

QueryCraft: Transformer-Guided Query Initialization for Enhanced Human-Object Interaction Detection

Yuxiao Wang^{1*}, Wolin Liang^{1*}, Yu Lei², Weiying Xue¹, Nan Zhuang³, Qi Liu^{1†}

¹School of Future Technology, South China University of Technology

²School of Information Science & Technology, Southwest Jiaotong University

³School of Software Technology, Zhejiang University

ftwangyuxiao@mail.scut.edu.cn, 202264690304@mail.scut.edu.cn, drliuqi@scut.edu.cn

Abstract

Human-Object Interaction (HOI) detection aims to localize human-object pairs and recognize their interactions in images. Although DETR-based methods have recently emerged as the mainstream framework for HOI detection, they still suffer from a key limitation: Randomly initialized queries lack explicit semantics, leading to suboptimal detection performance. To address this challenge, we propose QueryCraft, a novel plug-and-play HOI detection framework that incorporates semantic priors and guided feature learning through transformer-based query initialization. Central to our approach is **ACTOR** (Action-aware Cross-modal TransfORmer), a cross-modal Transformer encoder that jointly attends to visual regions and textual prompts to extract action-relevant features. Rather than merely aligning modalities, ACTOR leverages language-guided attention to infer interaction semantics and produce semantically meaningful query representations. To further enhance object-level query quality, we introduce a **Perceptual Distilled Query Decoder (PDQD)**, which distills object category awareness from a pre-trained detector to serve as object query initialization. This dual-branch query initialization enables the model to generate more interpretable and effective queries for HOI detection. Extensive experiments on HICO-Det and V-COCO benchmarks demonstrate that our method achieves state-of-the-art performance and strong generalization.

Code — <https://drliuqi.github.io/>

Extended version — <https://arxiv.org/abs/2508.08590>

introduction

Human-Object Interaction (HOI) detection aims to identify and localize human-object pairs in images while recognizing their interactions, generating structured (human, action, object) triplets. Beyond traditional detection tasks, HOI detection requires understanding semantic relationships between entities, demanding robust scene modeling capabilities. With applications spanning behavior recognition, image captioning, video analysis, and robotic perception, this task has attracted significant research attention (Liao et al. 2022; Wang et al. 2024c, 2025).

*These authors contributed equally.

†Corresponding author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

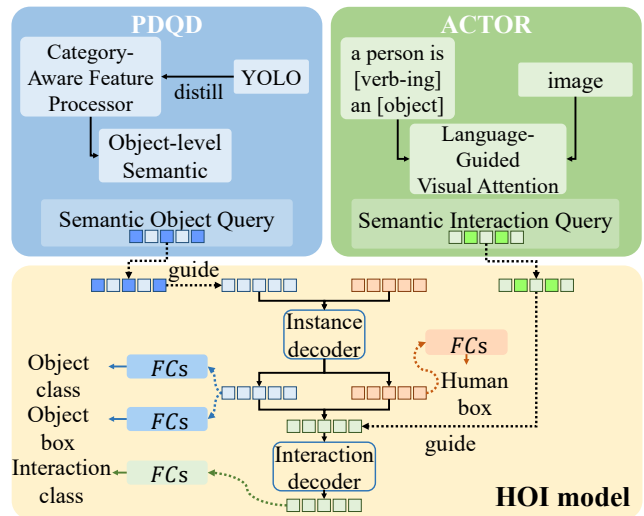


Figure 1: Overview of QueryCraft. PDQD (left) produces object-level semantic representations by distilling knowledge from YOLO. ACTOR (right) generates semantic interaction queries through language-guided visual attention using text prompts. These semantically-enriched queries guide the Instance and Interaction Decoders to predict human boxes, object boxes/classes, and interaction classes, forming the complete HOI detection model.

Early HOI detection methods follow a two-stage pipeline: detecting humans and objects via detectors like Faster R-CNN (Ren et al. 2015), then classifying interactions through feature fusion, spatial encodings, or graph-based reasoning (Chao et al. 2018; Wang et al. 2024b; Gao et al. 2021). Despite reasonable performance, these approaches suffer from computational inefficiency due to combinatorial pair explosion and limited end-to-end optimization capability (Wang et al. 2024a). The emergence of Transformer-based architectures, particularly DETR’s set prediction formulation, has revolutionized HOI detection by eliminating redundant proposals. DETR-inspired methods like GEN-VLKT (Liao et al. 2022), TED-Net (Wang, Liu, and Lei 2024), and KI2HOI (Xue et al. 2025) have advanced the field through interaction-aware queries, action embeddings,

and spatial priors, significantly improving detection performance. Despite advances in DETR-based HOI detection, key challenges remain in modeling complex interactions and semantic generalization (Yuan et al. 2023; Wang et al. 2024d). DETR’s random query initialization lacks explicit semantic grounding, limiting accurate representation of object categories and interaction semantics.

To address the limitations of randomly initialized queries in DETR-based HOI detection, we propose QueryCraft, a novel framework that incorporates semantic priors and guided feature learning through transformer-based query initialization, as shown in Figure 1. At the core of our approach is ACTOR (Action-aware Cross-modal TransFORMer), a cross-modal transformer encoder that jointly leverages semantic alignment and prior-guided mechanisms to enhance both the semantic understanding of interactions and the clarity of query representations. In addition, we introduce the Perceptual Distilled Query Decoder (PDQD), which learns object-aware semantic features by distilling object category awareness from a pre-trained detector. PDQD incorporates high-quality object detection results and uses auxiliary supervision during training to guide the decoder in learning object categories and spatial cues more effectively. During inference, this dual-branch query initialization strategy—combining ACTOR’s action-aware queries with PDQD’s object-aware queries—provides semantically grounded and structurally informative representations, enabling more accurate and stable prediction of interaction triplets. Main Contributions:

- A novel Action-aware Cross-modal TransFORMer (ACTOR) is proposed, which leverages language-guided attention to infer interaction semantics and generate semantically meaningful query representations.
- A Perceptual Distilled Query Decoder (PDQD) is introduced to address semantic ambiguity in query initialization by distilling object category awareness from pre-trained detectors.
- A dual-branch query initialization strategy is designed that combines action-aware and object-aware queries for more effective HOI detection.
- State-of-the-art performance is achieved on HICO-Det and V-COCO benchmarks, demonstrating the effectiveness of incorporating semantic priors into HOI detection.

Related Work

Two-stage HOI Approaches

Early approaches in HOI detection adopted a sequential methodology, initially employing established object detection architectures (Ren et al. 2015; Girshick 2015) to identify human and object instances, followed by a secondary phase that analyzed potential human-object combinations to determine interaction categories (Chao et al. 2018; Gao, Zou, and Huang 2018; Kim et al. 2020b; Zhang, Campbell, and Gould 2021; Xu et al. 2022). Notable examples include the HO-RCNN framework (Chao et al. 2018) and the iCAN architecture (Gao, Zou, and Huang 2018). The latter introduced an instance-focused attention mechanism using visual

features to highlight relevant regions for each detected entity. However, methods relying solely on visual features of isolated instances miss broader scene context, limiting interaction recognition. To overcome this, Wang et al. proposed an enhanced attention mechanism that integrates contextual understanding with appearance-based learning (Wang et al. 2019). Later, EfHOI (Xu et al. 2022) introduced a human-centric reasoning architecture that uses contextual relationships to improve interaction prediction accuracy.

Transformer-based HOI Detection

Unlike sequential approaches, one-stage architectures (transformer-based) enable direct image-to-triplet transformation (Chen et al. 2021; Liao et al. 2022; Peng et al. 2023; Wang, Liu, and Lei 2024; Xue et al. 2025). PPDM (Liao et al. 2020) introduced geometric centerpoints between human-object pairs as interaction anchors, improving efficiency and accuracy. Though IP-Net (Wang et al. 2020) used similar centerpoint strategies, it showed limited generalization (Zou et al. 2021). Inspired by the success of DETR (Carion et al. 2020) in object detection, recent works have explored Transformer-based architectures for end-to-end HOI detection (Zou et al. 2021; Liao et al. 2022; Xue et al. 2025), utilizing query embeddings for feature extraction. GEN-VLKT (Liao et al. 2022) addressed post-processing overhead through guided embeddings and integrated CLIP (Radford et al. 2021) for enhanced semantic understanding. For non-contact interactions, Wang et al. (Wang, Liu, and Lei 2024) developed distributed decoding to capture contextual information around instances. KI2HOI (Xue et al. 2025) improves HOI detection by integrating visual-language model knowledge, addressing the limitation of existing methods that rely on extensive manual annotations. However, existing DETR-based HOI methods use random queries to query for humans, objects, and interactions, which makes it difficult to learn query vectors for the challenging HOI task, affecting their overall performance (Wang et al. 2024a).

Method

The overall architecture of QueryCraft is shown in Figure 2. Designed as a plug-and-play module, QueryCraft can be integrated into any query-based HOI detection framework. Its core contribution lies in two complementary modules that inject semantic priors into query initialization. The Perceptual Distilled Query Decoder (PDQD) learns object-aware representations via an auxiliary multi-label classification task. These representations not only initialize instance decoder queries but are also added to decoder outputs to improve object localization and object classification. In parallel, the Action-aware Cross-modal TransFORMer (ACTOR) exploits language priors by performing cross-modal attention between visual features and textual action prompts, producing action-aware queries that enhance interaction decoder outputs and improve action classification.

Given an input image $\mathbf{X} \in \mathbb{R}^{3 \times H \times W}$, a backbone (e.g., ResNet-50 (He et al. 2016) or Swin Transformer (Liu et al. 2021)) is used to extract feature maps $\mathbf{F} \in \mathbb{R}^{C \times H' \times W'}$, where H' and W' represent the height and width of the

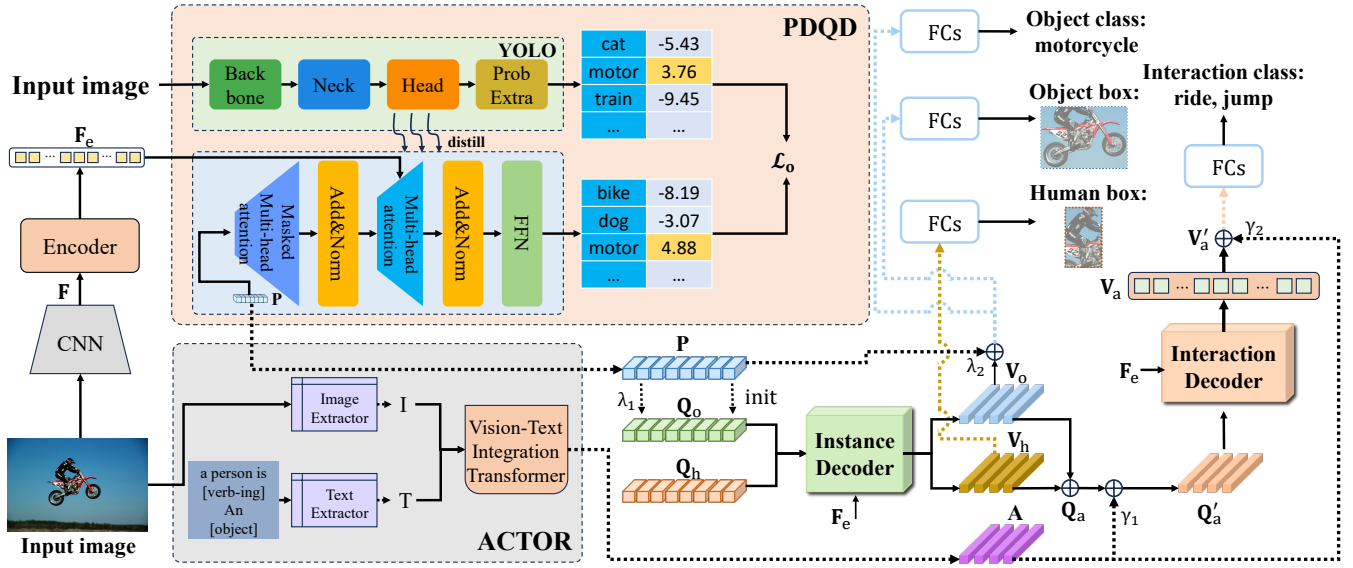


Figure 2: Detailed architecture of QueryCraft. The framework consists of three main components: (1) PDQD module (top) distills object knowledge from pre-trained YOLO to produce object queries P with semantic awareness; (2) ACTOR module (bottom left) extracts action-aware features by integrating visual and textual information through a Vision-Text Integration Transformer, generating interaction queries A ; (3) Instance and Interaction Decoders (right) process these queries to predict human boxes (V_h), object boxes/classes (V_o), and interaction classifications (V_a). The model jointly optimizes predictions through fully connected layers (FCs) to output object classes, bounding boxes, human boxes, and interaction classes.

extracted feature map, respectively. C denotes the number of feature channels. The input feature F is fed into the encoder network to extract global image features, denoted as F_e . The instance decoder takes as input the encoded features F_e and the query matrices for humans and objects, denoted as $Q_h \in \mathbb{R}^{N_q \times C'}$ and $Q_o \in \mathbb{R}^{N_q \times C'}$, respectively, where N_q is the number of queries and C' is the feature dimension. It outputs human features $V_h \in \mathbb{R}^{N_q \times C'}$ and object features $V_o \in \mathbb{R}^{N_q \times C'}$. After passing through a fully connected layers (FCS), V_h and V_o produce the human bounding boxes $B_h \in \mathbb{R}^{N_q \times 4}$, object bounding boxes $B_o \in \mathbb{R}^{N_q \times 4}$, and object classes $O_c \in \mathbb{R}^{N_q}$. Each 4-dimensional bounding box is represented as (x_1, y_1, x_2, y_2) , where (x_1, y_1) denotes the top-left corner and (x_2, y_2) denotes the bottom-right corner. Additionally, V_h and V_o are used as query matrices $Q_a \in \mathbb{R}^{N_q \times C'}$ for the action decoder, which, together with F_e , outputs action features $V_a \in \mathbb{R}^{N_q \times C'}$. V_a are passed through an FCS layer to predict the final action classes $C \in \mathbb{R}^{N_q}$. Ultimately, we obtain a set of quadruples $\mathcal{O} = \{\langle \mathbf{B}_h, \mathbf{B}_o, \mathbf{O}_c, \mathbf{C} \rangle\}$, each representing a HOI. The total number of HOI predictions is N_q . A post-processing step, such as Non-Maximum Suppression (NMS), is applied to filter out redundant interaction pairs. However, due to the use of randomly initialized query matrices, the above pipeline suffers from ambiguous query semantics, which negatively impacts the overall detection performance.

Perceptual Distilled Query Decoder (PDQD)

The PDQD module addresses the limitation of random query initialization in DETR-based HOI detectors by learn-

ing object-aware query representations through an auxiliary multi-label classification task. This module generates semantically meaningful queries that encode object category information, which subsequently enhances both object detection and classification performance.

Architecture The PDQD module includes a learnable projection tokens $P \in \mathbb{R}^{N_q \times C'}$. P is processed through a transformer decoder to interact with the encoded image features:

$$F_{obj} = \text{TransformerDecoder}(P, F_e), \quad (1)$$

where $F_e \in \mathbb{R}^{C \times H' \times W'}$ is the feature from the backbone encoder, and $F_{obj} \in \mathbb{R}^{N_q \times C'}$ represents the object feature.

To enforce the ability of object category awareness, we apply global average pooling followed by a multi-label classification head:

$$f_{pool} = \frac{1}{N_q} \sum_{i=1}^{N_q} F_{obj}^{(i)}, \quad f_{pool} \in \mathbb{R}^{C'}, \quad (2)$$

$$y^{cls} = \text{MLP}(f_{pool}), \quad y^{cls} \in \mathbb{R}^{N_{cls}}, \quad (3)$$

where $N_{cls} = 80$ is the number of object categories, and the MLP consists of two linear layers with ReLU activation: $\text{Linear}(C' \rightarrow 128) \rightarrow \text{ReLU} \rightarrow \text{Linear}(128 \rightarrow N_{cls})$.

Training Objective The PDQD module is trained to perform multi-label object classification by distilling knowledge from a pre-trained YOLO detector. For each training image, we first obtain YOLO's detection results and construct a pseudo ground-truth multi-label vector $y^L \in$

$\{0, 1\}^{N_{cls}}$, where $y_j^L = 1$ if YOLO detects at least one instance of class j with confidence score above threshold τ (we set $\tau = 0.5$ in our experiments).

The PDQD module is then trained with a multi-label binary cross-entropy loss using these YOLO-generated labels:

$$\mathcal{L}_o = -\frac{1}{N_{cls}} \sum_{j=1}^{N_{cls}} \left[y_j^L \log(\sigma(y_j^{cls})) + (1 - y_j^L) \log(1 - \sigma(y_j^{cls})) \right], \quad (4)$$

where $y_j^L \in \{0, 1\}$ indicates whether YOLO detected object class j in image i , y_j^{cls} is the logit predicted by our classification head, and $\sigma(\cdot)$ denotes the sigmoid function.

Through this design, the projection tokens \mathbf{P} learn to encode rich object category information from the training data, effectively distilling perceptual knowledge that benefits both the initialization and final prediction stages of HOI detection.

Query Enhancement The learned projection tokens \mathbf{P} serve a dual purpose in our framework. First, they initialize the object queries in the instance decoder:

$$\mathbf{Q}'_o = \mathbf{Q}_o + \lambda_1 \cdot \mathbf{P}, \quad (5)$$

where \mathbf{Q}_o represents the original random initialization, and λ_1 is a weighting factor. We specifically enhance \mathbf{Q}_o rather than \mathbf{Q}_h because \mathbf{Q}_h is solely responsible for human localization, while \mathbf{Q}_o handles both localization and recognition across all object categories. Through knowledge distillation, \mathbf{P} encapsulates rich object-centric information from the pre-trained detector, making it particularly suitable for initializing \mathbf{Q}_o to improve multi-class object detection performance.

Second, after the instance decoder produces its output features $\mathbf{V}_o \in \mathbb{R}^{N_q \times C'}$, we further enhance them with the learned object representations:

$$\mathbf{V}'_o = \mathbf{V}_o + \lambda_2 \cdot \mathbf{P}. \quad (6)$$

The \mathbf{V}'_o is then used for predicting object bounding boxes, and object categories.

Action-aware Cross-modal Transformer (ACTOR)

While PDQD provides object-aware initialization, accurate HOI detection also requires understanding the semantic space of possible interactions. The ACTOR module addresses this by leveraging language priors to generate action-aware query representations through cross-modal attention between visual and textual features.

Motivation and Design HOI are inherently tied to semantic concepts expressed in natural language. For instance, the visual pattern of "riding a bicycle" corresponds to specific spatial relationships that can be described textually. ACTOR exploits this correspondence by using textual descriptions of all possible interaction categories as a semantic anchor space to guide query initialization.

Given a set of N_T interaction categories (e.g., ⟨ride, bicycle⟩, ⟨eat, banana⟩, and so on.), we construct textual prompts for each action using templates \mathcal{T} . These prompts

are encoded using a pre-trained text extractor to obtain textual embeddings $\mathbf{T} \in \mathbb{R}^{N_T \times d_t}$:

$$\mathbf{T} = \text{TextExtractor}(\mathcal{T}(v_c, o_c)), \quad (7)$$

where \mathcal{T} represents text template such as "a person [action] a [object]", v_c denotes action, and o_c denotes object. In addition, we utilize an image extractor to extract high-level visual features $\mathbf{I} \in \mathbb{R}^{1 \times d_i}$.

We hypothesize that the visual manifestation of HOI can be effectively characterized through their linguistic descriptions. Formally, let \mathbf{I} denote the visual feature space and \mathbf{T} the linguistic embedding space. We seek to learn a mapping $\phi : (\mathbf{I}, \mathbf{T}) \rightarrow \mathbf{A}$ where \mathbf{A} represents a shared semantic space that captures the underlying interaction patterns. Our key insight is that pre-trained vision-language models (e.g., CLIP (Radford et al. 2021) or BLIP (Li et al. 2022, 2023)) have already learned rich alignments between \mathbf{I} and \mathbf{T} . ACTOR exploits this alignment by constructing a semantic dictionary of interaction prototypes in the linguistic space, which serves as anchors for grounding visual queries.

Cross-modal Semantic Alignment ACTOR employs a novel asymmetric cross-attention mechanism where visual features query the semantic dictionary to discover relevant interaction patterns. \mathbf{I} is expanded into N_q query seeds:

$$\mathbf{Q}^{(0)} = \mathbf{R}^\top \mathbf{I}, \quad (8)$$

where $\mathbf{I} \in \mathbb{R}^{1 \times N}$, $\mathbf{R} \in \mathbb{R}^{1 \times N_q}$ (with $r_i = 1, \forall i, r_i \in \mathbf{R}$), and $\mathbf{Q}^{(0)} \in \mathbb{R}^{N_q \times N}$. The cross-modal refinement then processes these features through L layers (we use $L = 3$), where each layer $l \in \{1, \dots, L\}$ consists of:

$$\mathbf{Q}^{(l+1)} = \mathcal{F}_{\text{ACTOR}}(\mathbf{Q}^{(l)}, \mathbf{T}), \quad (9)$$

where $\mathcal{F}_{\text{ACTOR}}$ implements a specialized transformer block that performs:

$$\begin{aligned} \mathbf{S}^{(l)} &= \mathcal{S} \left(\frac{\mathbf{Q}^{(l)} \mathbf{W}_Q (\mathbf{T} \mathbf{W}_K)^\top}{\sqrt{d_t}} \right), \\ \tilde{\mathbf{Q}}^{(l)} &= \mathbf{S}^{(l)} (\mathbf{T} \mathbf{W}_V), \\ \mathbf{Q}^{(l+1)} &= \text{FFN}(\text{LN}(\mathbf{Q}^{(l)} + \tilde{\mathbf{Q}}^{(l)})). \end{aligned} \quad (10)$$

The attention weights $\mathbf{S}^{(l)}$ can be interpreted as a soft assignment of visual queries to semantic interaction prototypes, effectively performing implicit reasoning about which interactions are plausible given the visual context.

Action Query Generation The output of ACTOR, $\mathbf{A} = \mathbf{Q}^{(L)} \in \mathbb{R}^{N_q \times C'}$, represents action-aware queries that encode interaction semantics. These queries capture which actions are likely to occur in the image by attending to relevant textual descriptions. Similar to PDQD, these action-aware representations serve dual purposes:

$$\mathbf{Q}'_a = \mathbf{Q}_a + \gamma_1 \cdot \mathbf{A}. \quad (11)$$

Additionally, after the interaction decoder processes these initialized queries, we apply residual enhancement:

$$\mathbf{V}'_a = \mathbf{V}_a + \gamma_2 \cdot \mathbf{A}, \quad (12)$$

Method	Backbone	Detector	mAP \uparrow		
			Full	Rare	Non-Rare
HO-RCNN (Chao et al. 2018)	CaffeNet	Fast R-CNN	7.81	5.37	8.54
InteractNet (Gkioxari et al. 2018)	ResNet-50	Faster R-CNN	9.94	7.16	10.77
iCAN (Gao, Zou, and Huang 2018)	ResNet-50	Faster R-CNN	14.84	10.45	16.15
UnionDet (Kim et al. 2020a)	ResNet-50-FPN	RetinaNet	17.58	11.72	19.33
IP-Net (Wang et al. 2020)	Hourglass-104	FPN	19.56	12.79	21.58
VSGNet (Ulutun, Iftekhar, and Manjunath 2020)	ResNet-152	Faster R-CNN	19.80	16.05	20.91
FCMNet (Liu, Chen, and Zisserman 2020)	ResNet-50	Faster R-CNN	20.41	17.34	21.56
ACP (Kim et al. 2020b)	Res-DCN-152	Faster R-CNN	20.59	15.92	21.98
PD-Net (Zhong et al. 2020)	ResNet-152-FPN	Faster R-CNN	20.81	15.90	22.28
SG2HOI (He et al. 2021)	ResNet-50	Faster R-CNN	20.93	18.24	21.78
DJ-RN (Li et al. 2020)	ResNet-50	Faster R-CNN	21.34	18.53	22.18
SCG (Zhang, Campbell, and Gould 2021)	ResNet-50	Faster R-CNN	21.85	18.11	22.97
HOTR (Kim et al. 2021)	ResNet-50	DETR	25.10	17.34	27.42
MSTR (Kim et al. 2022)	ResNet-50	DETR	31.17	25.31	33.92
GEN-VLKT (Liao et al. 2022)	ResNet-50	DETR	33.51	28.52	35.00
RLIPv2 (Yuan et al. 2023)	ResNet-50	DETR	35.38	29.61	37.10
RLIPv2 (Yuan et al. 2023)	Swin-L	DETR	45.09	43.23	45.64
TED-Net (Wang, Liu, and Lei 2024)	ResNet-50	DETR	34.00	29.88	35.24
LOGICHOI (Li et al. 2024)	ResNet-50	DETR	35.47	32.03	36.22
KI2HOI (Xue et al. 2025)	ResNet-50	DETR	34.20	32.26	36.10
GEN-VLKT + Ours	ResNet-50	DETR	34.63	30.24	35.95
			+1.12	+1.72	+0.95
RLIPv2 + Ours	ResNet-50	DETR	36.42	30.84	38.01
			+1.04	+1.23	+0.91
RLIPv2 + Ours	Swin-L	DETR	46.01	44.75	46.42
			+0.92	+1.52	+0.78
TED-Net + Ours	ResNet-50	DETR	35.17	31.29	36.16
			+1.17	+1.41	+0.92
LOGICHOI + Ours	ResNet-50	DETR	36.48	33.24	37.16
			+1.01	+1.21	+0.94
KI2HOI + Ours	ResNet-50	DETR	35.27	33.69	36.99
			+1.07	+1.43	+0.89

Table 1: Performance comparisons on HICO-Det Dataset.

where Q'_a initializes the interaction decoder queries.

The cross-modal semantic alignment in ACTOR effectively creates a soft retrieval process over the action vocabulary. By learning to attend to relevant action descriptions based on visual content, the model develops a semantic understanding of potential interactions. Through this design, ACTOR provides action-specific semantic query initialization, which, in conjunction with PDQD, forms a comprehensive semantic prior, enabling effective HOI detection.

Experiments

Datasets

We evaluate QueryCraft on two HOI detection benchmarks: HICO-DET and V-COCO. HICO-DET contains 47,776 images (38,118 training, 9,658 testing) with 600 HOI categories formed by 80 objects and 117 action verbs. Performance is measured by mAP under Full (all 600 categories), Rare (138 categories with fewer than 10 training instances) settings, and Non-Rare (the remaining categories). V-COCO comprises 10,346 images with 80 interaction categories cov-

ering 29 action types. HICO-DET tests scalability with its large vocabulary and long-tail distribution, while V-COCO evaluates fine-grained action understanding through role-based annotations.

Effectiveness for Regular HOI Detection

Results on HICO-DET Table 1 presents our results on HICO-DET. QueryCraft demonstrates consistent improvements across all baseline methods, with gains ranging from +0.92 to +1.17 mAP on the Full set. The improvements are most pronounced on the Rare subset, averaging +1.42 mAP across all methods. For instance, GEN-VLKT achieves +1.72 mAP on Rare interactions, while RLIPv2 with Swin-L backbone still gains +1.52 despite its strong baseline (43.23 mAP). These substantial improvements on rare categories validate that our semantic initialization provides crucial inductive biases for limited-data scenarios. The consistent gains across diverse architectures confirm that QueryCraft addresses a fundamental limitation in DETR-based HOI detection rather than being method-specific.

Method	Backbone	Detector	mAP \uparrow
GEN-VLKT	ResNet50	DETR	62.4
RLIPv2	ResNet50	DETR	65.9
RLIPv2	Swin-L	DETR	72.1
TED-Net	ResNet50	DETR	63.4
LOGICHOI	ResNet50	DETR	64.4
KI2HOI	ResNet50	DETR	63.9
GEN-VLKT + Ours	ResNet50	DETR	63.7 (+1.3)
RLIPv2 + Ours	ResNet50	DETR	67.3 (+1.4)
RLIPv2 + Ours	Swin-L	DETR	73.1 (+1.0)
TED-Net + Ours	ResNet50	DETR	65.1 (+1.7)
LOGICHOI + Ours	ResNet50	DETR	65.6 (+1.2)
K2HOI + Ours	ResNet50	DETR	65.0 (+1.1)

Table 2: Performance comparisons on V-COCO Dataset.

Method	Source	mAP \uparrow		
		Full	Unseen	Seen
GEN-VLKT	UV	28.74	20.96	30.23
RLIPv2	UV	-	-	-
KI2HOI	UV	31.85	25.20	32.95
GEN-VLKT + Ours	UV	30.08 (+1.34)	23.06 (+2.10)	31.37 (+1.14)
KI2HOI + Ours	UV	33.01 (+1.16)	26.76 (+1.56)	34.32 (+1.37)
GEN-VLKT	UO	25.63	10.51	28.92
RLIPv2	UO	-	-	-
KI2HOI	UO	28.84	16.50	31.70
GEN-VLKT + Ours	UO	27.21 (+1.58)	12.96 (+2.45)	30.65 (+1.73)
KI2HOI + Ours	UO	30.11 (+1.27)	18.37 (+1.87)	32.94 (+1.24)
GEN-VLKT	NF-UC	23.71	25.05	23.38
RLIPv2	NF-UC	36.94	22.65	40.51
KI2HOI	NF-UC	27.77	28.89	28.31
GEN-VLKT + Ours	NF-UC	24.69 (+0.98)	26.25 (+1.20)	24.27 (+0.89)
RLIPv2 + Ours	NF-UC	38.07 (+1.13)	24.73 (+2.08)	41.57 (+1.06)
KI2HOI + Ours	NF-UC	28.92 (+1.15)	30.13 (+1.24)	29.42 (+1.11)
GEN-VLKT	RF-UC	30.56	21.36	32.91
RLIPv2	RF-UC	42.26	31.23	45.01
KI2HOI	RF-UC	34.10	26.33	35.79
GEN-VLKT + Ours	RF-UC	31.61 (+1.05)	22.89 (+1.53)	33.84 (+0.93)
RLIPv2 + Ours	RF-UC	43.48 (+1.22)	33.02 (+1.79)	45.94 (+0.93)
KI2HOI + Ours	RF-UC	35.32 (+1.22)	27.85 (+1.52)	36.76 (+0.97)

* RLIPv2 uses Swin-L as backbone. GEN-VLKT and KI2HOI use ResNet-50 as backbone.

Table 3: Performance comparison for zero-shot HOI detection on HICO-Det.

Results on V-COCO Table 2 shows our evaluation on V-COCO dataset, where QueryCraft achieves improvements ranging from +1.0 mAP to +1.7 mAP points. TED-Net shows the largest gain (+1.7 mAP), and other methods demonstrate consistent improvements around +1.1 mAP to +1.4 mAP. Notably, even RLIPv2 with Swin-L backbone, which already achieves 72.1 mAP, benefits from our approach (+1.0 mAP). The effectiveness on V-COCO’s fine-grained action recognition task, which requires precise object localization and role understanding, further validates that both PDQD’s object-aware and ACTOR’s action-semantic initializations contribute meaningfully to interaction detection across different evaluation protocols.

Zero-Shot Generalization Table 3 presents zero-shot experiments across four protocols on HICO-DET. QueryCraft consistently enhances generalization capabilities across all settings. In Unseen Verb (UV), our method achieves substantial gains, particularly on unseen interactions (GEN-VLKT: +2.10, KI2HOI: +1.56 mAP), demonstrating ACTOR’s effectiveness in reasoning about novel actions through language-guided attention. For Unseen Object (UO), improvements are even more pronounced, with GEN-VLKT gaining +2.45 mAP on unseen, showcasing PDQD’s ability to transfer object knowledge to unseen categories. In compositional settings (NF-UC and RF-UC), QueryCraft maintains robust improvements, with RLIPv2 achieving +2.08 and +1.79 mAP on unseen subsets respectively. PDQD and ACTOR work synergistically—PDQD enables novel object recognition through distilled detection knowledge, while ACTOR facilitates understanding of new actions via cross-modal reasoning. This dual semantic grounding proves particularly valuable in zero-shot scenarios where random initialization lacks meaningful priors, fundamentally enhancing the model’s ability to generalize beyond its training distribution.

Intransitive Interaction Detection Table 4 presents results on intransitive (non-contact) interactions using HICO-DET-IC and V-COCO-IC benchmarks. On HICO-DET-IC, QueryCraft achieves substantial improvements, with GEN-VLKT gaining +1.84 mAP on Rare interactions and TED-Net showing consistent gains across all subsets (+1.33 mAP Full). Both methods achieve +1.22 mAP improvement on V-COCO-IC. The strong performance on intransitive interactions, which require understanding spatial relationships rather than physical contact, demonstrates the versatility of our approach. ACTOR’s language-guided initialization proves particularly valuable as intransitive actions (e.g., “watching”, “flying”) are better characterized linguistically than visually. Combined with PDQD’s accurate object localization for spatial reasoning, QueryCraft enables effective detection of both contact-based and abstract relationships, showcasing its applicability to comprehensive real-world scene understanding.

Training Efficiency Analysis Table 5 demonstrates that QueryCraft significantly accelerates training convergence. All evaluated methods show consistent epoch reductions: TED-Net (-18.6%), LOGICHOI (-13.9%), KI2HOI (-13.3%), and GEN-VLKT (-12.6%). This acceleration stems from improved initialization landscapes—while random initialization requires models to simultaneously learn query semantics and detection patterns, our semantic initialization provides meaningful starting points through PDQD’s object-awareness and ACTOR’s action-relevant information. This reduces the hypothesis space during training, leading to faster convergence and more stable optimization.

Ablation Studies

We conduct comprehensive ablation studies on HICO-DET and V-COCO datasets. All ablation experiments are performed using GEN-VLKT as the baseline method with ResNet-50 backbone.

Method	HICO-Det-IC			V-COCO-IC
	Full	Rare	Non-Rare	
GEN-VLKT	27.25	23.73	28.03	33.59
TED-Net	30.09	30.09	30.24	38.71
GEN-VLKT + Ours	28.45	25.57	29.13	34.81
TED-Net + Ours	31.42	31.22	31.36	39.93

Table 4: Performance comparisons on HICO-Det-IC and V-COCO-IC.

ACTOR	PDQD	HICO-Det			V-COCO
		Full	Rare	Non-Rare	
○	○	33.51	28.52	35.00	62.4
○	●	33.83	29.10	35.13	62.8
●	○	34.24	29.91	35.58	63.1
●	●	34.63	30.24	35.95	63.7

Table 6: Performance of different components.

γ_1	γ_2	HICO-Det			V-COCO
		Full	Rare	Non-Rare	
1	1	34.63	30.24	35.95	63.7
1	0.1	34.35	29.87	35.74	63.5
0.1	1	33.91	29.42	35.32	63.1
0.1	0.1	33.85	29.34	35.27	62.8

Table 8: Ablation study using different γ_1 and γ_2 .

Component Analysis Table 6 presents ablation results for ACTOR and PDQD modules. From the baseline of 33.51 mAP (GEN-VLKT without either module), enabling PDQD alone yields +0.32 mAP improvement, while ACTOR alone provides a larger +0.73 mAP gain, with particularly strong improvements on Rare interactions (+1.39 mAP). This suggests action semantics represent a primary bottleneck in HOI detection. Crucially, combining both modules achieves 34.63 mAP (+1.12 total), exceeding the sum of individual gains (1.05), indicating positive synergy.

Impact of Weighting Parameters Tables 7 and 8 analyze the weighting parameters for PDQD (λ_1, λ_2) and ACTOR (γ_1, γ_2). Both modules achieve optimal performance with all parameters set to 1.0. For PDQD, reducing λ_1 (initialization) to 0.1 causes a -0.43 mAP drop, while reducing λ_2 results in -0.25 mAP, indicating initialization is more critical. ACTOR shows even stronger sensitivity: reducing γ_1 to 0.1 yields a substantial -0.72 mAP decrease, nearly double PDQD’s impact. The Rare subset is particularly sensitive to ACTOR parameters (30.24→29.34 mAP when both are 0.1), confirming that action semantics are crucial for rare HOI.

Impact of text Templates in ACTOR Table 9 evaluates four template variations for generating action descriptions in ACTOR. $T_1, T_2, T_3,$ and T_4 correspond to the templates “person {1} {2}”, “someone {1} a/an {2} outdoors or indoors”, “a person is {1}ing a/an {2}”, and “person interacting with a/an {2} by {1}ing”, respectively. Overall, the im-

Method	Epochs	Epochs	Reduction
GEN-VLKT	87	76	↓12.6%
TED-Net	97	79	↓18.6%
LOGICHOI	79	68	↓13.9%
KI2HOI	83	72	↓13.3%

Table 5: Convergence epochs comparison. Best epoch numbers are obtained from our reproduction using official code implementations.

λ_1	λ_2	HICO-Det			V-COCO
		Full	Rare	Non-Rare	
1	1	34.63	30.24	35.95	63.7
1	0.1	34.38	29.93	35.82	63.4
0.1	1	34.20	29.92	35.49	63.2
0.1	0.1	34.21	29.84	35.58	63.1

Table 7: Ablation study using different λ_1 and λ_2 .

Templates	HICO-Det			V-COCO
	Full	Rare	Non-Rare	
T_1	34.61	30.14	35.75	63.7
T_2	34.57	30.03	35.83	63.6
T_3	34.63	30.24	35.95	63.7
T_4	34.37	29.89	35.68	63.4

Table 9: Performance of different text templates in ACTOR.

pact of using different language prompts on model performance is negligible. The progressive template T_3 achieves optimal performance (34.63 mAP), while the minimal template T_1 performs nearly identically (-0.02 mAP), demonstrating robustness to simplicity. Adding spatial context (T_2) actually decreases performance (-0.06 mAP), suggesting that unnecessary complexity dilutes essential action-object relationships. The small performance variations across different templates confirm ACTOR’s robustness to linguistic variations, attributed to the vision-language model’s ability to map diverse expressions to similar representations.

Conclusion

In this paper, we presented QueryCraft, a novel framework that addressed the fundamental limitation of random query initialization in DETR-based HOI detection methods. Through comprehensive semantic query initialization, we demonstrated that providing meaningful priors significantly improves both detection performance and training efficiency. QueryCraft introduces two modules: PDQD leverages knowledge distillation from pre-trained object detectors to generate object-aware queries, and ACTOR exploits vision-language alignment to create action-semantic queries through cross-modal attention. These modules work synergistically, with PDQD providing robust object understanding and ACTOR enabling compositional reasoning about interactions. Extensive experiments validate the effectiveness of QueryCraft across multiple dimensions.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62202174, in part by the GJYC program of Guangzhou under Grant 2024D01J0081, and in part by the ZJ program of Guangdong under Grant 2023QN10X455, and in part by the Fundamental Research Funds for the Central Universities under Grant 2025ZYGXZR053.

References

- Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 213–229. Springer.
- Chao, Y.-W.; Liu, Y.; Liu, X.; Zeng, H.; and Deng, J. 2018. Learning to detect human-object interactions. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 381–389.
- Chen, M.; Liao, Y.; Liu, S.; Chen, Z.; Wang, F.; and Qian, C. 2021. Reformulating HOI detection as adaptive set prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9004–9013.
- Gao, C.; Zou, Y.; and Huang, J.-B. 2018. ICAN: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*.
- Gao, Y.; Kuang, Z.; Li, G.; Zhang, W.; and Lin, L. 2021. Hierarchical Reasoning Network for Human-Object Interaction Detection. *IEEE Transactions on Image Processing*, 30: 8306–8317.
- Girshick, R. 2015. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448.
- Gkioxari, G.; Girshick, R.; Dollár, P.; and He, K. 2018. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8359–8367.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
- He, T.; Gao, L.; Song, J.; and Li, Y.-F. 2021. Exploiting scene graphs for human-object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15984–15993.
- Kim, B.; Choi, T.; Kang, J.; and Kim, H. J. 2020a. Union-Det: Union-level detector towards real-time human-object interaction detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 498–514. Springer.
- Kim, B.; Lee, J.; Kang, J.; Kim, E.-S.; and Kim, H. J. 2021. HOTR: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 74–83.
- Kim, B.; Mun, J.; On, K.-W.; Shin, M.; Lee, J.; and Kim, E.-S. 2022. Mstr: Multi-scale transformer for end-to-end human-object interaction detection. In *International Conference on Learning Representations*, 19578–19587.
- Kim, D.-J.; Sun, X.; Choi, J.; Lin, S.; and Kweon, I. S. 2020b. Detecting human-object interactions with action co-occurrence priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 718–736. Springer.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.
- Li, L.; Wei, J.; Wang, W.; and Yang, Y. 2024. Neural-logic human-object interaction detection. *Advances in Neural Information Processing Systems*, 36.
- Li, Y.-L.; Liu, X.; Lu, H.; Wang, S.; Liu, J.; Li, J.; and Lu, C. 2020. Detailed 2d-3d joint representation for human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10166–10175.
- Liao, Y.; Liu, S.; Wang, F.; Chen, Y.; Qian, C.; and Feng, J. 2020. PPDm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 482–490.
- Liao, Y.; Zhang, A.; Lu, M.; Wang, Y.; Li, X.; and Liu, S. 2022. GEN-VLKT: Simplify Association and Enhance Interaction Understanding for HOI Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20123–20132.
- Liu, Y.; Chen, Q.; and Zisserman, A. 2020. Amplifying key cues for human-object-interaction detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 248–265. Springer.
- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.
- Peng, H.; Liu, F.; Li, Y.; Huang, B.; Shao, J.; Sang, N.; and Gao, C. 2023. Parallel Reasoning Network for Human-Object Interaction Detection. *arXiv preprint arXiv:2301.03510*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.
- Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28.
- Ulutun, O.; Iftekhhar, A.; and Manjunath, B. S. 2020. VS-GNet: Spatial attention network for detecting human object interactions using graph convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13617–13626.

- Wang, T.; Anwer, R. M.; Khan, M. H.; Khan, F. S.; Pang, Y.; Shao, L.; and Laaksonen, J. 2019. Deep contextual attention for human-object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5694–5702.
- Wang, T.; Yang, T.; Danelljan, M.; Khan, F. S.; Zhang, X.; and Sun, J. 2020. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4116–4125.
- Wang, Y.; Lei, Y.; Cui, L.; Xue, W.; Liu, Q.; and Wei, Z. 2024a. A review of human-object interaction detection. In *2024 2nd International Conference on Computer, Vision and Intelligent Technology (ICCVIT)*, 1–6. IEEE.
- Wang, Y.; Lei, Y.; Wei, Z.; Xue, W.; Jiang, X.; Zhuang, N.; and Liu, Q. 2025. Prompt Guidance and Human Proximal Perception for HOT Prediction with Regional Joint Loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Wang, Y.; Lei, Y.; Xiong, Q.; Xue, W.; Liu, Q.; and Wei, Z. 2024b. DeHOT: Reconstructing Pseudo-3D Scenes for Human-Object Contact Detection. In *2024 5th International Conference on Computer, Big Data and Artificial Intelligence (ICCBD+ AI)*, 497–500. IEEE.
- Wang, Y.; Liu, Q.; and Lei, Y. 2024. Ted-net: Dispersal attention for perceiving interaction region in indirectly-contact hoi detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7): 5603–5615.
- Wang, Y.; Neng, W.; Wei, Z.; Lei, Y.; Xue, W.; Zhuang, N.; Xu, Y.; Jiang, X.; and Liu, Q. 2024c. Precision-Enhanced Human-Object Contact Detection via Depth-Aware Perspective Interaction and Object Texture Restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Wang, Y.; Wei, Z.; Jiang, X.; Lei, Y.; Xue, W.; Liu, J.; and Liu, Q. 2024d. FreeA: Human-object Interaction Detection using Free Annotation Labels. *arXiv preprint arXiv:2403.01840*.
- Xu, K.; Li, Z.; Zhang, Z.; Dong, L.; Xu, W.; Yan, L.; Zhong, S.; and Zou, X. 2022. Effective actor-centric human-object interaction detection. *Image and Vision Computing*, 121: 104422.
- Xue, W.; Liu, Q.; Wang, Y.; Wei, Z.; Xing, X.; and Xu, X. 2025. Towards zero-shot human-object interaction detection via vision-language integration. *Neural Networks*, 187: 107348.
- Yuan, H.; Zhang, S.; Wang, X.; Albanie, S.; Pan, Y.; Feng, T.; Jiang, J.; Ni, D.; Zhang, Y.; and Zhao, D. 2023. Rlipv2: Fast scaling of relational language-image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21649–21661.
- Zhang, F. Z.; Campbell, D.; and Gould, S. 2021. Spatially conditioned graphs for detecting human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13319–13327.
- Zhong, X.; Ding, C.; Qu, X.; and Tao, D. 2020. Polysemy deciphering network for human-object interaction detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 69–85. Springer.
- Zou, C.; Wang, B.; Hu, Y.; Liu, J.; Wu, Q.; Zhao, Y.; Li, B.; Zhang, C.; Zhang, C.; Wei, Y.; et al. 2021. End-to-end human object interaction detection with HOI transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11825–11834.