Pose-Guided Multi-Cue Explicit Query Construction for Disambiguating Human-Object Interactions

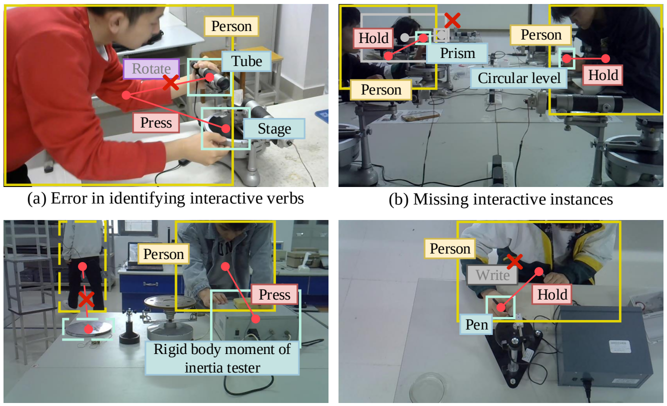
Minghao Zou, Shangkun Liu, Qingtian Zeng, Xue Zhang, *Member, IEEE*, Guiyuan Yuan, Xiaoshuai Hao, Jun Liu, *Senior Member, IEEE*, Wei Zhou, *Senior Member, IEEE*

***Abstract*—Human-Object Interaction (HOI) detection remains challenging due to the semantic ambiguity of interaction categories and the limited discriminability of their feature representations. Existing approaches often improve recognition by employing sophisticated models or auxiliary textual annotations. While effective in certain gains, these solutions incur additional computational or annotation costs and struggle to capture intrinsic interaction regularities. To address these issues, we propose Pose-Guided Multi-Cue Explicit Query Construction (PM-EQC), a unified Transformer-based framework that builds upon collaborative modeling of appearance, spatial, and pose cues for discriminative interaction reasoning. At its core, the Collaborative Multi-Cue Query Constructor (CM-CQC) jointly models dependencies among visual cues to generate explicit query embeddings. CM-CQC further incorporates a hierarchical pose contextualization mechanism: global body configurations adaptively guide attention to local critical joints, yielding fine-grained pose embeddings and more precise interaction disambiguation. Owing to its modular design, PM-EQC integrates seamlessly with diverse backbones and benefits from their advances. Extensive experiments on PhysLab, HICO-DET, and V-COCO datasets demonstrate that PM-EQC achieves state-of-the-art performance, and the code is publicly available at https://github.com/ZMH-SDUST/PM-EQC.**

***Index Terms*—Human-Object Interaction Detection, Query Learning, Multi-Cue Visual Reasoning, PM-EQC**

I. Introduction

Human-Object Interaction (HOI) detection is a human-centric visual parsing task that aims to capture interaction relationships between humans and objects [1]. An HOI instance is commonly formulated as a $< human, verb, object >$ triplet. Unlike conventional object detection, which localizes instances with category labels, HOI detection additionally characterizes their interactions with descriptive verbs. This enables richer scene understanding in applications such as visual reasoning and behavior analysis [2], [3].



(a) Error in identifying interactive verbs

(b) Missing interactive instances

(c) Misidentification of non-interactions

(d) Missing interactive verbs

Fig. 1. **Visualization of detection errors.** Correctly detected interactions are outlined with solid-colored lines, false-positive interactions with dotted-colored lines, missed interactions with solid gray lines, and misclassified interactions are highlighted with purple labels.

Despite rapid advances, practical HOI detection still faces challenges due to semantic ambiguities in interactivity and interaction types [4]. Figure 1 summarizes four common detection errors. Extensive case analysis reveals that these errors mainly arise from relying solely on appearance cues, which are insufficient to capture the underlying regularities of interactions [5]. Incorporating richer visual signals, such as human pose and relative spatial configuration, can offer more discriminative evidence [3], [6]. For instance, ‘Press’ consistently involves close contact and applied force between the human and the object, whereas ‘Write’ exhibits distinctive pen-holding postures and fine-grained finger dynamics. These observations highlight the necessity of integrating multi-cues for reliable HOI disambiguation.

Although recent methods have explored integrating multiple visual cues [2], [3], [5], [7], they still struggle with incomplete or oversimplified feature representations. For instance, methods such as STIP [2], PViC [5], and TED-Net [8] primarily model appearance and spatial relations while overlooking pose details, which increases the likelihood of verb misclassification. Subsequently, as detailed in [6], [9], some methods construct global posture representations from detected keypoints. While effective in some cases, global modeling often neglects fine-grained joint information, hindering the disambiguation

This work was supported in part by National Science and Technology Major Project of China under Grant 2022ZD0119501; in part by National Natural Science Foundation of China under Grants 52574256, 52374221, 62402261, and 72404177; in part by Sci. & Tech. Development Fund of Shandong Province of China under Grants ZR2022MF288 and ZR2023MF097; and in part by Taishan Scholar Program of Shandong Province under Grant TSTP20250506. (Minghao Zou and Shangkun Liu contributed equally to this work.) (Corresponding authors: Qingtian Zeng and Wei Zhou.)

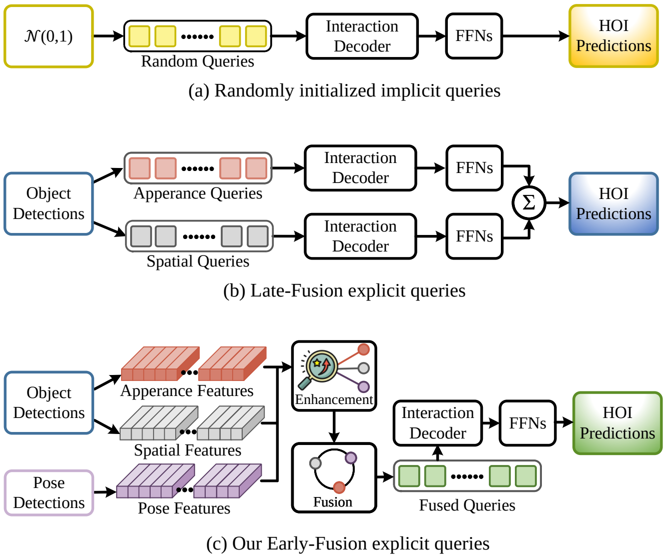
Minghao Zou is with the College of Computer Science and Engineering, Shandong University of Science and Technology, China, and also with the School of Computer Science and Informatics, Cardiff University, UK (email: zoum1@cardiff.ac.uk)

Shangkun Liu, Qingtian Zeng, Xue Zhang, and Guiyuan Yuan are with the College of Computer Science and Engineering, Shandong University of Science and Technology, China (email: liushangkun@sdust.edu.cn, qtzeng@sdust.edu.cn, xuezhang@sdust.edu.cn, gyuan_yuan@163.com)

Xiaoshuai Hao is with the Xiaomi EV, China (email: haoxiaoshuai714@163.com)

Jun Liu is with the School of Computing and Communications, Lancaster University, UK (email: j.liu81@lancaster.ac.uk)

Wei Zhou is with the School of Computer Science and Informatics, Cardiff University, UK (email: zhouw26@cardiff.ac.uk)



(a) Randomly initialized implicit queries

(b) Late-Fusion explicit queries

(c) Our Early-Fusion explicit queries

Fig. 2. **Comparison of query construction paradigms in HOI detection.**

of semantically similar interactions. In contrast, several zero-shot HOI approaches leverage techniques such as Vision-Language Models (VLMs) [10], caption-guided visual learning [11], and multi-knowledge fusion [12] to enhance generalization. However, these approaches shift the complexity to the language modeling component and tend to overlook the rich visual dynamics inherent in HOIs. Consequently, there remains significant room for improvement in developing richer and more discriminative visual representations for interaction reasoning.

These observations motivate us to reformulate HOI detection as a collaborative pose context-guided multi-cue explicit query learning task, centered on the principle of jointly modeling visual cues for discriminative interaction reasoning. Specifically, our framework leverages three complementary sources of visual cues:

(1) Fine-grained body pose features. This component captures both global body posture and fine-grained local joint configurations, offering critical cues for distinguishing subtle differences in interactive behaviors.
(2) Appearance features. This component characterizes the visual attributes of human and object instances, thereby constraining the space of plausible interaction types. It focuses on matching specific interaction verbs rather than exhaustively modeling all possible interactions.
(3) Relative spatial relationship. This component models the spatial configuration between human and object instances, which helps localize the interaction region, estimate interaction confidence, and identify interaction patterns.

To this end, we propose a novel interaction detection framework called Pose-Guided Multi-Cue Explicit Query Construction (PM-EQC). Its modular design enables seamless integration with existing backbones, achieving state-of-the-art performance across multiple datasets and supervised settings without additional text annotations or complex external knowledge. The core of PM-EQC is the Collaborative Multi-Cue Query Constructor (CM-CQC), which enhances, aligns, and fuses multiple visual features to construct discriminative query vectors for interaction instances. As shown in Figure 2, unlike traditional late-fusion approaches [3], [13] that process cues independently, CM-CQC employs early fusion during query construction to jointly model cross-cue dependencies. This strategy improves feature complementarity and expressiveness while reducing redundant post-processing. Furthermore, beyond generating context-aware appearance features and spatially invariant relative position features, CM-CQC integrates global body structure with fine-grained joint dynamics, where global pose features adaptively guide attention to key joints, enabling precise disambiguation of semantically similar interactions.

In summary, our contributions are threefold:

(1) We propose PM-EQC, a flexible and generalizable HOI detection framework that relies solely on visual feature encoding and fusion, eliminating the need for auxiliary textual annotations or complex external knowledge.
(2) We introduce CM-CQC, a pose context-guided explicit query constructor that jointly models the interdependencies among visual cues to enable discriminative and context-aware interaction reasoning.
(3) PM-EQC achieves state-of-the-art performance on PhysLab, HICO-DET, and V-COCO datasets, and demonstrates strong generalization to unseen interaction combinations, yielding improvements of up to 28.3% compared with prior approaches.

II. Related Works

Existing research on HOI detection primarily follows either one-stage or two-stage paradigms, while recent Transformer-based models further incorporate query learning to strengthen interaction reasoning. These studies have advanced architectural design, feature fusion, and semantic guidance.

A. *One-stage Methods*

One-stage methods have emerged as the mainstream paradigm in HOI detection owing to their end-to-end design and high inference efficiency. Early works typically relied on Transformer-based architectures to jointly localize Human-Object (HO) pairs and predict interactions, implicitly modeling pairwise relations through attention mechanisms [1], [7], [14]–[18]. However, most of these methods employ a shared predictor for both instance localization and interaction recognition, thereby limiting the decoupling of instance-level reasoning from relation-level modeling.

Building on these foundations, recent studies have introduced architectural refinements to alleviate this limitation, such as decoupled decoder branches [19]–[21], multi-scale feature fusion [15], and deformable attention mechanisms [22]. In parallel, some approaches incorporate external semantic signals, including pseudo-labels and caption-derived priors, to guide query learning and reduce ambiguity [14], [23], [24]. While effective, these text-driven strategies typically incur additional annotation costs and rely heavily on language supervision, which may limit their scalability and generalization in practical scenarios.

B. Two-stage Methods

Compared with one-stage methods, two-stage approaches follow a decoupled pipeline: instead of jointly localizing instances and predicting interactions, they first detect humans and objects with a pretrained detector and then infer their interactions based on these fixed regions [2], [3], [6], [7]. Such a decoupled formulation narrows the search space and simplifies interaction reasoning, which often leads to improved accuracy and reduced training complexity.

A key challenge in this paradigm lies in the fusion of heterogeneous features. Existing works tackle this by exploring appearance and spatial cues [2], [25], introducing pose-aware attention [3], or utilizing fine-grained body-part representations [6]. However, they mainly focus on selecting useful visual cues but do not sufficiently explore how to represent and integrate them effectively, especially for jointly encoding global posture and local joint details. In addition to visual signals, some methods introduce semantic embeddings of objects and verbs to enrich category-level representations [4], [7]. Nevertheless, this reliance on external textual supervision shifts part of the modeling burden to language inputs and may overlook visually grounded interaction patterns.

C. Query Learning

Queries constitute a central component of Transformer-based HOI detectors, serving as Decoder inputs for interaction parsing [3], [19], [22], [25]. One-stage approaches typically initialize queries randomly [15], [19]. While computationally simple, such initialization lacks informative priors, resulting in limited expressiveness and a greater tendency to converge to suboptimal solutions. To address this limitation, prior works have explored different strategies, such as sequentially linking instance and interaction Decoders [22], [26], incorporating relational text to improve query discriminability [14], [24], [26], and leveraging knowledge distillation or deeper architectures to enrich query embeddings [23], [27], [28]. Despite these efforts, queries remain lacking sufficient interaction expressiveness, thereby restricting detection accuracy [5].

In contrast, two-stage methods generally initialize queries from object detector outputs [2], [3], [5], [25], thereby narrowing the search space and providing stronger priors for interaction reasoning. Representative strategies include spatially structured queries for HO pairs [2], DETR-style Self-Attention [29] over unary and pairwise markers [25], and pose-aware initialization [3]. Although these designs improve query initialization, challenges remain in capturing fine-grained interaction cues and integrating heterogeneous context, highlighting the need for more expressive query construction mechanisms.

III. Method

This section introduces the PM-EQC framework, outlining its overall architecture and the core module called CM-CQC. By elucidating the underlying structure and functionality of PM-EQC, we aim to provide a comprehensive description of its design principles and mechanisms. Subsequent subsections will provide a detailed breakdown of CM-CQC, highlighting its key features and how it contributes to detecting HOIs.

A. Overall Architecture

As illustrated in Figure 3, PM-EQC comprises three primary components: object detection, pose estimation, and interaction reasoning.

Object Detection. The object detector follows a Transformer-based encoder-decoder architecture, leveraging its ability to capture long-range dependencies and global context [5], [15]. Given an input image $\mathbf{X} \in \mathbb{R}^{3\times H\times W}$, the Backbone extracts the primary visual features $\mathbf{V}_{\mathrm{b}} \in \mathbb{R}^{D_{\mathrm{b}}\times H_1\times W_1}$. These features, together with the sinusoidal positional embedding $\mathbf{P} \in \mathbb{R}^{D_{\mathrm{b}}\times H_1\times W_1}$ [30], are fed into the Image Encoder to obtain a context-aware representation $\mathbf{V}_{\mathrm{e}} \in \mathbb{R}^{D_{\mathrm{b}}\times H_1\times W_1}$. The computation of $\mathbf{P}$ is formally defined in Equations (1)–(3):

$$\mathbf{P}_{\delta}(x,i) = \begin{cases} \sin\left(\frac{x}{10000^{(2k/D_{\mathrm{p}})}}\right) & \text{if } i = 2k, \\ \cos\left(\frac{x}{10000^{(2k/D_{\mathrm{p}})}}\right) & \text{if } i = 2k+1, \end{cases} \tag{1}$$

$$\mathbf{P}_{\delta}(y,i) = \begin{cases} \sin\left(\frac{y}{10000^{(2k/D_{\mathrm{p}})}}\right) & \text{if } i = 2k, \\ \cos\left(\frac{y}{10000^{(2k/D_{\mathrm{p}})}}\right) & \text{if } i = 2k+1, \end{cases} \tag{2}$$

$$\mathbf{P}_{\delta} = \mathbf{P}_{\delta}(x) \oplus \mathbf{P}_{\delta}(y), \tag{3}$$

where $\mathbf{P}_{\delta}(x)$ and $\mathbf{P}_{\delta}(y)$ denote the positional embeddings of the feature point $\delta = (x,\, y)$ along the x- and y-axes, respectively. $\oplus$ denotes the concatenation operator and i indexes the embedding vector, ranging from 0 to D_{p}. Here, D_{p} represents the dimension of the vector, and its value is half of D_{b}.

The Instance Decoder takes as input the $\mathbf{V}_{\mathrm{e}}$ and learnable query vectors $\mathbf{Q}_{\mathrm{ins}} \in \mathbb{R}^{N_{\mathrm{ins}}\times D_{\mathrm{b}}}$. Here, N_{ins} denotes the number of queries, which is set to a sufficiently large value to ensure comprehensive coverage of potential instances [22]. Within the Decoder, $\mathbf{Q}_{\mathrm{ins}}$ is matched with the global features and undergoes a series of operations such as weighting, regularization, and non-linear mapping, to generate the instance embedding matrix $\mathbf{V}_{\mathrm{d}} \in \mathbb{R}^{N_{\mathrm{ins}}\times D_{\mathrm{b}}}$. These embeddings are further mapped via separate Feed-Forward Networks (FFNs) f_c and f_b into class predictions $\mathbf{C} \in \mathbb{R}^{N_{\mathrm{ins}}\times N_{\mathrm{c}}}$ and bounding box coordinates $\mathbf{B} \in \mathbb{R}^{N_{\mathrm{ins}}\times 4}$. For clarity, we categorize instances into action initiators (humans, denoted by the subscript h) and action responders (objects, denoted by the subscript o), facilitating an explicit representation for the detection process.

Pose Estimation. The second component focuses on extracting pose features $\mathbf{V}_{\mathrm{p}} \in \mathbb{R}^{N_{\mathrm{p}}\times 18\times 4}$ from the detected humans in the image, where N_{p} denotes the number of detected human bodies. Specifically, each body is represented by 18 keypoints, and each keypoint is further encoded as a four-dimensional vector that includes the keypoint type, its x-and y-coordinates, and a confidence score.

Interaction Reasoning. In the third component, the object detection results and extracted pose features are processed by the CM-CQC module for matching and filtering. This process yields a set of explicit HOI queries $\mathbf{Q}_{\mathrm{hoi}} \in \mathbb{R}^{N_{\mathrm{ho}}\times D_{\mathrm{q}}}$, which encompass various types of discriminative visual cues. Here, N_{ho} denotes the number of queries, and each element of $\mathbf{Q}_{\mathrm{hoi}}$ corresponds to a detected interaction instance. Through the multi-cue fusion conducted in CM-CQC, the constructed queries

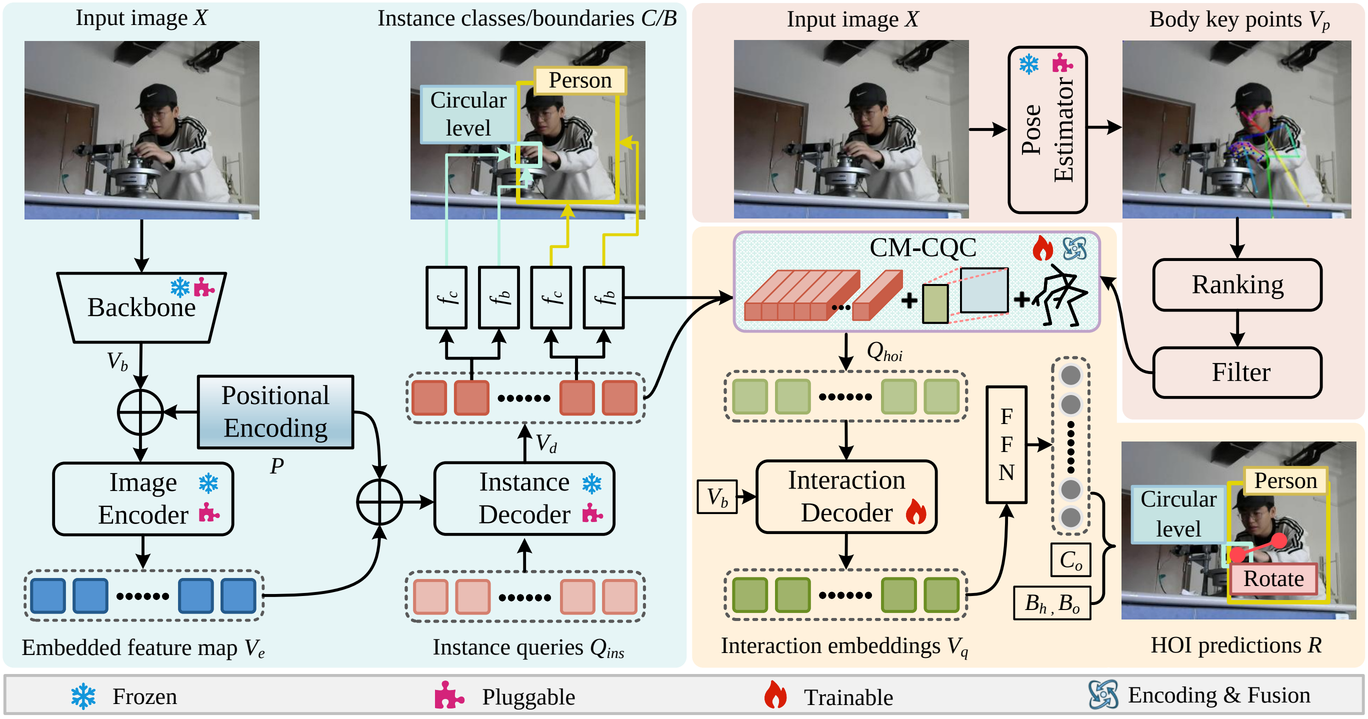


Fig. 3. **Overview of the PM-EQC framework.** The input image is processed through parallel branches for object detection and pose estimation to extract instance-level visual cues. These cues are fused via the CM-CQC module to generate interaction queries, which are subsequently decoded to predict interactions between detected humans and objects.

jointly encode appearance, spatial, and pose representations, enabling each query to capture the interaction context, i.e., the semantic, geometric, and motion dependencies between the human and the object within a pair.

During the reasoning stage, the Interaction Decoder performs further refinement using a Cross-Attention mechanism with the global feature memory $\mathbf{V}_{\text{b}}$. This operation enables each query to incorporate global contextual cues from the entire scene, such as background appearance and environmental semantics. The Cross-Attention operation is defined as (4):

$$\mathbf{V}_{\text{q}} = \text{CrossAttn}(\mathbf{Q}_{\text{hoi}}\mathbf{W}_{\text{CQ}}, \mathbf{V}_{\text{b}}\mathbf{W}_{\text{CK}}, \mathbf{V}_{\text{b}}\mathbf{W}_{\text{CV}}), \qquad (4)$$

where CrossAttn$(\cdot, \cdot, \cdot)$ denotes a Cross-Attention operation, and $\mathbf{W}_{\text{CQ}}$, $\mathbf{W}_{\text{CK}}$, and $\mathbf{W}_{\text{CV}}$ are trainable weight matrices [5]. This two-stage process enables the model to perform context-aware interaction reasoning by exploiting both the local dependencies encoded in $\mathbf{Q}_{\text{hoi}}$ and the global contextual features represented by $\mathbf{V}_{\text{b}}$.

The resulting interaction embeddings are then projected by a classifier f_c into verb logits $\mathbf{C}_{\text{v}} \in \mathbb{R}^{N_{\text{ho}} \times N_{\text{v}}}$, where N_{v} denotes the total number of interaction categories. Finally, detected HO instances are paired with their corresponding verb predictions [2] to obtain the final HOI prediction set $\mathbf{R} = \{\mathbf{r}_i \,|\, \mathbf{r}_i \in \mathbb{R}^{10}\}_{i=1}^{N_{\text{ho}}}$. Each prediction $\mathbf{r}_i$ encodes the bounding box coordinates of the human and object, the object class, and the corresponding verb category.

B. Collaborative Multi-Cue Query Constructor

This section details the Collaborative Multi-Cue Query Constructor (CM-CQC), which is designed to generate explicit queries [5] for interaction detection. As shown in Figure 4, CM-CQC consists of two complementary branches. The Instance Context (IC) branch integrates visual appearance and spatial configuration cues to encode instance-level semantics and geometric constraints. In parallel, the Pose-Contextualized Reasoning (PC-R) branch models human body dynamics, leveraging global pose patterns to guide local joint attention for fine-grained disambiguation of semantically similar interactions. By jointly modeling these heterogeneous cues, CM-CQC produces discriminative, context-aware query embeddings that substantially enhance HOI reasoning and recognition.

IC Branch. This branch integrates two complementary sources of visual cues: visual appearance and relative spatial configuration. The appearance stream captures how humans and objects co-occur in visual scenes, thereby clarifying how their appearances jointly contribute to observed interactions [31], [32]. Specifically, to capture contextual dependencies among all detected humans and objects, we apply a Self-Attention module over the instance features, producing the attention-enhanced appearance representations $\mathbf{V}_{\text{a}}$. The computation is given in Equation (5):

$$\mathbf{V}_{\text{a}} = \text{SelfAttn}(\mathbf{V}_{\text{d}}\mathbf{W}_{\text{SQ}}, \mathbf{V}_{\text{d}}\mathbf{W}_{\text{SK}}, \mathbf{V}_{\text{d}}\mathbf{W}_{\text{SV}}). \qquad (5)$$

where SelfAttn$(\cdot, \cdot, \cdot)$ denotes a Self-Attention operator, and $\mathbf{W}_{\text{SQ}}$, $\mathbf{W}_{\text{SK}}$, and $\mathbf{W}_{\text{SV}}$ are trainable projection matrices [5].

Next, we decompose $\mathbf{V}_{\text{a}}$ into human features $\mathbf{V}_{\text{a,h}} \in \mathbb{R}^{N_{\text{h}} \times D_{\text{b}}}$ and object features $\mathbf{V}_{\text{a,o}} \in \mathbb{R}^{N_{\text{o}} \times D_{\text{b}}}$, where N_{h} and N_{o} refer to the number of human and object instances, respectively. We then enumerate all possible HO pairs to obtain $\mathbf{V}_{\text{a,ho}} \in \mathbb{R}^{N_{\text{ho}} \times 2D_{\text{b}}}$, where each row represents the concatenated features of an HO pair, and $N_{\text{ho}} = N_{\text{h}} \times N_{\text{o}}$ denotes the total

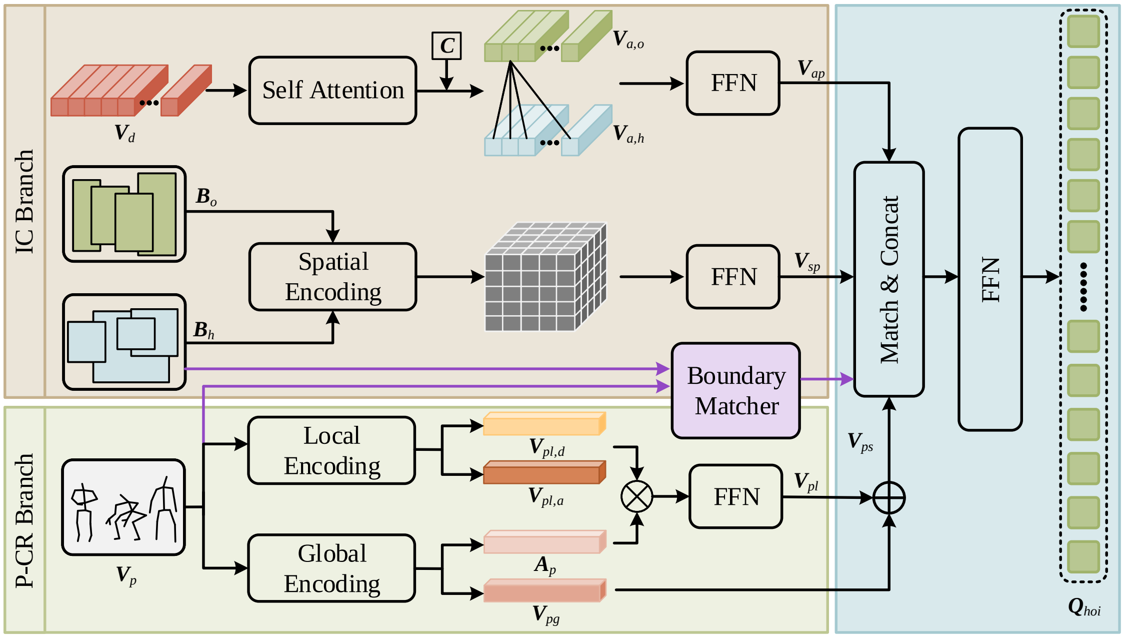


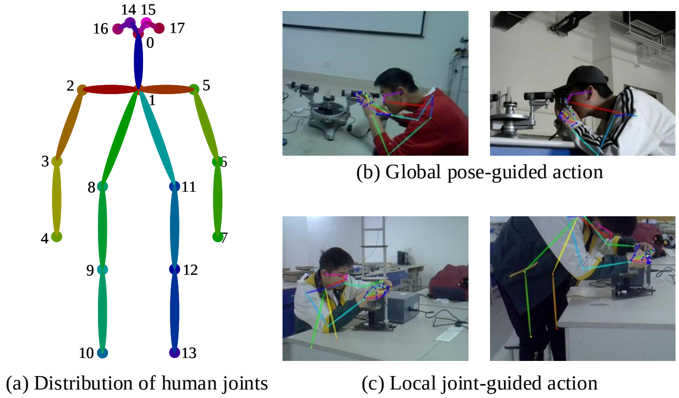
Fig. 4. **Overview of the CM-CQC.** Visual appearance, relative spatial, and human pose features are refined and then fused to generate interaction queries.

number of pairs. $\mathbf{V}_{\text{a,ho}}$ is finally passed through an FFN layer to generate refined appearance features $\mathbf{V}_{\text{ap}} \in \mathbb{R}^{N_{\text{ho}} \times D_{\text{b}}}$. This systematic modeling of intra- and inter-instance relationships enhances the semantic representation of HOIs, facilitating more reliable appearance-based interaction reasoning.

The spatial stream complements this process by encoding relative positional relationships between HO pairs, including center distance, relative area, intersection over union, aspect ratio, and relative orientation [5]. These features are embedded using translation-invariant encodings, ensuring robustness against changes in imaging parameters such as viewpoint or focal length [19]. The concatenated positional encodings are then processed by an FFN to produce refined spatial representations $\mathbf{V}_{\text{sp}} \in \mathbb{R}^{N_{\text{ho}} \times D_{\text{s}}}$, which capture discriminative information about HO spatial configurations.

P-CR Branch. This branch addresses the inherent limitations of relying solely on visual appearance features for action representation, which are often susceptible to noise interference and lack sufficient discriminability [3], [6]. To overcome these issues, we incorporate pose estimation algorithms, such as OpenPose [33], to extract the locations and connections of key body joints. This process provides a preliminary pose configuration that describes both the positions and orientations of joints. Building on this representation, we analyze human actions from two complementary perspectives: global pose features and local joint features.

Global pose features capture the overall state of the body, reflecting balance, stability, and coordination [34]. For example, Figure 5(b) illustrates the "eyepiece imaging" action, which requires full-body coordination: the eyes approach and focus on the tube, the hands stabilize it, the torso leans forward, and the body remains seated. Such actions demand systematic coordination across multiple joints. To obtain the global pose features $\mathbf{V}_{\text{pg}} \in \mathbb{R}^{N_{\text{p}} \times D_{\text{pg}}}$, we construct sequential connections among human keypoints and enhance them through a feed-forward network (FFN) layer f_{pg}:



(a) Distribution of human joints
(b) Global pose-guided action
(c) Local joint-guided action

Fig. 5. **Examples of interactions illustrating behavioral patterns captured through global pose or local joint features.**

$$\mathbf{V}_{\text{pg}} = f_{pg}(\text{reshape}(\mathbf{V}_{\text{p}}, (N_{\text{p}}, -1))). \tag{6}$$

In contrast, local joint features emphasize fine-grained coordination among specific joints, which is critical for certain actions [35]. As shown in Figure 5(c), the "adjusting the limit screw" action relies primarily on the precise configuration of the arm and hand joints, even though other body parts may vary significantly across performers. To capture these relationships, we compute joint distances $\mathbf{V}_{\text{pl,d}} \in \mathbb{R}^{N_{\text{p}} \times 18}$ and angles $\mathbf{V}_{\text{pl,a}} \in \mathbb{R}^{N_{\text{p}} \times 18}$, which together model local motion and coordination patterns among these joints.

Furthermore, since different actions place varying emphasis on particular joints, we introduce a Global Pose-Modulated Local Attention (GPLA) mechanism to construct the attention representations $\mathbf{A}_{\text{p}} \in \mathbb{R}^{N_{\text{p}} \times 18}$ for local joint features. Specifically, global pose embeddings provide contextual priors that guide attention toward discriminative local features, thereby facilitating the identification of subtle motion patterns and fine-grained spatial relationships. The underlying rationale is that global features encode coarse-grained action morphology and execution patterns, providing structure-aware contextual cues about coarse-grained action types [36]. In turn, these cues inform selective attention to interaction-relevant joints. The formulation of GPLA is given in Equations (7)–(8):

$$\mathbf{V}_{\text{pl,d}}, \mathbf{V}_{\text{pl,a}} = \mathcal{G}(\mathbf{V}_{\text{p}}), \tag{7}$$

$$\mathbf{A}_{\text{p}} = \text{softmax}(f_a(\mathbf{V}_{\text{pg}}) \times \mathbf{W} + \mathbf{b}), \tag{8}$$

where $\mathcal{G}(\cdot)$ denotes a geometric mapping function that transforms the raw pose feature $\mathbf{V}_{\text{p}}$ into structured geometric representations. For each human instance, it computes Euclidean distances and orientation angles between sequentially adjacent keypoints along the body's kinematic chain, yielding $\mathbf{V}_{\text{pl,d}}$ and $\mathbf{V}_{\text{pl,a}}$. Subsequently, the attention weights are computed by an FFN layer f_a, parameterized by the learnable weight matrix $\mathbf{W}$ and bias $\mathbf{b}$.

Then, we refine the attention-weighted local features using another FFN layer f_{pl} to obtain $\mathbf{V}_{\text{pl}} \in \mathbb{R}^{N_{\text{p}} \times D_{\text{pl}}}$:

$$\mathbf{V}_{\text{pl}} = f_{pl}(\mathbf{V}_{\text{pl,d}} \cdot \mathbf{A}_{\text{p}} \oplus \mathbf{V}_{\text{pl,a}} \cdot \mathbf{A}_{\text{p}}). \tag{9}$$

Finally, we concatenate $\mathbf{V}_{\text{pg}}$ with $\mathbf{V}_{\text{pl}}$ to form the comprehensive pose representation $\mathbf{V}_{\text{ps}} \in \mathbb{R}^{N_{\text{p}} \times D_{\text{p}}}$, where $D_{\text{p}} = D_{\text{pg}} + D_{\text{pl}}$. This crucial step aims to enhance the accuracy and robustness of interaction recognition by effectively integrating both global and local pose information.

Query Construction with Boundary Matching. Effective query construction requires integrating diverse visual cues. Appearance and spatial features are both derived from the object detector, and thus naturally aligned within the detected HO pairs. In contrast, pose features are produced by an external pose estimator and therefore must be explicitly associated with the detection results. A common practice is to localize human bounding boxes first and then perform pose estimation within each box [3], [6]. Although intuitive, this sequential strategy often yields redundant pose predictions within a single image, thereby incurring substantial computational costs. To address this issue, we employ a parallel detection scheme and design a Boundary Matcher (BM) to efficiently align pose features with object detector outputs.

Specifically, BM employs a straightforward yet effective maximum-IOU matching mechanism. For human poses in $\mathbf{V}_{\text{p}}$, we first compute their bounding boxes $\mathbf{B}_{\text{p}} \in \mathbb{R}^{N_{\text{p}} \times 4}$. Subsequently, we construct an IOU matrix between $\mathbf{B}_{\text{h}}$ (from the object detector) and $\mathbf{B}_{\text{p}}$ (from the pose estimator). For each human detection $\mathbf{b}_{\text{h}} \in \mathbf{B}_{\text{h}}$, the corresponding body pose with the highest IoU is selected as its unique match. The matching relationship is defined by the mapping function $\phi : \mathbf{B}_{\text{h}} \to \mathbf{B}_{\text{p}}$, as illustrated in Equation (10).

$$\phi = \text{argmax}(\text{IOU}(\mathbf{B}_{\text{h}}, \mathbf{B}_{\text{p}}), \text{axis} = 1). \tag{10}$$

Next, we utilize $\phi(\cdot)$ to match the pose features $\mathbf{V}_{\text{ps}}$ of each human instance within HO pairs. These features are then concatenated to form $\mathbf{V}_{\text{po}} \in \mathbb{R}^{N_{\text{ho}} \times D_{\text{p}}}$. The process is mathematically expressed by Equation (11):

$$\mathbf{V}_{\text{po}} = \text{concat}(\mathbf{V}_{\text{ps}}[\phi(\sigma_1)], \mathbf{V}_{\text{ps}}[\phi(\sigma_2)], \cdots, \mathbf{V}_{\text{ps}}[\phi(\sigma_{N_{\text{ho}}})]), \tag{11}$$

where σ_i denotes the index in $\mathbf{B}_{\text{h}}$ of the human instance associated with the i-th HO pair.

Finally, we concatenate the features extracted from the three branches and perform transformations to generate interaction queries $\mathbf{Q}_{\text{hoi}}$. This process is formalized in Equation (12):

$$\mathbf{Q}_{\text{hoi}} = \text{ReLU}(f_m(\text{LN}(\mathbf{V}_{\text{ap}}) \oplus \text{LN}(\mathbf{V}_{\text{sp}}) \oplus \text{LN}(\mathbf{V}_{\text{po}}))), \tag{12}$$

where f_m denotes a fully connected layer and LN represents layer normalization. By integrating multi-source visual cues, the proposed design yields discriminative query embeddings, thereby enhancing the robustness of interaction classification in the downstream Interaction Decoder [5].

IV. Experiments and Analysis

In this section, we conduct comprehensive experiments to evaluate the effectiveness of the proposed model and benchmark it against state-of-the-art methods. In addition, we perform ablation studies, parameter sensitivity analysis, qualitative visualizations, and limitation analysis. These evaluations provide insights into how different design choices influence performance, offering a deeper understanding of the model's strengths and limitations.

A. Datasets and Evaluation Metrics

1) Datasets: To evaluate the effectiveness of PM-EQC, we conduct extensive experiments on PhysLab [37], HICO-DET [38], and V-COCO [39] dataset. The PhysLab dataset captures the experimental processes of different students conducting four physics experiments, including spectrometer-based measurement, rigid-body inertia determination, surface tension measurement, and object density measurement. It contains 4,500 images, covering 34 experimental instruments and 24 interaction verbs. The HICO-DET dataset consists of 38,118 training images and 9,658 testing images. It encompasses 80 object classes, 117 action classes, and 600 HOI interaction classes defined as verb-object pairs. In contrast, V-COCO is derived from the MS-COCO [40] dataset. It comprises 5,400 training images and 4,946 test images, encompassing 80 object classes and 29 action classes.

2) Evaluation: We adopt the widely adopted mean Average Precision (mAP) metric to evaluate HOI detection performance [3], [5], [22]. A predicted triplet $< human, verb, object >$ is considered a True Positive (TP) if three conditions are met: 1) the human and object instances must be correctly classified, 2) the predicted bounding boxes for both instances achieve an IoU above 0.5 with their ground truths, and 3) the interaction verb is accurately classified.

For PhysLab and HICO-DET, following established conventions [2], [5], experimental results are reported across three subsets: Full (all HOI categories), Rare (categories with fewer than 10 training samples), and Non-Rare (categories with at least 10 training samples). Under the Default Setting, AP for each HOI category is computed using all test images, regardless of whether they contain the target object. For HICO-DET, we also report using the Known Object Setting, where AP is calculated only on images containing the relevant object. Regarding the V-COCO dataset, we evaluate under two

TABLE I
COMPARISON OF DETECTION PERFORMANCE (mAP ×100) ON THE HICO-DET AND V-COCO TEST SETS. 'A', 'S', 'P' AND 'L' REPRESENT THE APPEARANCE FEATURE, SPATIAL FEATURE, HUMAN POSE FEATURE AND LANGUAGE FEATURE, RESPECTIVELY.

			HICO-DET						**V-COCO**	
			Default Setting			Known Object Setting				
Method	**Backbone**	**Feature**	Full	Rare	Non-Rare	Full	Rare	Non-Rare	AP^{S1}_{role}	AP^{S2}_{role}
RCD_3 [9]	ResNet-50	A+S+P	20.93	18.95	21.32	23.02	20.96	23.42	-	-
STIP [2]	ResNet-50	A+S+L	32.22	28.15	33.43	35.29	31.43	36.45	66.0	$\underline{70.7}$
GEN-VLKT$_s$ [41]	ResNet-50	A+L	33.75	29.25	35.10	36.78	32.75	37.99	62.4	64.5
OCN [4]	ResNet-50	A+L	30.91	25.56	32.51	-	-	-	64.2	66.3
RLIP [24]	ResNet-50	A+L	32.84	26.85	34.63	-	-	-	61.9	64.2
Wu et al. [6] CDT	ResNet-50	A+S+P	**35.15**	**33.71**	35.58	37.56	35.87	38.06	63.0	65.1
[26] MUREN [19]	ResNet-50	A+L	30.48	25.48	32.37	-	-	-	61.4	65.4
VT-HOI [42]	ResNet-50	A	32.87	28.67	34.14	35.52	30.88	36.91	**68.8**	**71.0**
DiffHOI-S [14]	ResNet-50	A+L	31.88	27.52	34.03	34.95	30.38	37.20	65.4	68.0
TED-Net [8] PDN-	ResNet-50	A+L	34.41	31.07	35.40	37.31	34.56	38.14	61.1	63.5
M [43] SOV-STG-S	ResNet-50	A+S	34.00	29.88	35.24	37.13	33.63	38.18	63.4	65.0
[7] KI2HOI [10]	ResNet-50	A	31.84	26.38	33.47	34.45	29.34	35.97	62.9	64.8
PM-EQC-S (Ours)	ResNet-50	A+L	33.80	29.28	35.15	36.22	30.99	37.78	-	-
	ResNet-50	A+S+L	34.20	$\underline{32.26}$	**36.10**	$\underline{37.85}$	$\underline{35.89}$	$\underline{38.78}$	63.9	65.0
	ResNet-50	A+S+P	$\underline{34.58}$	31.25	$\underline{35.69}$	**38.28**	**36.03**	**39.04**	$\underline{66.1}$	68.4
OCN [4]	ResNet-101	A+L	31.43	25.80	33.11	-	-	-	65.3	67.1
UPT [25]	ResNet-101	A+S	32.31	28.55	33.44	35.65	31.60	36.86	60.7	66.2
GEN-VLKT$_l$ [41]	ResNet-101	A+L	34.95	31.18	36.08	38.28	34.36	39.37	63.6	65.9
CQL+GEN-VLKT$_l$ [44]	ResNet-101	A+L	$\underline{36.03}$	$\underline{33.16}$	$\underline{36.89}$	$\underline{38.82}$	$\underline{35.51}$	$\underline{39.81}$	$\underline{66.5}$	$\underline{69.9}$
OpenCat [45]	ResNet-101	A+L	32.68	28.42	33.75	-	-	-	61.9	63.2
PDN-L [43]	ResNet-101	A	33.18	27.95	34.75	35.86	20.57	37.43	64.7	66.7
SOV-STG-L [7]	ResNet-101	A+L	35.01	30.63	36.32	37.60	32.77	39.05	63.9	65.4
PM-EQC-M (Ours)	ResNet-101	A+S+P	**36.19**	**33.39**	**37.06**	**40.09**	**39.08**	**40.39**	**67.4**	**70.1**
ERNet [22] ViPLO$_s$	ENetV2-XL	A	35.92	30.13	38.29	-	-	-	-	64.2
[3] ViPLO$_l$ [3]	VIT-B/32	A+S+P	34.95	33.83	35.28	38.15	36.77	38.56	60.9	66.6
DiffHOI-L [14] PViC	VIT-B/16	A+S+P	37.22	35.45	37.75	40.51	38.82	41.15	62.2	68.0
[5]	Swin-L	A+L	41.50	39.96	41.96	43.62	41.41	44.28	$\underline{65.7}$	68.2
FGAHOI [15]	Swin-L	A+S	$\underline{44.32}$	**44.61**	$\underline{44.24}$	$\underline{47.81}$	**48.38**	$\underline{47.64}$	64.1	$\underline{70.2}$
VC-HOI [11]	Swin-L	A	37.18	20.71	39.11	38.93	31.93	41.02	60.5	61.2
SOV-STG-Swin-L [7]	ViT-B/16	A+S+L	35.60	35.21	35.71	-	-	-	59.2	65.0
TKCE [12]	Swin-L	A+L	43.35	42.25	43.69	45.53	43.62	46.11	63.9	65.4
PM-EQC-L (Ours)	ResNet-50+ViT-L	A+L	39.57	37.71	40.12	-	-	-	-	-
	Swin-L	A+S+P	**44.53**	$\underline{44.23}$	**44.80**	**48.01**	$\underline{47.67}$	**48.24**	**67.9**	**70.8**

standard protocols: Scenario 1 (AP^{S1}_{role}), where occluded object boxes must be predicted as [0,0,0,0], and Scenario 2 (AP^{S2}_{role}), where occluded objects can be omitted.

B. Implementation Details

To assess the scalability and effectiveness of PM-EQC, we conduct experiments with three backbone-object detector combinations: ResNet-50 [46] with DETR [30], ResNet-101 [46] with DETR [30], and Swin-Transformer [47] with Hybrid-DETR ($\mathcal{H}$-DETR) [48]. For pose estimation, we employ OpenPose [33], which has been demonstrated to accurately extract reliable keypoint configurations on the above datasets. For BM, we set an IoU threshold of 0.5 to establish reliable correspondence between human detections and pose estimates. During training, object detectors and pose estimators are initialized with pre-trained weights and kept frozen to simplify optimization. Following prior works [2], [5], [19], [22], we set the number of instance queries to 100 and the number of Interaction Decoder layers to 2.

To facilitate robust model learning, we apply several data augmentation techniques, including random horizontal flipping, scaling, cropping, and perturbations to hue, saturation, and brightness. The model is optimized using focal loss [49], while AdamW [50] is used with a learning rate of 1e-4 and weight decay of 1e-4. Training runs for 35 epochs, with the learning rate halved every 10 epochs. More details on the experimental environment and data preprocessing procedures are provided in Appendix A.

C. Comparisons with the state-of-the-art

To ensure a fair comparison with existing methods that group results by backbone architecture, we evaluate our method using three representative backbones: lightweight (ResNet-50), medium (ResNet-101), and heavy (Swin-Transformer-Large, Swin-L). Tables I and II report the HOI detection performance of our method on the HICO-DET, V-COCO, and PhysLab datasets, highlighting comparisons with state-of-the-art baselines.

Overall, our framework achieves highly competitive results across settings, ranking within the Top2 in 30 out of 33 experimental configurations. It is worth noting that while many recent HOI detectors improve performance with textual priors or VLM-based guidance, PM-EQC achieves comparable or superior results with purely visual cues. This indicates that our multi-cue visual modeling strategy can serve as an effective alternative to text-driven approaches, even in challenging zero-shot scenarios.

TABLE II
COMPARISON OF DETECTION PERFORMANCE (MAP ×100) ON THE PHYSLAB TEST SET.

Method	Backbone	Feature	PhysLab		
			Full	Rare	Non-Rare
STIP [2]	ResNet-50	A+S+L	62.62	61.00	62.73
GEN-VLKT$_s$ [41]	ResNet-50	A+L	58.71	**75.00**	57.58
OCN [4]	ResNet-50	A+L	52.19	68.01	51.10
LOGICHOI [51]	ResNet-50	A+L	53.34	50.97	53.50
TED-Net [8]	ResNet-50	A+S	60.97	69.89	60.35
SOV-STG-S [7]	ResNet-50	A+L	45.42	34.17	49.50
PM-EQC-S (Ours)	ResNet-50	A+S+P	**66.34**	64.22	**66.49**
OCN [4]	ResNet-101	A+L	49.50	50.00	49.46
UPT [25]	ResNet-101	A+S	65.28	63.47	65.43
GEN-VLKT$_l$ [41]	ResNet-101	A+L	60.19	64.25	59.91
SOV-STG-L [7]	ResNet-101	A+L	52.35	46.59	58.44
PM-EQC-M (Ours)	ResNet-101	A+S+P	**68.64**	**65.37**	**68.84**
ERNet [22]	ENetV2-XL	A	56.39	49.26	62.36
PViC [5]	Swin-L	A+S	71.42	68.47	71.62
FGAHOI [15] SOV-	Swin-L	A	60.78	54.89	67.23
STG-Swin-L [7] **PM-**	Swin-L	A+L	56.48	45.23	58.44
EQC-L (Ours)	Swin-L	A+S+P	**73.75**	**73.27**	**73.79**

For PhysLab, PM-EQC delivers comprehensive improvements of 5.9%, 5.1%, and 3.3% over the next-best models when using ResNet-50, ResNet-101, and Swin-L backbones, respectively. For HICO-DET, PM-EQC shows larger gains on the Non-Rare and Full subsets than on the Rare subset, suggesting that while our approach effectively leverages discriminative visual cues, challenges remain in capturing interaction patterns from limited samples. It highlights the potential of refining explicit interaction priors from a data-centric perspective to further boost detection performance.

For V-COCO, PM-EQC consistently performs better in Scenario 2 than in Scenario 1. This difference arises because our method encodes diverse relative positional relationships between humans and objects, making it sensitive to positional information. Consequently, it conflicts with enforcing regression of occlusion object boundaries to [0,0,0,0] in Scenario 1. In contrast, by relaxing the boundary constraints, PM-EQC achieves higher performance in Scenario 2. In addition, performance differences between the three settings further confirm that PM-EQC benefits from advances in backbone networks and object detectors.

When utilizing the 'A+S+P' feature set, the model proposed by Wu et al. [6] exhibits competitive performance. However, compared to this model, PM-EQC avoids dataset performance bias, i.e., the significant performance disparity between the HICO-DET and V-COCO datasets. This consistency highlights PM-EQC's robustness and generalization ability across varied data distributions. Moreover, PM-EQC supports zero-shot inference of unseen interactions, underscoring its superior capacity to capture diverse interaction patterns.

More specifically, we conduct experiments in two zero-shot settings, namely Rare First Unseen Combinations (RF-UC) and Non-Rare First Unseen Combinations (NF-UC), to further validate the generalization ability of our method. In these settings, the training set includes all objects and verbs but omits certain verb-object combinations. Specifically, 480 HOIs are seen during training, while 120 are held out as unseen. The RF-UC and NF-UC settings select unseen classes from the tail and head of the HOI distribution, respectively. Experimental results in Table III showcase the exceptional detection performance of PM-EQC across various backbones, zero-shot settings, and data subsets, demonstrating its robust generalization capabilities. In addition, PM-EQC notably enhances unseen interaction detection performance without additional language features compared to existing methods. In particular, PM-EQC-L surpasses VC-HOI [11] by more than 28.3%. This advancement can be attributed to the precise extraction of relative position features and human posture characteristics, coupled with the nuanced understanding and knowledge transfer of interaction patterns.

For instance, considering the action 'ride', regardless of scene or riding tools, humans typically adopt a seated position above the tool and use their hands for control. By capturing multifaceted behavioral patterns, PM-EQC discerns commonalities in interaction behaviors, enabling accurate predictions even for unseen combinations. Furthermore, scaling from PM-EQC-S to PM-EQC-L yields substantial gains, demonstrating that the framework effectively leverages improvements in backbones and detectors. Therefore, fine-tuning pre-trained modules will provide additional optimization capacity, highlighting the scalability of PM-EQC.

D. Ablation Study

PM-EQC demonstrates competitive performance across various model settings. To balance efficiency and accuracy, we adopt the configuration "ResNet-50+DETR+OpenPose" as a lightweight yet effective baseline for ablation and parameter analyses. This choice reduces computational cost and mitigates interference from overly complex architectures [5].

Table IV reports the ablation results of visual cues in CM-CQC. When utilizing a single visual cue, on both PhysLab and HICO-DET, appearance features outperform spatial features by 1.60 and 1.04 mAP, and outperform pose features by 1.54 and 0.77 mAP, respectively. These results align with prior studies underscoring the effectiveness of high-dimensional appearance representations [5], [15], [19], [20]. Notably, integrating two cues consistently improves over single-cue settings, with the contribution of pose being particularly notable. Moreover, integrating all three visual features yields the best performance across all subsets. This underscores the complementary and consistent interactive benefits of these visual cues.

Table V further presents the attention ablation results on CM-CQC. For clarity, we denote the attention mechanisms in the IC and P-CR branches as $\mathbf{A}_{\mathrm{a}}$ and $\mathbf{A}_{\mathrm{p}}$, respectively. The results demonstrate that the integration of $\mathbf{A}_{\mathrm{a}}$ improves mAP by 0.95 on PhysLab and 0.28 on HICO-DET, indicating that aggregating crucial contextual information from neighbors can enhance instance representation [19]. Comparing H3 with H1 further highlights the effectiveness of constructing attention values for local joint features using global pose cues. Employing both $\mathbf{A}_{\mathrm{a}}$ and $\mathbf{A}_{\mathrm{p}}$ simultaneously yields gains of 2.01 and 0.72 mAP, underscoring the importance of attention-driven cue integration.

In addition, we conduct feature ablation experiments within the P-CR branch, and the corresponding results are presented in Table VI. Overall, the inclusion of global pose features or

TABLE III
COMPARISON OF DETECTION PERFORMANCE (MAP ×100) ON THE HICO-DET TEST SET UNDER THE RF-UC AND NF-UC SETTINGS.

			RF-UC			**NF-UC**		
Method	**Backbone**	**Feature**	Full	Unseen	Seen	Full	Unseen	Seen
GEN-VLKT$_s$ [41]	ResNet-50	A+L	30.56	21.36	32.91	23.71	25.05	23.38
RLIP [24]	ResNet-50	A+L	30.52	19.19	33.35	26.16	20.27	27.67
LOGICHOI [51]	ResNet-50	A+S+L	33.17	25.97	34.93	27.95	26.84	27.86
CLIP4HOI [52]	ResNet-50	A+S+L	-	-	-	28.90	31.44	28.26
HOICLIP [53] VT-	ResNet-50	A+L	32.99	25.53	34.85	27.75	26.39	28.10
HOI [42] ZSHOI-	ResNet-50	A+L	30.20	21.74	32.31	29.27	32.75	28.39
VLT [54] KI2HOI	ResNet-50	A+L	24.51	16.50	26.90	20.46	19.47	20.93
[10] HOLa [55]	ResNet-50	A+S+L	34.10	26.33	**35.79**	27.77	28.89	28.31
PM-EQC-S (Ours)	ResNet-50	A+S+L	**34.19**	**30.61**	35.08	**32.36**	35.25	**31.64**
	ResNet-50	A+S+P	33.60	29.00	34.75	31.25	**35.57**	30.16
SCL [56]	ResNet-101	A	28.08	19.07	30.39	24.34	21.73	25.00
EoID [57] OpenCat	ResNet-101	A+L	29.52	22.04	31.39	26.69	26.77	26.66
[45]	ResNet-101	A+L	31.38	21.46	33.86	27.08	23.25	28.04
PM-EQC-M (Ours)	ResNet-101	A+S+P	**34.74**	**30.19**	**35.87**	**32.68**	**36.44**	**31.74**
ADA-CM [58]	ResNet-50+ViT-B/16	A+L	33.01	27.63	34.35	31.39	32.41	31.13
ContextHOI [59]	ViT-B/16	A+L	24.01	19.34	25.33	-	-	-
DHD [60]	Swin-B	A+S+L	28.53	23.32	30.09	23.14	27.35	22.09
EZ-HOI [61] TKCE	ResNet-50+ViT-B	A+L	33.13	29.02	34.15	31.17	33.66	30.55
[12]	ResNet-50+VIT-L	A+L	33.91	28.95	35.50	32.10	33.20	32.15
VC-HOI [11] **PM-**	ViT-B/16	A+S+L	34.49	32.34	35.02	32.28	35.81	31.40
EQC-L (Ours)	Swin-L	A+S+P	**40.28**	**40.25**	**40.29**	**41.20**	**45.95**	**40.02**

TABLE IV
THE ABLATION RESULTS (MAP ×100) OF MODEL VARIANTS WITH DIFFERENT VISUAL CUES ON PHYSLAB AND HICO-DET TEST SETS.

	Modality Fusion			**PhysLab**			**HICO-DET**		
#	Spatial	Appearance	Pose	Full	Rare	Non-Rare	Full	Rare	Non-Rare
M1	✓			63.19	60.77	63.25	32.85	29.44	34.00
M2		✓		64.79	61.34	65.03	33.89	30.47	34.94
M3			✓	63.25	61.49	63.37	33.12	30.20	34.00
M4	✓	✓		65.27	62.33	65.48	34.03	30.59	35.09
M5	✓		✓	64.05	62.56	64.15	33.16	30.38	34.25
M6		✓	✓	65.72	63.90	65.85	34.30	30.98	35.29
M7	✓	✓	✓	**66.34**	**64.22**	**66.49**	**34.58**	**31.25**	**35.69**

TABLE V
THE ABLATION RESULTS (MAP ×100) OF ATTENTION MECHANISMS ON PHYSLAB AND HICO-DET TEST SETS.

	Attention		**PhysLab**			**HICO-DET**		
#	$\mathbf{A}_a$	$\mathbf{A}_p$	Full	Rare	Non-Rare	Full	Rare	Non-Rare
H1			64.33	61.15	64.55	33.86	30.89	34.98
H2	✓		65.28	63.71	65.39	34.14	30.61	35.18
H3		✓	65.06	63.62	65.16	34.22	30.65	35.28
H4	✓	✓	**66.34**	**64.22**	**66.49**	**34.58**	**31.25**	**35.69**

TABLE VI
THE ABLATION RESULTS (MAP ×100) OF FEATURE COMPONENTS IN THE P-CR BRANCH ON PHYSLAB AND HICO-DET TEST SETS.

	Features		**PhysLab**			**HICO-DET**		
#	Global	Local	Full	Rare	Non-Rare	Full	Rare	Non-Rare
F1			65.27	62.33	65.48	34.29	30.59	35.39
F2	✓		65.80	63.62	65.95	34.42	30.43	35.61
F3		✓	66.07	63.73	66.23	34.47	30.81	35.62
F4	✓	✓	**66.34**	**64.22**	**66.49**	**34.58**	**31.25**	**35.69**

TABLE VII
DIFFERENT CHOICES OF FEATURES AS KEYS/VALUES IN THE CROSS-ATTENTION MECHANISM.

			PhysLab			HICO-DET		
#	Key/Values	Feature Head	Full	Rare	Non-Rare	Full	Rare	Non-Rare
D1	Backbone C3	C4,C5	65.24	62.89	65.40	33.88	30.75	35.07
D2	Backbone C4	C5	65.02	62.78	65.18	33.93	29.63	**35.22**
D3	Backbone C5	None	**65.54**	**63.66**	**65.67**	**34.04**	**31.08**	35.01
D4	Image Encoder	None	64.51	62.54	64.44	31.97	27.25	33.43
E1	Backbone C5	Self-Attention	66.09	64.01	66.23	34.28	**31.30**	35.28
E2	Backbone C5	Window-Attention	**66.34**	**64.22**	**66.49**	**34.58**	31.25	**35.69**

local joint features improves the HOI detection performance. However, we observe a slight decline in performance on the Rare subset of the HICO-DET dataset after incorporating global pose features, whereas integrating local joint features further improves performance by 0.22 mAP. This outcome suggests that global pose features struggle to capture the essential attributes of interaction when the number of training samples is limited, due to human behavioral variability. Conversely, the local joint features offer a more localized and detailed behavior representation [6], enabling the extraction of motion patterns from limited samples. Furthermore, as shown in F4, leveraging both features jointly leads to mAP improvements of 1.07 and 0.29 on the two datasets, respectively. This synergy arises from their distinct yet complementary perspectives on human posture, enabling a more comprehensive posture representation and thereby more accurate interaction recognition.

E. Parameter Study

To further assess how parameter choices affect PM-EQC's performance, we conduct a series of experiments. Given the openness of PM-EQC, the analysis centers on the Interaction Decoder and the structural configurations of CM-CQC. Specifically, we investigate three key aspects: (1) the source of cross-attention components in the Decoder, (2) the position embedding and the number of Decoder layers, and (3) the selection of detection function components.

Table VII presents the experimental results of using different feature sources for the Key/Values in the Interaction Decoder, with the Feature Head responsible for refining the global features. The findings from D1-D4 indicate that backbone outputs provide superior global features compared to encoder outputs, as they retain unbiased semantic and spatial information. In contrast, Image Encoder features are optimized for downstream object detection, which may introduce task-specific biases and limit their effectiveness as general global representations.

Furthermore, the comparison between D3, E1, and E2 reveals that adding attention mechanisms can improve global feature refinement. Specifically, incorporating the Self-Attention [29] mechanism improves the mAP by 0.55 (PhysLab) and 0.24 (HICO-DET), while the Window-Attention [47] mechanism yields larger gains of 0.80 and 0.54. This difference arises because Window-Attention emphasizes local dependencies [47], aggregating neighboring information to enhance local interaction patterns. The resulting refined features better align with the interaction query space, facilitating more accurate HOI detection.

In Table VIII, we report the results of parameter experiments on the selection of position embeddings and the number of Decoder layers. For two-stage HOI detection algorithms, as explicit instance boundaries are provided by the object detector, position embeddings are typically constructed from box centers, which we refer to as the 'Standard' approach. Furthermore, a novel method modulates position embeddings based on the width and height of the box, referred to as the 'Modulated' approach [65].

The results of experiments K1-K3 demonstrate the positive impact of incorporating position embeddings on detection accuracy. Notably, modulated position embeddings yield substantial improvements of 1.46 and 0.41 mAP on the PhysLab and HICO-DET datasets, respectively. Furthermore, experiments K3 and G1-G3 investigate the impact of the number of Decoder layers on detection performance. Taking PhysLab as an example, performance saturates when three decoder layers are employed. This phenomenon can be attributed to the effective fusion of multi-source visual cues in the CM-CQC. Furthermore, diminishing returns from additional layers may arise due to vanishing or exploding gradients. These results provide valuable insights into effective configurations for the Interaction Decoder within PM-EQC.

To validate the versatility of PM-EQC, we conduct a series of experiments that replace the functional modules, including the backbone, pose estimator, and object detector. Table IX presents the performance comparison of the modules in the "Compared module" column, where the benchmark performance of modules in related fields progressively improves from top to bottom. For instance, AlphaPose [62] outperforms Openpose [33] in terms of pose estimation. The experimental results demonstrate PM-EQC's adaptability, as it can integrate existing functional modules while maintaining a high level of interaction detection performance. Moreover, the performance improvements achieved in the three modules' respective fields positively impact PM-EQC's interaction detection capability. This observation highlights its ability to leverage advancements in feature extraction, pose estimation, and object detection. Such a partnership enhances the development potential of our framework, enabling it to adapt to future technological advancements and emerging challenges.

TABLE VIII
DIFFERENT CHOICES OF LAYER NUMBER AND POSITIONAL EMBEDDING OF THE INTERACTION DECODER.

			PhysLab			HICO-DET		
#	Positional Embed.	Layers	Full	Rare	Non-Rare	Full	Rare	Non-Rare
K1	None	1	62.97	60.98	63.11	33.55	30.41	34.48
K2	Standard	1	64.06	62.23	64.19	33.80	30.21	34.87
K3	Modulated	1	**64.43**	**62.67**	**64.56**	**33.96**	**30.87**	**34.88**
G1	Modulated	2	65.63	63.45	65.78	**34.58**	31.25	**35.69**
G2	Modulated	3	**66.34**	**64.22**	**66.49**	34.33	**31.77**	35.28
G3	Modulated	4	65.82	64.02	65.95	34.17	31.08	35.09

TABLE IX
COMPARISON OF DETECTION PERFORMANCE (mAP ×100) WITH DIFFERENT CHOICES OF THE BACKBONE, OBJECT DETECTOR, AND POSE ESTIMATOR.

		PhysLab			HICO-DET		
Fixed module	Compared module	Full	Rare	Non-Rare	Full	Rare	Non-Rare
DETR+Openpose	+ResNet-50 [46]	66.34	64.22	66.49	34.58	31.25	35.69
	+ResNet-101 [46]	**68.64**	**65.37**	**68.84**	**36.19**	**33.39**	**37.06**
DETR+ResNet-50	+Openpose [33]	66.34	64.22	66.49	34.58	31.25	35.69
	+AlphaPose	67.11	64.88	67.26	35.46	33.21	36.17
	+HR[62]Net-W32 [63]	**67.24**	**64.99**	**67.39**	**35.52**	**33.55**	**36.25**
ResNet-50+Openpose	+DETR [30]	66.34	64.22	66.49	34.58	31.25	35.69
	+D-DETR [64]	67.70	64.79	67.90	35.29	31.63	36.44
	+$\mathcal{H}$-DETR [48]	**68.21**	**65.34**	**68.41**	**37.32**	**32.14**	**38.85**

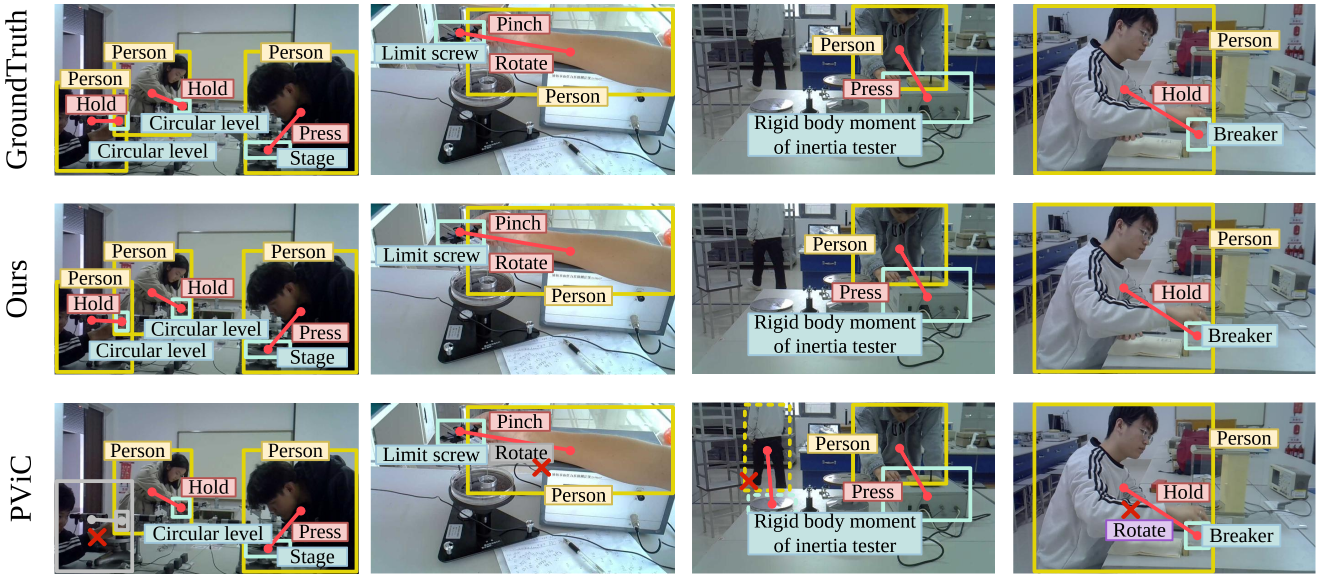


Fig. 6. **Qualitative results of HOI detection on the PhysLab dataset.**

F. Qualitative Analysis

To intuitively demonstrate PM-EQC's strengths, we compare it with another feature fusion model called PViC. Their visualization examples of PhysLab and HICO-DET are presented in Figures 6 and 7, respectively. It can be seen that PViC exhibits four types of detection errors: missing interaction instances, misclassified interaction categories, incorrect assessment of instance interactivity, and missing interaction labels. In contrast, PM-EQC reduces such errors by integrating and refining human pose characteristics and relative positional features. For instance, distinguishing between actions 'Rotate' and 'Hold' depends on subtle posture differences. In 'Rotate', the wrist typically exhibits a pronounced twist or deflection, while the fingers are flexibly bent to facilitate continuous force application during manipulation. In contrast, 'Hold' involves limited arm motion and a stable body posture, reflecting the need for sustained force and coordinated control to maintain a secure grip. Therefore, compared with PViC, our method leverages detailed pose cues to reduce ambiguities in interaction detection, thereby yielding more accurate results.

G. Computational Cost Analysis

In this study, we meticulously assess the balance between our method's detection performance and computational cost.

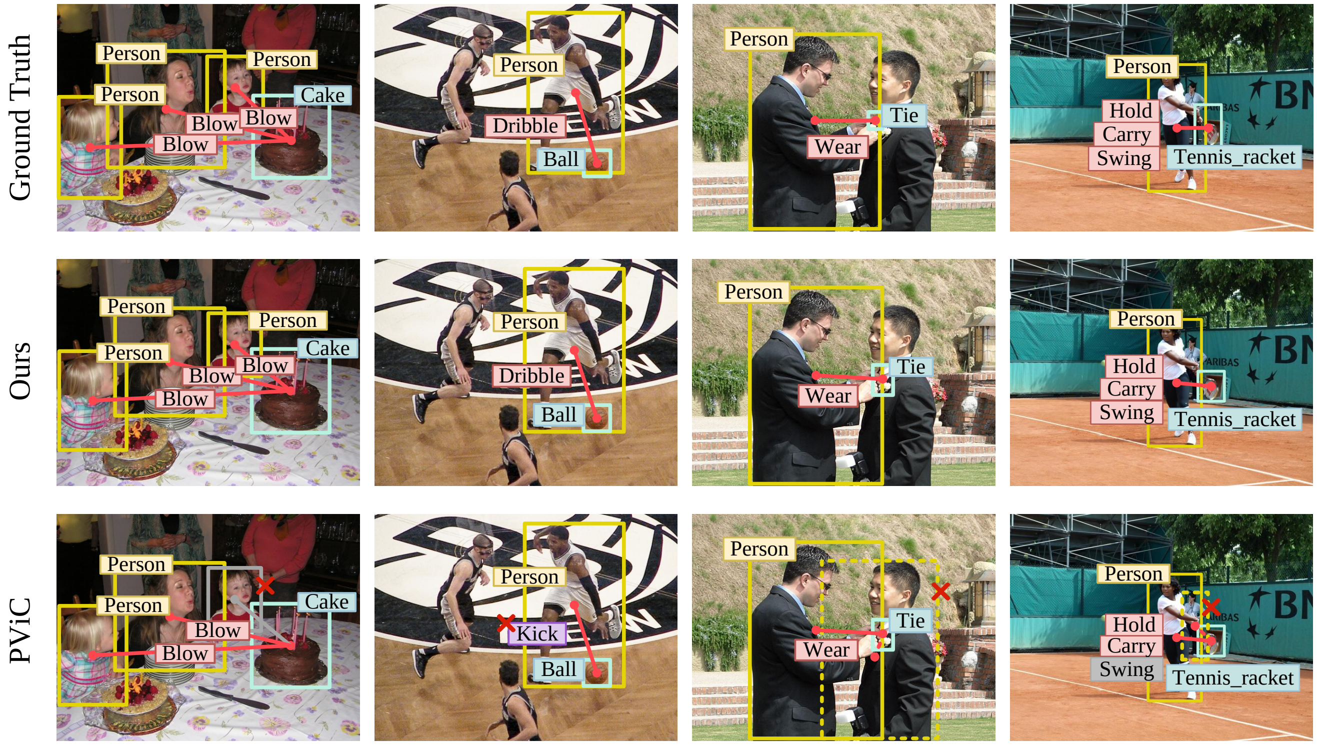


Fig. 7. **Qualitative results of HOI detection on the HICO-DET dataset.**

The experimental results are detailed in Table X, where the 'Params' column distinguishes the total count of model parameters (denoted as 'A') versus trainable parameters (denoted as 'T'), while the 'Time' column reports the mean inference time per image. It can be seen that PM-EQC maintains a favorable balance, achieving strong accuracy with moderate cost.

TABLE X
COMPARISON OF COMPUTATIONAL COST ON PE-HOI WITH NVIDIA GEFORCE RTX 3090.

Type	Method	mAP ↑	Params (A/T) ↓	Time ↓
One-stage	QPIC [1]	54.07	41.68/41.46M	39.13ms
	OCN [4]	52.19	43.77/43.55M	47.42ms
	CDN [17]	59.44	41.67/41.45M	39.95ms
	MUREN [19]	63.27	75.19/74.95M	99.85ms
Two-stage	STIP [2]	62.62	54.71/13.19M	71.38ms
	UPT [25]	65.28	73.70/31.31M	98.46ms
	PM-EQC-S (Ours)	66.34	96.46/30.70M	97.84ms

Specifically, compared with one-stage methods, PM-EQC-S requires fewer trainable parameters yet achieves larger accuracy gains. For example, relative to CDN [17], MUREN [19] improves mAP by 3.83 but with 33.5M more parameters and 59.9 ms longer inference. In contrast, PM-EQC-S improves mAP by 6.90 with less inference time and even reducing the model's training parameters by 10.75M. Compared with two-stage methods, PM-EQC also outperforms UPT [25], gaining 1.06 mAP with comparable efficiency. These findings highlight PM-EQC's practical effectiveness in balancing performance and computational cost.

V. CONCLUSION

In the domain of HOI detection, complex backgrounds and irrelevant objects often disrupt the detection process, leading to misidentification of interaction types and confusion in localizing the correct interaction targets. Upon extensive analysis of detection results, we have identified that such errors primarily stem from insufficient visual cues. To address this issue without increasing model depth or requiring additional annotations, we introduce PM-EQC, an HOI detection framework that integrates appearance features, relative spatial relationships, and human pose cues. Owing to its modular plug-and-play design, PM-EQC readily scales across diverse scenarios and benefits from advances in feature extraction, object detection, and pose estimation techniques.

Building on this framework, we further address the lack of visual context during query construction in Transformer-based approaches. To this end, we propose CM-CQC, which generates explicit and context-rich query representations by jointly modeling heterogeneous cues, with particular emphasis on pose features. Comprehensive ablation studies confirm the effectiveness of each component, while evaluations on the PhysLab, HICO-DET, and V-COCO benchmarks demonstrate strong performance under both fully supervised and zero-shot settings. Visualization results further highlight the model's ability to resolve interaction ambiguities. Overall, these results demonstrate the effectiveness of our design, and we hope the framework can support future advances in HOI detection and related visual reasoning tasks.

VI. FUTURE WORK

Although our method shows strong performance, it also presents two main limitations. First, although freezing the object detector and pose estimator helps reduce training costs, it restricts their adaptability to novel domains and consequently constrains detection performance. Second, while the use of a multi-stage design helps with model training and improves detection accuracy, it also incurs higher inference costs due to increased model complexity.

Looking ahead, we identify two promising directions for improvement. (1) Transfer learning: Incorporating incremental learning to fine-tune the detector and pose estimator could enhance the model's adaptability to new domains while controlling training overhead. (2) Model lightweighting: Techniques such as pruning and knowledge distillation can reduce computational demands while preserving accuracy, and may also help capture statistical interaction patterns to improve the model's robustness and reliability.

Beyond algorithmic improvements, another promising direction is to enrich the framework with new modalities, such as textual information or LLM-derived knowledge, and to extend it to real-world applications like human-robot collaboration, assistive perception systems, intelligent monitoring, and embodied artificial intelligence. Assessing how the multi-cue modeling strategy adapts to these domains is a valuable path for further exploration.

Appendix A
Experimental Environment and Data Preprocessing

Environment Setup. All experiments were conducted with Python 3.8 and PyTorch 1.11.1 on Ubuntu 20.04. The hardware environment consisted of four NVIDIA GeForce RTX 3090 GPUs (24 GB each), with a batch size of 12 images per GPU. All third-party dependencies are specified in the released *requirements.txt* file.

Data Preprocessing. Building upon the established environment setup, we performed data preprocessing for both the object detector and pose estimator components. For the object detector, following standard HOI pipelines [5], [25], all images were resized while preserving their aspect ratio. During training, the shorter image side was randomly selected from [480, 512, 544, 576, 608, 640, 672, 704, 736, 768, 800], and fixed at 800 during testing. Data augmentation included random horizontal flipping, color jittering (brightness, contrast, and saturation), random cropping, and normalization using ImageNet [66] statistics.

For the pose estimator, the default configurations provided by the official implementation were adopted. Input images were resized to 368 × 368 pixels and normalized to the range [-0.5, 0.5]. Heatmaps and Part Affinity Fields (PAFs) were generated and smoothed with a Gaussian filter to localize keypoints using a confidence threshold of 0.1. Valid limb connections were then established according to the predefined kinematic chain with a PAF threshold of 0.05. Incomplete poses (fewer than four valid joints or low average confidence) were discarded before matching with detected human bounding boxes. Missing joints were zero-padded, and pairs with confidence below 0.5 were filtered out during training. Further implementation details are available in the released code repository.

Reproducibility. All training scripts, pretrained weights, and configuration files are publicly available in the accompanying GitHub repository.

References

[1] M. Tamura, H. Ohashi, and T. Yoshinaga, "Qpic: Query-based pairwise human–object interaction detection with image-wide contextual information," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 410–10 419.

[2] 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*, 2022, pp. 19 548–19 557.

[3] J. Park, J.-W. Park, and J.-S. Lee, "Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 152–17 162.

[4] H. Yuan, M. Wang, D. Ni, and L. Xu, "Detecting human–object interactions with object-guided cross-modal calibrated semantics," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 3206–3214.

[5] F. Z. Zhang, Y. Yuan, D. Campbell, Z. Zhong, and S. Gould, "Exploring predicate visual context in detecting of human-object interactions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 411–10 421.

[6] 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*, 2022, pp. 121–136.

[7] J. Chen, Y. Wang, and K. Yanai, "Focusing on what to decode and what to train: Sov decoding with specific target guided denoising and vision language advisor," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025, pp. 9416–9425.

[8] Y. Wang, Q. Liu, and Y. Lei, "Ted-net: Dispersal attention for perceiving interaction region in indirectly-contact hoi detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 5603–5615, 2024.

[9] Y.-L. Li, X. Liu, X. Wu, X. Huang, L. Xu, and C. Lu, "Transferable interactiveness knowledge for human-object interaction detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3870–3882, 2022.

[10] W. Xue, Q. Liu, Y. Wang, Z. Wei, X. Xing, and X. Xu, "Towards zero-shot human–object interaction detection via vision–language integration," *Neural Networks*, vol. 187, p. 107348, 2025.

[11] Y. Zeng, Y. Mao, Z. Lu, W. Zhou, and H. Li, "Leveraging visual captions for enhanced zero-shot hoi detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025, pp. 1–5.

[12] F. Nan, N. Zhang, Q. Liu, W. Jing, G. Dai, Y. Chen, and F. Tian, "Exploring triple knowledge cues for zero-shot human-object interaction detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025, pp. 1–5.

[13] X. Zhong, C. Ding, Y. Hu, and D. Tao, "Disentangled interaction representation for one-stage human-object interaction detection," *arXiv preprint arXiv:2312.01713*, pp. 1–17, 2023.

[14] 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*, pp. 1–14, 2023.

[15] 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.

[16] Y. Cheng, Z. Wang, W. Zhan, and H. Duan, "Multi-scale human-object interaction detector," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 4, pp. 1827–1838, 2023.

[17] A. Zhang, Y. Liao, S. Liu, M. Lu, Y. Wang, C. Gao, and X. Li, "Mining cross-person cues for body-part interactiveness learning in hoi detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 209–17 220, 2021.

[18] D. Yang, Y. Zou, C. Zhang, M. Cao, and J. Chen, "Rr-net: Relation reasoning for end-to-end human-object interaction detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3853–3865, 2022.

[19] 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*, 2023, pp. 2925–2934.

[20] H. Peng, F. Liu, Y. Li, B. Huang, J. Shao, N. Sang, and C. Gao, "Parallel reasoning network for human-object interaction detection," *arXiv preprint arXiv:2301.03510*, pp. 1–9, 2023.

[21] Z. Liu and X. Zhang, "Object centric body part attention network for human-object interaction detection," in *Chinese Conference on Pattern Recognition and Computer Vision*, 2023, pp. 378–391.

[22] 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.

[23] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5353–5363.

[24] 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. 37 416–37 431, 2022.

[25] F. Z. Zhang, D. Campbell, and S. Gould, "Efficient two-stage detection of human–object interactions with a novel unary-pairwise transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 104–20 112.

[26] D. Zong and S. Sun, "Zero-shot human-object interaction detection via similarity propagation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 12, pp. 17 805–17 816, 2024.

[27] X. Qu, C. Ding, X. Li, X. Zhong, and D. Tao, "Distillation using oracle queries for transformer-based human–object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 558–19 567.

[28] X. Han, G. Niu, M. Zhou, and X. Zhang, "Knowledge distilled group prompts learning for hoi detection with large vision-language models," in *IEEE International Conference on Multimedia and Expo*, 2025, pp. 1–6.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[30] 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*, 2020, pp. 213–229.

[31] N. Wang, G. Zhu, H. Li, M. Feng, X. Zhao, L. Ni, P. Shen, L. Mei, and L. Zhang, "Exploring spatio–temporal graph convolution for video-based human–object interaction recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 5814–5827, 2023.

[32] ——, "Exploring spatio–temporal graph convolution for video-based human–object interaction recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 5814–5827, 2023.

[33] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.

[34] H. Zhou, Q. Liu, and Y. Wang, "Learning discriminative representations for skeleton based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 608–10 617.

[35] W. Song, T. Chu, S. Li, N. Li, A. Hao, and H. Qin, "Joints-centered spatial-temporal features fused skeleton convolution network for action recognition," *IEEE Transactions on Multimedia*, vol. 26, pp. 4602–4616, 2024.

[36] K. Gedamu, Y. Ji, L. Gao, Y. Yang, and H. T. Shen, "Relation-mining self-attention network for skeleton-based human action recognition," *Pattern Recognition*, vol. 139, p. 109455, 2023.

[37] M. Zou, Q. Zeng, Y. Miao, S. Liu, Z. Wang, H. Liu, and W. Zhou, "Physlab: A benchmark dataset for multi-granularity visual parsing of physics experiments," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 12 799–12 806.

[38] 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*, 2018, pp. 381–389.

[39] S. Gupta and J. Malik, "Visual semantic role labeling," *arXiv preprint arXiv:1505.04474*, pp. 1–11, 2015.

[40] 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 *European Conference on Computer Vision*, 2014, pp. 740–755.

[41] 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*, 2022, pp. 20 123–20 132.

[42] H. Wang, H. Yu, and Q. Zhang, "Detecting zero-shot human-object interaction with visual-text modeling," in *International Conference on Virtual Reality*, 2023, pp. 155–162.

[43] Y. Cheng, H. Duan, C. Wang, and Z. Chen, "Parallel disentangling network for human–object interaction detection," *Pattern Recognition*, vol. 146, p. 110021, 2024.

[44] C. Xie, F. Zeng, Y. Hu, S. Liang, and Y. Wei, "Category query learning for human-object interaction classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 275–15 284.

[45] S. Zheng, B. Xu, and Q. Jin, "Open-category human-object interaction pre-training via language modeling framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 392–19 402.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[47] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[48] D. Jia, Y. Yuan, H. He, X. Wu, H. Yu, W. Lin, L. Sun, C. Zhang, and H. Hu, "Detrs with hybrid matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 702–19 712.

[49] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[50] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, pp. 1–19, 2017.

[51] L. Li, J. Wei, W. Wang, and Y. Yang, "Neural-logic human-object interaction detection," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 21 158–21 171.

[52] 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*, vol. 36, pp. 45 895–45 906, 2023.

[53] 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*, 2023, pp. 23 507–23 517.

[54] S. Sarma, P. Kalkar, and A. Sur, "Boosting zero-shot human-object interaction detection with vision-language transfer," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 6355–6359.

[55] Q. Lei, B. Wang, and R. T. Tan, "Hola: Zero-shot hoi detection with low-rank decomposed vlm feature adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 1825–1835.

[56] Z. Hou, B. Yu, and D. Tao, "Discovering human–object interaction concepts via self-compositional learning," in *European Conference on Computer Vision*, 2022, pp. 461–478.

[57] M. Wu, J. Gu, Y. Shen, M. Lin, C. Chen, and X. Sun, "End-to-end zero-shot hoi detection via vision and language knowledge distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 2839–2846.

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

[59] J. Gao, K.-H. Yap, K. Wu, D. T. Phan, K. Garg, and B. S. Han, "Contextual human object interaction understanding from pre-trained large language model," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 13 436–13 440.

[60] M. Wu, Y. Liu, J. Ji, X. Sun, and R. Ji, "Toward open-set human object interaction detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 6066–6073.

[61] Q. Lei, B. Wang, and R. Tan, "Ez-hoi: Vlm adaptation via guided prompt learning for zero-shot hoi detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 55 831–55 857, 2024.

[62] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, "Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7157–7173, 2022.

[63] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[64] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[65] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "Dab-detr: Dynamic anchor boxes are better queries for detr," *arXiv preprint arXiv:2201.12329*, pp. 1–19, 2022.

[66] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.