

Benchmarking Multimodal Large Language Models for Zero-Shot Human-Object Interaction

Oriol Rabasseda^{1, 2, *}, Zennie Li¹, Kamal Nasrollahi^{1, 3}, and Sergio Escalera²

¹Milestone Systems A/S, Banemarksvej 50, 2605 Broendby, Denmark

²Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007, Barcelona, Spain

³Aalborg University, Rendsburggade 14, 9000 Aalborg, Denmark

This is a preprint of the accepted manuscript, not the final published version on Image and Vision Computing.

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated strong zero-shot reasoning across diverse domains, but their potential for Human-Object Interaction (HOI) remains unexplored. Unlike traditional HOI approaches that rely on extensive labeled datasets and fixed interaction taxonomies, MLLMs offer flexible reasoning, making them well-suited for dynamic, real-world scenarios. In this work, we present the first systematic study of MLLMs for HOI recognition and detection within a pipeline that combines an external object detector for entity localization with zero-shot multimodal reasoning for interaction understanding. We introduce a benchmark framework for principled zero-shot evaluation, consisting of (i) a unified taxonomy of zero-shot prompting strategies, and (ii) a methodology that transforms token-level logits into interaction scores, enabling fair comparison with specialized HOI models under standard metrics. To support prompt-based evaluation, we additionally construct new question-answer annotations for standard HOI datasets, transforming them into a format suitable for MLLMs. Building on this, we propose two visual prompting strategies, namely Objects as Context and Black Other Objects, that guide MLLMs to target objects while preserving scene context. Experiments on HICO, HICO-DET, and V-COCO demonstrate state-of-the-art zero-shot performance, achieving 53.50 mAP on HICO and 23.69 mAP on HICO-DET. Overall, our work provides a principled benchmarking framework for MLLMs in HOI, opening the door to new models and datasets and enabling flexible open-world interaction understanding. Code and adapted datasets are publicly available¹.

Keywords:

Zero-shot Learning, Multimodal Large Language Models, Human-Object Interaction Recognition, Human-Object Interaction Detection, Prompt Engineering

1 Introduction

Multimodal Large Language Models (MLLMs) have recently advanced in tasks that connect vision and language, including medical image understanding (Li et al. (2023), Moor et al. (2023)), document parsing

*Corresponding author (O. Rabasseda). Email address: oriol.rabasseda@gmail.com.

¹<https://github.com/oriol-rabasseda/mlms-zs-hoi>

© 2026. This manuscript version is made available under the CC-BY-NC-ND 4.0 license <https://creativecommons.org/licenses/by-nc-nd/4.0/>

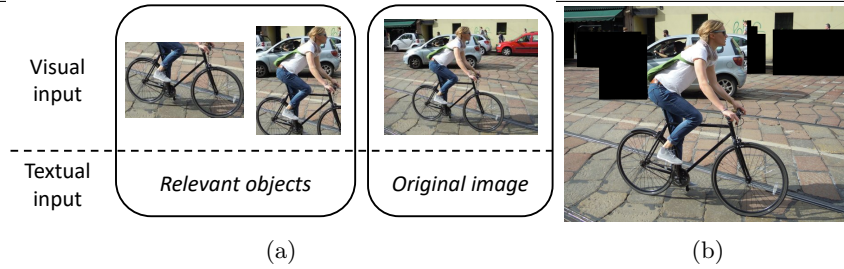


Figure 1: (a) Visual example of Objects as Context: cropped relevant objects are input alongside text prompts to enrich visual tokens. (b) Visual example of Black Other Objects: all non-target objects are masked to remove distractions.

(Liu et al. (2024), Hu et al. (2024)), and embodied agent interaction (Huang et al. (2023)). These successes highlight their versatility across diverse benchmarks and domains.

In this work, we investigate MLLMs for Human-Object Interaction (HOI), a core problem in scene understanding. HOI is particularly important for downstream real-world applications such as surveillance, human behavior analysis, and video understanding.

Traditional HOI methods (Yuan et al. (2023), Yang et al. (2023a), Kim et al. (2023)) rely on large labeled datasets and pre-defined sets of interaction classes. However, such closed-world assumptions limit their ability to generalize: real-world scenarios often contain unseen interactions, novel objects, or evolving vocabularies. For HOI to be broadly useful in practice models must adapt beyond fixed taxonomies.

Prior work on HOI generalization has attempted to address this challenge by withholding certain classes during training and evaluating transfer performance (Shen et al. (2018)). However, these approaches remain constrained by labeled training sets and by evaluation protocols that only partially capture open-world conditions. In contrast, MLLMs can operate without task-specific training: by combining multimodal reasoning with natural-language instructions, they enable flexible and interactive zero-shot recognition (Zhang et al. (2023b)).

Despite the promise of MLLMs for open-world HOI, a major obstacle is the absence of benchmarks that can fairly evaluate their capabilities. To close this gap, we design the first evaluation framework for MLLMs tailored to HOI recognition and detection in a zero-shot setting. These benchmarks enable systematic comparison with prior HOI methods and provide a foundation for studying MLLMs in realistic open-world conditions.

A key challenge in applying MLLMs to HOI is how to formulate prompts and measure their predictions. To address the prompt design issue, we propose a unified taxonomy of zero-shot prompting strategies for classification and adapt public datasets into a question-answering format suitable for MLLMs. To measure the predictions, we design a method that converts token-level logits into classification scores, allowing direct use of standard metrics such as mAP and ensuring rigorous and fair comparison with specialized HOI models.

Using this framework, we study a broad set of zero-shot prompting methods under our taxonomy and evaluate several lightweight MLLMs. In addition, we introduce two visual prompting methods, namely Objects as Context (Figure 1a) and Black Other Objects (Figure 1b), that guide attention to specified regions while preserving scene context. Overall, we establish strong zero-shot baselines, opening a path toward flexible open-world interaction understanding.

Our main contributions are:

- (1) **A benchmark framework for evaluating MLLMs in HOI**, including (i) a unified taxonomy of zero-shot prompting strategies for classification, and (ii) a method that derives classification scores from token logits, enabling the **first direct comparison with specialized HOI models**.

- (2) The construction of novel question-answering annotations for 3 of standard HOI datasets, enabling MLLMs to perform zero-shot classification with natural-language prompts.
- (3) A comprehensive benchmark of MLLMs and zero-shot strategies on HOI recognition and detection, obtaining state-of-the-art zero-shot results.
- (4) Two novel visual prompting methods, namely Objects as Context and Black Other Objects, that improve object specification.

The rest of the paper is organized as follows. Section 2 reviews related work on MLLMs, zero-shot strategies, and HOI. Section 3 presents the taxonomy of zero-shot strategies. Section 4 outlines our methodology and zero-shot strategies. Section 5 reports experiments. Section 6 highlights findings and future directions.

2 Related work

2.1 Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) extend Large Language Models (LLMs) to process multimodal inputs, typically text and images. Unlike earlier Vision–Language Models (VLMs), which commonly use fixed encoders and shallow fusion mechanisms for text–image alignment, MLLMs integrate an LLM backbone into the architecture (Yang et al. (2024), Cai et al. (2024)). This enables conversational interaction, instruction following, and interleaved multimodal reasoning (Lin et al. (2024)). Advances in high-resolution encoding (Dong et al. (2024), Dehghani et al. (2023)) and efficient backbones have led to competitive models (Yao et al. (2024), Li et al. (2024), Wang et al. (2024)).

These models have shown broad versatility across domains, including document parsing (Liu et al. (2024), Hu et al. (2024)), medical image interpretation (Li et al. (2023), Moor et al. (2023)), and visual reasoning (Liu et al. (2023), Zhu et al. (2023)). Their extensive generalist pretraining makes them effective in zero-shot and open-world settings.

Despite these strengths, MLLMs have not yet been systematically explored for Human–Object Interaction (HOI), a vision task that requires reasoning about localized entities and interaction vocabularies. Existing attempts with vision–language backbones use fixed matching paradigms and lack instruction-following capabilities (Ning et al. (2023), Mao et al. (2023)), leaving open the question of how MLLMs perform in zero-shot HOI recognition and detection.

2.2 Zero-shot prompting strategies for MLLMs

Early zero-shot work in vision–language models began with CLIP (Radford et al. (2021)), which showed that classification could be achieved without task-specific training. However, CLIP’s inability to follow detailed instructions or perform multi-step reasoning limited its applicability to complex tasks. These limitations motivated the use of MLLMs for zero-shot inference, particularly in Visual Question Answering (VQA).

Prompting strategies are central to unlocking zero-shot performance and fairly benchmarking MLLMs. They can be broadly divided into textual and visual approaches. Textual strategies aim to improve reasoning and reliability, for instance by generating diverse candidate prompts or constructing them automatically (Lan et al. (2023)), enriching them with external knowledge (Cao and Jiang (2024), Lin et al. (2022)), or adopting chain-of-thought templates for step-by-step reasoning (Zhang et al. (2023b)). Complementarily, visual prompting modifies the visual input to guide attention, using techniques such as cropping, masking, or altering background context (Shtedritski et al. (2023)).

Despite their promise, most evaluations of these strategies remain confined to VQA benchmarks (Marino et al. (2019), Lu et al. (2022), Goyal et al. (2017)). Although extensions to classification tasks are scarce, studies such as Gupta et al. (2024) have explored action recognition with MLLMs, but these efforts typically lack classification score distributions. Without such outputs, standard metrics like mean Average Precision (mAP) cannot be computed, hindering direct comparison with specialized baselines in HOI.

Finally, while several taxonomies of prompting strategies have been proposed (Chen et al. (2023), Wu et al. (2024b)), they primarily target textual reasoning or VQA. No prior work systematically unifies textual and visual prompting for classification settings. Our work addresses this gap by introducing a unified taxonomy and evaluation framework tailored to zero-shot HOI.

2.3 Zero-shot Human-Object Interaction recognition and detection

Human-Object Interaction (HOI) tasks study how humans and objects interact in images. While recognition focuses on classifying interactions from a fixed set, detection additionally requires localizing both humans and objects. These tasks are particularly challenging due to the large number of possible interactions and the long-tailed, skewed class distributions found in real-world data.

Early HOI recognition and detection methods have advanced through benchmarks such as HICO (Chao et al. (2015)) and HICO-DET/V-COCO (Chao et al. (2018), Gupta and Malik (2015)), driving progress in both fully supervised and partially zero-shot settings. Partially zero-shot protocols withhold subsets of HOI classes during training (Shen et al. (2018), Ma et al. (2022)), testing limited generalization to unseen interactions. More recently, CLIP-based approaches (Liao et al. (2022), Mao et al. (2023)) leverage vision-language pretraining for transfer to unseen categories, demonstrating the potential of multimodal representations for open-world HOI. However, these models still depend on task-specific training subsets and struggle with complex scene-level reasoning.

HOI detection methods are commonly two-stage or one-stage. Two-stage approaches detect humans and objects first, then classify interactions for candidate pairs (Liu et al. (2020), Zhang et al. (2022b)), whereas one-stage methods predict interactions and detect objects jointly in an end-to-end model (Liao et al. (2022), Zou et al. (2021)).

In contrast, our work applies instruction-following MLLMs to HOI in a *pure zero-shot* setting. By removing reliance on labeled HOI training data, our benchmarking setup directly tests models under open-world conditions.

3 Taxonomy of zero-shot strategies for MLLMs

This section categorizes zero-shot learning strategies to leverage MLLMs as classifiers. We aggregate these techniques based on the type of supplementary information they provide to the model and how it is incorporated into the input of the MLLM. Figure 2 shows the taxonomy, which is divided into Visual Prompting methods, Textual Prompting methods, and Hybrid methods. Table 1 categorizes relevant methods under the proposed taxonomy.

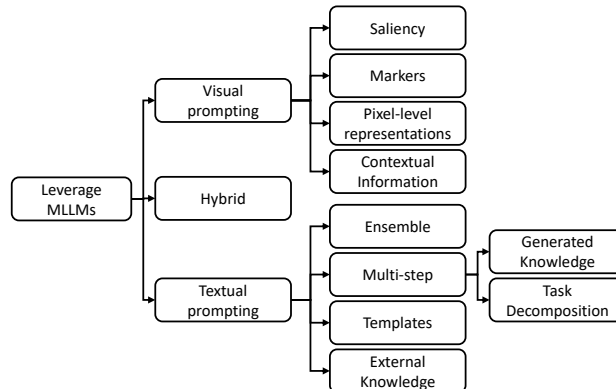


Figure 2: Taxonomy of zero-shot methods leveraging MLLMs divided into textual, visual, and hybrid strategies.

Taxonomy	Prompting methods
Ensemble	Resampling, Self-consistency (Wang et al. (2022)), Tree-of-thoughts (Yao et al. (2023)), Graph-of-thoughts (Besta et al. (2024))
Generated Knowledge	Multimodal Chain-of-thought (Zhang et al. (2023b)), Generated knowledge (Liu et al. (2021))
Task Decomposition	Prompt chaining, Least-to-most prompting (Zhou et al. (2022))
Templates	Role-prompting, Multimodal Chain-of-thought (Zhang et al. (2023b)), Meta prompting (Zhang et al. (2023a)), Prompt Pattern Catalog (White et al. (2023))
External Knowledge	FLARE (Jiang et al. (2023)), In-Context RALM (Ram et al. (2023))
Saliency	RedCircle (Shtedritski et al. (2023)), FGVP (Yang et al. (2023c))
Markers	RedCircle (Shtedritski et al. (2023)), AutoAD-Zero (Xie et al. (2024)), Set-of-Mark (Yang et al. (2023b)), OWG (Tziafas and Kasaei (2024)), Pivot (Nasiriany et al. (2024)), VTPrompt (Jiang et al. (2024)), TextCoT (Luan et al. (2024)), CRG (Wan et al. (2024))
Pixel-level representations	FGVP (Yang et al. (2023c)), SCAFFOLD (Lei et al. (2024)), AO-Planner (Chen et al. (2024a))
Contextual Information	Objects as Context (ours)

Table 1: Position in zero-shot taxonomy of relevant methods.

3.1 Textual prompting methods

Textual prompting focuses on applying specific techniques to the textual part of the multimodal input, so that the MLLM can improve classification by making use of the proper inherent knowledge or the necessary additional information. The prompting methods are categorized into:

- **Templates:** Textual templates consist of reusing specific patterns or sentences that have been deemed useful in similar tasks. Specific templates include role prompting, use of quotes to separate parts of the prompt, and Chain-of-Thought (Kojima et al. (2022)) prompts. Recent methods offer adaptability to multiple tasks, such as pattern catalog (White et al. (2023)) and meta prompting (Zhang et al. (2023a)). Pattern catalog provides an organized collection of prompt templates, while meta prompting focuses on structuring problem formulation.
- **External knowledge:** Concatenating updated or specialized data into the prompt as context is crucial for tasks where the model lacks the required knowledge or information. Providing or retrieving information from a corpus, the Internet or specialized tools fall into this category.
- **Ensemble:** Methods that query the LLM in a parallel way belong to this category. Widely used methods include resampling and parallel chain-of-thought-like methods. To fuse the different answers obtained voting strategies and frequency metrics are commonly used.
- **Multi-step:** Multi-step methods query the MLLM several times in a sequential way. We distinguish two strategies: Generated Knowledge and Task Decomposition. Generated Knowledge takes advantage of the knowledge of the MLLM obtained through previous queries and inputs it as context for the final query (*e.g.* image captioning or reasoning from Zhang et al. (2023b)). Task Decomposition splits the problem into sub-tasks and merges all the sub-tasks solutions to solve the main task.

3.2 Visual prompting methods

Visual prompting methods manipulate images and videos to provide additional information, enhancing the visual perception capabilities of the MLLM. Visual prompting strategies are categorized into Markers, Pixel-level Representations, Saliency methods and Contextual Information.

- **Markers:** Overlay markers such as bounding boxes, circles, arrows, and dots to highlight, identify, or draw attention to particular regions in the visual input. Markers help the MLLM understand the image content and correlate it with the corresponding text.
- **Pixel-level representations:** Overlay fine-grained pixel-level information on the visual input, in opposition to coarse markers or boxes. Object contour highlighting and semantic segmentation masking are examples of this category.
- **Saliency:** Modify or delete non-relevant parts of the visual input. Compared to other categories, saliency removes potentially misleading information instead of emphasizing relevant regions. Background adjustments fall into this category (*e.g.* blurring the background).
- **Contextual Information:** Provide the MLLM with additional visual inputs as context, encapsulating relevant information from the original visual input. In this category, the original visual input is not modified; instead, additional information is fed to the MLLM, such as crops of relevant objects.

3.3 Hybrid methods

Combinations of textual and/or visual prompting methods are possible, as most methods are not mutually exclusive. As an example of a hybrid method, [Zhang et al. \(2023b\)](#) proposes Multimodal Chain-of-thought, a two-stage method that uses chain-of-thought templates for the generation of reasoning and inputs it as generated knowledge.

4 Methodology

In this section, we present our methodology for leveraging MLLMs in zero-shot HOI recognition and detection. Section 4.1 presents our HOI recognition method, including object recognition, MLLM querying, and probability computation. Section 4.2 describes a pipeline that incorporates object detection and object specification strategies. We also introduce two visual prompting methods (Objects as Context and Black Other Objects) designed to guide the model’s attention while maintaining scene context.

4.1 HOI Recognition

Given an input image x , the task of HOI Recognition is to predict, for each object class $i \in \mathcal{O}$ (where $\mathcal{O}$ is the set of possible object categories) a probability distribution for each of its potential interactions. Each object i is associated with a predefined set of interactions $\mathcal{I}_i$ (*e.g.*, *holding*, *riding*, or *carrying* depending on the object). To account for cases where an object does not participate in any interaction, this set is sometimes extended with a *no_interaction* label, yielding $\mathcal{I}'_i = \{\text{no_interaction}\} \cup \mathcal{I}_i$.

For each object i , the model produces a vector of predicted probabilities $\hat{y}_i = (\hat{y}_{i1}, \hat{y}_{i2}, \dots, \hat{y}_{i|\mathcal{I}'_i|}) \in [0, 1]^{|\mathcal{I}'_i|}$, where $\hat{y}_{ij}$ represents the likelihood that interaction $j \in \mathcal{I}'_i$ occurs with object class i . Since multiple interactions can co-occur for the same object and across different objects in the image, HOI Recognition is inherently a multi-label problem. The total output dimensionality is therefore $\sum_{i \in \mathcal{O}} |\mathcal{I}'_i|$. Formally, the objective is to learn a mapping $f: \mathcal{X} \rightarrow \prod_{i \in \mathcal{O}} [0, 1]^{|\mathcal{I}'_i|}$ from the input space of images $\mathcal{X}$ to the interaction probability vectors, which can be evaluated against ground-truth annotations using standard multi-label metrics.

Our method, including object detection and MLLM-based interaction evaluation, is illustrated in Figure 3. To improve inference efficiency, an object detector is used to identify which object classes are present in the image and skip unnecessary queries to the MLLM. Specifically, for each detected object class i , the maximum detection score m_i among all instances of that class is used as an indicator of presence. If $m_i \leq \tau$,

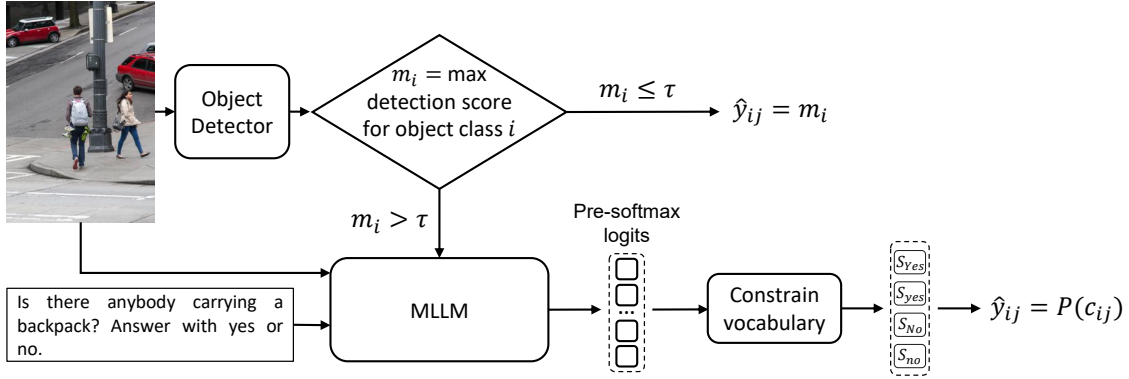


Figure 3: Zero-shot HOI recognition method using MLLMs. The output $\hat{y}_{ij}$ represents the classification score for object class i and interaction j , $P(c_{ij})$ denotes the classification probability computed according to Equation (1), and τ is the detection threshold for querying the MLLM. The image is taken from the HICO test set.

the object class is considered absent and all interactions in $\mathcal{I}'_i$ are assigned this low score ($\hat{y}_{ij} = m_i$), where τ is a hyperparameter of our method. Equivalently, if an object class i is not detected, $\hat{y}_{ij} = 0$ for all $j \in \mathcal{I}'_i$.

If $m_i > \tau$, the object class i is treated as present and the MLLM is queried to evaluate all interactions in $\mathcal{I}'_i$. Interaction queries are constructed by expressing each interaction in gerund form (e.g., *holding*, *carrying*) and inserting it into the template: “Is there anybody {gerund} a {object}? Answer with yes or no.”. Given this instruction, the MLLM is prompted to produce a single-word answer, either affirmative or negative. This formulation yields one independent prediction for each object–interaction pair, allowing multiple interactions to be recognized simultaneously for the same object class.

Although the MLLM produces a textual output, we do not use the actual generated text to compute classification scores. This is because the textual answer only provides a single binary label, which does not produce a probability distribution reflecting the underlying confidence of the model. Instead, we extract the pre-softmax logits of the first generated token, which is sufficient because this token determines whether the answer is affirmative or negative.

We then constrain the vocabulary to a set of affirmative tokens $\mathcal{A}$ and negative tokens $\mathcal{N}$ and compute probabilities over this restricted set using a standard softmax. This effectively treats the restricted logits as a complete vocabulary and allows us to produce classification scores from the likelihood of an affirmative answer versus a negative answer. The specific contents of the sets $\mathcal{A}$ and $\mathcal{N}$ are discussed in Section 5.2.1.

Formally, the classification score for interaction c_{ij} , where i is the object and j the interaction, is computed as:

$$\hat{y}_{ij} = P(c_{ij}) = \frac{\sum_{w \in \mathcal{A}} \exp(s_w)}{\sum_{w \in \mathcal{A}} \exp(s_w) + \sum_{w \in \mathcal{N}} \exp(s_w)}, \quad (1)$$

where s_w denotes the logit corresponding to token w . Although HOI recognition is a multi-label problem, sometimes object classes include a *no_interaction* label that should be mutually exclusive with its interactive labels. Estimating this label independently could therefore produce contradictory predictions in which the same object is classified as both interacting and not interacting. To avoid this issue, its probability is computed as the probability that none of the interactions associated with object class i occurs. Consequently, the *no_interaction* confidence decreases whenever the confidence of one or more interactive labels increases. Formally, where there is a *no_interaction* class,

$$P(c_{i,no_interaction}) = \prod_{j \in \mathcal{I}_i} (1 - P(c_{ij})). \quad (2)$$

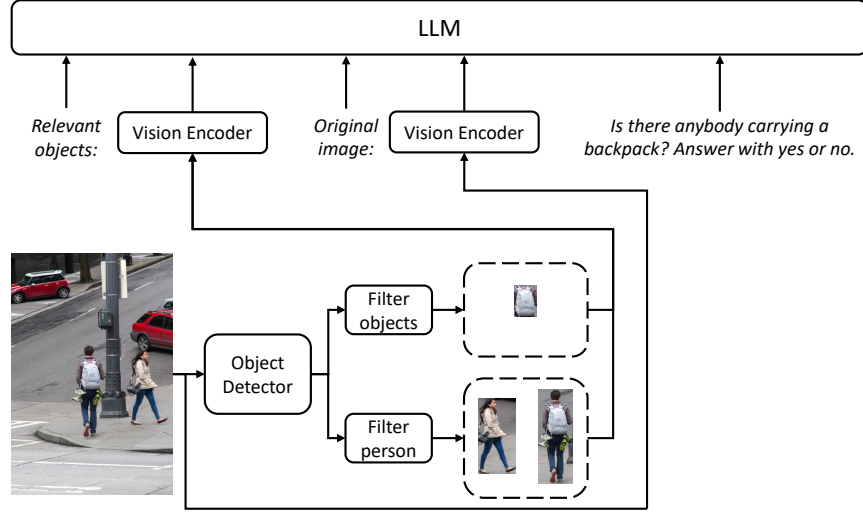


Figure 4: Objects as Context architecture. Step 1: Detect all objects. Step 2a: Filter relevant objects (*e.g.* “backpacks”) based on the question. Step 2b: Filter all persons. Step 3: Input the information into the LLM, specifying visual tokens related to relevant objects/persons and the original image.

Building on this pipeline we evaluate a set of zero-shot prompting strategies drawn from the taxonomy in Section 3. The complete list of strategies and their use in recognition is provided in Table 2, with the corresponding prompts in the Appendix. Multimodal Chain-of-Thought (MCoT) and Caption as Context are two-stage prompting approaches, where an initial prompt generates meaningful information that is incorporated as context for the final binary interaction prediction. On the other hand, Ensemble of Prompts averages the predicted scores obtained from multiple prompt templates. We compare all strategies against a baseline with no additional prompting strategy.

Furthermore, we introduce a visual prompting method, *Objects as Context*, which falls under Contextual Information. In this strategy, the detected regions corresponding to persons and objects of the target class are cropped and supplied as additional visual inputs to the MLLM. This approach aims to focus the model on the entities involved with the queried interaction while still retaining the overall scene context. To avoid overloading the input with less relevant detections, only the k_1 most confidently detected target objects and the k_2 largest persons (foreground) are included, where k_1 and k_2 are hyperparameters. Figure 4 illustrates the resulting architecture.

4.2 HOI Detection

HOI Detection extends HOI Recognition by requiring not only the classification of which interactions occur, but also the localization of the human–object pairs involved. In recognition, the object detector is used only for filtering and does not constrain the MLLM from handling the task end-to-end. In detection, however, it becomes an integral component of the pipeline, providing bounding boxes for both humans and objects.

Formally, given an input image $x \in \mathcal{X}$, let $\mathcal{H} = \{h_1, \dots, h_{N_h}\}$ denote the set of detected human bounding boxes and $\mathcal{B} = \{o_1, \dots, o_{N_o}\}$ the set of detected object bounding boxes provided by the external detector. The task is to predict, for each human–object pair $\langle h_n, o_m \rangle$ and each interaction $j \in \mathcal{I}'_{o_m}$, a probability indicating whether the interaction occurs. Since multiple interactions can co-occur for the same pair (*e.g.* *holding* and *carrying a bicycle*), the task is inherently multi-label at the pair level.

Our method follows a two-stage approach: the object detector first provides bounding boxes for humans and objects, after which the MLLM evaluates candidate pairs $\langle h_n, o_m \rangle$. To reduce inference cost and avoid including less relevant detections, we do not evaluate all possible pairs. Instead, we first discard humans and objects whose detection confidence falls below the threshold τ at instance level. Drawing inspiration from the strategy Objects as Context, we use the thresholds k_1 and k_2 , retaining only the k_1 most confidently detected objects and the k_2 largest detected persons.

The interaction classification stage mirrors the procedure described for HOI Recognition: queries are expressed in gerund form and appended with the instruction “Answer with yes or no.”, logits of the first token are extracted, and probabilities are computed over the affirmative ($\mathcal{A}$) and negative ($\mathcal{N}$) token sets. The score for *no_interaction* is computed following the same formulation as in Equation 2, where applicable.

The main challenge is then to specify to the MLLM which human and object are being queried. To address this, we explore different visual and textual prompting strategies for object specification.

Among these strategies, we propose two visual prompting methods. *Objects as Context* is adapted from recognition to only filter the specified object and person. *Black Other Objects* (a Saliency method) masks all detected objects other than the queried human and object, aiming to reduce distractions while preserving the scene background. In addition, we also test established object-specification strategies adapted from [Shtedritski et al. \(2023\)](#). A complete list of all strategies along with their taxonomy in Section 3 and their use in detection appears in Table 2.

Category (from Taxonomy)	Strategy	HOI Recognition	HOI Detection
Hybrid	Multimodal Chain-of-Thought (Zhang et al. (2023b))	✓	–
Textual – Generated Knowledge	Caption as Context	✓	–
Textual – Ensemble	Ensemble of Prompts	✓	–
Textual – External Knowledge	Textual Center Coordinates, Textual Boundary Coordinates	–	✓
Visual – Contextual Information	Objects as Context (ours)	✓	✓
Visual – Markers	Overlay Bounding Boxes, Overlay Elliptic Markers	–	✓
Visual – Saliency	Blur the Background, Gray the Background, Crop ² , Black Other Objects (ours)	–	✓

Table 2: Summary of prompting strategies evaluated in our benchmark, grouped by the taxonomy in Section 3.

All methods are compared against the baseline, which does not apply any object specification strategy. Examples of textual questions and visual prompts for each strategy are provided in the Appendix and illustrated in Figure 5.

²The relevant area cropped encompasses both the target object and person.



Figure 5: Example of all visual prompting methods side-by-side compared to the base image (a). (b) Overlay Bounding Boxes highlights humans and objects with rectangular boxes. (c) Overlay Elliptic Markers uses ellipses to indicate the queried regions. (d) Blur the Background removes contextual clutter by blurring everything outside the target entities. (e) Crop focuses only on the cropped human-object pair. (f) Black Other Objects masks all objects except the queried pair while preserving the background. (g) Gray the Background reduces distraction by desaturating non-target regions.

5 Experiments

5.1 Implementation details

The object detector used is the extra-large YOLO11 model (Khanam and Hussain (2024)), pre-trained on the MS-COCO dataset (Lin et al. (2014)).

For HOI recognition, the hyperparameters are set to $\tau = 0.1$, $k_1 = 10$, and $k_2 = 3$ (see Section 5.5). For HOI detection, k_1 and k_2 remain unchanged, while τ is increased to 0.5 to reduce the number of MLLM queries.

Inference uses greedy decoding with a single output token and is run on three NVIDIA GeForce RTX 3090 GPUs. Additional inference hyperparameters are listed in the Appendix.

5.2 HOI recognition

We construct a HICO Q&A dataset by adding question-answer annotations for each object-interaction pair, with an illustrative example shown in Figure 3. The dataset (Chao et al. (2015)) contains 47,776 images annotated with 600 HOI categories (combinations of 117 action verbs and 80 object categories from MS-COCO), and the original split provides 38,118 training and 9,658 test images. In the *pure zero-shot* evaluation protocol, only the test set is used, with no human-object interactions for fine-tuning. This setting effectively mimics an open-world environment and, to our knowledge, has not previously been applied to the HICO dataset.

Performance is measured by the mean Average Precision (mAP) over HOI classes, as proposed in the original paper, because HOI categories are not mutually exclusive.

5.2.1 Baseline and score soundness

We evaluate the lightweight versions of state-of-the-art MLLM families, namely MiniCPM-V (Yao et al. (2024)), LLaVA-OneVision (Li et al. (2024)), Phi-3.5-Vision (Abdin et al. (2024)), InternVL2.5 (Chen et al. (2024d;c;b)), and Qwen2-VL (Bai et al. (2023), Wang et al. (2024)). These models vary in how they answer the prompt “Answer with yes or no”, with differences mainly in capitalization (Yes” vs. yes” or No” vs. no”). To ensure consistency and fair comparison across models, we standardize evaluation by defining the affirmative and negative token sets as $\mathcal{A} = \{\text{“yes”, “Yes”}\}$ and $\mathcal{N} = \{\text{“no”, “No”}\}$.

Although the classification score in Equation (1) is computed from the logits restricted to $\mathcal{A} \cup \mathcal{N}$, it is important that the tokens in these sets also receive the highest probability under the model’s distribution. This condition ensures that the restricted softmax faithfully reflects the model’s intended prediction. To

verify this assumption, and to confirm the robustness gained by including both lowercase and uppercase variants, we evaluate two indicators of score soundness:

- **Different Answers:** frequency of output tokens outside $\mathcal{A} \cup \mathcal{N}$. These tokens are considered to deviate from the expected yes/no answer format.
- **Incoherences:** cases where the token with highest confidence is in $\mathcal{A}$ (i.e. “Yes” or “yes”) but the classification score (from Equation (1)) is below 0.5, or conversely when the token is in $\mathcal{N}$ (“No” or “no”) but the score exceeds 0.5. This indicates a mismatch between the expected model’s output and the obtained classification score.

Model	mAP $\uparrow$	Diff. Answers	Incoherences
MiniCPM-V 1	48.20	0%	10.06%
MiniCPM-V 2	47.48	0%	2.74%
LLaVA-OV-0.5B (OV)	46.15	0%	1.25%
LLaVA-OV-0.5B (SI)	44.61	0%	2.27%
Phi-3.5-Vision	40.44	7.05%	0.00%
InternVL2.5-1B	41.81	7.44%	0.48%
InternVL2.5-2B	43.48	7.45%	0.39%
Qwen2-VL 2B	51.39	0%	0.80%
Qwen2-VL 7B	52.89	0%	0.48%

Table 3: Comparison of MLLMs on the HICO dataset: performance and score soundness study.

Table 3 shows the mAP values for the baseline computations (without zero-shot strategies from Table 2) along with the soundness metrics (i.e., Different Answers and Incoherences) on the HICO dataset. The results reveal that, with few exceptions in InternVL2.5 and Phi-3.5-Vision, other tokens never receive the highest output score. This validates the focus on the defined tokens given the prompt instructions. Moreover, the low incidence of incoherences supports the use of both cases (“Yes”/“yes” and “No”/“no”) to enhance generalization across models.

The top-performing families are Qwen2-VL and MiniCPM-V. To streamline experiments on zero-shot strategies, we select MiniCPM-V 1 and Qwen2-VL 2B for subsequent tests. Additionally, LLaVA-OV-0.5B (OV), which is trained on multi-image data, is included in experiments that utilize multi-image prompting, namely Objects as Context.

5.2.2 Zero-shot HOI recognition

Strategy	MiniCPM-V 1	Qwen2-VL 2B	LLaVA-OV 0.5B (OV)
MCoT	39.45	30.94	-
Caption as Context	44.81	49.43	-
Objects as Context	44.54	53.50	47.79

Table 4: Evaluations on HICO dataset using zero-shot strategies.

Table 4 reports the performance of zero-shot strategies on HICO. MCoT and Caption as Context lead to lower performance compared to the baselines in Table 3, whereas Objects as Context improves results for models that support multi-image inputs (Qwen2-VL and LLaVA-OneVision). Figure 6 demonstrates that using Ensemble of Prompts mitigates the bias of a single default prompt and improves mAP, though gains plateau beyond three prompts. The lower mAP of the two-stage strategies can be attributed to their tendency to produce more confident predictions conditioned on the generated text. Because the second stage treats the context provided by the first stage as fixed, uncertainty in the initial generation is not propagated to the final score. Consequently, the model may produce low-uncertainty predictions even when the generated

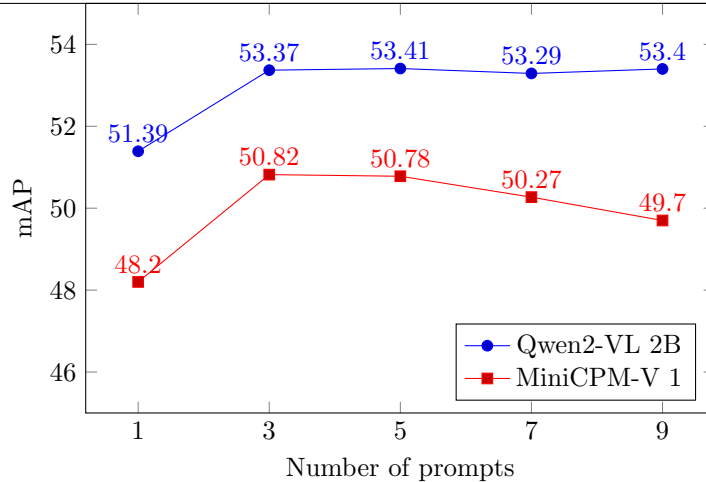


Figure 6: Evaluation on HICO dataset with different number of prompts using Ensemble of Prompts.

context is ambiguous or incorrect, penalizing misclassified samples more severely under the mAP metric. As analyzed in the Appendix, these methods achieve comparable performance after thresholding, indicating reasonable binary decision quality despite less favorable score rankings. When grouped by taxonomy, both Ensemble and Visual Prompting (Objects as Context) outperform the Multi-step with Generated Knowledge methods (MCoT and Caption as Context).

We analyze whether the combination of the two methods that improve performance yields even higher results. However, combining Ensemble of Prompts with Objects as Context (forming a Hybrid approach in our taxonomy) using Qwen2-VL 2B results in a mAP of 51.71, which is lower than when each method is used individually. Therefore, the best *pure zero-shot* result is obtained with Qwen2-VL 2B using Objects as Context, reaching a mAP of **53.50**. As summarized in Table 5, this result surpasses the strongest partially zero-shot method, which reports 42.0 mAP, highlighting the competitiveness of the proposed *pure zero-shot* evaluation protocol.

Method	mAP
<i>Supervised methods</i>	
R*CNN (Gkioxari et al. (2015))	28.5
PaStaNet (Li et al. (2020))	46.3
DEFR (Jin et al. (2022))	65.6
<i>Partially zero-shot</i>	
VCL [†] (Hou et al. (2020))	26.7
RelViT (Ma et al. (2022))	42.0
<i>Pure Zero-shot</i>	
Qwen2-VL 2B with Objects as Context	53.50

Table 5: Comparison with prior work for HOI recognition on HICO. *Note:* [†] results are reported in Ma et al. (2022).

5.3 HOI detection

We conduct experiments on two datasets for HOI detection: HICO-DET (Chao et al. (2018)) and V-COCO (Gupta and Malik (2015)). Based on these datasets, we construct HICO-DET Q&A and V-COCO Q&A

versions by adding question–answer annotations for each object–interaction pair, enabling evaluation with MLLMs in a prompt-based format (an illustrative example is shown in Figure 4).

HICO-DET extends the HICO dataset with 151,276 annotated human–object pairs across 47,776 images, providing bounding boxes for both humans and objects together with interaction labels. The split contains 38,118 training and 9,658 test images. Evaluation follows the standard mean Average Precision (mAP) metric: a detection is considered correct when both the human and object boxes have an Intersection-over-Union (IoU) above 0.5 with the ground truths. Results are reported for three category sets: Full (all 600 HOI categories), Rare (138 categories with fewer than ten training instances), and Non-Rare (462 categories with at least ten instances). Evaluations are conducted in a *pure zero-shot* setting, where no human–object interactions from HICO-DET are used for fine-tuning. *Pure zero-shot* results on HICO-DET have previously been reported (Yuan et al. (2022)), providing a useful reference point for our benchmark.

V-COCO is a subset of MS-COCO containing 10,346 images (2,533 train, 2,867 validation, 4,946 test) annotated with 26 action categories. Each action is associated with one or two object roles (e.g., “cut” may involve both knife and food). As in HICO-DET, Average Precision is computed for each action–role pair and the mean is reported. The official evaluation code (Gupta and Malik (2015)) defines two variants for handling missing object–role annotations: (1) $AP_{role}^{\#1}$, where the object is considered occluded and a dummy bounding box must be predicted, and (2) $AP_{role}^{\#2}$, where the object role is treated as out of scope and no bounding box is required. In our pipeline, the MLLM is evaluated without using any V-COCO supervision, while the object detector is pre-trained on MS-COCO. As a result, the *pure zero-shot* setting strictly applies to the MLLM, while the overall pipeline benefits from an object detector trained on data from the same distribution.

To enable MLLM evaluation, we generate a question–answer version of each dataset. For HICO-DET, we create a natural language question for each human–object pair and interaction, formatted to produce a yes/no answer. Because the pipeline requires a predefined set of possible HOI categories with specific object classes and interactions to prompt the MLLM and V-COCO defines only action–role pairs, we construct the set of HOI categories from the action–object combinations found in the *trainval* split. This adaptation does not alter the official evaluation metrics and ensures fair comparison with other methods.

Table 6 summarizes the performance of the prompting strategies listed in Table 2. Across both HICO-DET and V-COCO the trends are consistent: (1) passing object detections as text (Textual Center Coordinates and Boundary Coordinates) or overlaying bounding boxes/elliptic markers degrades performance; (2) saliency-based modifications have mixed effects: Gray the Background and Crop are neutral or negative, whereas Blur the Background and Black Other Objects consistently improve results; and (3) adding Objects as Context provides a modest boost. Overall, these three strategies are the only ones that show positive impact across all models, with Qwen2-VL 2B + Blur the Background emerging as the best configuration on both datasets.

Although the qualitative trends are similar across datasets, the magnitude of the improvements is substantially larger on V-COCO. We attribute this difference to the higher interaction density of V-COCO: on average, its images contain 6.04 detected objects and 4.46 interactions, compared with 5.25 detected objects and 2.10 interactions in HICO-DET. Isolating the target human–object region may be particularly helpful in scenes containing many simultaneous interactions, as it reduces ambiguity among multiple candidate interactions. Apart from the larger magnitude of the changes in V-COCO, some methods also behave slightly differently across the two datasets. Crop improves over the corresponding baseline on V-COCO, whereas it is neutral or detrimental on HICO-DET. Black Other Objects consistently improves performance by removing potentially distracting visual information. However, in interaction-dense scenes, masking several objects may introduce sharp artificial boundaries. One possible interpretation is that the vision encoders of different MLLMs respond differently to these artifacts. Although our experiments do not isolate this factor, it could explain the model-dependent performance differences observed on V-COCO for Black Other Objects.

Table 7 contextualizes the best configuration among representative HOI detection methods under different settings according to its supervision level. Results show that Qwen2-VL 2B with Blur the Background

Method	HICO-DET			V-COCO	
	Full	Rare	Non-rare	$AP_{role}^{\#1}$	$AP_{role}^{\#2}$
<i>MiniCPM-V 1</i>					
Baseline	20.78	26.63	19.03	19.95	21.98
Objects as Context	22.17	27.53	20.57	21.46	23.09
Overlay Bounding Boxes	15.90	19.55	14.81	5.65	6.11
Overlay Elliptic Marker	17.70	22.55	16.25	5.94	6.24
Blur Background	22.46	27.83	20.85	24.82	26.86
Crop	20.96	26.49	19.31	23.70	26.16
Black Other Objects	22.43	28.60	20.59	26.16	28.56
Gray Background	19.59	26.02	17.67	7.18	7.73
Textual Center Coords.	13.84	14.98	13.50	6.29	6.64
Textual Boundary Coords.	15.41	20.35	13.94	13.38	15.16
<i>Qwen2-VL 2B</i>					
Baseline	20.00	27.59	17.73	17.28	19.19
Objects as Context	20.56	27.50	18.48	18.03	19.95
Overlay Bounding Boxes	19.41	24.86	17.78	6.70	7.14
Overlay Elliptic Marker	19.86	25.70	18.12	5.31	5.53
Blur Background	23.69	30.04	21.79	32.27	34.93
Crop	18.46	25.56	16.34	19.19	21.81
Black Other Objects	23.01	29.87	20.97	22.89	25.17
Gray Background	19.65	24.75	18.13	11.05	11.79
Textual Center Coords.	12.84	14.74	12.27	4.95	6.03
Textual Boundary Coords.	15.76	20.40	14.37	18.07	20.60
<i>LLaVA-OV-0.5B (OV)</i>					
Baseline	18.52	26.82	16.04	16.31	18.61
Objects as Context	19.23	25.10	17.47	17.65	19.28

Table 6: *Pure zero-shot* evaluation on HICO-DET and V-COCO of the prompting strategies for object specification.

achieves 23.69 mAP on HICO-DET and 32.3/34.9 $AP_{role}^{\#1}/AP_{role}^{\#2}$ on V-COCO, surpassing the previously reported *pure zero-shot* method RLIPv2.

The *pure zero-shot* setting is particularly promising for rare categories. Since no model is trained on the evaluation data, the interaction reasoning module does not directly inherit the long-tail class distributions of the benchmark annotations. Consistent with this expectation, Table 7 shows that the best configuration performs better on Rare categories than on Non-rare categories on HICO-DET (30.04 vs. 21.79 mAP). Moreover, on the Rare split, it obtains a strong performance compared to the methods leveraging a pre-trained object detector.

5.4 HOI detection with ground truth objects

To further isolate the effect of the benchmarked prompting strategies, we introduce a *pure zero-shot* evaluation protocol that leverages ground-truth bounding boxes on HICO-DET. This protocol disentangles the contribution of object detection errors from interaction classification performance, allowing a more direct assessment of object specification strategies. While this evaluation setup has previously been applied to supervised HOI methods (Gao et al. (2018), Zhang et al. (2022b)), our work is the first to explore it in a *pure zero-shot* regime. In addition to reporting the quantitative results, we provide a qualitative analysis of the proposed visual prompting methods to better understand their behavior.

Method	HICO-DET			V-COCO	
	Full	Rare	Non-rare	$AP_{role}^{\#1}$	$AP_{role}^{\#2}$
<i>Fully supervised</i>					
GEN-VLKT (Liao et al. (2022))	34.95	31.18	36.08	63.6	65.9
SICHOI (Luo et al. (2024))	45.04	45.61	44.88	71.1	75.6
Pose-Aware (Wu et al. (2024a))	46.01	46.74	45.80	63.0	68.7
<i>Pre-trained object detector</i>					
ConsNet (Liu et al. (2020))	22.15	17.55	23.52	-	-
STIP (Zhang et al. (2022b))	28.81	27.55	29.18	-	-
<i>Partially zero-shot</i>					
CLIP4HOI (Mao et al. (2023))	32.11	-	-	-	-
KI2HOI (Xue et al. (2024))	34.56	-	-	-	-
<i>Pure Zero-shot</i>					
RLIPv2 (Yuan et al. (2023))	23.29	27.97	21.90	-	-
Qwen2-VL 2B with Blur the background	23.69	30.04	21.79	32.3	34.9

Table 7: Comparison with other methods on HICO-DET and V-COCO. Fully supervised methods are trained end-to-end with HOI annotations, while methods with a pre-trained object detector reuse a pre-trained object detector and train only the interaction head. Partially zero-shot methods hold out 120 action-object pairs during training and report HICO-DET performance over the full set of 600 classes, without separate Rare and Non-rare results. V-COCO does not define an analogous held-out setting. *Pure zero-shot* methods do not use any HOI training data. Not all methods are evaluated on both datasets; - denotes results that are not reported. We report the best overall configuration from Table 6 as our method.

Table 8 summarizes the results. Among the evaluated strategies, only Objects as Context and Black Other Objects consistently improve HOI classification across all models relative to their baselines. Overall, using ground-truth bounding boxes yields substantially higher mAP compared to Table 6, confirming that a significant portion of the performance gap arises from detector errors. The best-performing configuration is Objects as Context with Qwen2-VL 2B.

Qualitative analysis: the following observations help explain why the proposed visual prompting methods (Objects as Context and Black Other Objects) obtain superior results in this setting on HICO-DET. All qualitative analyses are performed using Qwen2-VL 2B, as it is the MLLM reporting better results.

- (i) Figure 7 shows that some HOI classes experience a steep increase in performance. Most of these classes (*Sign baseball bat*, *Wash Sheep*, and *Inspect backpack*) are within the set of Rare interactions and have



Figure 7: Image examples of the HOIs that exhibit the highest performance gains when applying the identified prompting methods. Objects as Context improves detection for interactions *Sign baseball bat* (+47.69 AP), *Wash sheep* (+29.44 AP) and *Repair oven* (+27.26 AP), while Black Other Objects enhances detection for *Inspect backpack* (+34.43 AP), *Sign baseball bat* (+28.83 AP) and *Eat cake* (+28.75 AP).

Method	mAP		
	Full	Rare	Non-rare
<i>MiniCPM-V 1</i>			
Baseline	46.35	47.79	45.32
Objects as Context	48.41	50.85	47.68
Overlay Bounding Boxes	39.31	42.06	38.49
Overlay Elliptic Marker	42.89	45.38	42.14
Blur Background	45.93	47.94	45.33
Crop	48.41	51.80	47.39
Black Other Objects	47.64	52.43	46.20
Gray Background	40.58	46.19	38.91
Textual Center Coordinates	36.80	31.57	38.37
Textual Boundary Coordinates	41.50	42.25	41.27
<i>Qwen2-VL 2B</i>			
Baseline	49.20	53.51	47.92
Objects as Context	51.41	56.57	49.86
Overlay Bounding Boxes	45.18	48.49	44.19
Overlay Elliptic Marker	45.55	47.00	45.12
Blur Background	46.43	51.51	44.91
Crop	48.83	53.86	47.34
Black Other Objects	50.01	55.14	48.48
Gray Background	42.65	47.52	41.20
Textual Center Coordinates	31.71	28.75	32.60
Textual Boundary Coordinates	39.03	42.25	38.07
<i>LLaVA-OV-0.5B (OV)</i>			
Baseline	44.06	49.15	42.53
Objects as Context	47.33	50.67	46.33

Table 8: *Pure zero-shot* evaluation on HICO-DET removing the effect of the object detector.

less than or equal to five ground truth images in the test set. The increase might be due to the benefits of properly specifying small or background objects (*e.g.* in Figures 7b,d the action of the HOI class mentioned occurs in the background) or occluded objects (*e.g.* the cake in Figure 7e is out of the image frame).

- (ii) Figures 8a,b,c show examples of the HOI classes that experience a larger performance drop when applying Objects as Context for object specification. We identify that the performance decline may occur when either there is no interaction (Figures 8a,b) or cropping the object deletes crucial information (Figure 8c). In the former case, crops of the objects lead to higher classification scores of interactive HOIs which, by the construction of the *no_interaction* score, imply a lower score assigned to it. In the latter case, crops of the object targeted by the action (*e.g.* wine glass) are not meaningful without crops of the instruments performing the action (*e.g.* wine bottle) in order to infer the action (*e.g.* fill a wine glass).
- (iii) Figures 8d,e,f show examples of the HOI classes that experience a larger performance drop when applying Black Other Objects for object specification. In all examples there are several people performing the same action. The drop in performance may stem from the removal of additional examples, which may help the MLLM reinforce the certainty of interaction and better understand the scene. The opposite effect happens when using this prompting strategy with several individuals performing different actions on the same object, as explained in observation (vii).



Figure 8: Image examples of the HOIs that exhibit the highest performance losses when applying the identified prompting methods. Objects as Context suffers detection decrease for interactions *No_interaction spoon* (-28.90 AP), *No_interaction carrot* (-21.41 AP) and *Fill wine glass* (-17.62 AP), while Black Other Objects downgrades detection for *Lie on surfboard* (-48.34 AP), *No_interaction bicycle* (-36.23 AP) and *Buy pizza* (-35.97 AP).

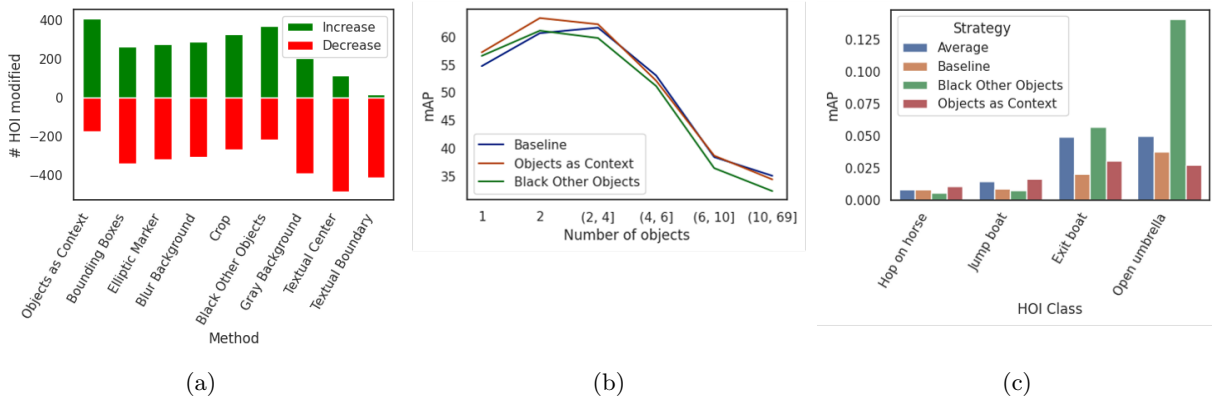


Figure 9: Impact of prompting methods on HICO-DET using ground truth bounding boxes. (a) Number of HOIs showing an increase or decrease in AP compared to the baseline for all prompting methods. (b) Performance of the proposed methods relative to the number of detected objects of the target class. (c) Performance gains of the proposed methods on the most challenging classes. Most challenging classes are considered those that have a lower average AP over all object specification approaches.

- (iv) Figure 9a shows the number of HOI classes whose performance either improves or worsens when introducing each prompting method, compared to the baseline of Qwen2-VL 2B. It is observed that Objects as Context and Black Other Objects, along with Crop, are the only strategies able to improve the performance for a majority of HOIs.
- (v) Figure 9c highlights the worst-performing HOI classes across all object specification prompting methods. For the classes *Hop on horse* and *Jump boat*, Objects as Context slightly improves performance relative to both the baseline and the average of the prompting methods. Similarly, for *Exit boat* and *Open umbrella*, Black Other Objects proves beneficial. Overall, performance on these classes remains very low, underscoring the inherent difficulty of the task for any prompting strategy.
- (vi) Figure 9b shows the performance according to the number of objects detected that belong to the target class. As illustrated, the higher the number of objects detected of the same class, the more likely the MLLM will be confused. To avoid specification confusion, MLLMs should be able to correctly understand prompting techniques to focus on the specified object instance.
- (vii) Figures 10a,b,c,d show examples of images and HOI classes that Objects as Context predict successfully but all other prompting methods fail to detect. Objects as Context may provide the MLLM with richer contextual cues on the specified objects by providing crops of the objects and people. These crops might

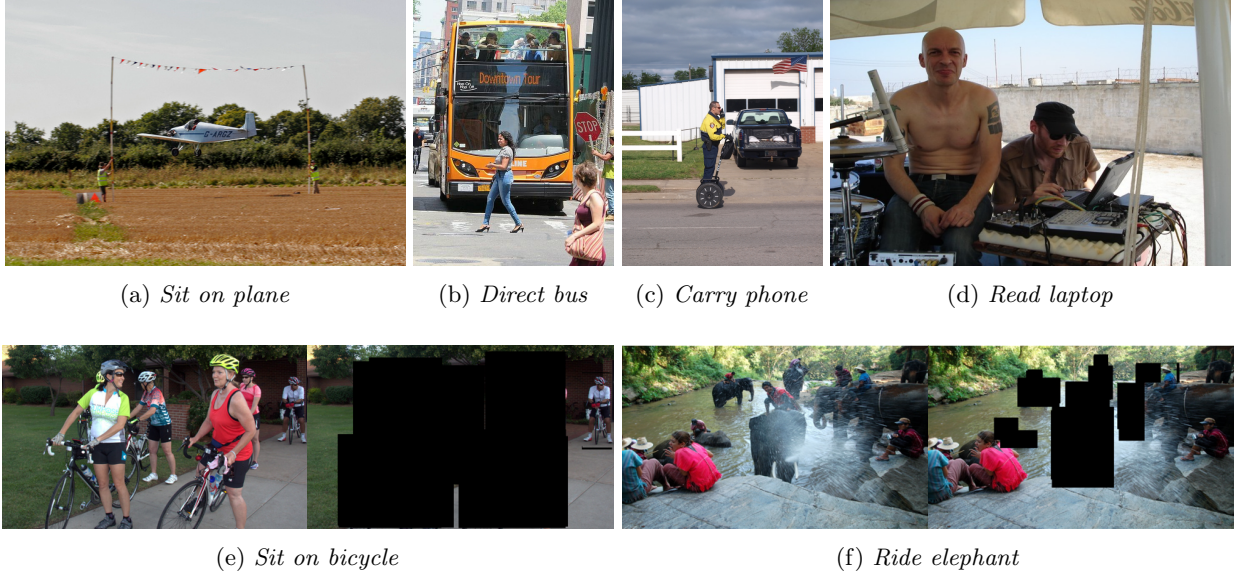


Figure 10: Examples where Objects as Context (a-d) and Black Other Objects (e-f) correctly classify a HOI that all other methods fail to detect: (a) a focused crop of the pilot, (b) a crop of the person holding the STOP sign, (c) a crop of the small and partially occluded cell phone, and (d) a crop of the laptop and the interacting person presumably help the model with more specific visual context. (e) Masking other cyclists and (f) masking irrelevant overlapping objects presumably direct model’s attention to the correct interaction.

be especially important when objects or people are either small (Figures 10a,b) or partially occluded (Figure 10c), since their visual tokens might be enhanced.

(viii) Figures 10e,f show examples of images and HOI classes, along with their representation after applying the visual prompting strategy, that Black Other Objects predict successfully while all other prompting methods tested fail to detect. In both cases, there are many people interacting differently with the same object class (bicycle for Figure 10e and elephant for Figure 10f), which may confuse the MLLM. Black Other Objects is able to erase misleading information from the scene while keeping the relevant context, thus helping the MLLM direct its attention to the correct area in the image.

In summary, isolating detection errors through this evaluation protocol not only clarifies the impact of visual prompting on object specification but also provides interpretable insights into how MLLMs handle this challenge, paving the way for more reliable zero-shot approaches to open-world HOI.

5.5 Ablation studies

To better understand the effects and robustness of the different components of the proposed HOI recognition pipeline, we perform a series of ablation studies on the HICO dataset using mainly Qwen2-VL 2B. These experiments analyze how different components and hyperparameters influence HOI recognition performance and evaluation computational cost. Specifically, we study the computation of the *no_interaction* score, the impact of the object detector model, the effect of object-class filtering through the threshold τ , and the role of the context-selection parameters k_1 and k_2 .

5.5.1 No-interaction score

We analyze the behavior of the *no_interaction* computation by splitting the 600 HICO classes into 520 interacting classes and 80 *no_interaction* classes. The interacting subset reaches 57.87 mAP, whereas the *no_interaction* subset obtains 9.32 mAP. This gap indicates that the product-based score in Equation 2 accumulates uncertainty for this class: although it makes the score sounder, ambiguous positive scores

across many interactions can quickly reduce the final *no_interaction* confidence. However, treating *no_interaction* as an independent class with its own questions could lead to inconsistent outputs, such as all scores being low, including *no_interaction*, or assigning high scores to *no_interaction* and interactive classes simultaneously.

5.5.2 Object Detector

Since our pipeline relies on an external object detector as the entry point, we analyze how sensitive the overall HOI recognition performance is to the choice of object detector backbone. Table 9 reports results for four established detectors, YOLOX-x (Ge et al. (2021)), YOLO11-x (Khanam and Hussain (2024)), Faster R-CNN (Ren et al. (2015)), and DINO (Zhang et al. (2022a)), evaluated with MiniCPM-V 1 and Qwen2-VL 2B. Performance differences across detectors are modest: DINO achieves the highest mAP (55.49 with Qwen2-VL 2B), but at the cost of retaining over 350k queries, which substantially increases evaluation overhead. YOLO11-x offers the most favorable trade-off, reaching competitive mAPs of 48.20 (MiniCPM-V 1) and 51.39 (Qwen2-VL 2B) while minimizing the number of retained queries ($\approx 252k$). For this reason, we adopt YOLO11-x as the default detector in our pipeline.

Object Detector	MiniCPM-V 1	Qwen2-VL 2B	#Questions
YOLOX-x (Ge et al. (2021))	48.59	52.45	275,746
YOLO11-x (Khanam and Hussain (2024))	48.20	51.39	251,593
Faster R-CNN (Ren et al. (2015))	48.31	52.28	351,466
DINO (Zhang et al. (2022a))	51.27	55.49	352,653

Table 9: Comparison of object detectors. Evaluation is conducted on HICO. # Questions denotes the number of retained queries after the object detector filtering step. All models are trained on MS-COCO.

To further illustrate detector robustness, we include in the Appendix an extended comparison of YOLO11 model sizes and an exploratory experiment where the MLLM itself is used as an object recognizer. Overall, these results show that our benchmark setup is not overly sensitive to the specific choice of detector, and that the main trends reported in Sections 5.2–5.4 hold across reasonable detection backbones.

5.5.3 Class-wise object detection threshold

We next analyze the effect of the object-class filtering stage, which determines which HOI categories are questioned to the MLLM. This experiment aims to quantify how much the pipeline depends on the object detector-based filtering step. Table 10 evaluates different values of the object class score threshold τ . Lowering τ increases the number of retained object classes, which consequently increases the number of questions evaluated by the MLLM. If the pipeline relied strongly on this filtering stage, retaining more object classes would be expected to produce substantial performance gains. We select $\tau = 0.10$ as the default operating point because it provides a favorable balance between recognition accuracy and computational cost. More importantly, the relatively small performance variation across thresholds shows that the object-class filtering stage does not dominate the final recognition accuracy, indicating that the proposed pipeline remains robust even when the number of retained object classes changes considerably.

τ	MiniCPM-V 1	Qwen2-VL 2B	#Questions
0.00	50.98	53.59	990,910
0.05	49.33	52.61	273,668
0.10	48.20	51.39	251,593
0.25	47.92	51.67	223,304
0.40	47.06	50.82	206,492
0.50	46.23	49.63	195,441

Table 10: Effect of varying class score thresholds in the pipeline. Evaluation conducted on HICO.

5.5.4 Number of objects and people retained

In Objects as Context, hyperparameters k_1 and k_2 control the maximum number of objects and people provided to the MLLM as context. Objects are filtered by target class and ranked by detection confidence, while people are ranked by bounding box size, prioritizing larger instances as they should represent more relevant people in the scene. Tables 11 and 12 show that there is no great difference between the values, but $k_1 = 10$ and $k_2 = 3$ are selected because of their slightly superior results.

# Objects (k_1)	mAP
1	53.35
5	53.45
10	53.50
20	53.47
50	53.47

Table 11: Effect of varying the number of objects kept for Objects as Context. Evaluation conducted on HICO using Qwen2-VL 2B.

# People (k_2)	mAP
1	53.45
3	53.50
5	53.44
10	53.47

Table 12: Effect of varying the number of people kept for Objects as Context. Evaluation conducted on HICO using Qwen2-VL 2B.

6 Conclusions, Limitations, and Future Work

We presented the first benchmark and prompting taxonomy for *pure zero-shot* HOI recognition and detection using MLLMs and proposed two effective visual prompting strategies, Objects as Context and Black Other Objects, which improve performance by guiding model attention and reducing distractions. Evaluations on the constructed Q&A datasets established a strong *pure zero-shot* baseline, achieving zero-shot state-of-the-art results (53.50 mAP on HICO, 23.69 mAP on HICO-DET, 32.27 mAP on V-COCO) without HOI-specific fine-tuning.

While these results are promising, several limitations should be considered. First, the reported evaluation performance inherits the biases of the underlying HOI datasets, including their annotation errors and fixed interaction categories. These biases may affect the evaluation protocol and the interpretation of the reported metrics. However, this limitation does not affect the model’s generalization capabilities, since it is evaluated in a *pure zero-shot* setting.

Second, our pipeline relies on pre-trained models, namely the external object detector and the MLLMs used for interaction reasoning, whose capabilities and errors may limit the overall performance. Because the detector is restricted to the object categories learned during training, in our case COCO classes, the complete pipeline is not fully open-world and may miss objects outside this vocabulary. This restriction does not limit the underlying HOI recognition capabilities of the MLLM, since the detector is used only to filter the questions evaluated for efficiency and can be removed at deployment. For HOI detection, future work could instead leverage the grounding capabilities of recent MLLMs to localize open-vocabulary objects directly.

Third, although the proposed pipeline incorporates several practical optimizations to reduce computational and memory requirements, such as object class filtering, evaluating MLLMs for HOI datasets remains computationally demanding due to the large number of image-question pairs that must be processed. These design choices introduce an inherent trade-off between computational efficiency and evaluation completeness. Developing more efficient evaluation strategies that preserve performance while reducing inference cost remains an important direction for future research.

Fourth, *no.interaction* labels remain challenging. Future work should improve their scoring while balancing output consistency with accurate estimation of interaction absence.

Overall, this benchmark enables systematic evaluation of MLLMs on open-world HOI and provides a basis

for extending the proposed methodology to other multi-label problems, particularly those characterized by long-tail class distributions, where *pure zero-shot* inference may reduce reliance on imbalanced annotations.

Ethical considerations

HOI datasets such as HICO, HICO-DET, and V-COCO often include images sourced from the internet that depict people in everyday activities. While our work focuses on technical benchmarking, methods developed here could potentially be applied in sensitive domains such as surveillance or security. We emphasize that such applications raise privacy and ethical concerns, and we encourage future work to carefully consider safeguards and responsible deployment.

Acknowledgements

This work has been partially supported by the Spanish project PID2025-170489NB-I00 and by ICREA under the ICREA Academia programme.

References

- M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.
- J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
- M. Besta, N. Blach, A. Kubicek, R. Gerstenberger, M. Podstawski, L. Gianinazzi, J. Gajda, T. Lehmann, H. Niewiadomski, P. Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690, 2024.
- Z. Cai, M. Cao, H. Chen, K. Chen, K. Chen, X. Chen, X. Chen, Z. Chen, Z. Chen, P. Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.
- R. Cao and J. Jiang. Knowledge generation for zero-shot knowledge-based vqa. *arXiv preprint arXiv:2402.02541*, 2024.
- Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE international conference on computer vision*, pages 1017–1025, 2015.
- Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to detect human-object interactions. In *2018 IEEE winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018.
- B. Chen, Z. Zhang, N. Langrené, and S. Zhu. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *arXiv preprint arXiv:2310.14735*, 2023.
- J. Chen, B. Lin, X. Liu, L. Ma, X. Liang, and K.-Y. K. Wong. Affordances-oriented planning using foundation models for continuous vision-language navigation. *arXiv preprint arXiv:2407.05890*, 2024a.
- Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.
- Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024c.

- Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. 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, 2024d.
- T. Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.
- M. Dehghani, B. Mustafa, J. Djolonga, J. Heek, M. Minderer, M. Caron, A. Steiner, J. Puigcerver, R. Geirhos, I. M. Alabdulmohsin, et al. Patch n’pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274, 2023.
- X. Dong, P. Zhang, Y. Zang, Y. Cao, B. Wang, L. Ouyang, S. Zhang, H. Duan, W. Zhang, Y. Li, et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *Advances in Neural Information Processing Systems*, 37:42566–42592, 2024.
- C. Gao, Y. Zou, and J.-B. Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018.
- Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. YoloX: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.
- G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with r* cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1080–1088, 2015.
- Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.
- A. Gupta, P. Vuillecard, A. Farkhondeh, and J.-M. Odobez. Exploring the zero-shot capabilities of vision-language models for improving gaze following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 615–624, 2024.
- S. Gupta and J. Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.
- Z. Hou, X. Peng, Y. Qiao, and D. Tao. Visual compositional learning for human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 584–600. Springer, 2020.
- A. Hu, Y. Shi, H. Xu, J. Ye, Q. Ye, M. Yan, C. Li, Q. Qian, J. Zhang, and F. Huang. mplug-paperowl: Scientific diagram analysis with the multimodal large language model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6929–6938, 2024.
- J. Huang, S. Yong, X. Ma, X. Linghu, P. Li, Y. Wang, Q. Li, S.-C. Zhu, B. Jia, and S. Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023.
- S. Jiang, Y. Zhang, C. Zhou, Y. Jin, Y. Feng, J. Wu, and Z. Liu. Joint visual and text prompting for improved object-centric perception with multimodal large language models. *arXiv preprint arXiv:2404.04514*, 2024.
- Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, and G. Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, 2023.
- Y. Jin, Y. Chen, L. Wang, J. Wang, P. Yu, L. Liang, J.-N. Hwang, and Z. Liu. The overlooked classifier in human-object interaction recognition. *arXiv preprint arXiv:2203.05676*, 2022.
- R. Khanam and M. Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

- S. Kim, D. Jung, and M. Cho. Relational context learning for human-object interaction detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2925–2934, 2023.
- T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
- Y. Lan, X. Li, X. Liu, Y. Li, W. Qin, and W. Qian. Improving zero-shot visual question answering via large language models with reasoning question prompts. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4389–4400, 2023.
- X. Lei, Z. Yang, X. Chen, P. Li, and Y. Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *arXiv preprint arXiv:2402.12058*, 2024.
- B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.
- Y.-L. Li, L. Xu, X. Liu, X. Huang, Y. Xu, S. Wang, H.-S. Fang, Z. Ma, M. Chen, and C. Lu. Pastanet: Toward human activity knowledge engine. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 382–391, 2020.
- Y. Liao, A. Zhang, M. Lu, Y. Wang, X. Li, and S. Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20123–20132, 2022.
- J. Lin, H. Yin, W. Ping, P. Molchanov, M. Shoybi, and S. Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699, 2024.
- T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. 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.
- Y. Lin, Y. Xie, D. Chen, Y. Xu, C. Zhu, and L. Yuan. Revive: Regional visual representation matters in knowledge-based visual question answering. *Advances in neural information processing systems*, 35: 10560–10571, 2022.
- H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- J. Liu, A. Liu, X. Lu, S. Welleck, P. West, R. L. Bras, Y. Choi, and H. Hajishirzi. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387*, 2021.
- Y. Liu, J. Yuan, and C. W. Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *Proceedings of the 28th ACM international conference on multimedia*, pages 4235–4243, 2020.
- Y. Liu, B. Yang, Q. Liu, Z. Li, Z. Ma, S. Zhang, and X. Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024.
- P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

- B. Luan, H. Feng, H. Chen, Y. Wang, W. Zhou, and H. Li. Textcot: Zoom in for enhanced multimodal text-rich image understanding. *arXiv preprint arXiv:2404.09797*, 2024.
- J. Luo, W. Ren, W. Jiang, X. Chen, Q. Wang, Z. Han, and H. Liu. Discovering syntactic interaction clues for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28212–28222, 2024.
- X. Ma, W. Nie, Z. Yu, H. Jiang, C. Xiao, Y. Zhu, S.-C. Zhu, and A. Anandkumar. Relvit: Concept-guided vision transformer for visual relational reasoning. *arXiv preprint arXiv:2204.11167*, 2022.
- Y. Mao, J. Deng, W. Zhou, L. Li, Y. Fang, and H. Li. Clip4hoi: Towards adapting clip for practical zero-shot hoi detection. *Advances in Neural Information Processing Systems*, 36:45895–45906, 2023.
- K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.
- M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, and P. Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.
- S. Nasiriany, F. Xia, W. Yu, T. Xiao, J. Liang, I. Dasgupta, A. Xie, D. Driess, A. Wahid, Z. Xu, et al. Pivot: Iterative visual prompting elicits actionable knowledge for vlms. *arXiv preprint arXiv:2402.07872*, 2024.
- S. Ning, L. Qiu, Y. Liu, and X. 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.
- A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11: 1316–1331, 2023.
- S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.
- L. Shen, S. Yeung, J. Hoffman, G. Mori, and L. Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In *2018 IEEE winter conference on applications of computer vision (wacv)*, pages 1568–1576. IEEE, 2018.
- A. Shtedritski, C. Rupprecht, and A. Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11987–11997, 2023.
- G. Tziafas and H. Kasaei. Towards open-world grasping with large vision-language models. *arXiv preprint arXiv:2406.18722*, 2024.
- D. Wan, J. Cho, E. Stengel-Eskin, and M. Bansal. Contrastive region guidance: Improving grounding in vision-language models without training. In *European Conference on Computer Vision*, pages 198–215. Springer, 2024.
- P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

- X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.
- E. Z. Wu, Y. Li, Y. Wang, and S. Wang. Exploring pose-aware human-object interaction via hybrid learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17815–17825, 2024a.
- J. Wu, Z. Zhang, Y. Xia, X. Li, Z. Xia, A. Chang, T. Yu, S. Kim, R. A. Rossi, R. Zhang, et al. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*, 2024b.
- J. Xie, T. Han, M. Bain, A. Nagrani, G. Varol, W. Xie, and A. Zisserman. Autoad-zero: A training-free framework for zero-shot audio description. In *Proceedings of the Asian Conference on Computer Vision*, pages 2265–2281, 2024.
- W. Xue, Q. Liu, Q. Xiong, Y. Wang, Z. Wei, X. Xing, and X. Xu. Towards zero-shot human-object interaction detection via vision-language integration. *arXiv preprint arXiv:2403.07246*, 2024.
- A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.
- J. Yang, B. Li, F. Yang, A. Zeng, L. Zhang, and R. Zhang. Boosting human-object interaction detection with text-to-image diffusion model. *arXiv preprint arXiv:2305.12252*, 2023a.
- J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023b.
- L. Yang, Y. Wang, X. Li, X. Wang, and J. Yang. Fine-grained visual prompting. *Advances in Neural Information Processing Systems*, 36:24993–25006, 2023c.
- S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36: 11809–11822, 2023.
- Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.
- H. Yuan, J. Jiang, S. Albanie, T. Feng, Z. Huang, D. Ni, and M. Tang. Rlip: Relational language-image pre-training for human-object interaction detection. *Advances in Neural Information Processing Systems*, 35:37416–37431, 2022.
- H. Yuan, S. Zhang, X. Wang, S. Albanie, Y. Pan, T. Feng, J. Jiang, D. Ni, Y. Zhang, and D. Zhao. Rlipv2: Fast scaling of relational language-image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21649–21661, 2023.
- H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022a.
- Y. Zhang, Y. Pan, T. Yao, R. Huang, T. Mei, and C.-W. Chen. Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19548–19557, 2022b.
- Y. Zhang, Y. Yuan, and A. C.-C. Yao. Meta prompting for agi systems. *arXiv preprint arXiv:2311.11482*, 2023a.
- Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023b.

- D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.
- D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- C. Zou, B. Wang, Y. Hu, J. Liu, Q. Wu, Y. Zhao, B. Li, C. Zhang, C. Zhang, Y. Wei, et al. End-to-end human object interaction detection with hoi transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11825–11834, 2021.

Appendices

For a better understanding of the main paper, we provide additional details in this appendix, including:

- Section A shows the prompts used to implement Multimodal Chain-of-Thought.
- Section B shows the prompts used to implement Caption as Context.
- Section C shows the procedure to generate prompts for Ensemble of Prompts.
- Section D shows examples of textual prompts used for each object specification prompting strategy.
- Section E further lists the hyperparameters used.
- Section F studies the effect of the size of the object detector on our pipeline.

A Multimodal Chain-of-Thought templates

The reasoning generation step adds the sentence “Think step-by-step.” (Kojima et al. (2022)) at the end of the question. The final question prompt is **Context:** <reasoning>. **Question:** <Question>. **Answer** with yes or no..

B Caption as Context templates

The caption generation step uses “Describe the image in detail.”. The final question prompt is **Context:** <caption>. **Question:** <Question>. **Answer** with yes or no..

C Ensemble of Prompts generation

We use the action synonyms provided in the action description of the HICO (Chao et al. (2015)) dataset to build new prompt templates for each HOI class. The generation of questions is shown in Figure C.1. The different templates for *person_option* are adapted from Gupta et al. (2024).

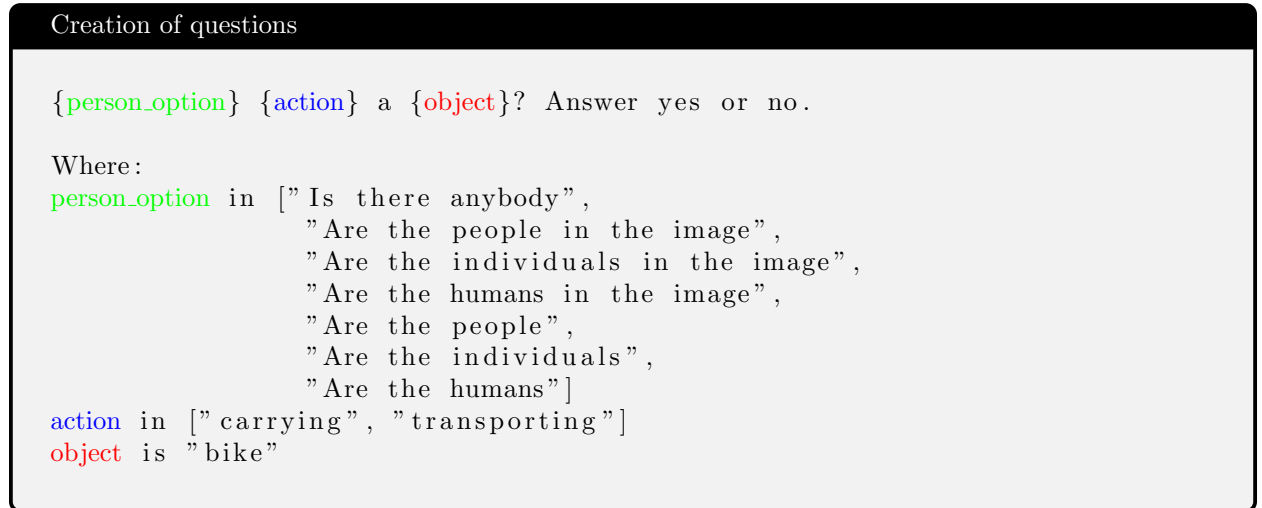


Figure C.1: Prompt generation for ensemble of prompts. A prompt is generated by substituting each part of the main question (top part) with combinations of person_option, action, and object. Each object-action pair has its own set of possible action synonyms.

In the cases where the desired number of questions is lower than the possible number of questions, a subset is retrieved at random. In the opposite case, when the ensemble number is higher than the possible number of questions, all are retrieved, and the first ones are computed more than once. Notice that the number of questions differs by the specific action since each verb has a different number of synonyms.

D Object specification prompting strategy textual templates

Following the visual example in Figure 5, the question for the HOI *carry backpack* for each object specification strategy tested is:

- Baseline: Is there anybody carrying a backpack? Answer with yes or no.
- Objects as Context: Is there anybody carrying a backpack? Answer with yes or no.
- Overlay Bounding Boxes: Is the person in the blue box carrying the backpack in the red box? Answer yes or no.
- Overlay Elliptic Markers: Is the person within the blue marker carrying the backpack within the red marker? Answer yes or no.
- Blur the Background: Is the foreground person carrying a backpack? Answer yes or no.
- Crop the relevant area: Is there anybody carrying a backpack? Answer with yes or no.
- Black Other Objects: Is there anybody carrying a backpack? Answer with yes or no.
- Gray the Background: Is the colored person carrying the colored backpack? Answer yes or no.
- Textual Center Coordinates: Is the person centered at (209, 368) carrying the backpack centered at (209, 320)? Answer yes or no.
- Textual Boundary Coordinates: Is the person within image coordinates (168, 243), (168, 493), (250, 243) and (250, 493) carrying the backpack within image coordinates (177, 278), (177, 363), (241, 278) and (241, 363)? Answer yes or no.

D.1 Thresholded accuracy analysis

One of the objectives of this work is to provide a benchmark for *pure zero-shot* HOI recognition that is directly comparable with existing methods. Therefore, the main paper reports results using mean Average Precision (mAP), the standard evaluation metric for HOI recognition. Since some prompting strategies tend to produce more confident predictions, we additionally analyze their binary classification performance after thresholding.

For each HOI query, the predicted probability is converted into a binary prediction using a threshold of 0.5. Predictions with a score greater than or equal to 0.5 are assigned the positive label, while the remaining predictions are assigned the negative label. The final accuracy is then computed as the percentage of correctly classified HOI queries over the entire evaluation set.

Strategy	MiniCPM-V 1	Qwen2-VL 2B	LLaVA-OneVision 7B
Baseline	98.88	99.10	98.68
MCoT	99.15	98.99	-
Caption as Context	99.03	99.21	-
Objects as Context	98.67	99.16	99.51

Table D.1: Thresholded accuracy (%) on HICO. Binary predictions are obtained by thresholding the predicted probabilities at 0.5.

Table D.1 reports the corresponding thresholded accuracy. The Multi-step with Generated Knowledge strategies achieve accuracy comparable to or slightly higher than the baseline, despite obtaining lower mAP in the main evaluation. This indicates that these strategies tend to produce more decisive predictions, improving the final classification after thresholding, while yielding score distributions that are less suitable for the ranking-based mAP metric. In contrast, Objects as Context provides the largest gains under mAP while maintaining a comparable thresholded accuracy, suggesting that its main benefit lies in improving the calibration and ranking of prediction scores rather than the binary decisions themselves.

E List of hyperparameters

Hyperparameter	Value
Attention implementation	Flash Attention 2 from Dao (2023)
Generation strategy	Greedy
Maximum number of new tokens	1
Maximum number of output tokens for generated knowledge	256
Minimum size of cropped objects ³	28 per side

Table E.1: List of hyperparameters for MLLMs in inference mode.

F Object detector size

Table F.1 shows the results of using different model sizes within the YOLO11 family. Performance scales smoothly with detector size up to the *large* variant. In addition, we report an approach that uses MLLMs as object recognizers based on the proposed classification score, thus removing the necessity of a pre-trained object detector. This approach directly queries the presence or absence of objects without localization, prompting one question per object class (*e.g.*, 80 per image for MS-COCO). However, this approach results in a very low performance for Qwen2-VL 2B, possibly because leveraging the same MLLM for both object recognition and HOI recognition propagates errors between tasks, highlighting the necessity of a dedicated detector.

Model	MiniCPM-V 1	Qwen2-VL 2B	#Questions
nano	43.37	47.16	240,581
small	45.86	49.73	249,280
medium	47.28	51.15	251,102
large	48.55	52.32	252,084
extra-large	48.20	51.39	251,593
MLLM (Qwen2-VL 2B)	50.96	14.45	264,782

Table F.1: Performance across YOLO11 model sizes. Evaluation conducted on HICO. All detectors are trained on MS-COCO.

³Only apply to Qwen2-VL and LLaVA-OneVision models.