Supplementary Material for SHOE - Semantic HOI Open-vocabulary Evaluation metric

Maja Noack[1] Qinqian Lei[2] Taipeng Tian[3] Bihan Dong[1]
Robby T. Tan[2,4] Yixin Chen[1] John Young[1] Saijun Zhang[1] Bo Wang[1]
[1]University of Mississippi [2]National University of Singapore
[3]Independent Researcher [4]ASUS Intelligent Cloud Services (AICS)
https://github.com/majnoa/SHOE

1. LLM Ablation for SHOE-Metric

To assess the sensitivity of SHOE to the specific set of LLMs (Qwen3-32B [25], DeepSeek-V3-0324 [17], Llama-4-Maverick-17B [22], Yi-1.5-34B-Chat [26], and Gemini-2.5-pro [6]) used to compute the final similarity score and whether a smaller or differently composed ensemble could achieve comparable alignment with human judgments, we evaluated all possible subsets of the five LLMs used in our ensemble, ranging from single-model configurations to the full five-model setup (Table 1). Across all configurations, agreement stays within a narrow 0.84–0.86 range. Single models already achieve strong alignment, though performance varies (e.g., DeepSeek-V3 is highest at 86.46 %; Maverick lowest at 84.11 %). Adding a second model provides only marginal benefit: most two-model combinations match the performance of the stronger individual model, indicating limited complementary information at this scale. Larger ensembles (3–4 models) modestly reduce variance but do not substantially increase the average agreement.

Although the gains in average agreement are modest, we retain the ensemble SHOE score as it may reduce the influence of model-specific biases. Our user study did not show strong trends of any of the included models but is limited to 500 pairs of HOIs. For datasets that focus on more specialized domains, individual model biases are more likely to surface, and the ensemble may provide a more balanced estimate. Overall, the ablation shows that SHOE is largely insensitive to the exact choice of LLMs. As long as models of comparable capability are included, the resulting similarity scores remain stable. As LLMs evolve, SHOE can adopt newer models without changing its characteristic behavior.

2. Open-Vocabulary Evaluation of HOI Models

We compare Vision-Language Models (VLMs) with state-of-the-art HOI methods in Table 2. As VLMs cannot reliably provide detection bounding boxes or confidence scores for HOI class prediction, we use the off-the-shelf object detection, DETR [3], commonly adopted in two-stage HOI detection methods, and VLMs' token probabilities for mAP-style ranked evaluation. However, not all VLMs provide token probabilities. Additionally, using DETR as an initial object detector omits the VLMs' capabilities to detect human, object and interaction, as it only has to find the interaction for the already predefined human-object pair. To extend our study, we report results for the confidence-free evaluation mode (i.e., category (2) confidence-free mode, see Sec. 3.2 in the main paper) for all models. We add Gemini-2.5-flash [6] and LLaVa-Onevision-72b [18] in this comparison as token-wise probabilities are not needed for this evaluation setting. We present the GT miss rate (false negatives), prediction miss rate (false positives), and SHOE mF1 for all the models included in our evaluation. On the single-HOI HICO-DET [4] subset (see Table 2). We define:

- **GT Miss Rate (FN %)**: The percentage of ground-truth HOI instances that are not matched by any prediction (false negatives):

$$\text{GT Miss Rate} = \frac{|\text{FN}|}{|\text{GT}|} \times 100$$

  where $|\text{GT}|$ is the number of ground-truth HOI instances, and $|\text{FN}|$ is the number of ground-truth instances with no matched prediction (e.g., under the chosen IoU and similarity threshold).
- **Prediction Miss Rate (FP %)**: The percentage of predicted HOI instances that do not match any ground-truth instance (false positives):

$$\text{Prediction Miss Rate} = \frac{|\text{FP}|}{|\text{Pred}|} \times 100$$

  where $|\text{Pred}|$ is the total number of predictions, and $|\text{FP}|$ is the number of predictions with no matched ground-truth.

#	DS	G.	Mv.	Qw.	Yi	Agree.%
1	1	0	0	0	0	86.46
1	0	1	0	0	0	84.56
1	0	0	1	0	0	84.11
1	0	0	0	1	0	85.60
1	0	0	0	0	1	84.87
Avg.						85.12±0.92
2	1	1	0	0	0	86.46
2	1	0	1	0	0	86.46
2	1	0	0	1	0	86.46
2	1	0	0	0	1	86.46
2	0	1	1	0	0	84.56
2	0	1	0	1	0	84.56
2	0	1	0	0	1	84.56
2	0	0	1	1	0	84.58
2	0	0	1	0	1	84.15
2	0	0	0	1	1	85.60
Avg.						85.39±0.99

#	DS	G.	Mv.	Qw.	Yi	Agree.%
3	1	1	1	0	0	85.55
3	1	1	0	1	0	85.76
3	1	1	0	0	1	86.12
3	1	0	1	1	0	85.40
3	1	0	1	0	1	85.52
3	1	0	0	1	1	85.93
3	0	1	1	1	0	84.71
3	0	1	1	0	1	84.53
3	0	1	0	1	1	84.88
3	0	0	1	1	1	85.12
Avg.						85.35±0.53
4	1	1	1	1	0	85.67
4	1	1	1	0	1	85.86
4	1	1	0	1	1	86.40
4	1	0	1	1	1	85.91
4	0	1	1	1	1	85.24
Avg.						85.82±0.42
5 (SHOE)	1	1	1	1	1	85.73

Table 1. Combined LLM ablation: left = sizes 1-2, right = sizes 3-5. Agreement scores for LLMs with human annotation. 1 = model included, 0 = excluded. # = number of models included. LLMs in order: DeepSeek-V3-0324, Gemini-2.5-pro, Llama-4-Maverick-17B, Qwen3-32B, Yi-1.5-34B-Chat.

Method Type	**Model**	**GT Miss Rate (FN %)**	**Prediction Miss Rate (FP %)**	**SHOE mF1**
Default	LAIN [8] (ViT-B)	21.04	37.77	45.80
	CMMP [14] (ViT-L)	20.06	37.64	50.47
	ADA-CM [12] (ViT-L)	19.18	36.21	51.93
	HOLA [11] (ViT-L)	19.45	35.83	52.21
Zero-shot RF-UC	CMMP [14] (ViT-B)	21.15	39.13	44.84
	EZ-HOI [10] (ViT-B)	20.45	38.10	46.94
	HOLA [11] (ViT-B)	20.00	37.29	49.07
	LAIN [8] (ViT-B)	20.83	38.20	46.30
Open-Vocabulary	CMD-SE [13]	36.93	60.67	37.07
	THID [24]	47.81	78.86	28.71
VLMs & MLLMs	Qwen2.5-32B-VL [1]	30.48	60.75	54.40
	InternVL3-38B [5]	55.09	86.10	40.04
	LLaVa-Onevision-72B [15]	32.04	64.13	54.50
	GPT-4.1 [20]	27.86	55.10	53.46
	Gemini-2.5-flash [6]	34.39	59.05	51.16

Table 2. Performance comparison across different HOI models on the single HOI HICO-DET [4] subset (575 classes) using GT Miss Rate, Prediction Miss Rate and SHOE mF1 scores on HICO-DET, along with GT miss rate (false negatives) and prediction miss rate (false positives). confidence score rank $i = 0.5$ was chosen for all models that provide confidence scores to make them comparable to the VLM and MLLM methods.

For models that provided confidence scores with their predictions, we set the confidence score rank to $i = 0.5$. We report that HOLA [11] (ViT-L) achieves the highest SHOE mF1 score (52.21) among existing fully supervised methods, while maintaining low miss rates for both ground-truth and predictions. HOLA (ViT-B) performs best in the zero-shot RF-UC setting, with a SHOE mF1 score of 49.07. Open-vocabulary models such as CMD-SE [13] and THID [24] have higher miss rates and lower mF1 scores, highlighting the challenge of open-world generalization. These findings are consistent with the model ratings observed on the full HICO-DET dataset. Notably, the VLMs and MLLMs perform comparably to or even outperform the existing HOI detection methods, despite lacking HOI-specific supervision, and exhibiting high GT and prediction miss rates. This is consistent with our observation that these models tend to make fewer predictions per ground-truth instance than standard HOI methods. The strong overall performance suggests that VLMs and MLLMs produce semantically plausible interaction predictions and generalize interaction concepts via language grounding.

3. Hardware Requirements and Runtime

In the following we provide an overview of the hardware used and the runtime required to compute the object and verb SHOE-similarity matrices on HICO-DET. These matrices have to be computed only once per dataset which enables all subsequent experiments to reuse them without any additional model inference. This section is meant as a practical reference for researchers designing new datasets, illustrating the one-time computational cost required to generate verb–verb and object–object similarity matrices.

Our setup combined an on-premise server with 4× NVIDIA RTX 6000 Ada (48 GB) GPUs, cloud inference via the Lambda service for selected models, and the Gemini-2.5-Pro API. We first evaluated the full set of ≈850k verb–verb pairs with Qwen3-32B, which required ≈6 days on the RTX 6000 Ada system. From these, we extracted the ≈120k non-zero similarity pairs and ran them on DeepSeek-V3-0324, Llama-4-Maverick-17B, and Yi-1.5-34B-Chat, each completing in ≈30 h. For the 40k object–object pairs, all models required ≈6 h each. DeepSeek and Llama were executed on Lambda cloud instances with runtimes matching the on-premise hardware, while Gemini-2.5-Pro completed both verb and object evaluations under batched API requests within ≈ 28h. To reduce wall-clock time, several runs were parallelized across cloud machines. The total cost of API usage and cloud inference was ≈ 200$ for the full HICO-DET similarity evaluation.

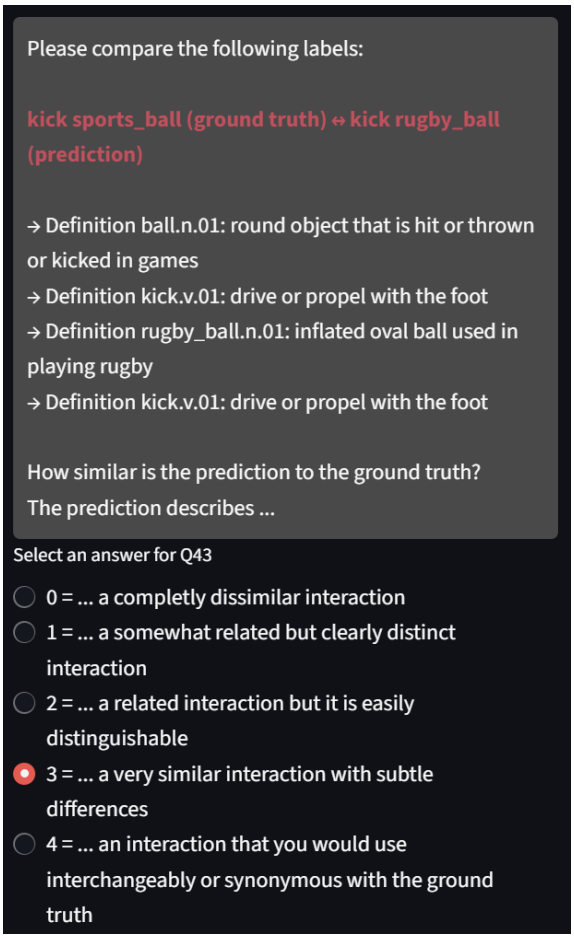


Figure 1. Example of the user study interface. Annotators are shown two HOI interactions along with their WordNet glosses and asked to rate their semantic similarity on a 5-point scale.

4. Synset Matching Procedure

Open-vocabulary predictions must first be mapped to their corresponding WordNet [19] synsets before they can be evaluated with the SHOE metric. This requires converting the model outputs into consistent lexical forms before matching them to WordNet. Predicted verbs that are not in base form are lemmatized (e.g., ”riding” to ”ride”), and object names containing underscores are split into their separate words. Each normalized lemma is then queried in WordNet to retrieve all associated synsets (i.e., possible senses). For multi-word objects such as “dining_table”, we query each component word separately and aggregate the valid synsets as a candidate pool. During evaluation, the synset from this pool with the highest semantic alignment to the ground-truth synset is selected, and this best-matching sense is used for the SHOE similarity computation.

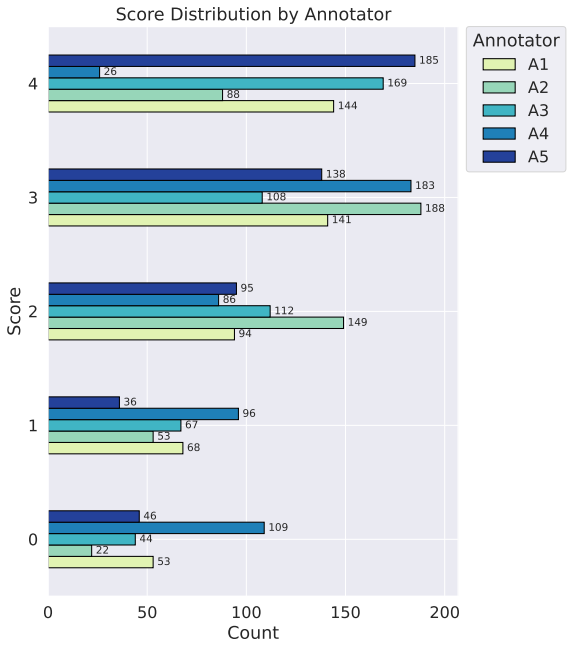


Figure 2. Per-annotator distribution of HOI similarity scores ranging from 0 (no similarity) to 4 (high similarity), showing individual rating tendencies across the annotation set.

5. User Study

5.1. User Study Setup and Examples

To better illustrate the design of our user study, we provide a representative example of the annotation interface and rating process. In each task, annotators were presented with a pair of HOI labels, along with the corresponding WordNet synsets and their glosses (definitions) for both the verb and object components. Annotators were asked to assess the semantic similarity between the two HOIs based on the information provided. Figure 1 shows an example of the annotation interface used in our study. The two HOIs are displayed along with their WordNet glosses to help annotators understand the meaning of each interaction. Annotators then assign a similarity score on a 5-point scale ranging from 0 (completely dissimilar) to 4 (interchangeable).

5.2. Inter-Annotator Score-Distribution

We analyzed the overall score distribution across annotators to examine how consistently they used the 0–4 rating scale (Figure 2). All annotators used the full range of scores, though each showed distinct rating tendencies. For example, A4 assigned score 0 more frequently than others, while A5 used score 4 most often. A2 and A3 tended to assign scores clustered around the middle of the scale. Although the evaluation criteria were consistent across annotators, their scoring patterns suggest varying degrees of strictness.

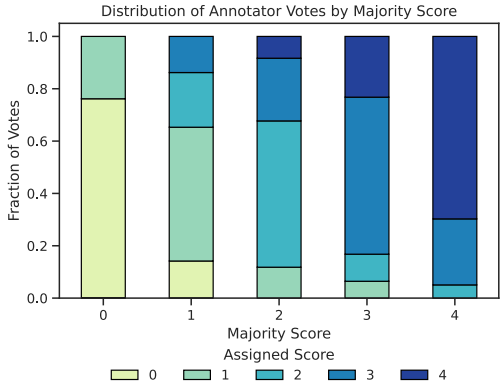


Figure 3. Distribution of Annotator Votes by Majority Scores. Each bar shows the number of times a given annotator (A1–A5) assigned each score from 0 to 4. The numeric labels on top of each bar indicate the raw count.

To better understand how annotators varied in their assignments depending on the overall consensus of an item, we grouped examples by their majority (most frequent) score and visualized the distribution of all annotator votes for each group (Figure 3). Items with a majority score of 0 or 4 showed strong agreement, with most annotators assigning the same score. In contrast, examples with mid-range majority scores (e.g., 2) showed more varied responses across neighboring scores, reflecting greater subjectivity in these cases.

5.3. Inter-Annotator Agreement Metrics

For the assessment of inter-annotator reliability we employed Quadratic Weighted Kappa (QWK) and Krippendorff's alpha that account for the ordinal structure of the data:

Quadratic Weighted Kappa. Quadratic Weighted Kappa (QWK) [7] measures agreement between two annotators, assigning partial credit for near matches. It is defined as:

$$\kappa = 1 - \frac{\sum_{i,j} w_{ij}\, O_{ij}}{\sum_{i,j} w_{ij}\, E_{ij}},$$

where O_{ij} is the observed count of examples where rater A assigned score i and rater B assigned score j, and E_{ij} is the expected count of such pairings under random chance. The weight matrix w_{ij} penalizes disagreements quadratically:

$$w_{ij} = \frac{(i-j)^2}{(k-1)^2},$$

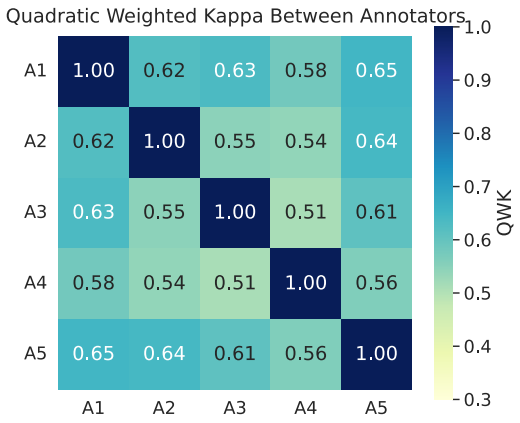


Figure 4. Pairwise Quadratic Weighted Kappa (QWK) between annotators. All values are between 0.51 and 0.65, indicating moderate agreement. Diagonal entries represent perfect self-agreement ($\kappa = 1.00$).

with $k = 5$ being the number of discrete rating levels. Figure 4 shows a moderate level of agreement overall, with stronger consistency between pairs such as A1 & A5 and A2 & A5, and slightly weaker alignment between A3 & A4. The average QWK score across all annotator pairs was 0.590, indicating moderate agreement.

Krippendorff's Alpha (Ordinal). Krippendorff's Alpha [9] generalizes agreement measurement to multiple annotators and missing values. For ordinal labels, it penalizes disagreement based on the normalized distance between scores:

$$\alpha = 1 - \frac{D_o}{D_e},$$

where D_o is the observed disagreement, and D_e is the expected disagreement under chance. The pairwise disagreement between scores i and j is calculated as:

$$\delta(i,j)^2 = \left(\frac{i-j}{k-1}\right)^2.$$

The Krippendorff's alpha for ordinal ratings was 0.566, supporting the finding of moderate consistency among annotators. These results suggest that while annotators may assign slightly different scores, they generally agree on the relative severity or quality of the examples. This indicates that the annotation process is reliable, as the ordinal structure of the labels is largely preserved across annotators.

Metric Model	Verb Agreement (%)	Object Agreement (%)
Standard SHOE	78.82	80.09
DeepSeek-V3-0324	79.89	81.21
Qwen3-32B	78.86	81.41
Yi-1.5-34B-Chat	79.01	80.61
Llama-4-Maverick-17B	77.92	80.99
Gemini-2.5-pro	78.34	81.34

Table 3. Average LLM agreement with human annotator ratings on HOI pairs where either the verb or object differs from the ground truth while the other is identical.

6. SHOE-Metric

6.1. Verb and Object based LLM agreement with human annotators

The two settings in Table 3 show that all LLMs achieve strong agreement with human annotators if either object or verb is changed while the other is the same as the GT. Object mismatches show slightly higher agreement over all LLMs than verb mismatches. DeepSeek-V3-0324 shows the highest agreement in both categories. Our SHOE score, which aggregates model predictions through majority voting, demonstrates competitive agreement in both settings and reliability for different types of prediction mismatches.

6.2. User Study with Images

Our primary user study was conducted without images in order to collect semantic similarity judgments that generalize across different scenes, not tied to any specific visual instance. Including images risks biasing annotators toward incidental scene details, such as lighting, occlusion, or unusual body poses, that have nothing to do with the semantic relationship between the interactions and are visual noise that HOI models are meant to filter out. However, as SHOE is ultimately applied to image-conditioned HOI predictions, we conducted a second user study to verify that semantic judgments remain stable when images are actually present. The annotators evaluated 500 HOI pairs, each accompanied by an image from the HICO-DET dataset corresponding to the ground-truth HOI class in the pair. They were tasked to decide if they wanted to update their original annotation or keep it the same with the image present. Across all 500 HOI pairs, only 7.65% of ratings were updated (95% CI: 6.29–9.00%). The magnitude of these changes was small, only a single point on the 0-4 point scale ($\approx$1.14 [95% CI 1.11, 1.17]). This demonstrates that the semantic judgments collected without images remain consistent even when annotators are later presented with visual context further supporting the reliability of our metric.

6.3. Statistical Basis for Participant and Annotation Counts

Our user study evaluates 500 HOI pairs through ratings from five independent annotators. To verify that this sample size is statistically adequate, we compute the Spearman correlation between model scores and human ratings. We obtain $\rho = 0.799$ with a 95% Fisher confidence interval (CI) of [0.765, 0.829]. This interval represents the range of plausible values for the true correlation if the study were repeated many times with the same sample size. The interval is narrow and strongly positive. This supports that 500 items provide a robust estimate of the true correlation. We also performed a permutation test with 10,000 resamples, which tests if such a high correlation could occur under random ratings. The result $p < 0.001$ indicates that the observed correlation is extremely unlikely to arise by chance.

To assess whether five annotators per item are sufficient, we estimate the within-item variability ($\sigma \approx 0.761$) and compute the standard error of the mean rating ($\mathrm{SE} = \frac{\sigma}{\sqrt{k}}$). Five raters achieved SE ≈ 0.34. This SE implies that the mean rating would vary by 0.34 of a point across different groups of annotators, which is small relative to the full rating range (0-4) and sufficient to distinguish meaningful differences between items. This shows that five annotators are enough to provide reliable similarity scores.

6.4. SHOE-algorithms

To make our work more reproducible we describe the implementation of the two SHOE-scoring algorithms. Our main SHOE-mAP-score algorithm is shown in Algorithm 0. For each image, predictions are filtered by a confidence threshold and matched to ground-truth instances based on the common IoU > 0.5 requirement on the human and object bounding boxes. Predictions count as soft true positives when they match spatially and have highest semantic similarity of all predictions for that GT. Predictions that do not match any ground truth get penalized through false positive penalties. For each unmatched prediction, the GT class with the highest similarity score is found. A FP is assigned to that class if the similarity score is larger than a certain threshold. The algorithm sorts predictions by confidence before it calculates true positives and false positives using soft labels for each class. Precision and recall get computed at each threshold before obtaining Average Precision (AP) through the calculation of area under the resulting precision-recall curve. The final SHOE mAP is the mean of AP scores across all classes.

The procedure for computing the SHOE mF1 score for HOI predictions using soft matching based on semantic similarity is described in Algorithm 0. If models provide confidence scores only predictions larger than a chosen threshold are included in the calculation ($s \geq \tau$). For each image the spatial alignment of each prediction is verified through bounding box IoU and the semantic similarity between predicted verb-object pairs and ground truth pairs is computed with our provided SHOE similarity maps. Predictions get matched identically to the SHOE-mAP-score. In the final aggregation step precision, recall and F1 scores for each class are calculated based on the accumulated FP, TP and FN before averaging over all classes for the mean F1 score (mF1).

7. Future Work

While large VLMs show strong semantic reasoning for HOI prediction, they currently rely on an external detector such as DETR due to their limited detection capabilities. VLMs require substantially more parameters (≈1.5B - ≈400B) compared to standard HOI detectors (≈40M), making them impractical for real-time deployment and HOI detection from videos. Their prompt sensitivity and non-deterministic outputs further complicate consistent predictions and integration into downstream systems. This highlights the need to develop true open-vocabulary, real-time HOI predictors that combine strong localization with true semantic understanding.

On the dataset side, existing HOI benchmarks often focus on relatively small subsets of interactions and do not exhaustively annotate all visible HOIs within each image often using broad action categories like ”cooking”, ”working”, or ”cleaning” but miss lower-level HOIs like ”cutting”, ”holding”, ”brushing”. As the field moves toward open-vocabulary HOI recognition, it will be essential to establish clear annotation guidelines and to develop more comprehensive datasets.

8. Benchmarking Datasets

We conducted our primary evaluation on HICO-DET [4], as it is widely used in the HOI community. Most HOI models provide pretrained checkpoints on HICO-DET making it suitable for comparison between standard, zero-shot and open-vocabulary HOI models. SWIG-HOI [23] is commonly used for open-vocabulary classification as it expands the label space to 1,000 object categories and 406 actions. However, its annotations inherit the scene-level abstraction of SWiG [21], which causes many fine-grained HOIs to be missed or collapsed into coarse descriptions. Fig. 5 shows a qualitative comparison between 3 common scenarios in both HICO-DET and SWIG-HOI dataset. Specific interactions such as eat, wash, or cut are annotated within HICO-DET. However, SWIG-HOI includes broader verbs like cook or clean as their ground-truth HOI classes, because it is derived from general scene understanding.

Moreover, SWIG-HOI includes only a single human annotation per image, even when multiple people interact with objects (see Fig. 5 second SWIG-HOI example ”cook veg-

Model	mAP	SHOE-mAP
GPT-4.1	**50.47**	57.82
Qwen2.5-VL-32B	41.12	**64.77**
InternVL3-38B	46.34	50.79

Table 4. Comparison of mAP and SHOE-mAP on V-COCO for three VLMs. VLMs token probabilities are used as a proxy for the confidence score and DETR is used for object detection.

etable/knife") where two people cutting vegetables are visible in the image but only 1 person and the corresponding HOI is annotated. This is acceptable for closed-set classifiers with fixed verb–object categories but becomes problematic in open-vocabulary evaluation with no predefined categories. In such settings, a model may correctly predict a valid but unannotated human–object interaction, which is then incorrectly counted as a false positive penalizing category-free predictors. SWIG-HOI's scene-level annotation style and incomplete human coverage makes it less suitable for open-vocabulary HOI evaluation. For this reason, we chose HICO-DET for our quantitative open-vocabulary evaluation.

V-COCO [16] is an older benchmark that offers substantially less interaction categories than the newer HICO-DET (26 actions vs 117 actions). Many HOI models do not provide checkpoints for V-COCO and focus on evaluating on HICO-DET or SWIG-HOI. To demonstrate that the SHOE metric transfers seamlessly across benchmarks we report some exemplary results on this dataset (Table 4). The trends of the evaluated LLMs mirror those observed on HICO-DET. GPT-4.1 shows strongest performance for the standard mAP score while Qwen2.5-VL-32B achieves the highest SHOE mAP on V-COCO..

9. Qualitative results

To better understand the results of Table 2, we present qualitative examples of GPT-4.1's predictions and their alignment with HICO-DET ground truth based on our instance-level SHOE score (see Figure 6). The examples show that GPT-4.1 often generates semantically plausible outputs that remain close in meaning to the intended action, even when the exact ground-truth interaction is missed.

For example, the ground-truth action "pet a giraffe" receives a high similarity score of 0.775 when matched with the prediction "touch a giraffe," which is semantically plausible given the close relation between these actions. Similarly, "wash a train" is matched with "clean a train" with a similarity score of 0.85. These examples highlight the ability of the VLMs and MLLMs to perform robust conceptual reasoning and demonstrate that our SHOE score can capture a broad range of semantically aligned open-vocabulary interactions.

10. Limitations

Our proposed SHOE-Metric relies on WordNet sense disambiguation and LLM similarity scoring, both of which can introduce biases into the metric. As WordNet senses were developed by linguists primarily from English corpora it may miss coverage in domain-specific or culturally diverse verb senses. Similarly, LLM judgments can reflect biases present in their training data. As existing HOI benchmarks focus on common general interactions, we did not find any apparent biases for these datasets during testing. For building highly domain-specific HOI datasets with low WordNet coverage (e.g., medical or laboratory settings) we suggest using definitions from common ontologies in the field such as UMLS [2] and reducing risks of domain biases by conducting an expert user study, similar to the methodology adopted in this work.

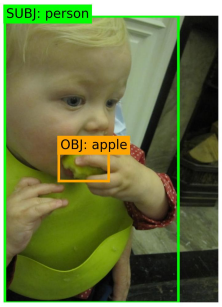
Algorithm 1: SHOE mAP Evaluation for HOI Predictions

Load ground truth set $\mathcal{G} = \{(v, o, b_h, b_o)\}$
Load predictions $p = (v', o', b'_h, b'_o, s) \in \mathcal{P}$
Load similarity maps S_v, S_o
Initialize per-class lists $\mathcal{S}_k$, $\mathcal{L}_k$, and GT counts $\mathcal{N}_{\text{GT}}[k] = 0$
for each image I **do**
    $\mathcal{G}_I \leftarrow$ ground truth for I
    $\mathcal{P}_I \leftarrow$ predictions for I
    Initialize matched flags for $\mathcal{P}_I$
    for each $(v, o, b_h, b_o) \in \mathcal{G}_I$ **do**
        $k \leftarrow$ class ID of (v, o)
        $\mathcal{N}_{\text{GT}}[k]$ += 1
        $\mathcal{M} \leftarrow$ { predictions in $\mathcal{P}_I$ where $\text{IoU}(b_h, b'_h) \geq \theta$ and $\text{IoU}(b_o, b'_o) \geq \theta$ and prediction not already matched }
        if $\mathcal{M} \neq \emptyset$ **then**
            Select $p^* = \arg\max_{p \in \mathcal{M}} \frac{1}{2}(S_v(v, v') + S_o(o, o'))$
            Let $p^* = (v'^*, o'^*, b'^*_h, b'^*_o, s^*)$, where $p^* \in \mathcal{M} \subseteq \mathcal{P}$
            Compute similarity score $\sigma = \frac{1}{2}(S_v(v, v'^*) + S_o(o, o'^*))$
            Append confidence and similarity score (s^*, σ) to $\mathcal{S}_k, \mathcal{L}_k$
            Mark p^* as matched
        else
            Append empty pair $(0.0, 0.0)$ to $\mathcal{S}_k, \mathcal{L}_k$
        end if
    end for
    for each unmatched $p = (v', o', b'_h, b'_o, s) \in \mathcal{P}_I$ **do**
        Find $(v^*, o^*) \in \mathcal{G}_I$ maximizing similarity
        $\sigma = \frac{1}{2}(S_v(v^*, v') + S_o(o^*, o'))$
        if $\sigma \geq \delta$ **then**
            $k^* \leftarrow$ class ID of (v^*, o^*)
            Append $(s, 0.0)$ to $\mathcal{S}_{k^*}, \mathcal{L}_{k^*}$
        end if
    end for
end for
for each class $k = 1 \ldots K$ **do**
    Sort $\mathcal{S}_k, \mathcal{L}_k$ by descending score
    Let ϵ be a small constant (e.g., 10^{-8}) to avoid division by zero
    Compute cumulative true positives: $\text{TP}_i = \sum_{j=1}^{i} \mathcal{L}_k[j]$
    Compute cumulative false positives: $\text{FP}_i = i - \text{TP}_i$
    $\text{Prec}_i = \text{TP}_i / (\text{TP}_i + \text{FP}_i + \epsilon)$
    $\text{Rec}_i = \text{TP}_i / (\mathcal{N}_{\text{GT}}[k] + \epsilon)$
    Compute AP_k as area under the PR-curve
end for
Report $\text{mAP} = \frac{1}{K} \sum_{k=1}^{K} \text{AP}_k$

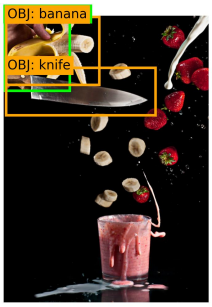
Algorithm 2: SHOE mF1 Evaluation for HOI Predictions

Load ground truth set $\mathcal{G} = \{(v, o, b_h, b_o)\}$
Load predictions $p = (v', o', b'_h, b'_o, s) \in \mathcal{P}$ with confidence score rank $s \geq \tau$
Load similarity maps S_v, S_o
Initialize TP = FP = FN = 0 per class
for each image I **do**
    $\mathcal{G}_I \leftarrow$ ground truth for I
    $\mathcal{P}_I \leftarrow$ predictions for I
    Initialize matched flags for $\mathcal{P}_I$
    for each $(v, o, b_h, b_o) \in \mathcal{G}_I$ **do**
        $\mathcal{M} \leftarrow$ { predictions in $\mathcal{P}_I$ where $\text{IoU}(b_h, b'_h) \geq \theta$ and $\text{IoU}(b_o, b'_o) \geq \theta$ and prediction not already matched }
        if $\mathcal{M} \neq \emptyset$ **then**
            Select $p^* = \arg\max_{p \in \mathcal{M}} \frac{1}{2}(S_v(v, v') + S_o(o, o'))$
            Let $p^* = (v'^*, o'^*, b'^*_h, b'^*_o, s^*)$, where $p^* \in \mathcal{M} \subseteq \mathcal{P}$
            Compute similarity score $\sigma = \frac{1}{2}(S_v(v, v'^*) + S_o(o, o'^*))$
            Accumulate TP for class (v, o) as σ
            Accumulate FP for class (v, o) as $1 - \sigma$
            Mark p^* as matched
        else
            Accumulate FN for class (v, o) as 1.0
        end if
    end for
    for each unmatched $p = (v', o', b'_h, b'_o) \in \mathcal{P}_I$ **do**
        Find best $(v, o) \in \mathcal{G}_I$ maximizing similarity
        $\sigma = \frac{1}{2}(S_v(v, v') + S_o(o, o'))$
        if $\sigma \geq \delta$ **then**
            Accumulate FP for class (v, o) as 1.0
        end if
    end for
end for
for each class (v, o) **do**
    Let ϵ be a small constant (e.g., 10^{-8}) to avoid division by zero
    Compute precision = $\frac{\text{TP}}{\text{TP+FP}+\epsilon}$
    Compute recall = $\frac{\text{TP}}{\text{TP+FN}+\epsilon}$
    Compute $F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision+recall}+\epsilon}$
end for
Report mean $F1$ over all classes

(a) HICO-DET



Object(s): Apple
Interactions: eat, hold



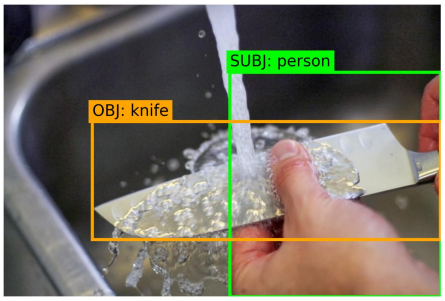
Object(s): Banana, Knife
Interactions: hold, cut with, wield

Object(s): Knife
Interactions: hold, wash

(b) SWIG-HOI

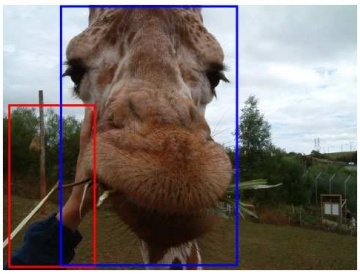
Object(s): Apple
Interactions: lick

Object(s): Knife, Vegetable
Interactions: cook



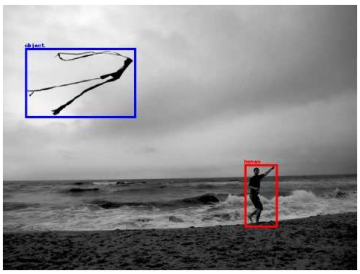
Object(s): Knife
Interactions: clean

Figure 5. Qualitative comparison of interaction annotations. HICO-DET captures fine-grained, specific interactions (e.g., ”eat”, ”wash”, ”cut with”), whereas SWIG-HOI primarily provides broad scene descriptions (e.g., ”cook”, ”clean”).



GT: **pet a/an giraffe**
Synsets: verb=pet.v.01, obj=giraffe.n.01
✅ matched prediction:
Prediction: touch a/an giraffe
Similarity: verb=0.550, obj=1.000, sim total=0.775

❌ unmatched prediction
Prediction: feed a/an giraffe
Most similar class: feed a/an giraffe with sim total=1.000



GT: **fly a/an kite**
Synsets: verb=fly.v.03, obj=kite.n.03
✅ matched prediction:
Prediction: fly a/an kite
Similarity: verb=1.000, obj=1.000, sim total=1.000

GT: **pull a/an kite**
Synsets: verb=pull.v.01, obj=kite.n.03
✅ matched prediction:
Prediction: hold a/an kite
Similarity: verb=0.350, obj=1.000, sim total=0.675

GT: **cut a/an cake**
Synsets: verb=cut.v.01, obj=cake.n.03
✅ matched prediction:
Prediction: cut a/an cake
Similarity: verb=1.000, obj=1.000, sim total=1.000

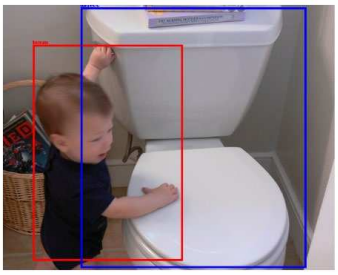
❌ unmatched prediction
Prediction: hold a/an knife
Most similar class: hold a/an knife with sim total=1.000

GT: **hold a/an motorcycle**
Synsets: verb=hold.v.02, obj=motorcycle.n.01
✅ matched prediction:
Prediction: hold a/an motorcycle handlebar
Similarity: verb=1.000, obj=1.000, sim total=1.000

GT: **ride a/an motorcycle**
Synsets: verb=ride.v.10, obj=motorcycle.n.01
✅ matched prediction:
Prediction: ride a/an motorcycle
Similarity: verb=1.000, obj=1.000, sim total=1.000

GT: sit_on a/an motorcycle
Synsets: verb=sit.v.01, obj=motorcycle.n.01
✅ matched prediction:
Prediction: ride a/an motorcycle
Similarity: verb=0.400, obj=1.000, sim total=0.700

GT: **wash a/an train**
Synsets: verb=wash.v.03, obj=train.n.01
✅ matched prediction:
Prediction: clean a/an train
Similarity: verb=0.700, obj=1.000, sim total=0.850



GT: **flush a/an toilet**
Synsets: verb=flush.v.05, obj=toilet.n.02
✅ matched prediction:
Prediction: touch a/an toilet
Similarity: verb=0.100, obj=1.000, sim total=0.550

Figure 6. Qualitative results for GPT4.1 on HICO-DET single HOI subset.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[2] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270, 2004. 7

[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. 1

[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. 1, 2, 6

[5] 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. 2

[6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1, 2

[7] Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619, 1973. 4

[8] Sanghyun Kim, Deunsol Jung, and Minsu Cho. Locality-aware zero-shot human-object interaction detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20190–20200, 2025. 2

[9] Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2018. 5

[10] 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. 2

[11] Qinqian Lei, Bo Wang, and Tan Robby T. Hola: Zero-shot hoi detection with low-rank decomposed vlm feature adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2025. 2, 3

[12] 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

[13] 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. 2, 3

[14] 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. 2

[15] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. 2

[16] 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. 7

[17] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 1

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[19] George Miller and Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998. 3

[20] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee,

Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. 2

[21] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *European Conference on Computer Vision*, pages 314–332. Springer, 2020. 6

[22] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[23] 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)*, pages 13455–13464, 2021. 6

[24] 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

[25] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1

[26] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024. 1