Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Full Length Article

Towards zero-shot human–object interaction detection via vision–language integration

Weiying Xue, Qi Liu *, Yuxiao Wang, Zhenao Wei, Xiaofen Xing, Xiangmin Xu

School of Future Technology, South China University of Technology, Guangdong Guangzhou, 511400, PR China

ARTICLE INFO

Keywords:
Human–object interaction
Multimodal integration
Zero-shot
Weakly supervision

ABSTRACT

Human–object interaction (HOI) detection aims to locate human–object pairs and identify their interaction categories in images. Most existing methods primarily focus on supervised learning, which relies on extensive manual HOI annotations. Such heavy reliance on closed-set supervised learning limits their generalization capabilities to unseen object categories. Inspired by the remarkable zero-shot capabilities of VLM, we propose a novel framework, termed Knowledge Integration to HOI (KI2HOI), that effectively integrates the knowledge of the visual–language model to improve zero-shot HOI detection. Specifically, we propose a ho-pair encoder to supplement contextual and interaction-specific semantic representation decoder into our model. Additionally, we propose two fusion strategies to facilitate prior knowledge transfer of VLM. One is visual-level fusion, producing more global context interaction features; another is language-level fusion, further enhancing the capability of VLM for HOI detection. Extensive experiments conducted on the mainstream HICO-DET and V-COCO datasets demonstrate that our model outperforms the previous methods in various zero-shot and full-supervised settings. The source code is available in https://github.com/xwyscut/K2HOI.

1. Introduction

Human–object interaction (HOI) detection is a process of detecting interaction between a human and an object in an image (Qin, Gu, & Tan, 2022). Precisely estimating human–object interactions can greatly improve various visual understanding tasks, such as image retrieval (Ji et al., 2024; Qin et al., 2022), visual question answering (Kim, Lee, Wu, Jung & Lee, 2021), and scene graph generation (Fu et al., 2023; Liu & Liu, 2024). Given a series of ⟨"Human", "Object", "Verb"⟩ triples, an HOI detector is needed to locate human–object pairs and identify their interactions. However, most HOI detectors typically require a significant number of pre-defined HOI categories. Considering the diversity and complexity of human–object interaction in the real world, it is time-consuming and laborious to define all-natural interaction annotations in advance manually.

In recent years, HOI detectors (Antoun & Asmar, 2022; Hou, Yu, Qiao, Peng, & Tao, 2021b; Jia & Ma, 2023; Kim, Lee, Kang, Kim and Kim, 2021) have led to enhanced performance by successively introducing the transformer (Carion et al., 2020), as shown in Fig. 1. Nevertheless, the query is initialized directly using the visual features derived from object detection in most transformer-based HOI detectors. After that, this inadequate query is input into the decoder for the decoding of features, which leads to a shortage of diverse and characteristic interaction features. The development of pre-trained visual–language models on large-scale data has significantly accelerated the progress in studying zero-shot learning (Du et al., 2022; Kim, Angelova and Kuo, 2023), especially CLIP (Radford et al., 2021) demonstrates remarkable transfer ability in various subsequent tasks. The most recent zero-shot HOI detectors leverage the comprehensive visual and linguistic knowledge of CLIP to detect novel HOIs. For example, the knowledge of pre-trained visual language model is transferred to EOID (Wu et al., 2023) and DOQ (Qu, Ding, Li, Zhong, & Tao, 2022) via knowledge distillation to achieve zero-shot HOI detection. Previous approaches employ the semantic word embeddings of the HOI labels to create a semantic space and then align with visual space, failing to fully capitalize on the potential of cross-modal information and Language Models in the field of Human–Object Interaction (HOI) detection. Knowledge distillation depends on the quality of the teacher model and the used training data. When the training process does not include unknown categories, the distillation process may be biased towards known category samples, thereby the generalization ability is limited.

To address these challenges, we propose a novel one-stage zero-shot framework for Human–Object Interaction detection, named Knowledge Integration to HOI (KI2HOI). We harness the capabilities of foundational models to enhance the comprehension of complex interactive

* Corresponding author.
E-mail addresses: 202320163283@mail.scut.edu.cn (W. Xue), drliuqi@scut.edu.cn (Q. Liu), ftwangyuxiao@mail.scut.edu.cn (Y. Wang), wza@scut.edu.cn (Z. Wei), xfxing@scut.edu.cn (X. Xing), xmxu@scut.edu.cn (X. Xu).

https://doi.org/10.1016/j.neunet.2025.107348
Received 23 October 2024; Received in revised form 22 January 2025; Accepted 3 March 2025
Available online 10 March 2025

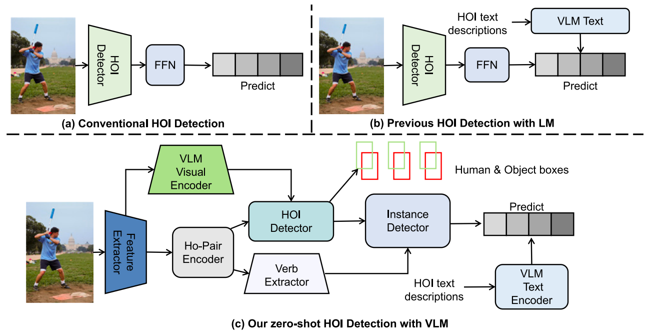

Fig. 1. Comparison of HOI detection. Conventional HOI detection required manually annotated datasets for training. Previous HOI detection with language model (LM) employed limited knowledge distillation to visual detectors, but it is limited to handling potential interactions among unseen human–object pairs. Our model fully leverages visual–language model (VLM) and verb queries for effective knowledge integration, to promote unseen interaction recognition.

semantics within visual data. Instead of using knowledge distillation on CLIP for detectors, KI2HOI learns *a priori* knowledge from VLM and integrates visual–linguistic knowledge in HOI detection. Specifically, we propose two fusion strategies to facilitate prior knowledge transfer of VLM. One is the visual level, we first obtain the instances generated by the off-the-shelf object detector, then put them into the ho-pair encoder to extract rich local and global visual features; another is the language level, motivated by the Query2label (Liu, Zhang, Yang, Su, & Zhu, 2021), a transformer-based model utilized for multi-label image classification, we first extract verb representations through verb feature learning and then combine them with vision–language knowledge through interaction representation decoder. While traditional interaction classification heads work well in handling visible interactions, they fall short in tackling invisible ones. Thus, to leverage the zero-shot recognition capability of the pre-trained VLM, we first use the text encoder of VLM to obtain text embeddings of the HOI labels. These embeddings then serve as a train-free classifier to perform extra interaction classification on the vision–language knowledge-enhanced instance queries. We combine the prompt-based verb classification head with the redesigned verb representation head to enhance the HOI prediction.

To summarize, our contributions are as follows:

- We propose a novel framework, named KI2HOI, for zero-shot HOI detection that directly retrieves visual and linguistic knowledge of VLM. Our KI2HOI effectively utilizes *a priori* knowledge and achieves superior zero-shot transferability.
- We develop visual and linguistic level strategies to fuse spatial information and semantic information for generating more expressive and intricate representations between verbs and their associated interactions.
- We conduct extensive experiments on HICO-DET for the zero-shot learning task and perform additional comprehensive experiments on HICO-DET and V-COCO for the supervised learning task. Our model outperforms the state-of-the-art methods in zero-shot and full-supervised settings, establishing a new state-of-the-art.

2. Related works

2.1. Human–object interaction detection

HOI detection methods can be roughly categorized as two-stage or one-stage solutions. Two-stage methods (Cao et al., 2023; Cheng, Wang, Zhan, & Duan, 2023; Li, Zou, Zhao, Li and Zhong, 2022; Liu, Chen & Zisserman, 2020; Wan, Liu, Zhou, Tuytelaars, & He, 2023; Wang, Yu, & Zhang, 2024) first detect all candidate interaction pairs and then feed to CNN network to predict the interaction relationships between candidate human–object pairs. Most two-stage detection models use existing object detection models and prioritize improving interaction prediction models. FCL (Hou et al., 2021b) proposed an object fabricator to generate effective object representations, which were then combined with verbs to compose new HOI samples, thus increasing the diversity of training data. UPT (Zhang, Campbell, & Gould, 2022) applied a unary-pairwise transformer to represent each target's instance details as unary and pairwise representations. In comparison to two-stage methods, one-stage solutions (Kim, Jung & Cho, 2023; Kim, Lee, Kang et al., 2021; Liao et al., 2022; Liu et al., 2022; Wu et al., 2023; Yang, Zou, Zhang, Cao, & Chen, 2022; Yuan et al., 2022; Zhou et al., 2022) captured context information during the early stage of feature extraction, leading to improved HOI detection performance. The success of DETR (Carion et al., 2020) has inspired many researchers in studying HOI detection QPIC (Tamura, Ohashi, & Yoshinaga, 2021) applied additional detection heads and relied on a bipartite graph matching algorithm to locate HOI instances and identify interactions. EOID (Wu et al., 2023) developed a teacher-student model and designed a two-stage Hungarian matching algorithm. RR-Net (Yang et al., 2022) introduced a relation-aware frame to build progressive structure for interaction inference, which imitates the human visual mechanism of recognizing HOI by comprehending visual instances and interactions coherently. HOICLIP (Ning, Qiu, Liu, & He, 2023) proposed a new transfer strategy that used visual semantic algorithms to represent verbs. Our work belongs to a one-stage end-to-end approach to study HOI detection.

2.2. Vision-and-language pre-training

The advanced vision-and-language pre-training (VLP) multimodal learning framework can acquire generalized multimodal representations from large-scale image and text data (Li, Zhang et al., 2022). It has a wide application in the fields of multimodal retrieval (Dzabraev, Kalashnikov, Komkov, & Petiushko, 2021; Li et al., 2024), visual and language navigation (Anderson et al., 2018), image description (Jin, Cheng, Shen, Chen, & Ren, 2021), and so on. Through effective cross-modal semantic alignment, particularly fine-grained semantic alignment, VLP contributes to cross-modal learning and generalization. Visual–language models succeed in enabling zero-shot open-vocabulary tasks from natural language supervision (Gu, Lin, Kuo, & Cui, 2021), which inspires us to apply visual–language models for zero-shot HOI detection tasks.

2.3. Zero-shot HOI

Zero-shot HOI detection aims to generalize to unseen HOI categories during training effectively. Since the majority of HOI exhibit a long-tail distribution, attributed to the compositional nature of HOIs, prior research (Hou, Yu, Qiao, Peng, & Tao, 2021a; Hou et al., 2021b; Liu, Yuan and Chen, 2020; Peyre, Sivic, Laptev, & Schmid, 2019; Radford et al., 2021; Yuan et al., 2022) on zero-shot HOI detection focuses on transferring knowledge from known HOI concepts to unseen classes. They can be categorized into three scenarios: unseen object, unseen action, and unseen combination. There exist primarily two research streams for addressing this task. One stream (Hou et al., 2021a, 2021b; Liu, Yuan et al., 2020) employed a combined learning approach for zero-shot HOI detection, which entailed separating HOI representations and combining known features to identify unseen HOI concepts. ConsNet (Liu, Yuan et al., 2020), for instance, constructed a consistency graph with both visual features of potential human–object pairs and word embeddings of HOI labels. With the advancement of multimodal learning, there is a growing interest in transferring knowledge from pre-trained visual language models, *e.g.* CLIP is used to extract text embeddings of HOI descriptions for HOI detection tasks (Radford et al., 2021). RLIP (Yuan et al., 2022) proposed a transferable HOI detector

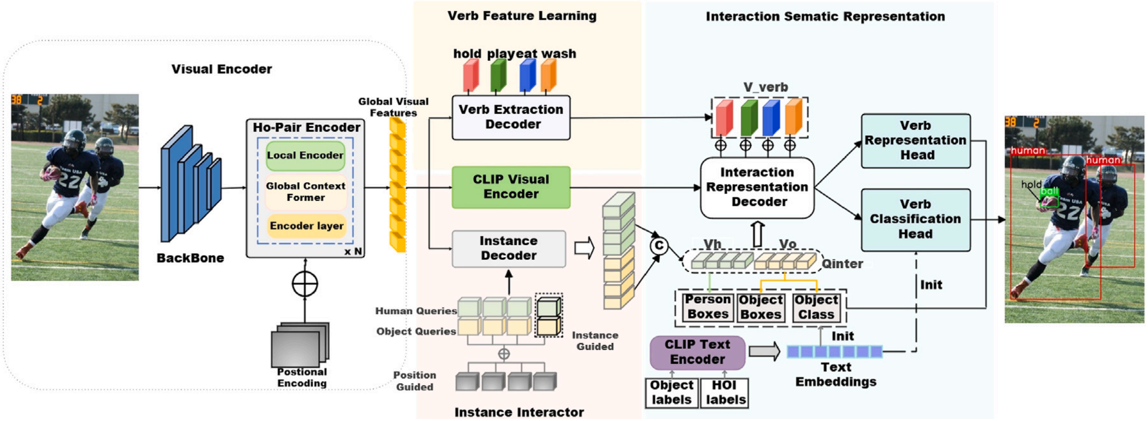

Fig. 2. Overview of KI2HOI pipeline. It consists of four parts: visual encoder, verb feature learning, instance interactor, and interaction semantic representation (ISR). Given an image, firstly, we obtain the feature map through the backbone and then use our dedicated visual encoder to extract contextual global features. The instance interactor injects CLIP spatial information and global features to locate human–object pairs and classify object categories. In the verb feature learning module, associated verb queries are fed to the verb extraction decoder to obtain fine-grained verb features. The interaction semantic representation model inputs the verb features and the interaction features from encoders to extract the interaction representation.

via natural language supervision. Building upon GEN-VLKT (Liao et al., 2022), HOICLIP (Ning et al., 2023) mapped image and text encodings to a joint visual-semantic space, to capture their correlations and effectively transfer knowledge from CLIP. Our work seeks to explore a more effective framework to make full use of CLIP for improving zero-shot HOI detection performance.

3. Methods

In this section, we introduce the details of the proposed framework that utilizes a pre-trained VLM for vision–language integration in zero-shot HOI detection. In Section 3.1, the overall architecture is presented. In Section 3.2, we introduce our visual encoder and a strategy for enhancing global feature extraction. Section 3.3 presents verb feature learning based on verb queries. In Section 3.4, we design an interaction semantic representation for transferring knowledge to HOI detection. Section 3.5 describes the training and inference procedures of our model.

3.1. Overall architecture

As shown in Fig. 2, our model consists of four primary components: visual encoder, verb feature learning, instance interactor, and interaction semantic representation. Given an input image $\boldsymbol{I}$, we initially extract feature maps via the backbone DETR (Carion et al., 2020). The feature maps are subsequently input into the ho-pair encoder to generate global visual feature $\boldsymbol{V}_g$, similar to GEN-VLKT (Liao et al., 2022). Human query $\boldsymbol{Q}_h$ and object query $\boldsymbol{Q}_o$ are inputted into the instance interactor to compute the mean for both types of queries in the corresponding decoder layer. These outputs queries in the last decoder are then fed to classifiers which initialize by label's text weights from CLIP to predict the interacting human bounding box $\boldsymbol{B}_h \in \mathbb{R}^{N\times 4}$ and object bounding box $\boldsymbol{B}_o \in \mathbb{R}^{N\times 4}$, where N is the number of queries, and object class $\boldsymbol{C}_o \in \mathbb{R}^{N\times C}$, where $\boldsymbol{C}$ denotes the object category.

Furthermore, verb queries are associated to interaction categories. For example, humans are more likely to catch or play sports ball than to bite a sports ball. To reflect this characteristic of HOIs, such queries interact with global visual features and become interaction-specific queries, as auxiliary information. We extract the spatial features $\boldsymbol{V}_{sp}$ from the pre-trained CLIP visual encoder as memory and feed into the interaction representation decoder by the cross-attention mechanism to augment the interaction representation and recognition. Finally, the HOI prediction categories are generated by the output of a linear classifier. The details of each component are explained in the following sections.

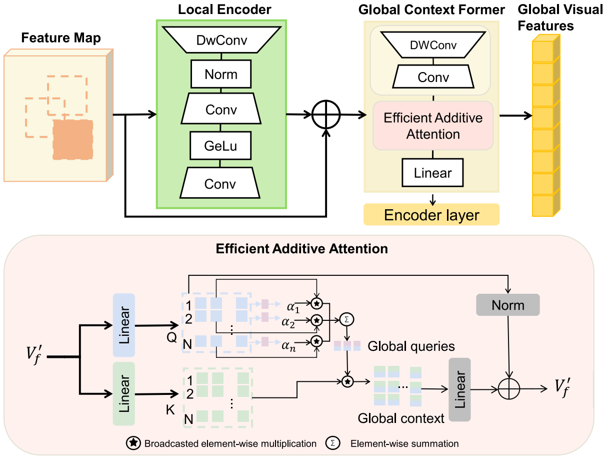

Fig. 3. Structure of ho-pair encoder. The local encoder is specifically engineered to encode efficient local characteristics, followed by 3 × 3 depth-wise convolution and two 1 × 1 convolutions for channel blending. The global context former is intended to capture comprehensive local–global representations by extracting local features from the local convolutional layers, an efficient additive attention module, and linear layers.

3.2. Visual encoder

Visual encoder employs the shared frozen backbone of DETR (Carion et al., 2020). Given an image input, the visual feature map from the ResNet-50 is projected via a convolutional layer to obtain a C channel feature map. Then, we embed the feature map adding cosine position embeddings following exit works for subsequent tasks. Existing encoders are limited to processing feature maps and position embeddings, often overlooking regions that contain contextual features essential for reasoning interactions. To fully harness the benefits of capturing global features, we propose the ho-pair encoder module, as depicted in Fig. 3. The ho-pair encoder consists of a local encoder and a global context former. This module incorporates a global context saliency-enhanced token mixing method, which aims to improve the model's ability to integrate and process visual context information.

We resize the cropped feature map to a 7 × 7 grid using ROI-Align and then feed them into the ho-pair encoder to acquire comprehensive local–global contextual information, Firstly, the feature map $\boldsymbol{V}_f$ is input into the local encoder for local feature extraction. The resulting feature map is fed into an efficient additive attention block after passing

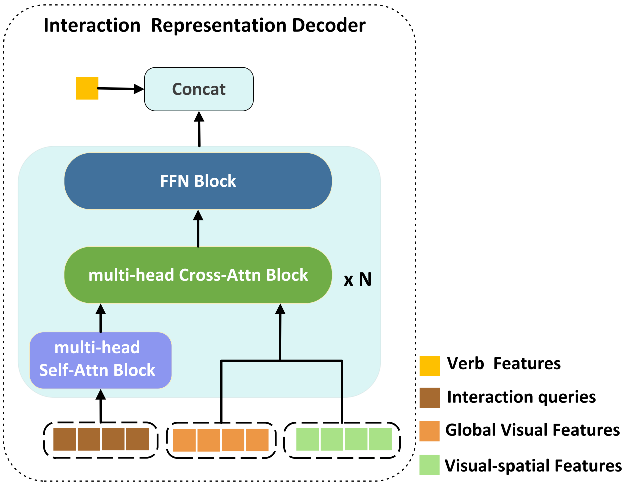

Fig. 4. Structure of interaction representation decoder. The interaction representation decoder commences with interaction queries Q_{inter}. These are initially processed by a multi-head self-attention block to capture internal relationships. Then, in a multi-head cross-attention block, the queries are combined with global visual features V_g and visual–spatial features V_{sp} to enrich the interaction representation. Finally, the outputs are concatenated with the verb features V_{verb} before being fed into the feed-forward network.

through convolutional layers, thereby enriching the feature representation with contextual awareness. To be specific, $\boldsymbol{V}_f$ is transformed into a query (Q) and key (K) with matrices $W_q, W_k \in \mathbb{R}^{(d\times d)}$ (where n is the token length and d is the embedding vector dimension). The query matrix Q multiplies a learnable vector $W_a \in \mathbb{R}^d$ to get the attention weights of the query, yielding a global attention query vector $a = Q \cdot W_a/\sqrt{d}$ in $\mathbb{R}^n$. Next, Q is multiplied by trainable weights and pooled for global queries. Then, it is element-wise multiplied by the broadcasted global queries, generating the global context representation. Inspired by the transformer architecture, we use a linear transformation layer for query–key interactions to obtain the hidden token representation. The output of the efficient additive attention can be described as:

$$\boldsymbol{V}'_f = \hat{Q} + L(K * q) \tag{1}$$

where $\hat{Q}$ denotes to the normalized query matrix, L denotes to the linear transformation.

Finally, the output global features $\boldsymbol{V}_g$ is processed via a linear block, comprising two 1 × 1 Conv layers, a normalization layer, and GeLU activation. The ho-pair encoder is defined as:

$$\boldsymbol{V}'_f = Conv_1(DWConv_3(BN(\boldsymbol{V}_f))) + \boldsymbol{V}_f, \tag{2}$$

$$\boldsymbol{V}'_f = AAttn(\boldsymbol{V}'_f) + \boldsymbol{V}'_f \tag{3}$$

$$\boldsymbol{V}_g = Linear(QK(\boldsymbol{V}'_f) + \boldsymbol{V}'_f). \tag{4}$$

where $Conv_{1,3}$ denotes 1×1, 3×3 Conv layer, BN denotes to Batch Normalization, followed by GeLU, and AAttn denotes the efficient additive attention.

3.3. Verb feature learning

Inspired by the Query2Label (Liu et al., 2021), we establish a novel approach termed learnable interaction-specific query, where $\boldsymbol{Q}_v \in \mathbb{R}^{N\times A}$, A is the number of interaction categories. $\boldsymbol{Q}_v$ differs from traditional transformer queries by establishing a one-to-one correspondence between each query and a particular instance throughout the training and inference phases. Such queries interact with global visual feature $\boldsymbol{V}_g$ through a verb extraction decoder and become interaction-specific queries. We design a self-attention and multi-head attention combination module and a feed-forward network (FFN) layer as verb extraction decoder, consisting of two layers. A set of verb queries $\boldsymbol{Q}_v$ aggregates the global feature information $\boldsymbol{V}_g$ through verb extraction decoder and is updated to $\boldsymbol{V}_{verb}$. In this way, the queries are learned to capture the verb *priors* and become good feature representations for these interactions. namely:

$$\hat{\boldsymbol{V}}_g = \text{selfAtten}(\boldsymbol{V}_g), \tag{5}$$

$$\boldsymbol{V}_{verb} = \text{MultiheadAttn}(\hat{\boldsymbol{V}}_g, \boldsymbol{Q}_v), \tag{6}$$

$$\boldsymbol{V}_{verb} = \text{FFN}(\boldsymbol{V}_{verb}). \tag{7}$$

3.4. Interaction Semantic Representation (ISR)

Conventional HOI detection uses an Interaction Classification Head to predict the confidence of each action for each pair of interactions and to judge the interaction. If the training sample for one of the actions is assumed to be too few or non-existent, the prediction for that action will be very inaccurate. This situation is more obvious in the zero-shot learning tasks, which is one of the reasons why traditional methods fail in the field of zero-shot learning. HOI detector maps visual features and tag-generated text features into the same space through CLIP, which performs quite well in the zero-shot learning tasks. We design the HOI labels and object labels are represented as “A photo of a person [verb] a/an [object]” and “A photo of a/an [object]” to assign different tokens to each HOI instance. To fully explore the CLIP knowledge, we propose to retrieve the text embeddings from the CLIP to better align them with the prior knowledge in the classifier weights.

Interaction Representation Decoder. We introduce an interaction semantic representation (ISR) module to extract the interactive representations of human–object interaction pairs. First, We add a learnable position guided embedding $\boldsymbol{P} \in \mathbb{R}^{N\times C}$ for human queries and object queries, *viz.*, $\boldsymbol{Q}_h \in \mathbb{R}^{N\times C}$ and $\boldsymbol{Q}_o \in \mathbb{R}^{N\times C}$ at the same position as interaction pairs. We compute the interaction queries $\boldsymbol{Q}_{inter} \in \mathbb{R}^{N\times C}$ by taking the average of the concatenation of $\boldsymbol{Q}_h$ and $\boldsymbol{Q}_o$. That is:

$$\boldsymbol{Q}_{inter} = \text{Cat}(\boldsymbol{Q}_h + \boldsymbol{P}, \boldsymbol{Q}_o + \boldsymbol{P})/2. \tag{8}$$

To guide interaction queries $\boldsymbol{Q}_{inter}$ to explore informative regions in both $\boldsymbol{V}_g$ and $\boldsymbol{V}_{sp}$, where $\boldsymbol{V}_{sp}$ represents the semantically aligned visual–spatial features obtained by feeding the image into the image encoder of CLIP, we design an interaction representation decoder with multiple cross-attention, as shown in Fig. 4. Each decoder consists of an attention block, a self-attention block, and a forward feedback network. $\boldsymbol{Q}_{inter}$ is first input to the self-attention block and the corresponding output is fed into the cross-attention mechanism with $\boldsymbol{V}_g$ and $\boldsymbol{V}_{sp}$. Subsequently, the final output is:

$$\boldsymbol{Q}_{inter} = \text{MHSA}(\boldsymbol{Q}_{inter}), \tag{9}$$

$$\boldsymbol{Q}'_{inter} = \text{MHCA}(\boldsymbol{Q}_{inter}, \boldsymbol{V}_{sp}), \tag{10}$$

$$\boldsymbol{Q}_{inter''} = \text{MHCA}(\boldsymbol{Q}_{inter}, \boldsymbol{V}_g), \tag{11}$$

$$\boldsymbol{Q}_{inter} = \text{FFN}(\boldsymbol{Q}'_{inter} + \boldsymbol{Q}''_{inter}), \tag{12}$$

where MHSA denotes a multi-head self-attention operation and MHCA denotes a multi-head cross-attention operation. Since, the obtained $\boldsymbol{Q}_{inter} \in \mathbb{R}^{N_q\times C}$ integrates the knowledge of CLIP and visual features, enabling the detection of fine-grained HOI.

By leveraging object and human information from the instance Interactor, we can concatenate interaction representations from the spatial feature map of CLIP and visual features from the detector

Table 1
Comparison with state-of-the-art methods on HICODET. All methods utilize the ResNet-50 backbone network. The letters in the Extra column indicate the extra input features: T (Linguistic features of label semantic embeddings), X (No extra features) (see Lei et al., 2023; Li, Liu, Wu, Li, & Lu, 2020; Tu et al., 2022; Zhang, Campbell & Gould, 2021; Zhang, Liao et al., 2021; Zheng, Xu, & Jin, 2023).

Method	Extra	mAP default			mAP know object		
		Non-Rare	Full	Rare	Non-Rare	Full	Rare
IDN (Li et al., 2020)	x	23.36	22.47	23.63	26.43	25.01	26.85
HOTR (Kim, Lee, Kang et al., 2021)	x	25.10	17.34	27.42	–	–	–
ATL (Hou et al., 2021a)	x	28.53	21.64	30.59	31.18	24.15	33.29
QPIC (Tamura et al., 2021)	x	29.07	21.85	31.23	31.68	24.14	33.93
FCL (Hou et al., 2021b)	x	29.12	23.67	30.75	31.31	25.62	33.02
SCG (Zhang, Campbell et al., 2021)	x	31.33	24.72	33.31	34.37	27.18	36.52
UPT (Zhang et al., 2022)	x	31.66	25.94	33.36	35.05	29.27	36.77
CDN (Zhang, Liao et al., 2021)	x	31.78	27.55	33.05	34.53	29.73	35.96
Iwin (Tu et al., 2022)	x	32.03	27.62	34.14	35.17	28.79	35.91
Liu et al. (2022)	x	33.51	30.30	34.46	36.28	33.16	37.21
GEN-VLKT (Liao et al., 2022)	T	33.75	29.25	35.10	36.78	32.75	37.99
ADA-CM (Lei et al., 2023)	x	33.80	31.72	34.42	–	–	–
OpenCat (Zheng et al., 2023)	T	32.68	28.42	33.75	–	–	–
HOICLIP (Ning et al., 2023)	T	**34.69**	31.12	35.74	37.61	34.47	38.54
KI2HOI	T	34.20	**32.26**	**36.10**	**37.85**	**35.89**	**38.78**

Table 2
Comparison with state-of-the-art methods on V-COCO dataset. – representative cannot be found from te original paper.

Method	Detector	Backbone	V-COCO	
			AP^{S1}_{role}	AP^{S2}_{role}
IDN (Li et al., 2020)	COCO	ResNet-50	53.3	60.3
HOTR (Kim, Lee, Kang et al., 2021)	HICO-Det	ResNet-50	55.2	64.4
ATL (Hou et al., 2021a)	COCO	ResNet-50	–	–
QPIC (Tamura et al., 2021)	HICO-Det	ResNet-50	58.8	61.0
FCL (Hou et al., 2021b)	COCO	ResNet-50	52.4	–
SCG (Zhang, Campbell et al., 2021)	COCO	ResNet-50-FPN	54.2	60.9
UPT (Zhang et al., 2022)	COCO	ResNet-50	59.0	64.5
CDN (Zhang, Liao et al., 2021)	HICO-DET	ResNet-50	62.3	64.4
Iwin (Tu et al., 2022)	HICO-Det	ResNet-50-FPN	60.5	–
Liu et al. (2022)	COCO	ResNet-50	63.0	65.2
GEN-VLKT (Liao et al., 2022)	HICO-Det	ResNet-50+ViT-B	62.4	64.4
ADA-CM (Lei et al., 2023)	COCO	ResNet-50+ViT-B	56.1	61.5
OpenCat (Zheng et al., 2023)	–		61.9	63.2
HOICLIP (Ning et al., 2023)	HICO-Det	ResNet-50+ViT-B	63.5	64.8
KI2HOI	HICO-Det	ResNet-50+ViT-B	**63.9**	**65.0**

to efficiently retrieve corresponding interaction representations and achieve strong generalization capabilities.

Verb Predictor via Knowledge Retrieval. To align verb features V_{verb} with Q_{inter}, we first use a projection operation to map both into the CLIP feature space. Subsequently, Q_{inter} is passed through a lightweight adapter designed for alignment with the verb features. Then, we connect Q_{inter} and V_{verb} to form a comprehensive interaction feature representation. Finally, the verb score is obtained as:

$$\boldsymbol{Q}_{inter} = \mathrm{Proj}(\boldsymbol{Q}_{inter}), \boldsymbol{V}_{verb} = \mathrm{Proj}(\boldsymbol{V}_{verb}), \tag{13}$$

$$\boldsymbol{C}_{verb} = \mathrm{MLP}(\boldsymbol{V}_{verb}), \boldsymbol{D}_{verb} = \mathrm{MLP}(\boldsymbol{Q}_{inter}), \tag{14}$$

$$\boldsymbol{C}_{verb} = \mathrm{Cat}(\boldsymbol{C}_{verb}, \boldsymbol{D}_{verb}), \tag{15}$$

$$\boldsymbol{S}_{verb} = \boldsymbol{C}_{verb}\boldsymbol{W}_{v}^{\boldsymbol{T}}. \tag{16}$$

where the verb score is computed by the cosine similarity between the verb features and the text weight of the verb representations. As well, a reconstruction loss function that quantifies the dissimilarity between features. That is:

$$L_{re} = L_1(\boldsymbol{Q}_{inter}, \boldsymbol{V}_{sp}). \tag{17}$$

where L_1 loss is used to minimize the distance between features and visual embeddings.

3.5. Training and inference

Training. We employ the Hungarian algorithm (Liao et al., 2022; Ning et al., 2023; Wu et al., 2023) for bipartite matching between predictions and ground truths. The matching cost consists of human bounding box regression loss $\boldsymbol{L}_{bh}$, object bounding box regression loss $\boldsymbol{L}_{bo}$, interaction-over-union loss $\boldsymbol{L}_u$, and classification loss $\boldsymbol{L}_c$. Combined with the reconstruction loss function $\boldsymbol{L}_{re}$, the final loss function is as follows:

$$\boldsymbol{L} = \lambda_{bh}\boldsymbol{L}_{bh} + \lambda_{bo}\boldsymbol{L}_{bo} + \lambda_u\boldsymbol{L}_u + \lambda_c\boldsymbol{L}_c + \lambda_{re}\boldsymbol{L}_{re}, \tag{18}$$

where λ_{bh}, λ_{bo}, λ_u, λ_c, λ_{re} are hyper-parameters for adjusting the weights of all losses.

Inference. Reconstruction loss is only used for training, $\boldsymbol{S}_h \in [0,1]^N$, $\boldsymbol{S}_o \in [0,1]^N$. In inference, the final score $\boldsymbol{S}_{final} \in [0,1]^N$ is summed by $\boldsymbol{S}_h$, $\boldsymbol{S}_O$ and $\boldsymbol{S}_{verb}$.

$$\boldsymbol{S}^i_{final} = \boldsymbol{S}_h + \boldsymbol{S}_o + \boldsymbol{S}_{verb}, i \in [1, c]. \tag{19}$$

4. Experiments

4.1. Experimental setup

Datasets. We evaluate our model on two widely-uesd benchmarks, HICO-DET (Chao, Liu, Liu, Zeng, & Deng, 2018) and V-COCO (Gupta & Malik, 2015). HICO-Det contains 47,776 images, of which 38,118 and

Table 3
Performance comparison for zero-shot HOI detection on HICO-DET. RF-UC indicates rare first setting, and NF-UC represents non-rare first unseen combination setting. UC, UO, and UV denote unseen composition, unseen object, and unseen verb settings, respectively (see Bansal, Rambhatla, Shrivastava, & Chellappa, 2020; Hou, Peng, Qiao, & Tao, 2020; Hou, Yu, & Tao, 2022).

Method	Type	Full	Seen	Unseen
VCL (Hou et al., 2020)	RF-UC	21.43	24.28	10.06
ATL (Hou et al., 2021a)	RF-UC	21.57	24.67	9.18
FCL (Hou et al., 2021b)	RF-UC	22.01	24.23	13.16
SCL (Hou et al., 2022)	RF-UC	28.08	30.39	19.07
RLIP (Yuan et al., 2022)	RF-UC	30.52	33.35	19.19
EoID (Wu et al., 2023)	RF-UC	29.52	31.39	22.04
GEN-VLKT (Liao et al., 2022)	RF-UC	30.56	32.91	21.36
HOICLIP (Ning et al., 2023)	RF-UC	32.99	34.85	25.53
KI2HOI	RF-UC	**34.10**	**35.79**	**26.33**
VCL (Hou et al., 2020)	NF-UC	16.22	18.52	18.06
ATL (Hou et al., 2021a)	NF-UC	18.25	18.78	18.67
FCL (Hou et al., 2021b)	NF-UC	18.66	19.55	19.37
SCL (Hou et al., 2022)	NF-UC	24.34	25.00	21.73
RLIP (Yuan et al., 2022)	NF-UC	26.19	27.67	20.27
GEN-VLKT (Liao et al., 2022)	NF-UC	23.71	23.38	25.05
EoID (Wu et al., 2023)	NF-UC	26.69	26.66	26.77
HOICLIP (Ning et al., 2023)	NF-UC	27.75	28.10	26.39
KI2HOI	NF-UC	**27.77**	**28.31**	**28.89**
FG (Bansal et al., 2020)	UC	12.26	12.60	10.93
ConsNet (Liu, Yuan et al., 2020)	UC	19.81	20.51	16.99
EoID (Wu et al., 2023)	UC	28.91	30.39	23.01
HOICLIP (Ning et al., 2023)	UC	32.99	34.85	25.53
KI2HOI	UC	**34.56**	**35.76**	**27.43**
FCL (Hou et al., 2021b)	UO	19.87	20.74	15.54
ATL (Hou et al., 2021a)	UO	20.47	21.54	15.11
GEN-VLKT (Liao et al., 2022)	UO	25.63	28.92	10.51
HOICLIP (Ning et al., 2023)	UO	28.53	30.99	16.20
KI2HOI	UO	**28.84**	**31.70**	**16.50**
ConsNet (Liu, Yuan et al., 2020)	UV	19.04	20.02	14.12
GEN-VLKT (Liao et al., 2022)	UV	28.74	30.23	20.96
EoID (Wu et al., 2023)	UV	29.61	30.73	22.71
HOICLIP (Ning et al., 2023)	UV	31.09	32.19	24.30
KI2HOI	UV	**31.85**	**32.95**	**25.20**

9658 images are used for training. It includes 600 HOI triplets, where 138 triplets are rare categories less than 10 training instances, and the remaining 462 categories are non-rare. HICO-Det also provides a zero-shot detection setting by holding out 120 rare interactions. V-COCO is a subset of the COCO dataset and consists of 10,396 images, with 5400 for training and 4964 for testing. It has 29 action categories, including 4 annotations without any interaction with objects.

Zero-shot Data Setups. We conduct experiments on the HICO-Det for zero-shot HOI detection, mainly using the following approaches: Rare First Unseen Combination (RF-UC), Non-rare First Unseen Combination (NF-UC), Unseen Verb (UV), Unseen Object (UO) and Unseen Combination(UC). In the UC setting, the training data includes all categories of objects and verbs but lacks some HOI triplet categories. We assess 120 unseen categories and 480 seen categories out of a total of 600 categories. The RF-UC selects the tail HOIs as unseen categories, whereas the NF-UC prefers head categories. In the UO setting, we select 12 unseen objects out of a total of 80 objects to define unseen HOIs. Additionally, we introduce a UV setting, where 20 verbs from a total of 117 verbs are randomly selected to construct 84 unseen and 516 seen HOIs.

Evaluation Metric. We evaluate our model using the mean Average Precision (mAP) as metric, a prediction is considered as true positive if the predicted human and object bounding boxes have an IoU of at least 0.5 with the ground truth, and the predicted interaction category matches the correct category. We report the standard mAP for HOI detection, dividing interactions into non-rare, rare, and unseen cases based on their occurrences in the training set. The mAP measures HOI detection performance with a threshold of 0.5 for the IoU between predicted and ground-truth bounding boxes.

Implementation Details. We use pre-trained DETR with ResNet-50 as the backbone network. The visual encoder is based on ViT-32/B CLIP, and during training, the parameters of CLIP remain unchanged. Our model's encoder and decoder have 3 layers with 64 queries, except the verb extraction decoder has 1 layer. We train the model using the AdamW optimizer with a specified interval for adjusting the learning rate. The training epochs is set to 90, with a gradual decrease in the learning rate after the 60th epoch. All experiments are conducted with a batch size of 32 on 8 NVIDIA A6000 GPUs.

4.2. Effectiveness for HOI detection

Zero-shot Detection. We conduct various zero-shot setting experiments on the HICO-DET dataset, and the results are shown in Table 3. Our model outperforms existing state-of-the-art methods, demonstrating strong performance competitiveness. Specifically, under the RF-UC and NF-UC settings, our model's relative mean average precision (mAP) for unseen categories exceeds that of EoID (Wu et al., 2023) by 23.26% and 7.91%, respectively. In the UA setting, our model outperforms the latest work, HOICLIP (Ning et al., 2023), with a 1.03 mAP improvement for unseen types. In both full and unseen categories, our model achieves 1.07 and 2.5 mAP improvements relative to EOID (Wu et al., 2023). In the case of unseen objects, our model's performance surpasses that of GEN-VLKT (Liao et al., 2022) by 3.21 mAP.

Fully Supervised Detection. To verify the generalization ability of the model, we conducted fully supervised experiments on HICO-DET and V-COCO. As tabulated in Table 1, our model achieves remarkable performance, exceeding GEN-VLKT and HOICLIP (Ning et al., 2023) by 3.01 mAP and 1.14 mAP for full categories, and by 1 mAP and 0.36 mAP for rare categories. This indicates that our model can handle the long-tailed distribution of HOI well. As tabulated in Table 2, we achieve 63.9 role AP on Scenario 1 and 65.0 role AP on Scenario 2 surpassing previous methods for the V-COCO dataset.

Robustness to Distributed Data. We investigate the robustness of the proposal under different data quantities. We decrease the proportion of training data from 100% to 15%, and we seek to achieve less performance loss in HOI detection. Compared to the state-of-the-art GEN-VLKT in Table 4, the proposed method achieves competitive performance in detecting both non-rare and rare categories. Furthermore, Table 4 highlights the improvements achieved by our model at various volumes of data. At 25% training data, our model exhibits a 78.41% increase in mAP gain for rare HICO-DET.

4.3. Ablation studies

Network Architecture Analysis. We conduct ablation experiments for each module on the HICO-DET dataset under the UV setting and the results are shown in Table 5. The GEN-VLKT without the knowledge distillation component.is regarded as the baseline. First, we examine the effect of CLIP and the result shows a 13.8 mAP improvement in HOI detection for Unseen categories. Thus, the visual and linguistic knowledge extracted by CLIP enable to learn deeper interaction understanding. Next, we replace the encoder with the proposed ho-pair encoder for using fine-grained visual features. As a result, we observe that the results come up to 30.99 mAP and 32.02 mAP in the Full and Seen categories, respectively. Finally, we add a verb feature learning to capture verb-related features, and the performance is further advanced to 31.85 mAP in full categories, which demonstrates the necessity of verb feature learning in the zero-shot HOI task.

Reconstruction Loss Setting. As shown in Table 6, we compare two loss types, *i.e.*, L_1 loss and L_2 loss, as the reconstruction loss. It demonstrates that the reconstruction loss is indispensable for making knowledge transfer effective. If exclusively employing L_1 loss, our model has demonstrated superior performance and achieved 1.05 mAP gain, surpassing the performance exhibited by only employing the L_2

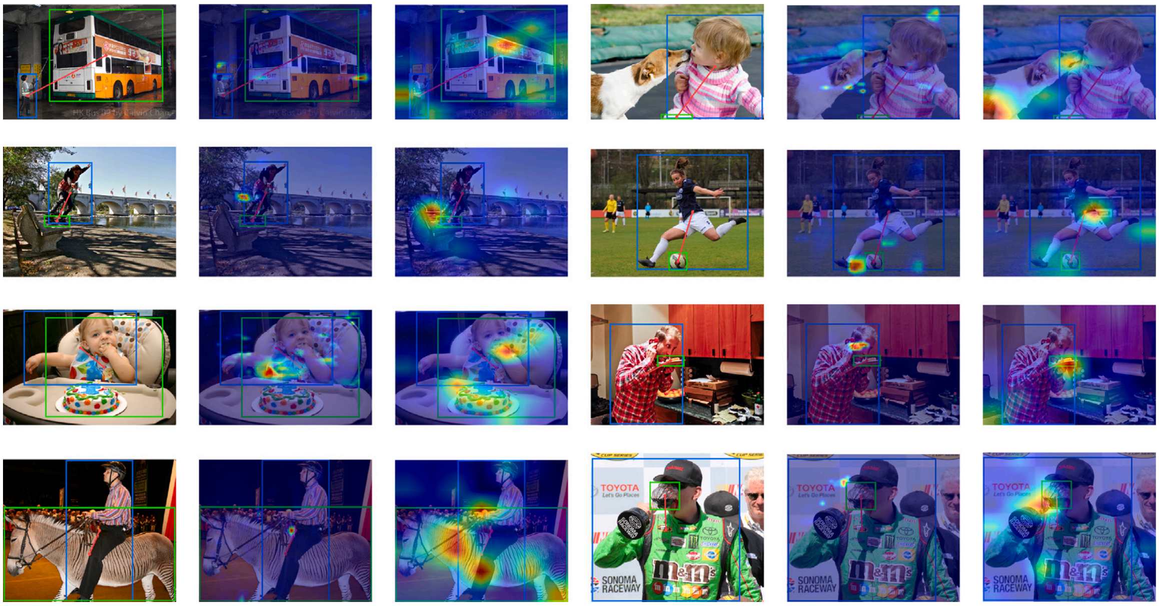

Fig. 5. Visualization of the HOI detection results. From left to right, **column 1**: HOI prediction results; **column 2**: attention maps from Verb Extraction Decoder; **column 3**: attention maps from Interaction Representation Decoder. Images are sampled from the HICO-DET dataset in UV test set.

Table 4
Robustness to different distribution data.

	Non-Rare HICO-DET			
Percentage	100%	50%	25%	15%
GEN-VLKT	33.75	26.55	22.14	20.40
KI2HOI	34.20	31.25	30.06	27.20
mAP gain (%)	+1.3	+19.2	+35.77	+33.3
	Rare HICO-DET			
Percentage	100%	50%	25%	15%
GEN-VLKT	29.25	18.94	14.04	13.84
KI2HOI	36.10	26.68	25.05	22.51
mAP gain (%)	+23.41	+40.86	+78.41	+38.51

Table 5
Network architecture analysis. Ablation studies are conducted on HICO-DET under the Unseen Verb (UV) setting.

Method	Full	Seen	Unseen
Baseline	28.20	30.49	9.57
+CLIP	30.45	31.65	23.37
+Ho-Pair encoder	30.99	32.02	23.96
+Verb feature learning	**31.85**	**32.95**	**25.20**

Table 6
Reconstruction loss setting.

L_1	L_2	Full	Rare	Non-Rare
–	–	33.31	31.83	35.05
✓	–	34.20	32.26	36.10
–	✓	34.06	30.65	35.08
✓	✓	34.08	30.18	35.24

loss. When combining L_1 and L_2 losses through average summation, the performance is still inferior compared to only applying L_1 loss.

Verb Extraction Decoder Layer Selection. We examine the effect of verb extraction decoder layer number on the verb features update and conduct experiments in the NF-UC setting. In Table 7, we find that a single layer exhibits superior performance. While increasing the number of layers, there is no consistent improvement.

Table 7
The impact of different verb extraction decoder layer numbers.

Layers	Full	Seen	Unseen
1	27.77	28.31	28.89
2	26.30	26.70	24.71
3	26.53	26.70	25.07

4.4. *Qualitative visualization results*

As shown in Fig. 5, we illustrate the characteristics of our model by visualizing the attention feature maps of the decoder. Our framework can effectively infer human–object interaction relationships in unseen interaction categories. We observe the attention maps from the Interaction Representation Decoder focus on a broader range of interaction-related regions, while the Verb Extraction Decoder mainly emphasizes the interaction target area. Additionally, taking the detection of “wash-bus”, as an example, it indicates that our model is capable of perceiving interactions between indirectly connected humans and objects. In Fig. 6, we note that our model has the capability to detect various types of interactions.

5. Conclusion and limitations

We present a new one-stage framework called K2IHOI to improve zero-shot HOI detection through the incorporation of visual–linguistic prior knowledge. Within this framework, contextual spatial details concerning human–object–human interaction triples are extracted using the ho-pair encoder, and knowledge-enriched semantic content is integrated into the visual model. Moreover, pertinent verb inquiries are converted into category representations specific to the interaction, and an interaction semantic representation module is integrated to enhance the comprehension of interactions. This approach has produced significant outcomes on commonly utilized benchmarks. Nevertheless, it does possess specific constraints. The incorporation of sizable language models such as CLIP during the training phase results in considerable computational expenses. Furthermore, the extraction of features from multiple branches results in an escalation in memory utilization. To tackle these challenges, forthcoming research could investigate methods for extracting concise yet meaningful features for HOI detection. It

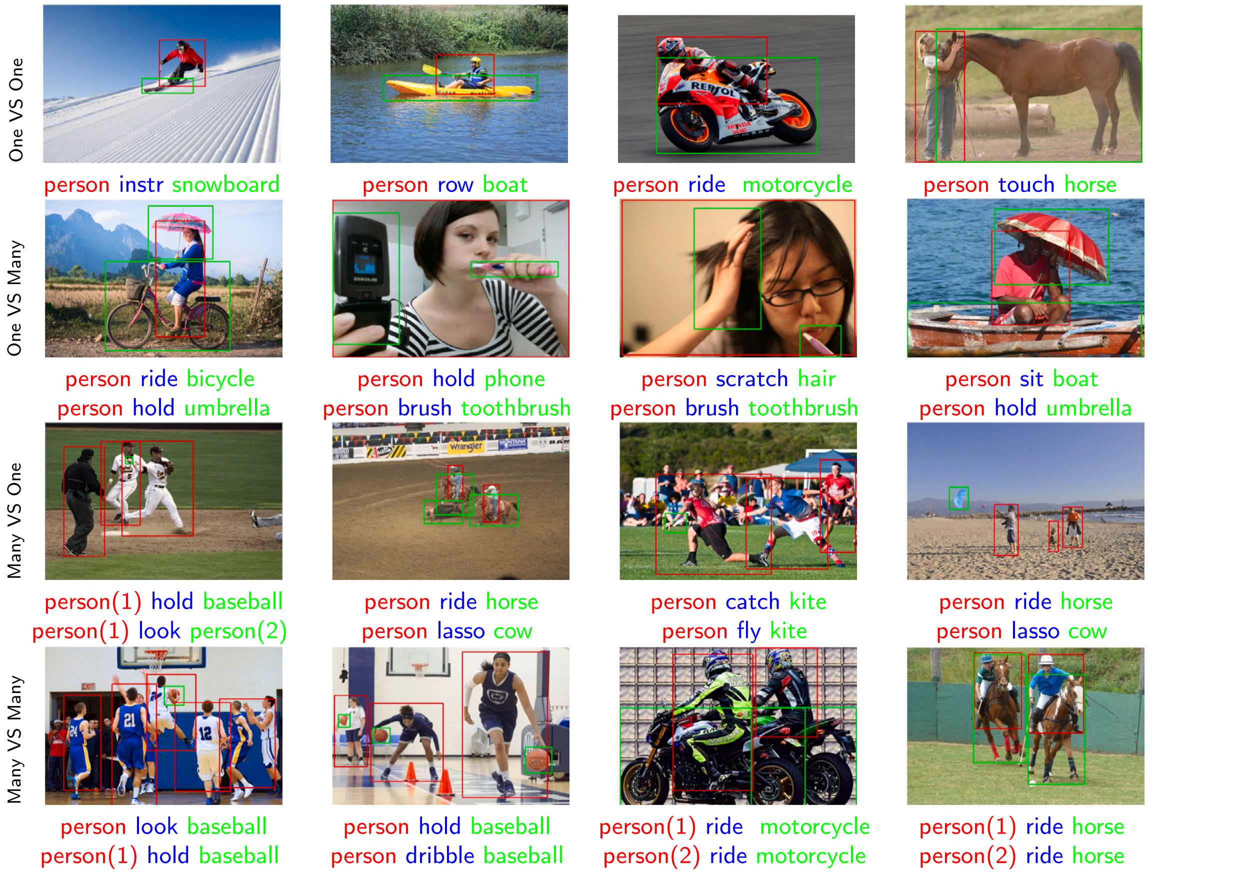

Fig. 6. Visualization of different interaction relationships. **One vs. One**: person interact with a single objects; **One vs. Many**: person interact with multiple objects; **Many vs. One**: person interact with a single object; **Many vs. Many**: person interact with multiple objects.

would be beneficial to modify the VLM in various ways to enhance its suitability for this task.

CRediT authorship contribution statement

Weiying Xue: Writing – review & editing, Writing – original draft, Validation, Data curation. **Qi Liu:** Supervision. **Yuxiao Wang:** Writing – review & editing, Writing – original draft. **Zhenao Wei:** Visualization. **Xiaofen Xing:** Supervision, Conceptualization. **Xiangmin Xu:** Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62202174, in part by the Basic and Applied Basic Research Foundation of Guangzhou under Grant 2023A04J1674, and in part by The Taihu Lake Innovation Fund for the School of Future Technology of South China University of Technology under Grant 2024B105611004.

Data availability

No data was used for the research described in the article.

References

Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., et al. (2018). Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3674–3683).

Antoun, M., & Asmar, D. (2022). Human object interaction detection: Design and survey. *Image and Vision Computing*, Article 104617.

Bansal, A., Rambhatla, S. S., Shrivastava, A., & Chellappa, R. (2020). Detecting human-object interactions via functional generalization. *vol. 34*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 10460–10469). 07.

Cao, Y., Tang, Q., Yang, F., Su, X., You, S., Lu, X., et al. (2023). Re-mine, learn and reason: Exploring the cross-modal semantic correlations for language-guided hoi detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 23492–23503).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213–229). Springer.

Chao, Y.-W., Liu, Y., Liu, X., Zeng, H., & Deng, J. (2018). Learning to detect human-object interactions. In *2018 IEEE winter conference on applications of computer vision* (pp. 381–389). IEEE.

Cheng, Y., Wang, Z., Zhan, W., & Duan, H. (2023). Multi-scale human-object interaction detector. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(4), 1827–1838. http://dx.doi.org/10.1109/TCSVT.2022.3216663.

Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., & Li, G. (2022). Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14084–14093).

Dzabraev, M., Kalashnikov, M., Komkov, S., & Petiushko, A. (2021). Mdmmt: Multidomain multimodal transformer for video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3354–3363).

Fu, Z., Zheng, C., Feng, J., Cai, Y., Wei, X.-Y., Wang, Y., et al. (2023). DRAKE: Deep pair-wise relation alignment for knowledge-enhanced multimodal scene graph generation in social media posts. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(7), 3199–3213. http://dx.doi.org/10.1109/TCSVT.2022.3231437.

Gu, X., Lin, T.-Y., Kuo, W., & Cui, Y. (2021). Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921.

Gupta, S., & Malik, J. (2015). Visual semantic role labeling. arXiv preprint arXiv:1505.04474.

Hou, Z., Peng, X., Qiao, Y., & Tao, D. (2020). Visual compositional learning for human-object interaction detection. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XV 16* (pp. 584–600). Springer.

Hou, Z., Yu, B., Qiao, Y., Peng, X., & Tao, D. (2021a). Affordance transfer learning for human-object interaction detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 495–504).

Hou, Z., Yu, B., Qiao, Y., Peng, X., & Tao, D. (2021b). Detecting human-object interaction via fabricated compositional learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14646–14655).

Hou, Z., Yu, B., & Tao, D. (2022). Discovering human-object interaction concepts via self-compositional learning. In *European conference on computer vision* (pp. 461–478). Springer.

Ji, Z., Li, Z., Zhang, Y., Wang, H., Pang, Y., & Li, X. (2024). Hierarchical matching and reasoning for multi-query image retrieval. *Neural Networks*, *173*, Article 106200.

Jia, W., & Ma, S. (2023). Query preference analysis on cascade inference human–object interaction detection transformer. *International Journal of Pattern Recognition and Artificial Intelligence*, *37*(13), Article 2356021.

Jin, W., Cheng, Y., Shen, Y., Chen, W., & Ren, X. (2021). A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. arXiv preprint arXiv:2110.08484.

Kim, D., Angelova, A., & Kuo, W. (2023). Region-aware pretraining for open-vocabulary object detection with vision transformers. CoRR abs/2305.07011. arXiv:2305.07011. URL: https://doi.org/10.48550/arXiv.2305.07011.

Kim, S., Jung, D., & Cho, M. (2023). Relational context learning for human-object interaction detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2925–2934).

Kim, B., Lee, J., Kang, J., Kim, E.-S., & 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* (pp. 74–83).

Kim, J.-J., Lee, D.-G., Wu, J., Jung, H.-G., & Lee, S.-W. (2021). Visual question answering based on local-scene-aware referring expression generation. *Neural Networks*, *139*, 158–167.

Lei, T., Caba, F., Chen, Q., Jin, H., Peng, Y., & Liu, Y. (2023). Efficient adaptive human-object interaction detection with concept-guided memory. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6480–6490).

Li, Y.-L., Liu, X., Wu, X., Li, Y., & Lu, C. (2020). Hoi analysis: Integrating and decomposing human-object interaction. *Advances in Neural Information Processing Systems*, *33*, 5011–5022.

Li, J., Wong, W. K., Jiang, L., Fang, X., Xie, S., & Xu, Y. (2024). CKDH: CLIP-based knowledge distillation hashing for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, http://dx.doi.org/10.1109/TCSVT.2024.3350695, 1–1.

Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., et al. (2022). Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10965–10975).

Li, Z., Zou, C., Zhao, Y., Li, B., & Zhong, S. (2022). Improving human-object interaction detection via phrase learning and label composition. *vol. 36*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 1509–1517). 2.

Liao, Y., Zhang, A., Lu, M., Wang, Y., Li, X., & Liu, S. (2022). GEN-VLKT: Simplify association and enhance interaction understanding for HOI detection. In *2022 IEEE/CVF conference on computer vision and pattern recognition* (pp. 20091–20100). http://dx.doi.org/10.1109/CVPR52688.2022.01949.

Liu, Y., Chen, Q., & Zisserman, A. (2020). Amplifying key cues for human-object-interaction detection. In *European conference on computer vision* (pp. 248–265). Springer.

Liu, X., Li, Y.-L., Wu, X., Tai, Y.-W., Lu, C., & Tang, C.-K. (2022). Interactiveness field in human-object interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20113–20122).

Liu, J., & Liu, Q. (2024). R3CD: Scene graph to image generation with relation-aware compositional contrastive control Diffusion. In *Proceedings of the AAAI conference on artificial intelligence.*

Liu, Y., Yuan, J., & Chen, C. W. (2020). Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 4235–4243).

Liu, S., Zhang, L., Yang, X., Su, H., & Zhu, J. (2021). Query2label: A simple transformer way to multi-label classification. arXiv preprint arXiv:2107.10834.

Ning, S., Qiu, L., Liu, Y., & He, X. (2023). HOICLIP: Efficient knowledge transfer for HOI detection with vision-language models. In *2023 IEEE/CVF conference on computer vision and pattern recognition* (pp. 23507–23517). http://dx.doi.org/10.1109/CVPR52729.2023.02251.

Peyre, J., Sivic, J., Laptev, I., & Schmid, C. (2019). Detecting unseen visual relations using analogies. In *2019 IEEE/CVF international conference on computer vision* (pp. 1981–1990). http://dx.doi.org/10.1109/ICCV.2019.00207.

Qin, Y., Gu, X., & Tan, Z. (2022). Visual context learning based on textual knowledge for image–text retrieval. *Neural Networks*, *152*, 434–449.

Qu, X., Ding, C., Li, X., Zhong, X., & Tao, D. (2022). Distillation using oracle queries for transformer-based human-object interaction detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19558–19567).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMLR.

Tamura, M., Ohashi, H., & Yoshinaga, T. (2021). 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* (pp. 10410–10419).

Tu, D., Min, X., Duan, H., Guo, G., Zhai, G., & Shen, W. (2022). Iwin: Human-object interaction detection via transformer with irregular windows. In *European conference on computer vision* (pp. 87–103). Springer.

Wan, B., Liu, Y., Zhou, D., Tuytelaars, T., & He, X. (2023). Weakly-supervised HOI detection via prior-guided bi-level representation learning. arXiv preprint arXiv:2303.01313.

Wang, H., Yu, H., & Zhang, Q. (2024). Human–object interaction detection via global context and pairwise-level fusion features integration. *Neural Networks*, *170*, 242–253.

Wu, M., Gu, J., Shen, Y., Lin, M., Chen, C., & Sun, X. (2023). End-to-end zero-shot hoi detection via vision and language knowledge distillation. *vol. 37*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 2839–2846). 3.

Yang, D., Zou, Y., Zhang, C., Cao, M., & Chen, J. (2022). RR-net: Relation reasoning for end-to-end human-object interaction detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(6), 3853–3865. http://dx.doi.org/10.1109/TCSVT.2021.3119892.

Yuan, H., Jiang, J., Albanie, S., Feng, T., Huang, Z., Ni, D., et al. (2022). Rlip: Relational language-image pre-training for human-object interaction detection. *Advances in Neural Information Processing Systems*, *35*, 37416–37431.

Zhang, F. Z., Campbell, D., & Gould, S. (2021). Spatially conditioned graphs for detecting human-object interactions. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 13319–13327).

Zhang, F. Z., Campbell, D., & Gould, S. (2022). 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* (pp. 20104–20112).

Zhang, A., Liao, Y., Liu, S., Lu, M., Wang, Y., Gao, C., et al. (2021). Mining the benefits of two-stage and one-stage hoi detection. *Advances in Neural Information Processing Systems*, *34*, 17209–17220.

Zheng, S., Xu, B., & Jin, Q. (2023). Open-category human-object interaction pre-training via language modeling framework. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19392–19402).

Zhou, D., Liu, Z., Wang, J., Wang, L., Hu, T., Ding, E., et al. (2022). Human-object interaction detection via disentangled transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19568–19577).