head-class bias after fine-tuning. In contrast, the pre-trained VLM without fine-tuning shows no obvious head-class bias, as recall remains relatively flat across classes (Fig. 8(a)). Overall, while class imbalance has long been recognized as a challenge for HOI-specific methods, our benchmark demonstrates that fine-tuned VLMs are not immune to this issue.

Evaluation among all HICO-DET images Tables 17, 18, 19 present results evaluated on the 9,546 questions derived from the entire HICO-DET test set, corresponding to the same three settings as in Tables 1,2,3 (main paper) of our benchmark. Across all metrics in Tables 17, 18, 19 both HOI-specific methods and VLMs achieve consistently much higher scores, typically by around 8 – 12 points in Instance-F1 and Micro-F1, compared with their performance on our main benchmark. This substantial margin confirms that our benchmark is significantly more challenging than the original HICO-DET benchmark. Unlike HICO-DET, which contains many simple, single-person single-object scenes, our benchmark emphasizes multi-person and confusing interactions. The noticeable performance gap validates the rationale of constructing a smaller but more challenging benchmark based on HICO-DET.

2.3. Implementation Details

Prompts for Benchmarking VLMs For general-purpose VLMs, we provide the question prompt together with explicit answer-format instructions.

In *Setting 1* and *Setting 2*, we obtain choice selection results with the following prompt for target-human or target human-object pair accordingly:

```
Context: You are given an image
⟨image⟩ and a target person with
a bounding box ⟨ human box ⟩.
Question: Which of the following
describes the interactions between
the target person and any object in
the image? Choices: (A) ...,(B)
...,(C) ...,(D)... IMPORTANT: Reply
with the letter(s) ONLY, separated
by commas if multiple (e.g. A,B).
For example, if correct answers are
(A) and (B), your output must be:
A,B Do NOT include any brackets or
other symbols.
```

```
Context: You are given an image
⟨image⟩ and and two bounding boxes:
- Person bbox: ⟨ human box ⟩;
- Object bbox: ⟨ object box ⟩.
Question: Which of the following
describes the interactions between
the target person and the object?
Choices: (A) ...,(B) ...,(C)
...,(D)... IMPORTANT: Reply with
the letter(s) ONLY, separated by
commas if multiple (e.g. A,B). For
example, if correct answers are (A)
and (B), your output must be: A,B
Do NOT include any brackets or other
symbols.
```

Method	Dataset	Macro-F1 (%)	Instance-F1 (%)	Micro-F1 (%)	EM (%)	Avg. Prec. (%)	Avg. Rec. (%)
HOI-specific methods Out-of-distribution evaluation							
ADA-CM	-	-	-	-	-	-	-
CMMP	V-COCO	25.57	28.39	37.07	2.81	80.56	24.07
LAIN	V-COCO	11.62	17.04	23.99	3.01	68.26	14.55
HOLa	-	-	-	-	-	-	-
CMD-SE	SWiG-HOI	9.09	7.37	10.81	0.00	100.0	5.71
VLM Zero-shot evaluation: V-COCO-based sub-benchmark							
InternVL2.5-38B	V-COCO	22.73	27.07	34.41	5.01	84.35	21.62
InternVL3-38B	V-COCO	31.56	32.98	42.49	11.02	**86.11**	28.20
Qwen3-VL-30B-instruct	V-COCO	-	-	-	-	-	-
Qwen2.5-VL-32B	V-COCO	**43.97**	**52.73**	**60.93**	**21.04**	80.59	**48.98**
LLaVA-OV-7B	V-COCO	-	-	-	-	-	-
InternVL3-8B	V-COCO	5.51	4.39	6.76	0.80	77.63	3.53
Qwen2-VL-7B	V-COCO	-	-	-	-	-	-
Qwen2.5-VL-7B	V-COCO	**11.88**	**12.66**	**17.43**	**0.80**	**81.50**	**9.76**
VLM Zero-shot evaluation: SWiG-HOI-based sub-benchmark							
InternVL2.5-38B	SWiG-HOI	13.69	14.65	20.91	8.86	77.93	12.07
InternVL3-38B	SWiG-HOI	28.11	30.25	38.47	18.31	**81.70**	25.16
Qwen3-VL-30B-instruct	SWiG-HOI	1.97	1.35	2.19	0.76	76.60	1.11
Qwen2.5-VL-32B	SWiG-HOI	**45.62**	**52.66**	**59.92**	**25.66**	68.31	**53.37**
LLaVA-OV-7B	SWiG-HOI	-	-	-	-	-	-
InternVL3-8B	SWiG-HOI	5.33	4.55	7.32	2.38	74.85	3.85
Qwen2-VL-7B	SWiG-HOI	-	-	-	-	-	-
Qwen2.5-VL-7B	SWiG-HOI	**39.36**	**47.95**	**54.08**	**24.53**	**70.91**	**43.70**
VLM Zero-shot evaluation: Combined (main + two sub-) benchmarks							
InternVL2.5-38B	Combined	20.67	18.58	26.82	8.44	82.34	16.02
InternVL3-38B	Combined	34.93	33.56	42.61	18.02	**83.81**	28.56
Qwen3-VL-30B-instruct	Combined	-	-	-	-	-	-
Qwen2.5-VL-32B	Combined	**48.65**	**52.77**	**60.61**	**25.17**	72.82	**51.91**
LLaVA-OV-7B	Combined	-	-	-	-	-	-
InternVL3-8B	Combined	6.42	4.66	7.56	2.04	**75.85**	3.99
Qwen2-VL-7B	Combined	-	-	-	-	-	-
Qwen2.5-VL-7B	Combined	**29.67**	**36.02**	**42.34**	**17.66**	72.92	**29.83**

Table 11. *Setting 1* experiment results comparison under two sub-benchmarks. Results are reported for VLMs and HOI-specific methods. Best performance within each group is highlighted in **bold**. "Avg. Prec." means the precision averaged across test set and "Avg. Rec." means the recall averaged across test set.

In *Setting 3*, the prompt used is

```
Question: Which of the following
properly describes the interactions
in the image ⟨image⟩? Choices:
(A) ...,(B) ...,(C) ...,(D)...
IMPORTANT: Reply with the letter(s)
ONLY, separated by commas if
multiple (e.g. A,B). For example,
if correct answers are (A) and
(B), your output must be: A,B Do
NOT include any brackets or other
symbols.
```

In *Setting 1*, a VLM is required to predict the target person or human-object bounding boxes before the choice selection. We use prompt to localize humans or objects before HOI prediction. Specifically, we provide an image to the model together with a text prompt asking it to output bounding boxes in JSON format. For person detection only, the following prompt is used:

```
Provide the bounding box
coordinates for every single person
in the input image.
The box coordinates represent as
[x1, y1, x2, y2], where x is the
horizontal pixel coordinate from
the left edge, and y is the vertical
pixel coordinate from the top edge.
Return the detection results in JSON
format strictly.
For example:
{ "boxes": [[32, 109, 644, 418],
[517, 0, 644, 23], [100, 50, 160,
200]] }
```

For detecting both persons and objects, we modify the prompt to:

Method	Dataset	Macro-F1 (%)	Instance-F1 (%)	Micro-F1 (%)	EM (%)	Avg. Prec. (%)	Avg. Rec. (%)
HOI-specific methods Out-of-distribution evaluation							
ADA-CM	-	-	-	-	-	-	-
CMMP	V-COCO	24.93	22.61	34.84	4.01	80.76	22.21
LAIN	V-COCO	12.42	14.28	23.61	3.21	66.54	14.35
HOLa	-	-	-	-	-	-	-
CMD-SE	SWiG-HOI	16..67	5.26	5.56	0.05	100.0	2.86
VLM Zero-shot evaluation: V-COCO-based sub-benchmark							
InternVL2.5-38B	V-COCO	5.83	5.88	6.58	2.00	71.67	3.45
InternVL3-38B	V-COCO	**10.87**	**10.07**	**14.57**	**4.01**	**72.66**	**8.10**
Qwen3-VL-30B-instruct	V-COCO	-	-	-	-	-	-
Qwen2.5-VL-32B	V-COCO	1.30	3.93	3.68	0.80	41.38	1.92
LLaVA-OV-7B	V-COCO	-	-	-	-	-	-
InternVL3-8B	V-COCO	-	-	-	-	-	-
Qwen2-VL-7B	V-COCO	-	-	-	-	-	-
Qwen2.5-VL-7B	V-COCO	**3.05**	**2.87**	**2.19**	**0.40**	**48.28**	**1.12**
VLM Zero-shot evaluation: SWiG-HOI-based sub-benchmark							
InternVL2.5-38B	SWiG-HOI	10.39	7.82	12.10	4.75	78.02	6.56
InternVL3-38B	SWiG-HOI	20.12	22.75	31.69	13.61	78.28	19.86
Qwen3-VL-30B-instruct	SWiG-HOI	-	-	-	-	-	-
Qwen2.5-VL-32B	SWiG-HOI	**38.48**	**42.72**	**52.81**	**21.02**	**66.92**	**43.61**
LLaVA-OV-7B	SWiG-HOI	-	-	-	-	-	-
InternVL3-8B	SWiG-HOI	2.43	2.43	4.07	1.24	**73.12**	2.09
Qwen2-VL-7B	SWiG-HOI	-	-	-	-	-	-
Qwen2.5-VL-7B	SWiG-HOI	**32.43**	**36.61**	**45.62**	**18.04**	70.63	**33.69**
VLM Zero-shot evaluation: Combined (main + two sub-) benchmarks							
InternVL2.5-38B	Combined	6.14	6.36	9.43	3.53	77.06	5.02
InternVL3-38B	Combined	18.70	18.46	26.49	10.60	**77.88**	15.96
Qwen3-VL-30B-instruct	Combined	-	-	-	-	-	-
Qwen2.5-VL-32B	Combined	**19.78**	**25.52**	**36.63**	**12.64**	67.58	**25.12**
LLaVA-OV-7B	Combined	-	-	-	-	-	-
InternVL3-8B	Combined	-	-	-	-	-	-
Qwen2-VL-7B	Combined	-	-	-	-	-	-
Qwen2.5-VL-7B	Combined	**16.17**	**21.91**	**30.26**	**10.93**	**69.97**	**19.31**

Table 12. Extended *Setting 1* experiment results comparison, derived from the V-COCO and SWiG-HOI images with 2499 questions in total. Results are reported for VLMs and HOI-specific methods. Best performance within each group is highlighted in **bold**. "Avg. Prec." means the precision averaged across test set and "Avg. Rec." means the recall averaged across test set.

```
 Provide the bounding box
coordinates for all visible objects
and humans in the input image
based on the following object list:
{OBJ_IDX_TO_OBJ_NAME}.
The box coordinates represent as
[x1, y1, x2, y2], where x is the
horizontal pixel coordinate from
the left edge, and y is the vertical
pixel coordinate from the top edge.
Return the detection results in JSON
format strictly.
For example:
{ "boxes": [[32, 109, 644, 418],
[517, 0, 644, 23], [100, 50, 160,
200]],
"labels": ["person", "bench",
"cup"] }
Only include objects from the given
list. Ensure the output is a valid
JSON dictionary without additional
comments, and that the lengths of
the boxes and labels arrays are
equal.
```

Here, `OBJ_IDX_TO_OBJ_NAME` denotes the 80 predefined object classes in HICO-DET. After obtaining the detection results, we compute the Intersection-over-Union (IoU) between the predicted boxes and the ground-truth boxes. Predictions with IoU greater than 0.5 are considered correct and passed to the HOI choice selection step, where the detected boxes are used to localize the corresponding human or human–object pair.

Bounding Box Process for VLMs In *Setting 1* and *Setting 2*, the input requires bounding boxes of the target person. Since different VLMs preprocess images in different ways, we adapt the bounding boxes accordingly to ensure consis-

Method	Dataset	Macro-F1 (%)	Instance-F1 (%)	Micro-F1 (%)	EM (%)	Avg. Prec. (%)	Avg. Rec. (%)
VLM Zero-shot evaluation: V-COCO-based sub-benchmark							
InternVL2.5-38B	V-COCO	38.46	48.43	52.39	8.42	83.91	38.08
InternVL3-38B	V-COCO	48.48	62.19	64.03	17.23	**85.10**	51.32
Qwen3-VL-30B-instruct	V-COCO	41.36	55.99	57.48	8.42	84.36	43.59
Qwen2.5-VL-32B	V-COCO	**55.61**	**71.18**	**72.85**	**28.86**	82.34	**65.33**
LLaVA-OV-7B	V-COCO	30.82	35.75	37.79	0.60	82.00	24.55
InternVL3-8B	V-COCO	37.78	48.73	51.12	5.41	83.24	36.89
Qwen2-VL-7B	V-COCO	34.49	48.76	48.07	1.80	**85.21**	33.47
Qwen2.5-VL-7B	V-COCO	**42.65**	**54.76**	**55.56**	**6.21**	82.35	**41.92**
VLM Zero-shot evaluation:SWiG-HOI-based sub-benchmark							
InternVL2.5-38B	SWiG-HOI	42.64	51.45	55.39	31.06	81.51	41.95
InternVL3-38B	SWiG-HOI	51.78	69.73	67.96	43.11	**82.30**	57.87
Qwen3-VL-30B-instruct	SWiG-HOI	**57.16**	**74.87**	**73.62**	**45.60**	82.22	66.65
Qwen2.5-VL-32B	SWiG-HOI	52.74	69.30	69.69	32.52	68.90	**70.50**
LLaVA-OV-7B	SWiG-HOI	43.23	62.26	58.61	37.93	80.71	46.01
InternVL3-8B	SWiG-HOI	44.83	52.76	55.96	31.98	78.06	43.61
Qwen2-VL-7B	SWiG-HOI	46.34	62.80	60.84	**38.36**	**84.63**	47.49
Qwen2.5-VL-7B	SWiG-HOI	**50.38**	**67.08**	**66.36**	34.52	72.13	**61.44**
VLM Zero-shot evaluation: Combined (main + two sub-) benchmarks							
InternVL2.5-38B	Combined	42.52	49.24	53.48	24.28	80.72	39.99
InternVL3-38B	Combined	55.53	67.66	67.04	**36.92**	**82.74**	56.35
Qwen3-VL-30B-instruct	Combined	54.22	67.84	67.34	35.46	81.79	57.22
Qwen2.5-VL-32B	Combined	**59.10**	**69.70**	**70.68**	32.89	73.52	**68.06**
LLaVA-OV-7B	Combined	44.76	55.78	53.19	**28.28**	79.74	39.91
InternVL3-8B	Combined	45.83	52.73	55.43	26.05	78.33	42.89
Qwen2-VL-7B	Combined	44.56	57.40	55.77	28.01	**82.13**	42.22
Qwen2.5-VL-7B	Combined	**48.27**	**61.65**	**61.43**	27.62	74.52	**52.26**

Table 13. *Setting 2* experiment results comparison, derived from the V-COCO and SWiG-HOI images with 2499 questions in total. Results are reported for VLMs and HOI-specific methods. Best performance within each group is highlighted in **bold**. "Avg. Prec." means the precision averaged across test set and "Avg. Rec." means the recall averaged across test set.

tency with the model input. Specifically, Qwen2/2.5-VL resizes input images such that both height and width are multiples of 14, while Qwen3-VL adjusts images to multiples of 16. We therefore resize the bounding boxes proportionally to the resized image coordinates. InternVL2.5/3 does not fix the image size but internally normalizes it. To align with its view of the image, we first query the model with a prompt asking for the perceived input resolution:

```
Please provide the coordinates
for the bottom-right point of the
input image. Assume the coordinate
system origin is at the top-left of
the image, with x increasing to the
right and y increasing downward.
Return the coordinates as [width,
height] in JSON format strictly.
For example: [638, 415].
```

Based on its returned size, we then rescale the bounding boxes into that coordinate system. LLaVA-OV takes the original image size directly as input. In this case, we use the original bounding boxes without additional processing. This preprocessing ensures that the bounding boxes we provide are always aligned with how each model internally processes the input image.

Prompt for Coarse Screening We provide the prompt template used for the coarse screening stage. This stage serves as an initial screening to identify negative candidates before applying the fine-grained manual refinement.

Below is the general template we used for GPT-4.1 to separate semantically consistent or inconsistent candidates:

```
You are given an image and a human
bounding box: ⟨human box⟩.
A list of candidate interactions is
provided: ⟨HOI candidates⟩.
The ground-truth interaction is:
⟨annotated ground-truth HOI classes⟩.
Find interactions that are clearly
different and unrelated to the
ground-truth interaction in the
image.
Any visually or semantically similar
interactions (e.g., synonyms,
paraphrases, same action with
different wording) must NOT be
selected.
Return ONLY the NEGATIVE group as a
JSON list of action+object phrases.
```

Method	Dataset	Macro-F1 (%)	Instance-F1 (%)	Micro-F1 (%)	EM (%)	Avg. Prec. (%)	Avg. Rec. (%)
VLM Zero-shot evaluation: V-COCO-based sub-benchmark							
InternVL2.5-38B	V-COCO	40.72	45.36	48.77	20.84	**73.51**	36.49
InternVL3-38B	V-COCO	**47.22**	60.55	62.45	**29.66**	72.21	55.01
Qwen3-VL-30B-instruct	V-COCO	39.72	53.33	55.65	23.25	67.39	47.39
Qwen2.5-VL-32B	V-COCO	44.76	**63.14**	**64.80**	28.26	67.78	**62.07**
LLaVA-OV-7B	V-COCO	26.10	34.17	37.64	12.02	65.67	26.38
InternVL3-8B	V-COCO	**39.20**	**49.51**	**52.55**	**24.05**	66.63	**43.38**
Qwen2-VL-7B	V-COCO	31.32	45.55	46.60	16.03	**68.00**	35.45
Qwen2.5-VL-7B	V-COCO	38.00	48.97	50.61	18.04	64.75	41.54
VLM Zero-shot evaluation: SWiG-HOI-sub-benchmark							
InternVL2.5-38B	SWiG-HOI	42.84	53.97	57.19	31.98	79.42	44.69
InternVL3-38B	SWiG-HOI	48.94	68.83	67.80	40.90	**80.46**	58.58
Qwen3-VL-30B-instruct	SWiG-HOI	**58.51**	**75.70**	**74.53**	**46.03**	81.42	68.71
Qwen2.5-VL-32B	SWiG-HOI	55.48	69.37	70.14	32.47	68.49	**71.88**
LLaVA-OV-7B	SWiG-HOI	44.22	62.77	59.03	38.36	81.34	46.32
InternVL3-8B	SWiG-HOI	45.22	52.56	56.78	30.96	78.52	44.47
Qwen2-VL-7B	SWiG-HOI	47.93	**67.50**	63.29	**42.57**	**84.99**	50.42
Qwen2.5-VL-7B	SWiG-HOI	**50.44**	66.67	**65.75**	32.52	71.12	**61.13**
VLM Zero-shot evaluation: Combined (main + two sub-) benchmarks							
InternVL2.5-38B	Combined	42.60	50.73	54.28	28.42	77.46	41.78
InternVL3-38B	Combined	53.23	66.19	66.17	37.03	**77.59**	57.68
Qwen3-VL-30B-instruct	Combined	54.21	**67.14**	67.27	**37.53**	76.86	59.80
Qwen2.5-VL-32B	Combined	**56.75**	67.09	**68.14**	31.40	68.93	**67.37**
LLaVA-OV-7B	Combined	45.18	55.87	54.18	31.26	77.07	41.77
InternVL3-8B	Combined	46.38	52.39	55.86	27.90	73.61	45.01
Qwen2-VL-7B	Combined	45.49	59.26	57.03	**33.11**	**77.92**	44.97
Qwen2.5-VL-7B	Combined	**47.36**	**60.03**	**60.18**	28.37	69.91	**52.82**

Table 14. Extended *Setting 2* experiment results comparison, derived from the V-COCO and SWiG-HOI images with 2499 questions in total. Results are reported for VLMs and HOI-specific methods. Best performance within each group is highlighted in **bold**. "Avg. Prec." means the precision averaged across test set and "Avg. Rec." means the recall averaged across test set.

```
Do not include explanations. Limit
to under 50 words.
Examples:
- GT: "ride a/an bike"
- Negative: ["inspect a/an bike",
"repair a/an bike", "wash a/an
bike"]
Now, based on the image and the
ground-truth interaction, return the
negative group from the candidate
list.
```

Then, we construct prompts for Qwen2.5-VL-32B and GPT-4o to re-evaluate the inconsistent candidates returned by GPT-4.1, ensuring that these cases are indeed true negatives.

```
You are given an image and a human
bounding box: ⟨human box⟩.
Definitions:
- Positive: Interactions that
appear in the image or are
semantically/visually related or
they may occur simultaneously.
- Negative: Interactions that are
clearly different or unrelated.
Question:
Is the interaction ⟨an HOI negative
candidate⟩ positive or negative with
respect to the image?
Please answer in the following
format:
Answer: Positive
or
Answer: Negative
```

This careful two-stage verification ensures that our coarse screening does not introduce bias to VLMs. As shown in Table 4 of the main paper, HOI-specific models even more agree with the negatives identified through this process than other VLM baselines, confirming that only clear negative HOIs, validated independently by three VLMs, are selected.

3. Evaluation Metrics Details

Let Q denote the set of all evaluation questions. For each question $q \in Q$, let P_q be the set of predicted interaction labels and G_q the ground-truth set of positive choices. Macro-F1 evaluates performance in a class-balanced manner. Let

Method	Dataset	Macro-F1 (%)	Instance-F1 (%)	Micro-F1 (%)	EM (%)	Avg. Prec. (%)	Avg. Rec. (%)
HOI-specific methods Out-of-distribution evaluation							
ADA-CM	-	-	-	-	-	-	-
CMMP	V-COCO	29.68	46.50	52.58	2.81	77.46	39.80
LAIN	V-COCO	15.35	28.49	35.11	3.81	65.49	23.99
HOLa	-	-	-	-	-	-	-
CMD-SE	SWiG-HOI	6.67	7.37	10.81	0.00	100.0	5.71
VLM Zero-shot evaluation: V-COCO-based sub-benchmark							
InternVL2.5-38B	V-COCO	38.58	50.76	55.05	8.02	78.31	42.45
InternVL3-38B	V-COCO	45.43	59.07	61.49	11.62	78.03	50.73
Qwen3-VL-30B-instruct	V-COCO	45.31	60.57	61.66	11.22	**81.75**	49.50
Qwen2.5-VL-32B	V-COCO	**47.17**	**66.83**	**68.81**	**16.83**	75.02	**63.55**
LLaVA-OV-7B	V-COCO	29.67	36.55	38.26	0.60	**84.20**	24.75
InternVL3-8B	V-COCO	**45.43**	**59.07**	**61.49**	**11.62**	78.03	**50.73**
Qwen2-VL-7B	V-COCO	38.27	51.66	52.11	5.01	84.14	37.74
Qwen2.5-VL-7B	V-COCO	43.34	55.59	57.38	8.62	79.79	44.80
VLM Zero-shot evaluation: SWiG-HOI-based sub-benchmark							
InternVL2.5-38B	SWiG-HOI	45.82	59.27	61.26	34.41	79.82	49.70
InternVL3-38B	SWiG-HOI	49.15	69.22	67.49	**41.11**	**82.50**	57.11
Qwen3-VL-30B-instruct	SWiG-HOI	**56.94**	**75.08**	**73.91**	40.73	78.05	70.19
Qwen2.5-VL-32B	SWiG-HOI	53.05	69.60	70.42	32.36	68.93	**71.99**
LLaVA-OV-7B	SWiG-HOI	43.11	64.12	59.66	38.68	**83.61**	46.37
InternVL3-8B	SWiG-HOI	**49.15**	**69.22**	**67.49**	**41.11**	82.50	**57.11**
Qwen2-VL-7B	SWiG-HOI	38.40	44.80	52.93	22.42	80.21	39.50
Qwen2.5-VL-7B	SWiG-HOI	48.97	67.34	65.94	34.47	72.90	60.20
VLM Zero-shot evaluation: Combined (main + two sub-) benchmarks							
InternVL2.5-38B	Combined	47.41	55.29	57.86	24.75	80.99	45.01
InternVL3-38B	Combined	54.77	65.44	64.72	30.57	**83.01**	53.04
Qwen3-VL-30B-instruct	Combined	57.01	**69.87**	69.18	**30.71**	81.29	60.21
Qwen2.5-VL-32B	Combined	**59.15**	68.31	**69.54**	26.99	72.97	**66.41**
LLaVA-OV-7B	Combined	43.79	55.99	52.88	24.39	**84.28**	38.53
InternVL3-8B	Combined	**52.17**	**63.37**	**62.78**	**28.42**	81.77	**50.95**
Qwen2-VL-7B	Combined	34.94	43.31	48.89	14.65	81.83	34.86
Qwen2.5-VL-7B	Combined	50.16	61.50	60.87	24.01	76.29	50.63

Table 15. *Setting 3* experiment results comparison, derived from the V-COCO and SWiG-HOI images with 2499 questions in total. Results are reported for VLMs and HOI-specific methods. Best performance within each group is highlighted in **bold**. "Avg. Prec." means the precision averaged across test set and "Avg. Rec." means the recall averaged across test set.

$\mathcal{C}$ denote the set of HOI classes. For each class $c \in \mathcal{C}$, we compute the F1-score over all questions involving c, denoted as F1_c. Macro-F1 is then obtained by averaging F1_c across all classes.

$$\text{Macro-F1} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{2 \sum_q \mathbf{1}[c \in P_q \cap G_q]}{\sum_q \mathbf{1}[c \in P_q] + \sum_q \mathbf{1}[c \in G_q]}, \tag{1}$$

where $\mathbf{1}[\cdot]$ is the indicator function, which equals 1 if the condition is true and 0 otherwise.

Instance-F1 measures performance at the question level. For each q, we compute the F1-score between P_q and G_q, and then average over all questions to obtain the overall score:

$$\text{Instance-F1} = \frac{1}{|Q|} \sum_{q \in Q} \text{F1}(q) = \frac{1}{|Q|} \sum_{q \in Q} \frac{2|P_q \cap G_q|}{|P_q| + |G_q|}. \tag{2}$$

Here $|\cdot|$ denotes set cardinality, and $P_q \cap G_q$ is the set of correctly predicted labels for question q.

Micro-F1 measures overall performance by aggregating predictions across all questions and computing a single F1-score from the total number of predicted and ground-truth labels:

$$\text{Micro-F1} = \frac{2 \sum_q |P_q \cap G_q|}{\sum_q |P_q| + \sum_q |G_q|}. \tag{3}$$

Finally, we adopt Exact Match Accuracy (EM), which checks whether the predicted interaction set for a question exactly matches the ground-truth set. Unlike the exact-match mAP metric in traditional HOI benchmarks, which is often affected by incomplete annotations and penalizes unlabeled interactions, our multiple-choice design mitigates this issue through curated negatives. Thus, EM provides a complementary measure of strict correctness: it reports how

Method	Setting	Macro-F1 (%)	Instance-F1 (%)	Micro-F1 (%)	EM (%)	Avg. Prec. (%)	Avg. Rec. (%)
Main Benchmark							
SFT Baseline	1	61.12	61.70	69.63	26.22	67.69	71.67
SFT Baseline	Extended 1	41.86	34.90	53.17	14.21	66.04	44.51
SFT Baseline	2	63.64	68.93	72.57	28.89	66.88	79.31
SFT Baseline	Extended 2	62.90	67.85	71.28	27.71	62.75	82.49
SFT Baseline	3	70.14	75.39	78.62	32.10	77.08	80.72
Sub Benchmarks: V-COCO-based							
SFT Baseline	1	55.44	71.31	77.60	34.27	74.46	81.02
SFT Baseline	Extended 1	24.91	24.92	44.10	8.42	62.72	34.00
SFT Baseline	2	58.87	80.29	81.60	36.87	73.82	91.20
SFT Baseline	Extended 2	46.72	67.06	70.60	25.05	56.89	93.02
SFT Baseline	3	57.18	80.98	82.17	36.47	74.00	92.36
Sub Benchmarks: SWiG-HOI-based							
SFT Baseline	1	49.63	55.78	64.81	21.23	57.55	74.16
SFT Baseline	Extended 1	40.55	44.67	57.71	15.88	56.84	58.61
SFT Baseline	2	57.48	70.32	72.39	27.28	59.24	93.04
SFT Baseline	Extended 2	57.16	70.37	72.44	27.82	59.32	93.01
SFT Baseline	3	59.72	70.27	72.52	26.04	59.28	93.37

Table 16. Experiment results comparison when Qwen2.5-VL-7B is fine-tuned on HOI datasets [3] using Supervised Fine-Tuning (SFT). "Avg. Prec." means the precision averaged across test set and "Avg. Rec." means the recall averaged across test set.

Method	Macro-F1 (%)	Instance-F1 (%)	Micro-F1 (%)	EM (%)	Avg. Prec. (%)	Avg. Rec. (%)
HOI-specific methods						
ADA-CM	58.54	68.49	**79.95**	26.15	90.80	71.41
CMMP	58.49	68.00	79.51	25.84	89.66	71.43
LAIN	55.51	66.32	77.84	25.79	87.50	70.11
HOLa	**58.83**	68.41	79.53	26.57	89.51	**71.56**
CMD-SE	52.36	**69.73**	79.80	**32.83**	**90.81**	71.18
VLM zero-shot evaluation						
InternVL2.5-38B	65.41	68.08	68.58	41.55	93.30	54.22
InternVL3-38B	71.32	76.05	72.41	46.55	**93.45**	59.10
Qwen2.5-VL-32B	**74.89**	**82.15**	**80.05**	**53.15**	89.25	**72.57**
LLaVA-OV-7B	62.48	69.04	63.87	36.93	92.57	48.76
InternVL3-8B	**71.92**	**78.21**	**75.49**	**50.74**	**93.86**	**63.13**
Qwen2-VL-7B	44.77	42.84	49.05	19.24	93.57	33.24
Qwen2.5-VL-7B	68.49	74.10	69.53	42.55	93.11	55.48

Table 17. *Setting 1* experiment results comparison, derived from the whole HICO-DET images with 9546 questions in total. Results are reported for VLMs and HOI-specific methods. Best performance within each group is highlighted in **bold**. "Avg. Prec." means the precision averaged across test set and "Avg. Rec." means the recall averaged across test set.

often the model's predictions are entirely correct.

$$\text{EM} = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}[P_q = G_q]. \tag{4}$$

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 8

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan 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,

Method	Macro-F1 (%)	Instance-F1 (%)	Micro-F1 (%)	EM (%)	Avg. Prec. (%)	Avg. Rec. (%)
VLM zero-shot evaluation						
InternVL2.5-38B	59.91	61.82	63.97	38.07	92.18	48.98
InternVL3-38B	71.41	78.72	76.99	51.99	**93.58**	65.39
Qwen2.5-VL-32B	**75.76**	**83.85**	**83.01**	**58.27**	91.02	**76.29**
LLaVA-OV-7B	64.28	70.75	65.35	40.60	92.42	50.54
InternVL3-8B	64.40	67.55	67.63	42.34	90.26	54.07
Qwen2-VL-7B	61.29	67.06	63.94	37.94	92.23	48.94
Qwen2.5-VL-7B	**67.23**	**74.33**	**70.59**	**45.06**	**92.69**	**57.00**

Table 18. *Setting 2* experiment results comparison, derived from the whole HICO-DET images with 9546 questions in total. Results are reported for VLMs and HOI-specific methods. Best performance within each group is highlighted in **bold**. "Avg. Prec." means the precision averaged across test set and "Avg. Rec." means the recall averaged across test set.

Method	Macro-F1 (%)	Instance-F1 (%)	Micro-F1 (%)	EM (%)	Avg. Prec. (%)	Avg. Rec. (%)
HOI-specific methods						
ADA-CM	56.03	62.08	**76.66**	38.68	90.32	66.59
CMMP	55.87	61.47	76.11	38.35	89.03	66.46
LAIN	53.33	60.42	74.98	37.80	87.24	65.73
HOLa	**56.69**	**62.32**	76.54	**39.69**	88.87	**67.22**
CMD-SE	51.82	58.81	73.72	38.60	**90.99**	61.97
VLM zero-shot evaluation						
InternVL2.5-38B	34.43	28.29	36.54	17.31	93.09	22.73
InternVL3-38B	49.39	43.18	52.87	28.82	**94.27**	36.73
Qwen2.5-VL-32B	**64.24**	**65.18**	**73.00**	**45.09**	91.39	**60.77**
LLaVA-OV-7B	-	-	-	-	-	-
InternVL3-8B	12.98	8.68	13.36	5.48	93.32	7.20
Qwen2-VL-7B	-	-	-	-	-	-
Qwen2.5-VL-7B	**43.37**	**38.94**	**45.81**	**23.30**	**93.32**	**30.36**

Table 19. *Setting 3* experiment results comparison, derived from the whole HICO-DET images with 9546 questions in total. Results are reported for VLMs and HOI-specific methods. Best performance within each group is highlighted in **bold**. "Avg. Prec." means the precision averaged across test set and "Avg. Rec." means the recall averaged across test set.

and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025. 6, 7, 8, 10, 13

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 9, 11, 12, 19, 21

[4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 2, 3, 5, 6

[5] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. On scaling up a multilingual vision and language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14432–14444, 2024. 8

[6] Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. Multi-object hallucination in vision language models. *Advances in Neural Information Processing Systems*, 37:44393–44418, 2024. 7

[7] 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*, pages 24185–24198, 2024. 6, 10

[8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023. 8

[9] Shounak Datta and Dhanasekar Sundararaman. Evaluating hallucination in large vision-language models based on context-aware object similarities. *arXiv preprint arXiv:2501.15046*, 2025. 7

[10] Xueqing Deng, Linjie Yang, Qihang Yu, Ali Athar,

Train set	Setting	Macro-F1 (%)	Instance-F1 (%)	Micro-F1 (%)	EM (%)	Avg. Prec. (%)	Avg. Rec. (%)
Ours	1	87.75	94.60	94.64	85.02	94.58	94.69
V-COCO-based	1	75.12	83.41	81.85	53.45	87.35	77.00
Ours	2	85.88	92.87	93.57	83.56	92.48	94.70
V-COCO-based	2	72.94	81.54	81.57	52.91	85.40	78.08
Ours	3	81.72	81.78	89.51	73.64	93.12	86.18
V-COCO-based	3	68.32	69.83	76.81	44.91	86.26	69.22

Table 20. Experiment results comparison, where the test set is derived from the whole HICO-DET images with 9546 questions in total. Qwen2.5-VL-7B is fine-tuned on HOI datasets [3] using Supervised Fine-Tuning (SFT). "Avg. Prec." means the precision averaged across test set and "Avg. Rec." means the recall averaged across test set.

Chenglin Yang, Xiaojie Jin, Xiaohui Shen, and Liang-Chieh Chen. COCONut-pancap: Joint panoptic segmentation and grounded captions for fine-grained understanding and generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. 2

[11] Hongyuan Dong, Zijian Kang, Weijie Yin, Xiao Liang, Chao Feng, and Jiao Ran. Scalable vision language model training via high quality data curation. *arXiv preprint arXiv:2501.05952*, 2025. 8

[12] Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 7

[13] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 6

[14] 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)*, pages 19034–19043, 2022. 2, 3

[15] 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)*, pages 74–83, 2021. 7

[16] Sanghyun Kim, Deunsol Jung, and Minsu Cho. Locality-aware zero-shot human-object interaction detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 20190–20200. Computer Vision Foundation / IEEE, 2025. 6, 7

[17] Qinqian Lei, Bo Wang, and Robby T. Tan. Ez-hoi: Vlm adaptation via guided prompt learning for zero-shot hoi detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 7

[18] Qinqian Lei, Bo Wang, and Tan Robby T. Hola: Zero-shot hoi detection with low-rank decomposed vlm feature adaptation. In *In Proceedings of the IEEE/CVF international conference on computer vision*, 2025. 6

[19] Ting Lei, Fabian Caba, Qingchao Chen, Hailin Jin, Yuxin Peng, and Yang Liu. Efficient adaptive human-object interaction detection with concept-guided memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6480–6490, 2023. 2, 6, 7, 9, 10

[20] Ting Lei, Shaofeng Yin, and Yang Liu. Exploring the potential of large foundation models for open-vocabulary hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16657–16667, 2024. 7

[21] Ting Lei, Shaofeng Yin, and Yang Liu. Exploring the potential of large foundation models for open-vocabulary hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16657–16667, 2024. 6

[22] 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*, pages 1–19. Springer, 2024. 6, 7, 8, 10

[23] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025, 2025. 6, 10

[24] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 3, 5

[25] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025. 8

[26] Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23507–23517, 2023. 7

[27] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 8

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning

transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[29] Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. Scaling vision-language models with sparse mixture of experts. *arXiv preprint arXiv:2303.07226*, 2023. 8

[30] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021. 7

[31] Yen-Linh Vu, Dinh-Thang Duong, Truong-Binh Duong, Anh-Khoi Nguyen, Thanh-Huy Nguyen, Le Thien Phuc Nguyen, Jianhua Xing, Xingjian Li, Tianyang Wang, Ulas Bagci, and Min Xu. Describe anything model for visual question answering on text-rich images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 7484–7494, 2025. 2

[32] Guangzhi Wang, Yangyang Guo, Ziwei Xu, and Mohan Kankanhalli. Bilateral adaptation for human-object interaction detection with occlusion-robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27970–27980, 2024. 9

[33] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 10

[34] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2022. 2, 3, 5, 6

[35] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[36] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024. 6

[37] Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. Vision-flan: Scaling human-labeled tasks in visual instruction tuning. *arXiv preprint arXiv:2402.11690*, 2024. 8

[38] Zhenlin Xu, Yi Zhu, Siqi Deng, Abhay Mittal, Yanbei Chen, Manchen Wang, Paolo Favaro, Joseph Tighe, and Davide Modolo. Benchmarking zero-shot recognition with vision-language models: Challenges on granularity and specificity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1827–1836, 2024. 7

[39] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13319–13327, 2021. 10

[40] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20104–20112, 2022. 2, 9, 10

[41] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*, abs/2504.10479, 2025. 6, 7, 10