InstructHOI: Context-Aware Instruction for Multi-Modal Reasoning in Human-Object Interaction Detection

Jinguo Luo[1,2], **Weihong Ren**[1,*], **Quanlong Zheng**[2,†], **Yanhao Zhang**[2,†],
Zhenlong Yuan[3], **Zhiyong Wang**[1], **Haonan Lu**[2], **Honghai Liu**[1]

[1]Harbin Institute of Technology, Shenzhen [2]OPPO AI Center

[3]Institute of Computing Technology, Chinese Academy of Sciences

{23s153135, weihongren, zhiyongwang, honghai.liu}@hit.edu.cn
{zhengquanlong, zhangyanhao, luhaonan}@oppo.com
yuanzhenlong21b@ict.ac.cn

Abstract

Recently, Large Foundation Models (LFMs), e.g., CLIP and GPT, have significantly advanced the Human-Object Interaction (HOI) detection, due to their superior generalization and transferability. Prior HOI detectors typically employ single- or multi-modal prompts to generate discriminative representations for HOIs from pretrained LFMs. However, such prompt-based approaches focus on transferring HOI-specific knowledge, but unexplore the potential reasoning capabilities of LFMs, which can provide informative context for ambiguous and open-world interaction recognition. In this paper, we propose InstructHOI, a novel method that leverages context-aware instructions to guide multi-modal reasoning for HOI detection. Specifically, to bridge knowledge gap and enhance reasoning abilities, we first perform HOI-domain fine-tuning on a pretrained multi-modal LFM, using a generated dataset with 140K interaction-reasoning image-text pairs. Then, we develop a Context-aware Instruction Generator (CIG) to guide interaction reasoning. Unlike traditional language-only instructions, CIG first mines visual interactive context at the human-object level, which is then fused with linguistic instructions, forming multi-modal reasoning guidance. Furthermore, an Interest Token Selector (ITS) is adopted to adaptively filter image tokens based on context-aware instructions, thereby aligning reasoning process with interaction regions. Extensive experiments on two public benchmarks demonstrate that our proposed method outperforms the state-of-the-art ones, under both supervised and zero-shot settings.

1 Introduction

Human-Object Interaction (HOI) detection plays a crucial role in high-level human-centric understanding, with applications across various domains [1, 2, 3]. The purpose of HOI detection is to detect a series of interactive triplets (i.e., $\langle$*human*, *action*, *object*$\rangle$) in open-world scenarios. This task can be specifically divided into two sub-tasks: localizing interactive human-object pairs and recognizing their interaction relationships.

Traditional HOI detectors can be primarily classified into one-stage and two-stage approaches. One-stage methods [4, 5, 6, 7] treat HOI detection as a unified multi-task learning problem, utilizing

*Corresponding Author
†Project Leader

39th Conference on Neural Information Processing Systems (NeurIPS 2025)

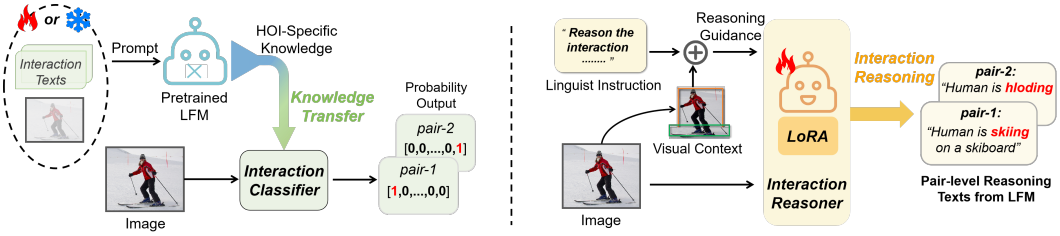

Figure 1: (a) Traditional LFM-based methods typically employ predefined or learnable prompts to transfer HOI-specific knowledge from pretrained LFM for interaction classification. (b) Our InstructHOI integrates visual interactive context with linguistic instructions to guide HOI-domain LFM in performing multi-modal interaction reasoning, generating pair-level interaction texts.

a multi-branch network to simultaneously perform human-object pair detection and interaction prediction. In contrast, two-stage ones [8, 9] first detect human-object instances using an off-the-shelf object detector, subsequently predicting interaction categories based on the visual features extracted from instance areas. Despite significant efforts in feature extraction strategies [10, 11, 12] and architecture improvements [13, 14, 15, 16], accurately identifying complex HOIs within open-world context remains challenging when relying solely on visual representation learning.

To further explore discriminative interaction representations, recent researches transfer HOI-specific knowledge from pretrained LFMs, including Vision-Language Models (VLMs) and Large Language Models (LLMs), using predefined or learnable prompts, as illustrated in Fig. 1(a). For the VLM-style methods, some earlier works [17, 18, 19] leverage static template prompts (e.g., "a photo of a person [action] a/an [object]") to derive linguistic prior knowledge from CLIP [20] at the category level. Subsequently, CMMP [21] introduces learnable multi-modal prompts, facilitating the adaptive transfer of semantic knowledge from CLIP at the instance level. Furthermore, the LLM-style methods [22, 23, 24] adopt language foundation models (e.g., ChatGPT) to generate finer-grained descriptive texts as interactive clue prompts, thereby transferring the generalizable knowledge of LLMs for HOI detection. However, these prompt-based methods primarily focus on knowledge transfer, but fail to exploit LFMs' reasoning capabilities which can provide informative context for ambiguous and open-world interaction recognition.

According to the aforementioned challenges, we propose InstructHOI, which leverages context-aware instructions to direct LFM in performing multi-modal reasoning for HOI detection, as depicted in Fig. 1(b). Specifically, for a pretrained LFM, we first perform HOI-domain fine-tuning to bridge the inherent knowledge gap between general and HOI domains [25] and enhance its interaction reasoning capability, using a light-weight strategy, i.e., LoRA [26]. Due to the limited availability of HOI reasoning data [25], we created a large-scale dataset containing 140K image-text pairs by aggregating five existing image-only HOI datasets and transforming the one-hot labels into interaction-reasoning conversations. Then, we develop a Context-aware Instruction Generator (CIG) to guide interaction reasoning. Unlike traditional language-only instructions [27, 28], CIG first mines visual interactive context (i.e., appearance and spatial context) at the human-object level. Next, the visual context is projected into linguistic space using a two-layer instruction projector, and then is fused with linguistic instructions, providing pair-level context guidance for multi-modal interaction reasoning. Furthermore, an Interest Token Selector (ITS) is adopted to adaptively filter informative image tokens based on the context-aware instructions and reorganize the reasoning token sequences, thereby aligning the reasoning process with interaction regions.

In this paper, our motivation is to explore the potential reasoning capability of LFMs to improve HOI detection. Unlike previous LFM-based approaches, our work directly leverages tailored instructions to guide LFM in facilitating multi-modal reasoning, thereby achieving open-world interaction recognition. Besides, we enhance traditional linguistic instructions by incorporating visual interactive context at the human-object level, thus providing pair-level multi-modal reasoning guidance. To summarize, our contributions are as follows:

- For a pretrained LFM, to bridge the gap between general and HOI-domain knowledge, we build a high-quality interaction-reasoning dataset and perform supervised fine-tuning using a lightweight strategy.

- We develop a Context-aware Instruction Generator (CIG) to enhance linguistic instructions by incorporating informative visual context at the human-object level, providing multi-modal reasoning guidance.

- To align the reasoning process with interaction regions, an Interest Token Selector (ITS) is adopted to adaptively filter and reorganize reasoning token sequences based on context-aware instructions.

- We evaluate our InstructHOI on two benchmarks: HICO-DET and V-COCO, and it outperforms the state-of-the-art methods, achieving superior performance in both supervised and zero-shot settings.

2 Related Work

Traditional HOI Detectors: Traditional HOI detectors can be primarily classified into one-stage and two-stage approaches. One-stage methods regard HOI detection as a multi-task learning, aiming to simultaneously perform object detection and interaction prediction. Earlier methods [29, 5, 4] typically adopt a multi-branch CNN architecture for parallel human-object instance localization and interaction recognition. Then, some auxiliary priors (e.g., interaction points [4] and union boxes [30]) are introduced to align instances with their corresponding interactions. Recently, Transformer-based methods [7, 6, 31] take a prominent position, due to their exceptional context capture ability. However, such a disentangled architecture may suffer from insufficient context exchange between the branches, leading to inferior prediction performance.

Two-stage methods treat HOI detection as two sequential sub-tasks. They initially localize human-object instances with an off-the-shelf detector and then identify interactions leveraging the visual features extracted from the instance regions. The early CNN-based methods [32, 33, 34, 35, 36, 9] strive to extract rich visual interaction representations, e.g., spatial relationship [37], gaze attention [11] and pose feature [33] to assist HOI detection. Recent Transformer-based methods [38, 39, 40, 41, 42] attempt to improve the vanilla Transformer for enhancing feature extraction of HOI detection. Despite significant efforts in feature extraction strategies and architecture enhancements, accurately distinguishing complex HOIs in open world remains challenging when relying solely on visual representation learning.

LFM-based HOI Detectors: LFM-based HOI detectors can be primarily classified into VLM-style and LLM-style. To further explore discriminative HOI representations, recent approaches [43, 44, 45] seek to extract prior knowledge from VLMs [46, 47] by leveraging their distinctive ability to unify visual and linguistic features. Among VLM-style methods, the pioneering works [44, 19] typically transform one-hot labels into annotation texts via a static prompt template, e.g., "a photo of a person [action] a/an [object]". These annotations are then encoded as linguistic priors using CLIP, enabling category-level knowledge transfer. In addition, MP-HOI [43] utilizes extra visual prompts to provide fine-grained visual priors, and aims to eliminate the ambiguity in linguistic descriptions. Furthermore, CMMP [21] introduces learnable multi-modal prompts, facilitating the adaptive transfer of semantic knowledge from CLIP at the instance level.

LLM-style methods [22, 23, 24, 48] usually employ language foundation models to generate finer-grained descriptive texts as interactive prompts, which can transfer the generalizable knowledge from LLMs to improve HOI detection. E.g., UniHOI [22] designs a knowledge retrieval process for ChatGPT to acquire comprehensive explanations for each HOI category, which provides rich contextual information for interaction prediction. CMD-SE [24] introduces a two-step GPT-querying mechanism to produce descriptions of human body, and thus generate general body-part prompts, which is helpful for recognizing ambiguous actions. However, the existing LFM-based methods primarily focus on transferring HOI-specific knowledge, but fail to explore the reasoning capabilities of LFMs, leading to incomplete exploration of their full potential, particularly for open-world interaction recognition.

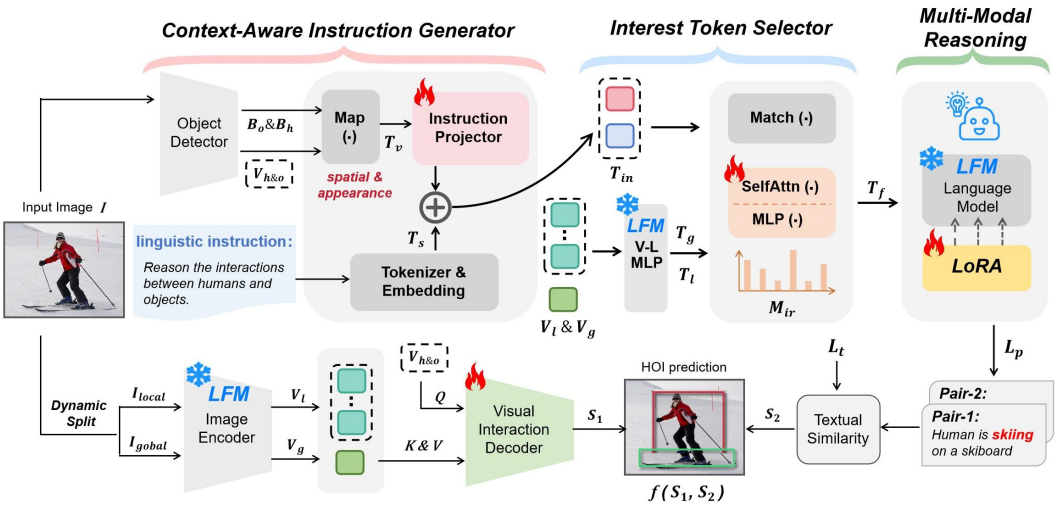

Figure 2: InstructHOI involves two interaction prediction branches: the multi-modal reasoning branch and the visual interaction decoder. The former includes Context-aware Instruction Generator (CIG), Interest Token Selector (ITS), and Multi-Modal Reasoning (MMR). CIG combines visual context with linguistic instructions to generate context-aware instructions T_{in} (Sec. 3.3). ITS then adaptively filters image tokens and reorganizes the reasoning token sequences T_f to align the reasoning process with interaction regions (Sec. 3.4). Finally, the LFM's language model in MMR, fine-tuned with LoRA, uses T_f to conduct multi-modal reasoning, generating pair-level interaction texts (Sec. 3.5). Meanwhile, the visual interaction decoder utilizes pair and global image features to perform interaction decoding (Sec. 3.5).

3 Method

3.1 Overall Architecture

The overall architecture of InstructHOI is illustrated in Fig. 2. Given an image I, an off-the-shelf object detector (i.e., DETR [49]) is first employed to localize human and object instances (B_h, B_o), and then obtain the Human-Object (H-O) pair features $V_{h\&o}$ by concatenating the instance features from DETR for each H-O pair. Meanwhile, following the dynamic image encoding strategy in [50], we dynamically split the image I and obtain the global and local images (I_{global}, I_{local}), which are then separately encoded into the global and local image features (V_g, V_l), using the pretrained image encoder of LFM (i.e., InternVL2 [50]).

InstructHOI involves two interaction prediction branches: the multi-modal reasoning branch and the visual interaction decoder. The former branch mainly includes three components: Context-aware Instruction Generator (CIG), Interest Token Selector (ITS), and Multi-Modal Reasoning (MMR). To direct LFM in facilitating multi-modal reasoning and achieving pair-level interaction prediction, CIG first extracts the appearance and spatial context embedding T_v of each H-O pair, which is then inserted into linguistic instructions, forming pair-specific context-aware instructions T_{in} (Sec. 3.3). Furthermore, to align reasoning process with interaction regions, ITS filters informative tokens from local image tokens T_l based on the instructions T_{in}, and then reorganizes them into reasoning token sequences T_f (Sec. 3.4). The LFM's language model in MMR, fine-tuned with **LoRA**, utilizes the filtered token sequences T_f to achieve pair-level interaction reasoning and acquire interaction-reasoning probability distribution S_2 (Sec. 3.5). Meanwhile, in the visual interaction decoder, the pair features $V_{h\&o}$ act as *Query*, while the global image features V_g act as *Key* and *Value*, performing interaction decoding and yielding interaction-decoding probability distribution S_1 (Sec. 3.5). Finally, both distributions S_1 and S_2 are combined to yield the final interaction score.

3.2 HOI-domain Fine-tuning

Different from task-specific models, Large Foundation Models (LFMs) are typically pretrained on vast and diverse datasets, acquiring general-domain knowledge across both visual and linguistic modalities. However, such general models often struggle to achieve accurate zero-shot interaction prediction, due to the gap between general knowledge and that specific to HOI domain [25]. To bridge the knowledge gap and enhance the interaction reasoning capability for HOI detection, we

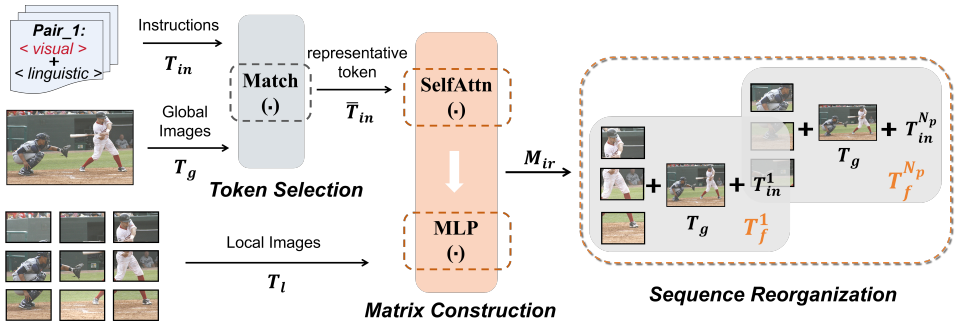

Figure 3: The illustration of Interest Token Selector. It contains three steps: token selection, matrix construction and sequence reorganization.

introduce a multi-modal LFM (i.e., InternVL2) and conduct HOI-domain fine-tuning on a generated high-quality dataset.

Specifically, we employ a light-weight adaptation strategy (i.e., LoRA) to facilitate efficient and low-consumption fine-tuning. As shown in Fig. 2, we freeze the entire pretrained LFM model, including the image encoder, the Vision-Language MLP (V-L MLP), and the language model, while only training a few injectable parameters (about **0.8%** of the pretrained model's parameters) for the language model of InternVL2, acquiring fine-tuned model $\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\boldsymbol{lora}}$. Additionally, due to the limited availability of HOI reasoning data [25], we aggregate five existing image-only HOI datasets [51, 52, 53, 54, 55] and build a high-quality dataset containing **140K** image-text pairs across 1K object categories, 600 action categories, and 15K HOI categories. To acquire interaction reasoning texts, we transform original one-hot labels into *Question&Answer* conversations. E.g., for a human-bike pair with 'sit' and 'ride' interactions, the conversation is formulated as: {*Question: Reason the interaction relationships between humans and objects in the image; Answer: Human is sitting on and riding a bike*}. Moreover, to avoid potential data leakage during evaluation, all testing data used in the experiments are excluded from the fine-tuning dataset. The visualization of the dataset is presented in the Supplementary Material.

3.3 Context-aware Instruction Generator (CIG)

Recent studies [56, 57] explore the potential reasoning capabilities of LFMs for tackling complex visual tasks using different strategies, e.g., chain-of-thought language instruction [27]. However, reasoning about the ambiguous interactions in complex scenarios needs accurate guidance to assist LFM in understanding the spatial relationship and instance appearance of each H-O pair, while language-only instructions can hardly provide these visual context clues. To remedy this, we develop a Context-aware Instruction Generator (CIG), which incorporates the informative visual context of each H-O pair in linguistic instructions, providing multi-modal guidance for pair-level interaction reasoning.

As shown in Fig. 2, we first derive the pair features $V_{h\&o}$ by combining the instance features for each H-O pair, which are extracted from the hierarchical backbone of the object detector and contain discriminative characteristics of each instance. Then, the multi-level $V_{h\&o}$ are flattened and concatenated, serving as the visual appearance representations of each H-O pair. To further encourage LFM to be aware of spatial relationships, we encode H-O spatial context based on the localization boxes $B_h\&B_o$. Following the previous works [58, 21], we extract spatial features from $B_h\&B_o$, by calculating the intersection-over-union, scaled distance, spatial direction, etc. Subsequently, both appearance and spatial features are mapped to a dimensionality of d, forming the visual context embedding T_v. The process can be formulated as:

$$T_v = \mathrm{Map}\left(\mathrm{SPEnc}\left(B_h, B_o\right), V_{h\&o}\right), \tag{1}$$

where $\mathrm{SPEnc}(\cdot)$ represents spatial feature encoding, and $\mathrm{Map}(\cdot)$ indicates feature mapping operation. Meanwhile, the linguistic instructions are also encoded into linguistic context embedding T_s using the tokenizer of InternVL2.

Inspired by the "ViT $\rightarrow$ V-L MLP $\rightarrow$ LLM" architecture in the existing studies (e.g., LLaVA [59]), we leverage a two-layer instruction projector to map the visual context embedding T_v into the linguistic embedding space, eliminating the knowledge gap between visual and linguistic modalities. Finally, the pair-level visual context embedding T_v is fused with the linguistic context embedding T_s, forming

pair-level context-aware instructions T_{in}. Overall, the process can be formulated as:

$$T_{in} = \mathrm{Concat}\left(\mathrm{Proj}_I\left(T_v\right), T_s\right), \tag{2}$$

where $\mathrm{Proj}_I(\cdot)$ indicates the instruction projector, which consists of a two-layer MLP followed by a GeLU layer and $\mathrm{Concat}(\cdot)$ denotes the concatenation operation.

3.4 Interest Token Selector (ITS)

According to the dynamic image encoding strategy [50], input image I is typically dynamically split, acquiring global and local images (I_{global}, I_{local}), where the global image I_{global} provides a holistic image context, and the local images I_{local} offer region-level context information. Considering that, within an image, the interactive regions of H-O pairs also dynamically vary based on their locations and interaction categories. E.g., for ⟨*human*, *hold*, *apple*⟩, the interactive region mainly focuses on the hand regions, while for ⟨*human*, *kick*, *football*⟩, the interactive region shifts to the leg regions. To align pair-level reasoning with the corresponding interaction regions, we develop an Interest Token Selector (ITS), which evaluates the interaction relevance of local images I_{local} for each H-O pair, based on the context-aware instructions T_{in}. Thus, ITS can adaptively select the informative image tokens and then reorganizes the reasoning token sequence for each H-O pair.

Following the multi-modal reasoning mechanism in InternVL2, the image features (V_l, V_g) are projected into linguistic space using the V-L MLP of InternVL2, acquiring image tokens (T_l, T_g). Additionally, as depicted in Fig. 3, the Interest Token Selector contains three steps: token selection, matrix construction and sequence reorganization. Firstly, to extract the representative tokens from the instructions T_{in}, we calculate the cosine similarity between T_{in} and global image tokens T_g, and select the *top-n* most similar ones as representative instruction tokens $\overline{T}_{in}$, as follows:

$$\overline{T}_{in} = \mathrm{Match}\left(T_{in}, T_g\right) \in \mathbb{R}^{N_p \times n \times d}, \tag{3}$$

where $\mathrm{Match}(\cdot)$ represents the cosine similarity operation and selection, N_p and d indicate the number of H-O pairs and token dimension, respectively. Afterwards, a self-attention layer is employed to facilitate feature fusion and context propagation between $\overline{T}_{in}$ and the local image tokens T_l. An MLP is then applied to predict the interaction relevance of local images for each H-O pair, constructing interaction-relevance matrix M_{ir}, as follows:

$$M_{ir} = \mathrm{MLP}(\mathrm{SelfAttn}(Q, K, V : \mathrm{concat}(\overline{T}_{in}, T_l))) \in \mathbb{R}^{N_p \times N_l}, \tag{4}$$

where $\mathrm{SelfAttn}(\cdot)$ represents a self-attention layer and N_l indicates the number of local images.

Finally, we use softmax operation to calculate the relevance probability distribution based on the matrix M_{ir}, subsequently selecting the informative tokens from T_l and reorganizing the reasoning token sequence T_f for each H-O pair in the format of [⟨*selected local images tokens*⟩, ⟨*global image tokens*⟩, ⟨*instruction tokens*⟩], as follows:

$$T_f = \mathrm{Concat}(\widetilde{T}_l, T_g, T_{in}), \quad \text{where } \widetilde{T}_l = \mathrm{Filter}\left(T_l, M_{ir}\right), \tag{5}$$

$\widetilde{T}_l$ represents the selected local images tokens and $\mathrm{Filter}(\cdot)$ indicates the image token filtering operation based on the interaction-relevance matrix. The visualization of ITS are provided in the Supplementary Material.

3.5 Inference and Training

Inference. The inference process is illustrated in Fig. 2. InstructHOI involves two interaction prediction branches: the visual interaction decoder and the multi-modal reasoning branch. In the visual interaction decoder, the interaction decoding is performed based on the visual representations, taking pair features $V_{h\&o}$ as ***Query*** and global image features V_g as ***Key*** and ***Value***, and yields interaction-decoding probability distribution S_1:

$$S_1 = \mathrm{Proj}_d(\mathrm{CrossAttn}(Q : V_{h\&o}; K, V : V_g)) \in \mathbb{R}^{N_p \times N_c}, \tag{6}$$

where $\mathrm{CrossAttn}(\cdot)$ indicates the cross-attention operation, and $\mathrm{Proj}_d(\cdot)$ represents the distribution projector, which consists of a MLP followed by a sigmoid operation, and N_c represents the number of interaction categories.

For the multi-modal reasoning branch, the fine-tuned InternVL2, $\boldsymbol{\pi_{\theta}^{lora}}$, utilizes the refined reasoning token sequence T_f to conduct multi-modal reasoning and generate pair-level interaction text descriptions. Next, we calculate textual cosine similarity between L_p and textual HOI labels L_t, and then compute interaction-reasoning probability distribution S_2, following [43]:

$$S_2 = \mathrm{Softmax}\left(\mathrm{F}_{\cos}\left(L_p, L_t\right)\right), \quad \text{where } L_p = \boldsymbol{\pi_{\theta}^{lora}}\left(T_f\right) \tag{7}$$

where $\mathrm{F}_{\cos}(\cdot)$ indicates the cosine similarity operation, and $S_2 \in \mathbb{R}^{N_p \times N_c}$. Finally, the total interaction prediction score is obtained by combining distributions S_1 and S_2:

$$S_{hoi} = (S_h)^{\lambda} \cdot (S_o)^{\lambda} \cdot S_1 \cdot S_2, \tag{8}$$

where S_h and S_o indicate the detection scores of human and object instances from the object detector, respectively.

Training. To supervise the visual interaction decoder, we employ the following Focal Loss:

$$\mathcal{L}_v = \frac{1}{\sum_{i=1}^{N_p}\sum_{c=1}^{N_c}\mathbf{y}^{i,c}} \sum_{i=1}^{N_p}\sum_{c=1}^{N_c} \mathrm{FocalLoss}(\mathbf{y}^{i,c}, S_1^{i,c}), \tag{9}$$

where $\mathbf{y}^{i,c} \in \{0,1\}$ in $\mathbf{y}$ indicates whether the groundtruth of the i-th human-object pair contains the c-th interaction class and $S_1^{i,c}$ in S_1 is the corresponding predicted probability. In addition, to supervise the multi-modal reasoning branch, we adopt similarity constraint loss, as follows:

$$\mathcal{L}_{sim} = -\frac{1}{N_p}\sum_{i=1}^{N_p} \log \frac{\sum_{c=1}^{N_c} \mathbf{y}^{i,c} \cdot \mathbf{Z}(i,c)}{\sum_{j=1}^{N_c} \mathbf{Z}(i,j)}, \quad \text{where } \mathbf{Z}(i,j) = \exp(\mathrm{F}_{\cos}(L_p^i, L_t^j)). \tag{10}$$

Overall, the total loss function is formulated as: $\mathcal{L} = \mathcal{L}_v + \alpha\mathcal{L}_{sim}$.

4 Experiments

4.1 Experimental Setting

Datasets. Following previous works, we conduct experiments on two commonly used HOI datasets: HICO-DET [51] and V-COCO [52]. The HICO-DET dataset comprises 47776 images, with 38118 for training and 9658 for testing, covering 117 actions, 80 objects, and 600 HOIs. Additionally, the 600 HOIs are divided into 138 Rare and 462 Non-Rare categories based on the sample distribution. The V-COCO dataset, contains 10346 images, including 5400 in the trainval set, and 4946 in the test set, across 29 actions, 80 objects, and 259 HOIs.

Evaluation Metric. Following the standard metric, the mean Average Precision (mAP) is adopted to evaluate the performance of InstructHOI. During the evaluation, a true positive HOI triplet needs to meet two criteria: 1) the predicted human and object bounding boxes should have Intersection over Union (IoU) values greater than 0.5 with the ground truth, and 2) the HOI classification is correct.

Implementation Details. We take the DETR for object detection and adopt the pretrained InternVL2$_{1b}$ as the foundation model. During training, we freeze the external models (DETR and InternVL2) and update the parameters of the remaining components in InstructHOI. The entire InstructHOI model is trained on four Tesla A800 GPUs with a batch size of 16 for 20 epochs, using the AdamW [60] optimizer.

4.2 Comparisons with the State-of-the-Arts

4.2.1 Supervised Setting

In Table 1, we present the quantitative results for the supervised setting on the HICO-DET and V-COCO datasets, respectively. Notably, our method outperforms all the existing state-of-the-art methods on both datasets. For the HICO-DET dataset, InstructHOI achieves remarkable mAPs of **47.68** and **49.89** in the default and known object full settings. Compared to recent state-of-the-art HOI detectors RLIPv2 [61] and Pose-Aware [62], our model obtains significant performance gains of 2.59 mAP (relatively 5.74%) over RLIPv2 and 1.67 mAP (relatively 3.63%) over Pose-Aware,

Table 1: Performance comparison on HICO-DET and V-COCO datasets. For results on HICO-DET, we follow commonly used experimental setting to fine-tune the object detector on its training set.

Method	Backbone	HICO-DET						V-COCO	
		Default			Known Object				
		Full	Rare	Non-Rare	Full	Rare	Non-Rare	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
One-stage methods									
FGAHOI [63]	Swin-L	37.18	30.71	39.11	38.93	31.93	41.02	-	-
RLIPv2 [61]	Swin-L	45.09	43.23	45.64	-	-	-	72.1	74.1
Two-stage methods									
PViC [64]	Swin-L	44.32	44.61	44.24	47.81	48.38	47.64	64.1	70.2
Pose-Aware [62]	Swin-L	46.01	46.74	45.80	49.50	50.59	49.18	63.0	68.7
LFM-based methods									
EZ-HOI [48]	R50+ViT-L	38.61	37.70	38.89	-	-	-	60.5	66.2
UniHOI-l [22]	R101+ViT-L	40.95	40.27	41.32	43.26	43.12	43.25	68.1	70.8
DiffusionHOI [65]	ViT-L	42.54	42.95	42.35	44.91	45.18	44.83	67.1	71.1
MP-HOI [43]	Swin-L+ViT	44.53	44.48	44.55	-	-	-	66.2	67.6
SICHOI [23]	R101+ViT-L	45.04	45.61	44.88	48.16	48.37	48.09	71.1	75.6
InstructHOI (Ours)	R50+ViT-L	45.95	46.51	45.78	48.57	49.23	48.37	70.8	74.2
InstructHOI (Ours)	R101+ViT-L	**47.68**	**47.97**	**47.59**	**49.89**	**50.92**	**49.58**	**72.4**	**76.1**

respectively. By introducing HOI-domain LFM, InstructHOI can extract rich contextual clues and discriminative interaction representations, to tackle ambiguous interaction detection in complex scenarios. Additionally, compared to the state-of-the-art LFM-based methods SICHOI [23] and MP-HOI [43], InstructHOI also achieves significant performance improvements, outperforming SICHOI by 2.64 mAP (relatively 5.86%) and MP-HOI by 3.15 mAP (relatively 7.07%), respectively, in the commonly used default full setting. Specifically, VLM-style methods (e.g., ADA-CM and MP-HOI) adopt single- or multi-modal prompts to transfer HOI-specific knowledge from LFM, while LLM-style approaches (e.g., UniHOI and SICHOI) generate comprehensive descriptions as interactive prompt based on language foundation models. However, all these LFM-based methods primarily focus on transferring HOI-specific knowledge, without exploring the potential reasoning capabilities of LFMs. Unlike existing knowledge transfer methods, our InstructHOI directly leverages context-aware instructions to guide LFM in facilitating pair-level multi-modal reasoning, acquiring discriminative interaction representations for ambiguous and open-world interaction recognition.

For V-COCO dataset, as reported in the right part of Table 1, InstructHOI also performs the best among all the state-of-the-art methods, achieving $AP^{\#1}_{role}$ of **72.4** in scenario #1 and $AP^{\#2}_{role}$ of **76.1** in scenario #2. Specifically, comparing to the recent HOI detectors RLIPv2 and SICHOI, InstructHOI demonstrates superior performance, e.g., $AP^{\#1}_{role}$ of 72.4 vs 72.1 and 71.1, and $AP^{\#2}_{role}$ of 76.1 vs 74.1 and 75.6. Even using "Resnet50" backbone, our method still performs better than most of the state-of-the-art approaches. The superiority of our proposed InstructHOI comes from the fact that we fully exploit the reasoning ability of large foundation models rather than simply transferring knowledge.

4.2.2 Zero-shot Setting

Consistent with previous zero-shot experiments [17, 66, 22], we evaluate our method on HICO-DET under four zero-shot settings: 1) Rare First Unseen Combination (RF-UC) constructs training set with all the object and verb categories but excludes a certain number of rare HOI categories. 2) Non-rare First Unseen Combination (NF-UC) prioritizes non-rare interactions when selecting the held-out HOI categories. 3) Unseen Object (UO) is designed to assess interaction recognition with novel object categories. 4) Unseen Verb (UV) focuses on discovering novel action categories. For a fair comparison, we present recent LFM-based zero-shot HOI detectors with the same "ResNet50" backbone in Table 2, where our InstructHOI surpasses all other methods across four zero-shot settings.

For RF-UC and NF-UC settings, InstructHOI achieves **36.82** mAP and **36.42** mAP for unseen HOI categories, respectively. Compared to the latest method SICHOI, our approach achieves gains of 3.14 mAP and 2.58 mAP in the RF-UC full and unseen settings, respectively, as well as gains of 2.59 mAP and 1.90 mAP in the NF-UC full and unseen settings, respectively. The reason is that our proposed InstructHOI has interactive reasoning capabilities, rather than simply borrowing general knowledge from the LFMs. As for UO and UV settings, InstructHOI attains mAPs of **39.92** and **31.64** for unseen HOI categories, respectively. Compared to the latest method CMMP, our approach achieves improvements of 0.79 mAP and 0.25 mAP in the UO full and unseen settings, respectively,

Table 2: Zero-shot generalization on HICO-DET [51].

Methods	Type	Full	Seen	Unseen
DiffusionHOI [65]	RF-UC	35.89	36.77	32.06
EZ-HOI [48]	RF-UC	36.73	37.35	34.24
CMMP [21]	RF-UC	37.13	37.42	35.98
SICHOI [23]	RF-UC	40.11	41.58	34.24
InstructHOI	RF-UC	**43.25**	**44.86**	**36.82**
BCOM [67]	NF-UC	32.03	31.76	33.12
HOIGen [68]	NF-UC	33.08	32.86	33.98
CMMP [21]	NF-UC	35.13	35.53	33.52
SICHOI [23]	NF-UC	35.75	36.06	34.52
InstructHOI	NF-UC	**38.34**	**38.82**	**36.42**
UniHOI [22]	UO	31.56	34.76	19.72
HOIGen [68]	UO	33.48	32.90	36.35
EZ-HOI [48]	UO	36.38	36.02	38.17
CMMP [21]	UO	36.74	36.15	39.67
InstructHOI	UO	**37.53**	**37.05**	**39.92**
HOIGen [68]	UV	32.34	34.31	20.27
UniHOI [22]	UV	34.68	36.78	26.05
CMMP [21]	UV	36.38	37.28	30.84
EZ-HOI [48]	UV	36.84	38.15	28.82
InstructHOI	UV	**38.12**	**39.17**	**31.64**

Table 3: Performance contribution of each component.

Method	HICO-DET (Default)			V-COCO	
	Full	Rare	Non-Rare	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
Base	36.21	32.84	37.22	64.8	70.1
+MMR	43.06	43.40	42.96	68.9	72.4
+MMR+CIG	44.98	44.69	45.07	70.2	73.9
+MMR+CIG+ITS	**45.95**	**46.51**	**45.78**	**70.8**	**74.2**

Table 4: Effect of Context-aware Instruction Generator.

Method	HICO-DET (Default)			V-COCO	
	Full	Rare	Non-Rare	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
Base+MMR	43.06	43.40	42.96	68.9	72.4
+SC	43.84	43.87	43.83	69.3	72.9
+AC	44.43	44.10	44.53	69.8	73.6
+AC+SC	**44.98**	**44.69**	**45.07**	**70.2**	**73.9**

Table 5: Effect of Interest Token Selector.

Image Token	HICO-DET (Default)			V-COCO	
	Full	Rare	Non-Rare	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
$\boldsymbol{T_g}$	43.62	42.29	44.02	69.3	72.7
$\boldsymbol{T_g}$ + $\boldsymbol{T_l}$	44.98	44.69	45.07	70.2	73.9
$\boldsymbol{T_g}$ + $\widetilde{\boldsymbol{T_l}}$	**45.95**	**46.51**	**45.78**	**70.8**	**74.2**

and 1.74 mAP and 0.80 mAP in the UV full and unseen settings, respectively. All the four zero-shot experimental results consistently demonstrate the effectiveness of our InstructHOI in detecting unseen and novel HOIs. By leveraging the interaction reasoning capabilities of HOI-domain LFMs, InstructHOI exhibits superior open-world interaction detection performance and generalization.

4.3 Ablation Study

In this subsection, we evaluate the effects of Multi-Modal Reasoning (MMR), Context-aware Instruction Generator (CIG), and Interest Token Selector (ITS) components in InstructHOI on both the HICO-DET and V-COCO datasets. For a fair comparison, we create a baseline mode (denoted as "Base") by simply combining the DETR (using Resnet50 as backbone) and visual interaction decoder branch (using ViT-L as backbone), which represents a degraded version of InstructHOI without MMR, CIG, and ITS. Here, the standalone 'MMR' refers to multi-modal reasoning with language-only instructions, while 'MMR+CIG' denotes multi-modal reasoning with context-aware instructions (see subsection 3.3). Additional ablation studies on advanced object detector, HOI-domain fine-tuning, and the number of representative tokens are provided in the Supplementary Material.

Component Ablation. As shown in Table 3, each component of InstructHOI significantly enhances the baseline model. Specifically, MMR improves the baseline by 6.85 mAP in full setting on HICO-DET, while CIG provides an additional improvement of 1.92 mAP. Ultimately, the combination of all the three components results in a total improvement of 9.74 mAP. The above results demonstrate that the reasoning capabilities of LFM can significantly improve HOI detection, as well as highlight the effectiveness of CIG and ITS in further improving LFM's reasoning abilities.

Context-aware Instruction Generator. Within the Context-aware Instruction Generator (CIG), we integrate visual context into linguistic instructions to enhance the spatial and appearance understanding of LFM. In Table 4, we evaluate the Spatial Context (SC) and the Appearance Context (AC) in CIG separately, based on the "Base + MMR" model (i.e., reasoning with language-only instructions). The results demonstrate that both AC and SC can enhance the language-only instructions, providing context guidance for interaction reasoning.

Interest Token Selector. As shown in Table 5, we evaluate the effectiveness of the Interest Token Selector (ITS) by using different combinations of image tokens: $\boldsymbol{T_g}$ (global image tokens), $\boldsymbol{T_l}$ (local image tokens), and $\widetilde{\boldsymbol{T_l}}$ (selected local image tokens). Specifically, the combination of ($\boldsymbol{T_g}$ + $\widetilde{\boldsymbol{T_l}}$) outperforms ($\boldsymbol{T_g}$ + $\boldsymbol{T_l}$) by 0.97 mAP in the full setting of HICO-DET. This indicates that the ITS effectively filters informative image tokens from $\boldsymbol{T_l}$, aligning the reasoning with interaction areas.

5 Conclusion

In this paper, we propose a novel LFM-based HOI detector, InstructHOI. Different from the existing LFM-based approaches, InstructHOI directly learns tailored instructions to guide LFM in facilitating multi-modal reasoning, and thus can improve the open-world interaction recognition. Specifically, we develop a Context-aware Instruction Generator (CIG) to enhance linguistic instructions by incorporating visual interactive context, forming pair-level reasoning guidance. Furthermore, an Interest Token Selector (ITS) is adopted to align reasoning process with interaction regions. Extensive experiments on two public benchmarks demonstrate that our proposed method outperforms the state-of-the-art ones, under both supervised and zero-shot settings. Ablation studies also prove the effectiveness of each component in our proposed InstructHOI.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants 62206075, 62573163, 62503139, and 62261160652, in part by the GuangDong Basic and Applied Basic Research Foundation under Grant 2024A1515012028, in part by the Shenzhen Science and Technology Program under Grant GXWD20231129125006001, in part by the Science and Technology Development Fund (FDCT), Macau SAR, under Grant 0095/2022/AFJ.

References

[1] B. Pang, K. Zha, Y. Zhang, and C. Lu, "Further understanding videos through adverbs: A new video task," in *AAAI Conference on Artificial Intelligence*, pp. 11823–11830, 2020.

[2] Y. Liu, W. Chen, Y. Bai, J. Luo, X. Song, K. Jiang, Z. Li, G. Zhao, J. Lin, G. Li, *et al.*, "Aligning cyber space with physical world: A comprehensive survey on embodied ai," *arXiv preprint arXiv:2407.06886*, 2024.

[3] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10164–10183, 2024.

[4] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng, "Ppdm: Parallel point detection and matching for real-time human-object interaction detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 482–490, 2020.

[5] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, "Learning human-object interaction detection using interaction points," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4116–4125, 2020.

[6] M. Tamura, H. Ohashi, and T. Yoshinaga, "Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10410–10419, 2021.

[7] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11825–11834, 2021.

[8] T. Gupta, A. Schwing, and D. Hoiem, "No-frills human-object interaction detection: Factorization, layout encodings, and training techniques," in *IEEE International Conference on Computer Vision*, pp. 9677–9685, 2019.

[9] C. Gao, J. Xu, Y. Zou, and J.-B. Huang, "Drg: Dual relation graph for human-object interaction detection," in *European Conference on Computer Vision*, pp. 696–712, Springer, 2020.

[10] J. Park, J.-W. Park, and J.-S. Lee, "Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 17152–17162, 2023.

[11] B. Xu, J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Interact as you intend: Intention-driven human-object interaction detection," *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1423–1432, 2019.

[12] L. Zhang, W. Suo, P. Wang, and Y. Zhang, "A plug-and-play method for rare human-object interactions detection by bridging domain gap," in *ACM International Conference on Multimedia*, pp. 8613–8622, 2024.

[13] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 19548–19557, 2022.

[14] J. Lim, V. M. Baskaran, J. M.-Y. Lim, K. Wong, J. See, and M. Tistarelli, "Ernet: An efficient and reliable human-object interaction detection network," *IEEE Transactions on Image Processing*, vol. 32, pp. 964–979, 2023.

[15] 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*, vol. 35, pp. 37416–37431, 2022.

[16] Z. Li, X. Li, C. Ding, and X. Xu, "Disentangled pre-training for human-object interaction detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 28191–28201, 2024.

[17] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 20123–20132, 2022.

[18] A. Iftekhar, H. Chen, K. Kundu, X. Li, J. Tighe, and D. Modolo, "What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5353–5363, 2022.

[19] X. Qu, C. Ding, X. Li, X. Zhong, and D. Tao, "Distillation using oracle queries for transformer-based human-object interaction detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 19558–19567, 2022.

[20] 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*, pp. 8748–8763, PMLR, 2021.

[21] T. Lei, S. Yin, Y. Peng, and Y. Liu, "Exploring conditional multi-modal prompts for zero-shot hoi detection," in *European Conference on Computer Vision*, pp. 1–19, Springer, 2025.

[22] Y. Cao, Q. Tang, X. Su, S. Chen, S. You, X. Lu, and C. Xu, "Detecting any human-object interaction relationship: Universal hoi detector with spatial prompt learning on foundation models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 739–751, 2023.

[23] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 28212–28222, 2024.

[24] T. Lei, S. Yin, and Y. Liu, "Exploring the potential of large foundation models for open-vocabulary hoi detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16657–16667, 2024.

[25] J. Gao, C. Cai, R. Wang, W. Liu, K.-H. Yap, K. Garg, and B.-S. Han, "Cl-hoi: Cross-level human-object interaction distillation from vision large language models," *arXiv preprint arXiv:2410.15657*, 2024.

[26] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[27] Y. Hu, O. Stretcu, C.-T. Lu, K. Viswanathan, K. Hata, E. Luo, R. Krishna, and A. Fuxman, "Visual program distillation: Distilling tools and programmatic reasoning into vision-language models," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9590–9601, 2024.

[28] S. Lee, W. J. Kim, J. Chang, and J. C. Ye, "LLM-CXR: Instruction-finetuned LLM for CXR image understanding and generation," in *International Conference on Learning Representations*, 2024.

[29] X. Wu, Y.-L. Li, X. Liu, J. Zhang, Y. Wu, and C. Lu, "Mining cross-person cues for body-part interactiveness learning in hoi detection," in *European Conference on Computer Vision*, pp. 121–136, Springer, 2022.

[30] B. Kim, T. Choi, J. Kang, and H. J. Kim, "Uniondet: Union-level detector towards real-time human-object interaction detection," in *European Conference on Computer Vision*, pp. 498–514, Springer, 2020.

[31] D. Zhou, Z. Liu, J. Wang, L. Wang, T. Hu, E. Ding, and J. Wang, "Human-object interaction detection via disentangled transformer," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 19568–19577, 2022.

[32] A. Bansal, S. S. Rambhatla, A. Shrivastava, and R. Chellappa, "Detecting human-object interactions via functional generalization," in *AAAI Conference on Artificial Intelligence*, vol. 34, pp. 10460–10469, 2020.

[33] D.-J. Kim, X. Sun, J. Choi, S. Lin, and I. S. Kweon, "Detecting human-object interactions with action co-occurrence priors," in *European Conference on Computer Vision*, pp. 718–736, Springer, 2020.

[34] X. Zhong, C. Ding, X. Qu, and D. Tao, "Polysemy deciphering network for robust human–object interaction detection," *International Journal of Computer Vision*, vol. 129, pp. 1910–1929, 2021.

[35] O. Ulutan, A. S. M. Iftekhar, and B. S. Manjunath, "Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2020.

[36] H. Wang, W.-s. Zheng, and L. Yingbiao, "Contextual heterogeneous graph network for human-object interaction detection," in *European Conference on Computer Vision*, pp. 248–264, Springer, 2020.

[37] Y. Liu, Q. Chen, and A. Zisserman, "Amplifying key cues for human-object-interaction detection," in *European Conference on Computer Vision*, pp. 248–265, Springer, 2020.

[38] B. Kim, J. Mun, K.-W. On, M. Shin, J. Lee, and E.-S. Kim, "Mstr: Multi-scale transformer for end-to-end human-object interaction detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 19578–19587, 2022.

[39] X. Liu, Y.-L. Li, X. Wu, Y.-W. Tai, C. Lu, and C.-K. Tang, "Interactiveness field in human-object interactions," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 20113–20122, 2022.

[40] C. Xie, F. Zeng, Y. Hu, S. Liang, and Y. Wei, "Category query learning for human-object interaction classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 15275–15284, 2023.

[41] F. Z. Zhang, D. Campbell, and S. Gould, "Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 20104–20112, 2022.

[42] L. Li, J. Wei, W. Wang, and Y. Yang, "Neural-logic human-object interaction detection," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[43] J. Yang, B. Li, A. Zeng, L. Zhang, and R. Zhang, "Open-world human-object interaction detection via multi-modal prompts," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16954–16964, 2024.

[44] S. Ning, L. Qiu, Y. Liu, and X. He, "Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 23507–23517, 2023.

[45] S. Zheng, B. Xu, and Q. Jin, "Open-category human-object interaction pre-training via language modeling framework," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 19392–19402, 2023.

[46] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*, pp. 12888–12900, PMLR, 2022.

[47] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International Conference on Machine Learning*, pp. 19730–19742, PMLR, 2023.

[48] Q. Lei, B. Wang, and T. Robby T., "Ez-hoi: Vlm adaptation via guided prompt learning for zero-shot hoi detection," *Advances in Neural Information Processing Systems*, 2024.

[49] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, pp. 213–229, Springer, 2020.

[50] 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," *arXiv preprint arXiv:2404.16821*, 2024.

[51] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *IEEE Winter Conference on Applications of Computer Vision*, pp. 381–389, 2018.

[52] S. Gupta and J. Malik, "Visual semantic role labeling," *arXiv preprint arXiv:1505.04474*, 2015.

[53] S. Wang, K.-H. Yap, H. Ding, J. Wu, J. Yuan, and Y.-P. Tan, "Discovering human interactions with large-vocabulary objects via query and multi-scale detection," in *IEEE International Conference on Computer Vision*, pp. 13475–13484, 2021.

[54] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *International journal of computer vision*, vol. 128, no. 7, pp. 1956–1981, 2020.

[55] Y.-L. Li, L. Xu, X. Huang, X. Liu, Z. Ma, M. Chen, S. Wang, H.-S. Fang, and C. L. Hake, "Human activity knowledge engine," *arXiv preprint arXiv:1904.06539*, vol. 2, no. 6, 2019.

[56] X. Wu, Y.-L. Li, J. Sun, and C. Lu, "Symbol-llm: leverage language models for symbolic system in visual human activity reasoning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[57] Z. Liu, H. Hu, S. Zhang, H. Guo, S. Ke, B. Liu, and Z. Wang, "Reason for future, act for now: A principled architecture for autonomous llm agents," in *International Conference on Machine Learning*, 2023.

[58] W. Jiang, W. Ren, J. Tian, L. Qu, Z. Wang, and H. Liu, "Exploring self-and cross-triplet correlations for human-object interaction detection," in *AAAI Conference on Artificial Intelligence*, vol. 38, pp. 2543–2551, 2024.

[59] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34892–34916, 2023.

[60] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[61] 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 *IEEE International Conference on Computer Vision*, pp. 21649–21661, 2023.

[62] E. Z. Wu, Y. Li, Y. Wang, and S. Wang, "Exploring pose-aware human-object interaction via hybrid learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 17815–17825, 2024.

[63] S. Ma, Y. Wang, S. Wang, and Y. Wei, "Fgahoi: Fine-grained anchors for human-object interaction detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2415–2429, 2024.

[64] F. Z. Zhang, Y. Yuan, D. Campbell, Z. Zhong, and S. Gould, "Exploring predicate visual context in detecting of human-object interactions," in *IEEE International Conference on Computer Vision*, pp. 10411–10421, 2023.

[65] L. Li, W. Wang, and Y. Yang, "Human-object interaction detection collaborated with large relation-driven diffusion models," *Advances in Neural Information Processing Systems*, 2024.

[66] T. Lei, F. Caba, Q. Chen, H. Jin, Y. Peng, and Y. Liu, "Efficient adaptive human-object interaction detection with concept-guided memory," in *IEEE International Conference on Computer Vision*, pp. 6480–6490, 2023.

[67] G. Wang, Y. Guo, Z. Xu, and M. Kankanhalli, "Bilateral adaptation for human-object interaction detection with occlusion-robustness," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 27970–27980, 2024.

[68] Y. Guo, Y. Liu, J. Li, W. Wang, and Q. Jia, "Unseen no more: Unlocking the potential of clip for generative zero-shot hoi detection," in *ACM International Conference on Multimedia*, pp. 1711–1720, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We emphasize the contributions and scope in the Introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitation of the proposed algorithm has been discussed in the supplementary material.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: There is no theoretical result.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide comprehensive implementation details both in main paper and in supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: As we promised, the data and code will be released upon the publication of our paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: : The experimental setup, including data splits, training and testing detailed, are provided in Method and Experiments sections.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We follow the default evaluations in the HOI detection field, which doesn't require error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide them in implementation details of main paper and supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work conforms the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The border impacts is provided in supplementary material.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The proposed method uses pre-trained models. This proposed methods is safe under the safeguards of adopted pre-trained models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cited the original paper that produced the code package or dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: There is no new assets released in this work.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: There is no research with human subjects in this work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLM does not impact the core methodology, scientific rigorousness, or originality of the research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.