Discovering Syntactic Interaction Clues for Human-Object Interaction Detection

Jinguo Luo[1†] Weihong Ren[1,2*] Weibo Jiang[1†] Xi'ai Chen[2,3] Qiang Wang[4] Zhi Han[2,3] Honghai Liu[1]
[1]Harbin Institute of Technology, Shenzhen [4]Shenyang University
[2] State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences
[3] Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences
{23s153135, jiangweibo}@stu.hit.edu.cn, {renweihong, honghai.liu}@hit.edu.cn
{chenxiai, wangqiang, hanzhi}@sia.cn

Abstract

*Recently, Vision-Language Model (VLM) has greatly advanced the Human-Object Interaction (HOI) detection. The existing VLM-based HOI detectors typically adopt a handcrafted template (e.g., a photo of a person [**action**] a/an [**object**]) to acquire text knowledge through the VLM text encoder. However, such approaches, only encoding the action-specific text prompts in vocabulary level, may suffer from learning ambiguity without exploring the fine-grained clues from the perspective of interaction context. In this paper, we propose a novel method to discover Syntactic Interaction Clues for HOI detection (SICHOI) by using VLM. Specifically, we first investigate what are the essential elements for an interaction context, and then establish a syntactic interaction bank from three levels: spatial relationship, action-oriented posture and situational condition. Further, to align visual features with the syntactic interaction bank, we adopt a multi-view extractor to jointly aggregate visual features from instance, interaction, and image levels accordingly. In addition, we also introduce a dual cross-attention decoder to perform context propagation between text knowledge and visual features, thereby enhancing the HOI detection. Experimental results demonstrate that our proposed method achieves state-of-the-art performance on HICO-DET and V-COCO.*

1. Introduction

Human-Object Interaction (HOI) detection aims to localize humans and objects from a given image, and also predicts the semantic relationships between them. A HOI instance can be represented as a triplet <human, action, object>. In recent years, HOI detection has attracted enormous attention, due to its significant role in a wide range of high-level computer vision tasks, such as video analysis [8], human action recognition [37] and image retrieval [50].

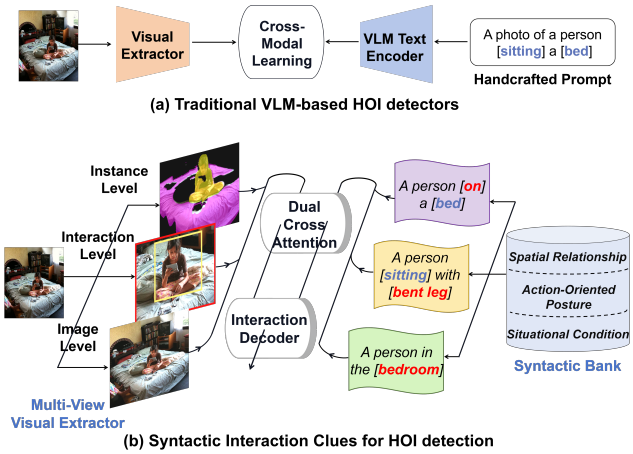

Figure 1. Comparison between previous VLM-based HOI detectors and our SICHOI. (a) Previous VLM-based methods adopt a handcrafted template to acquire action-specific knowledge in vocabulary level. (b) Our SICHOI model establishes a syntactic bank from three levels: *spatial relationship*, *action-oriented posture* and *situational condition*, which can provide informative and clear clues to distinguish different interactions.

The existing HOI detection methods can be roughly divided into one-stage and two-stage methods. The early one-stage ones try to perform object detection and interaction classification simultaneously, and then introduce additional auxiliary priors, e.g., interaction points [29, 49] and union boxes [20] to group HOI pairs. The recent transformer-based methods [43, 65, 69] directly predict HOI triplets without explicitly modelling human-object context, but may suffer from insufficient exchange between contextual clues [39]. Contrarily, the two-stage methods [11, 14, 61] first perform object detection and then explicitly infer the semantic relationships between each pair of human and object. They usually adopt off-the-shelf object detectors, e.g., DERT [3] to obtain object detections, and pay more atten-

*Corresponding author.
†Both authors contributed equally to this work.

tion on extracting interaction context from HOI pairs. Despite a lot of efforts, the HOI detection performance can't be further improved by only considering visual features.

Recent researches [9, 36, 62] have found that Vision-Language Model (VLM) [27, 41] has excellent performance in addressing open-vocabulary problem, due to its success in unifying visual representation and linguistic/text knowledge. Inspired by this, some methods [55–57, 63] apply the VLM to HOI detection and have achieved significant improvement. However, the existing VLM-based HOI detectors typically adopt a hand-crafted template (e.g., a photo of a person [**action**] a/an [**object**]) to acquire text knowledge, and they may suffer from learning ambiguity by only encoding the action-specific text prompt in vocabulary level.

As shown in Fig. 1(a), the VLM-based method encodes visual feature and text prompt separately, and then conducts cross-modal learning for HOI classification. Here, the HOI text prompt is generated by only using "**action**" and "**object**", which may cause learning ambiguity for similar interactions, e.g., "a person sitting a bed", "a person lying a bed" and "a person sitting a chair". The three text prompts have high similarity in the embedding space, and can't provide discriminative guidance for the visual encoder. Thus, we first investigate what are the essential elements for an interaction context, and then establish a syntactic interaction bank from three levels: *spatial relationship*, *action-oriented posture* and *situational condition*, as shown in Fig. 1(b). For the learning ambiguity, the interaction bank can provide informative and clear clues to distinguish different interactions. E.g., for "a person sitting a bed", the *spatial relationship* supplies preposition prompt "a person on a bed", the *action-oriented posture* offers posture prompt "a person sitting with bent leg", and the *situational condition* provides the environment prompt "a person in the bedroom". To align visual features with the syntactic interaction bank, we also propose a multi-view extractor to jointly aggregate visual features from instance, interaction, and image levels accordingly. Besides, we introduce a dual cross-attention decoder to facilitate context propagation and exchange between prior text prompts and the visual features.

Thus, in this paper, our motivation is to discover the most valuable text prompts to enhance the HOI detection. First, we adopt different well-designed templates to generate HOI text prompts from three levels for a specific interaction: spatial relationship ("a person [**preposition**] a/an [**object**]"), action-oriented posture ("a person [**action**] with [**action-oriented posture**]") and situational condition ("a person in the [**environment**]"). Unlike the previous methods, our work aims to exploring fine-grained text clues from the perspective of interaction context. It minimizes the representation gap between visual feature and text knowledge, and thus can promote the HOI detection. To summarize, our contributions are four-fold:

- To eliminate the learning ambiguity of handcrafted text prompt, we establish a syntactic interaction bank from three levels: spatial relationship, action-oriented posture and situational condition.
- Guided by the syntactic interaction bank, we introduce a multi-view extractor to jointly aggregate visual features from instance, interaction, and image levels accordingly.
- We propose a dual cross-attention decoder to facilitate context propagation and exchange between prior text prompts and the visual features, thereby enhancing the HOI detection.
- We evaluate our proposed SICHOI on two public benchmarks: V-COCO and HICO-DET, and it can achieve superior performance than other state-of-the-art methods (V-COCO of 71.1, HICO-DET of 45.04).

2. Related Work

2.1. CNN-Based HOI Detection

Early HOI detection methods are typically relied on CNN architectures, which can be divided into one-stage and two-stage methods. The early one-stage methods [20, 29, 49, 52] attempt to perform object detection and interaction classification simultaneously, by introducing auxiliary priors. E.g., UnionDet [20] groups the pairs of humans and objects by a union region, while PPDM [29] additionally predicts the interaction point to regularize the human and object detection. In other hand, the two-stage ones [1, 10, 11, 14, 23, 34, 45, 47, 54, 64] firstly localize humans and objects, and then reason the semantic relationships between each pair of human and object. To learn discriminative interaction feature for HOI detection, they further explore spatial relationship [34], human pose [14], gaze attention [54] and semantic feature [1, 11, 23, 34, 64] with extra branches. Additionally, VSGNet [45] adopts Graph Convolutional Network (GNN) [42] to aggregate interaction feature by regarding humans/objects as nodes and their actions as edges. In contrast, CHGN [47] models different humans and objects as different kinds of nodes, and incorporates interactions both from intra-class nodes and inter-class nodes, respectively.

2.2. Transformer-Based HOI Detection

Transformer [46], with a good capability of capturing the long-range dependency, has advanced many computer vision tasks. Similar to the CNN-based counterparts, the transformer-based HOI detectors can also be categorized into one-stage [2, 19, 24, 31, 43, 66, 69] and two-stage methods [22, 33, 53, 59]. For one-stage ones, HOIFormer [68] pioneeringly predicts the HOI triplets in an *end-to-end* manner where a quintuple matching loss is proposed to enable a unified supervision. HOTR [21] designs

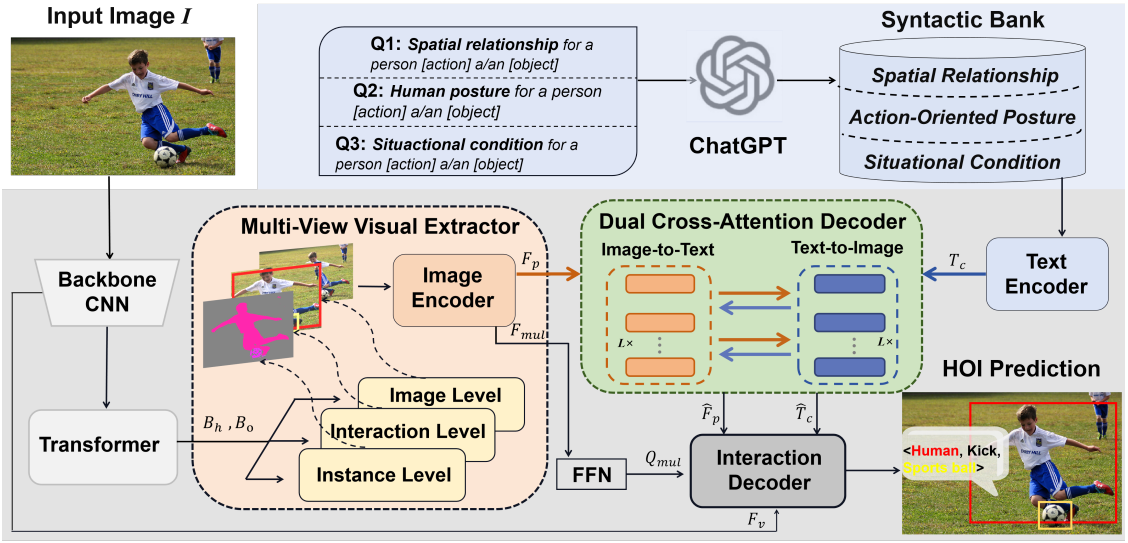

Figure 2. The pipeline of SICHOI model. It mainly includes three components: Multi-View Visual Extractor, Syntactic Bank and Dual Cross-Attention Decoder. Firstly, an input image I is fed into CNN and transformer to obtain the human and object bounding boxes, and then the visual features are extracted from instance, interaction and image levels, respectively. Meanwhile, we generate the syntactic prompts by asking ChatGPT three well-defined questions, which are then encoded into textual features. Finally, a Dual Cross-Attention Decoder is introduced to facilitate context propagation between visual and textual features to enhance HOI detection.

two parallel decoders: the instance decoder responsible for object detection, while the interaction decoder focuses on the interaction representations. However, such *end-to-end* methods may suffer from insufficient exchange between contextual clues, leading to inferior performance. Thus, MUREN [24] proposes a multiplex relation network to perform context exchange between three decoder branches by modelling relations of human, object and interaction features. As for the two-stage methods, they usually utilize pre-trained detectors to obtain object detections, and pay more attention on extracting interaction context from candidate HOI pairs. E.g., STIP [61] first produces candidate HOI pairs, and then exploit structure-aware priors to further enhance HOI detection. However, this method ignores the cross-triplet correlations. To enrich HOI features, ERNet [31] uses a multi-scale deformable transformer to refine instance and interaction tokens, respectively, and then adopts a post-processing to group HOI triplets.

2.3. VLM-Based HOI Detection

Recently, breakthroughs in VLM exhibit promising transfer ability for many downstream tasks. Existing VLM-based methods can be classified into three categories based on their ways to leverage VLM. One category is to only adopt the image encoder of VLM to extract visual knowledge. E.g., ViPLIO [39] replaces the traditional CNN extractor with the VLM image encoder due to its inherent advantage in language-image alignment. Besides, HOICLIP [38] utilizes a query-based retrieval to transfer prior visual knowledge from CLIP to visual feature map via a cross-attention mechanism. In contrast, the approaches [19, 40, 44, 51, 56, 57, 63] only utilize the VLM text encoder to integrate linguistic prior knowledge for HOI detection. E.g., EoID [51] transfers distribution of action probability from CLIP to the visual classification through knowledge distillation, and thus it can perform zero-shot HOI detection. AGER [44] uses textual prior to guide the learning of the instance encoder by enforcing a similarity between the textual representation and the instance token representation. Recently, some methods [7, 26, 30] apply both image and text encoders together to HOI detection. Specifically, ADA-CM [26] constructs a balanced concept-guided memory that jointly leverages domain-specific visual knowledge and domain-agnostic text knowledge to adaptively inject instance knowledge.

Though the VLM-based HOI detectors have achieved significant progress, such approaches typically encode a hand-crafted template to acquire prior interaction knowledge, which may suffer from learning ambiguity. To address the problem, we introduce a syntactic interaction bank containing fine-grained text prompts, to provide discriminative and clear interaction clues for HOI detection. Furthermore, we introduce a multi-view extractor to jointly aggregate visual knowledge and a dual cross-attention decoder to facilitate context propagation.

3. Method

3.1. Overall Architecture

The overall architecture of our proposed SICHOI is illustrated in Fig. 2. Firstly, for a given image I, we utilize ResNet [15] as backbone to obtain a spatial image

Table 1. Examples of fine-grained prompts in the syntactic bank. **AOP**, **SP** and **SC** represent the *action-oriented posture*, the *spatial relationship* and the *situational condition respectively*.

HOI	AOP	SR	SC
<ride, skis>	Bent leg	On a skis	Snowfield
<lie , bed>	Horizontal body	On a bed	Bedroom
<sit , bed>	Bent leg	On a bed	Bedroom
<lift up, fork>	Straight arm Clenched hand	A fork in hand	✗
<point, laptop>	Finger towards	A laptop near	✗
<kick, sports ball>	Swinged leg	Foot next to a sport ball	Sports field
<wave, bus>	Swinged arm	A bus near	Street

feature F_v, followed by a transformer to localize humans $B_h \in R^{N_h \times 4}$ and objects $B_o \in R^{N_o \times 4}$ (i.e., the DETR detector), where N_h and N_o represent the number of humans and objects, respectively. We then generate instance-level image I_{in} and interaction-level image I_{un} with B_h, B_o and I. The image encoder of BLIP [27] is subsequently adopted to jointly aggregate visual knowledge from instance, interaction and image levels, forming multi-view visual features F_{mul}. The fine-grained interaction Query Q_{mul} is thus obtained through imposing a Feed Forward Networks (**FFN**) on F_{mul}. Secondly, we apply the text encoder of BLIP to encode the text prompts from the syntactic text bank, and generate the contextual features T_c. The dual cross-attention decoder is then introduced to facilitate context propagation between T_c and the whole image patch tokens F_p to obtain $\hat{T}_c$ and $\hat{F}_p$. Finally, we perform interaction recognition through a decoder using the fine-grained Q_{mul} as the query, while considering $\hat{T}_c$, $\hat{F}_p$ and F_v collectively as the key and value.

3.2. Syntactic Interaction Bank (SIB)

The existing VLM-based HOI detectors typically adopt a hand-crafted template (e.g., a photo of a person [**action**] a/an [**object**]) to obtain text knowledge. However, such approaches, only encoding the action-specific text prompt in vocabulary level, may suffer from learning ambiguity when two interactions closely resemble each other (e.g., "a person sitting a bed" and "a person lying a bed"). In this work, we find that *spatial relationship*, *action-oriented posture* and *situational condition* are three main factors for a HOI context. Thus, we aim to explore the syntactic interaction clues rather than the simple action prompt for HOI detection.

Generation of Syntactic Prompts. By analyzing a large number of HOI triplets, we believe that the occurrence of an interaction depends on three factors: *spatial relationship*, *action-oriented posture* and *situational condition*. Besides, we ask ChatGPT* "How to judgement whether "a person action a/an object" in an image from visual perspective", and the response can also be summarized as the above three

*https://chat.openai.com

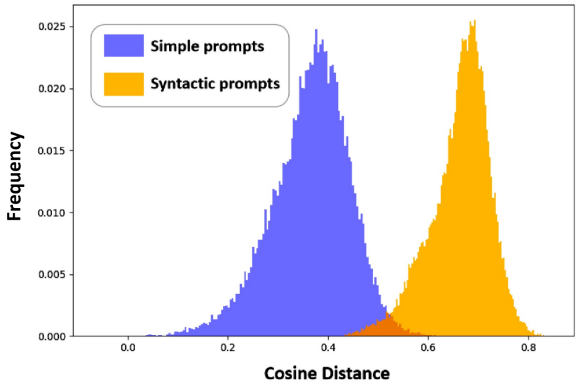

Figure 3. Distance distributions for simple prompts and syntactic prompts, respectively. On HICO-DET dataset, we calculate the cosine distances between different HOI textual features for each kind of text prompts.

factors. Thus, for each HOI category, we design three independent question templates accordingly:

Q1: When a person [**action**] a/an [**object**], what is the *spatial relationship* between the person and the object?

Q2: How to judgement whether a person [**action**] a/an [**object**] from *human posture perspective*?

Q3: What is the *situational condition* for a person [**action**] a/an [**object**]?

Through ChatGPT's answers, we can achieve syntactic and fine-grained text prompts from *spatial relationship*, *action-oriented posture* and *situational condition*, respectively. In Tab. 1, we list some typical interactions. Taking the triplet < human, ride, skis > for example, the syntactic text prompt can be summarized as "a person riding **on a skis** with **bent leg** in the **snowfield**" that considering all the three factors, and this is different from the handcrafted prompt "a photo of a person riding a skis". Each HOI category can generate a syntactic text prompt, and thus forming a Syntactic Interaction Bank on the whole dataset.

Fine-Grained Interactive Priors. For a given syntactic interaction bank, it consists of syntactic text prompt for each HOI category, and we denote it as $P_b = \left(p_b^1, p_b^2, \cdots p_b^i \cdots, p_b^{N_b}\right)$, where p_b^i represents the text prompt of the i^{th} HOI category, and N_b is the number of HOI instances in a given dataset. We apply the text encoder **TextEnc** $(\cdot)$ of BLIP to extract text knowledge $T_e = \left(t_e^1, t_e^2, \cdots t_e^i \cdots, t_e^{N_b}\right)$ from P_b, where t_e^i represents the text embedding of the i^{th} HOI category. Then, we further project T_e with a linear projection layer **Proj** $(\cdot)$ to acquire the contextual features $T_c = \left(t_c^1, t_c^2, \cdots t_c^i \cdots, t_c^{N_b}\right)$ as follows:

$$T_e = \textbf{TextEnc}\left(P_b\right), \tag{1}$$

$$T_c = \textbf{Proj}\left(T_e\right). \tag{2}$$

For HICO-DET [4] dataset, we first generate the handcrafted prompts and the syntactic prompts for all the HOI

categories, respectively. Then, we calculate the distances between different HOI textual features for each kind of text prompts. As shown in Fig. 3, the average cosine distance for syntactic prompts is about 0.7, which is larger than that of simple prompts (the average distance is about 0.4). This implies that the syntactic prompts are powerful to discriminate each HOI category, and thus can eliminate learning ambiguity for similar interactions.

3.3. Multi-View Visual Extractor (MVVE)

For the syntactic prompts, their interactive priors are from three levels: *spatial relationship*, *action-oriented posture* and *situational condition*. To align the visual features with the textual features, we adopt a multi-view extractor to aggregate visual knowledge from instance level, interaction level and image level accordingly. Specifically, we introduce SAM [25] to recognize instance segmentation for humans and objects within I to generate instance-level image I_{in}. Then, the BLIP image encoder **ViEnc(·)** encodes I_{in} to extract the spatial relationship for each human-object pair. In addition, the interaction-level image I_{un} is the union box that covers a candidate pair of human box and object box, which contains interaction clues, and can be used to extract action posture features. Further, to obtain the situational condition features, we take the image-level image I as input to **ViEnc(·)**, undertaking whole scene understanding.

Finally, we adopt a **FFN** to jointly fuse the three-level visual knowledge as follows:

$$F_{mul} = \textbf{Contact}(\textbf{ViEnc}(I_{in}), \textbf{ViEnc}(I_{un}), \textbf{ViEnc}(I)), \tag{3}$$

$$Q_{mul} = \textbf{FFN}(F_{mul}), \tag{4}$$

where **Contact** denotes the concatenation operator, and Q_{mul} is the human-object interaction query from the multi-view visual features.

3.4. Dual Cross-Attention Decoder (DCAD)

Dual Cross-attention Propagation. To facilitate context propagation between text knowledge and the visual features, we introduce a dual cross-attention decoder to perform bidirectional attentions for mutual enhancement of the two modalities. As shown in Fig. 4, from the multi-view visual extractor, we can also generate two regional image masks: instance mask M_m and union mask M_u, which can be used to strengthen the awareness of image potential structures. For each dual cross-attention block, we perform text-to-image cross attention by

$$\hat{T}_c = \textbf{Softmax}\left(\frac{T_c^T(F_p + \phi_m(M_m) + \phi_u(M_u))}{\sqrt{d}}\right) F_p, \tag{5}$$

where the contextual features T_c and the image patch tokens F_p = **ViEnc**(I) are taken as queries and values, respectively, and the image patch tokens with regional embeddings

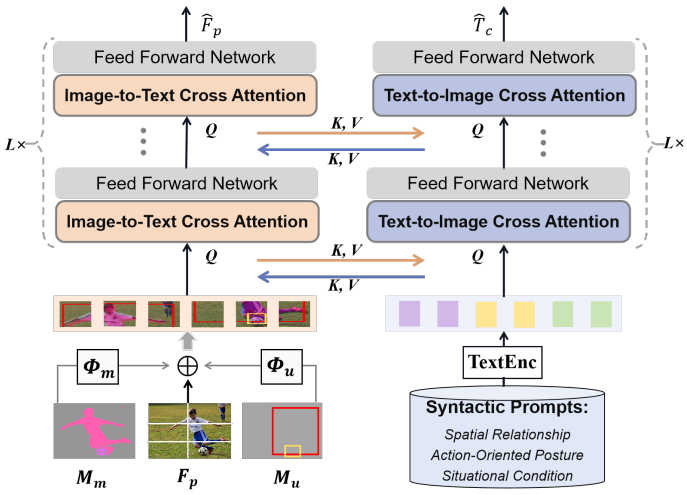

Figure 4. The architecture of dual cross-attention decoder. For each dual cross-attention block, we simultaneously conduct text-to-image and image-to-text cross attention using patch image tokens F_p, contextual features T_c and regional embeddings $\phi_m(M_m)$ and $\phi_u(M_u)$.

$F_p + \phi_m(M_m) + \phi_u(M_u)$ are used as keys. Here, ϕ_m and ϕ_u denote the instance mask projection and the union mask projection, respectively. Meanwhile, we can also conduct image-to-text cross attention by

$$\hat{F}_p = \textbf{Softmax}\left(\frac{(F_p + \phi_m(M_m) + \phi_u(M_u))^T T_c}{\sqrt{d}}\right) T_c, \tag{6}$$

where the image patch tokens with regional embeddigns $F_p + \phi_m(M_m) + \phi_u(M_u)$ are regarded as queries, and the contextual features T_c are applied as both keys and values.

Interaction Decoder. Finally, we perform interaction recognition through a decoder using the fine-grained Q_{mul} as queries, while considering the image patch tokens $\hat{F}_p$, contextual features $\hat{T}_c$ and image features F_v from ResNet jointly as the keys and values. The final interaction scores $\hat{\mathbf{y}}$ can be obtained as follows:

$$\hat{\mathbf{y}} = \textbf{Decoder}(Q_{mul}, \textbf{Concat}(\hat{T}_c, \hat{F}_p, (F_v + Pos))), \tag{7}$$

where *Pos* indicates the position embeddings for feature map F_v. To train this network, we apply the following Focal Loss (**FL**):

$$\mathcal{L}_{sic} = \frac{1}{\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbf{y}_{i,c}} \sum_{i=1}^{N}\sum_{c=1}^{C} \textbf{FL}(\hat{\mathbf{y}}_{i,c}, \mathbf{y}_{i,c}), \tag{8}$$

where N is the number of candidate human-object pairs, C is the number of interaction classes, $\mathbf{y}_{i,c} \in \{0, 1\}$ in $\mathbf{y}$ indicates whether the groundtruth of the i-th human-object pair contains the c-th interaction class and $\hat{\mathbf{y}}_{i,c}$ is the corresponding predicted probability from the interaction decoder.

Table 2. Performance comparison on HICO-DET and V-COCO datasets. The best result is marked with bold and the second best result is underlined. For results on HICO-DET, we follow commonly used experimental setting to finetune the object detector on its training set.

Method	Backbone	HICO-DET						V-COCO	
		Default			Known Object				
		Full	Rare	Non-Rare	Full	Rare	Non-Rare	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
CNN-based methods									
InteractNet [12]	R50-FPN	9.94	7.16	10.77	-	-	-	40.0	48.0
UnionDet [20]	R50-FPN	17.58	11.72	19.33	19.76	14.68	21.27	47.5	56.2
IP-Net [49]	HG-104	19.56	12.79	21.58	22.05	15.77	23.92	51.0	-
GPNN [67]	R50	19.42	13.98	20.91	22.01	15.73	22.80	50.4	-
ACP [23]	R152	20.59	15.92	21.98	-	-	-	53.2	-
Transformer-based methods									
HOI-Trans [69]	R101	26.61	19.15	28.84	29.13	20.98	31.57	52.9	-
AS-Net [5]	R50	28.87	24.25	30.25	31.74	27.07	33.14	53.9	-
QPIC [43]	R50	29.07	21.85	31.23	31.68	24.14	33.93	58.8	61.0
PhraseHOI [28]	R50	29.29	22.03	31.46	31.97	23.99	34.36	57.4	-
MSTR [22]	R50	31.17	25.31	32.92	34.02	28.83	35.57	62.0	65.2
DT [65]	R50	31.75	27.45	33.03	34.50	30.13	35.81	66.2	68.5
CDN [58]	R50	31.78	27.55	33.05	34.53	29.73	35.96	62.3	64.4
CATN [6]	R50	31.86	25.15	33.84	34.44	27.69	36.45	60.1	-
STIP [61]	R50	32.22	28.15	33.43	35.29	31.43	36.45	66.0	70.7
UPT [59]	R50	31.66	25.94	33.36	35.65	31.60	36.86	59.0	64.5
ParMap [52]	R50	35.15	33.71	35.58	37.56	35.87	38.06	63.0	65.1
ERNet [31]	EfficientNetV2-XL	35.92	30.12	38.29	-	-	-	64.2	-
CQL [53]	R101	36.03	33.16	36.89	38.82	35.51	39.81	66.5	69.9
RmLR [2]	R101	37.41	28.81	39.97	38.69	31.27	40.91	64.2	70.2
PViC [60]	R50	34.69	32.14	35.45	38.14	35.38	38.97	62.8	67.8
PViC [60]	Swin-L	44.32	44.61	44.24	47.81	48.38	47.64	64.1	70.2
VLM-based methods									
OpenCat [63]	R101+ViT-B/16	32.68	28.42	33.75	-	-	-	61.9	63.2
GEN-VLKT [30]	R50+ViT-B/16	33.75	29.25	35.10	36.78	32.75	37.99	62.4	64.5
RLIPv2 [57]	Swin-T	33.66	40.07	38.60	-	-	-	68.8	70.8
HOICLIP [38]	R50+ViT-B/32	34.69	31.12	35.74	37.61	34.47	38.54	63.5	64.8
DiffHOI [55]	R50+ViT	34.41	31.07	35.40	37.31	34.56	38.14	61.1	63.5
AGER [44]	R50	36.75	33.53	37.71	39.84	35.58	40.23	65.7	69.7
ViPLO [39]	R50+ViT-B/16	37.22	35.45	37.75	40.61	38.82	41.15	62.2	68.0
ADA-CM [26]	R50+ViT-L	38.40	37.52	38.66	-	-	-	58.6	64.0
DiffHOI [55]	Swin-L+ViT	41.50	39.96	41.96	43.62	41.41	44.28	65.7	68.2
SICHOI (Ours)	R50+ViT-B/16	41.79	42.38	41.61	44.27	43.64	44.46	67.9	72.8
SICHOI (Ours)	R101+ViT-L/16	**45.04**	**45.61**	**44.88**	**48.16**	**48.37**	**48.09**	**71.1**	**75.6**

4. Experiments

4.1. Experimental Setting

Datasets. V-COCO [13] is a subset of MS-COCO [32], consisting of 5,400 images in the trainval set and 4,946 images in the test set. It has 259 HOI categories over 29 actions and 80 objects. HICO-DET [4] consists of 38,118 images in training set and 9,658 in test set, and has 600 HOI categories over 117 actions and 80 objects. The 600 HOI categories are split into 138 Rare and 462 Non-Rare based on the number of instances.

Evaluation Metric. Following the standard metric, we use the mean Average Precision (mAP) to evaluate the model performance for the two benchmarks. A HOI triplet is considered as true positive if it localizes the human and object accurately (i.e., the Interaction-over-Union (IOU) between the predicted bounding boxes and ground truth is greater than 0.5) and also predicts the action correctly.

Implementation Details. We take the DETR for object detection, and leverage the pre-trained BLIP [27] on its official data as the VLM. We keep the external models (i.e., DETR and VLM) fixed during training, and other parts of the SICHOI model are trained on four Nvidia 3090 GPUs in an end-to-end way, with a batch size 16. The AdamW [35] optimizer is used for training with 30 epochs, where the starting learning rate is 5×10^{-5}, and then decays with the Cosine annealing training strategy.

4.2. Comparisons with the State-of-the-Arts

We report the quantitative results in terms of AP on HICO-DET and V-COCO dataset, respectively.

Tab. 2 shows the performance comparison of SICHOI and other state-of-the-art methods on HICO-DET dataset. It can be observed that SICHOI outperforms all existing methods. Specifically, SICHOI achieves **45.04** mAP in the default full setting, obtaining a performance gain of 0.72 mAP (relatively 1.62%) compared to the most recent approach PViC [60]. Also, compared with DiffHOI [55] and

RmLR [2], which are the state-of-the-art VLM-based and transformer-based methods, our model achieves a significant performance gain of 3.54 (relatively 8.53%) mAP and 7.63 mAP (relatively 20.40%), respectively. The results validate the superiority of our SICHOI. HOICLIP [38] and ViPLO [39] only employ the visual encoder of CLIP to encode the image features, but leave the textual information that contains semantic knowledge unexplored. ADA-CM [26] and DiffHOI [55] integrate the visual and text knowledge into one framework and facilitate the knowledge propagation between different modalities. However, they solely encode the action-specific text prompts in vocabulary level, e.g., "a photo of a person [action] a/an [object]", and may suffer from learning ambiguity without exploring the fine-grained clues from the perspective of interaction context. For our SICHOI, it establish syntactic interaction descriptions from three levels: spatial relationship, action-oriented posture and situational condition. Further, to align visual features with the syntactic interaction bank, we adopt the multi-view visual extractor to aggregate visual knowledge from instance, interaction and image level accordingly, and thus can enhance the HOI detection performance.

For V-COCO dataset, as reported in the right part of Tab. 2, SICHOI also performs the best among all the state-of-the-art methods. E.g., SICHOI works better than two recent HOI detectors RLIPv2 [57], and CQL [53] ($AP^{\#1}_{role}$ of 71.1 vs 68.8 and 66.5). It is noted that $AP^{\#2}_{role}$ is significantly improved by the SICHOI (4.8 higher than RLIPv2), and the improvement may be attributed to our model's ability to fully utilize scene semantics to infer missed or occluded information. E.g., when certain parts of humans or objects are occluded, SICHOI can rely on the context knowledge from the syntactic interaction bank to infer missing information or predict the likely actions.

In Tab. 3, we further conduct zero-shot comparison with the state-of-the-arts on HICO-DET dataset, following the experimental settings [30, 51]. Using ResNet50 as backbone, the proposed SICHOI model achieves gains of 6.61 mAP (relatively 23.92%) and 2.11 mAP (relatively 6.51%) on the two zero-shot settings, respectively, compared to the best performing approach ADA-CM [26]. The SICHOI can be further enhanced when it is equipped with ResNet101 as backbone (E.g., for NF, mAP is improved from 35.75 to 39.07). These improvements demonstrate the good generalization ability of our model for detecting HOIs belonging to unseen combinations. It is worth noting that under the Rare First setting, our model shows great superiority on unseen samples, which are corner cases that are challenging to be detected, over other methods. This proves that our syntactic interaction bank is very helpful when encountering the rare and unseen samples.

Table 3. Zero-shot comparison on HICO-DET. This table compares our model with state-of-the-art methods on the Zero-shot setting of HICO-DET. RF: Rare First. NF: Non-rare First.

Method	Backbone	Type	Unseen	Seen	Full
VCL [16]	R50	RF	10.6	24.28	21.43
ATL [17]	R50-FPN	RF	9.18	24.67	21.57
FCL [18]	R50	RF	13.16	24.12	22.01
THID [48]	ViT-B/16	RF	15.53	24.32	22.96
RLIPv1 [56]	R50	RF	19.19	33.35	30.52
GEN-VLKT [30]	R50+ViT-B/16	RF	21.36	32.91	30.56
HOICLIP [38]	R50+ViT-B/32	RF	25.53	34.85	32.99
RLIPv2 [57]	Swin-T	RF	26.95	39.92	37.32
ADA-CM [26]	R50+ViT-B/16	RF	27.63	34.35	33.01
SICHOI	R50+ViT-B/16	RF	34.24	41.58	40.11
SICHOI	R101+ViT-L/16	RF	**36.27**	**44.71**	**43.02**
VCL [16]	R50	NF	16.22	18.52	18.06
ATL [17]	R50-FPN	NF	18.25	18.78	18.67
FCL [18]	R50	NF	18.66	19.55	19.37
RLIPv1 [56]	R50	NF	20.27	27.67	26.19
RLIPv2 [57]	Swin-T	NF	21.07	35.07	32.27
GEN-VLKT [30]	R50+ViT-B/16	NF	25.05	23.38	23.71
HOICLIP [38]	R50+ViT-B/32	NF	26.39	28.10	27.75
ADA-CM [26]	R50+ViT-B/16	NF	32.41	31.13	31.39
SICHOI	R50+ViT-B/16	NF	34.52	36.06	35.75
SICHOI	R101+ViT-L/16	NF	**36.44**	**39.73**	**39.07**

4.3. Ablation Studies

In this subsection, we explore how the Syntactic Interaction Bank, Multi-View Visual Extractor, and Dual Cross-attention Decoder affect the HOI detection performance. For simplicity, we adopt R50 as backbone, and all the ablation studies are conducted on HICO-DET and V-COCO datasets.

To evaluate the effect of each component of SICHOI, we create a baseline mode (denoted as "Base") by only using plain DETR and transformer (i.e., without SIB, MVVE and DCAD). As summarized in Tab. 4, each component of our SICHOI can significantly advance the baseline model. E.g., SIB can improve the baseline model by 3.63 mAP and 1.81 mAP in the Full category of HICO-DET and $AP^{\#1}_{role}$ of V-COCO, respectively. In the following subsections, we will conduct in-depth analysis of the impact for each component on the model performance.

4.3.1 Syntactic Interaction Bank

We design the Syntactic Interaction Bank (SIB) to eliminate the learning ability of handcrafted text prompt. It is composed of three levels of prompts: *Action-Oriented Posture* (AOP), *Spatial Relationship* (SR), and *Situational Condition* (SC). As shown in Tab. 5, we explore the effects of employing different prompts to provide textual knowledge. For fair comparison, all the variants are based on the "Base+MVVE" model in the first row. As observed, using each level of syntactic prompts performs better than using the simple prompt (denoted as "+SP"). E.g., the Action-oriented Posture ("+AOP") achieves gains of 1.16 mAP and 0.31 mPA on HICO-DET and V-COCO, respectively. The

Table 4. Performance contribution of each component in our SICHOI. **SIB**: Syntactic Interaction Bank. **MVVE**: Multi-View Visual Extractor. **DCAD**: Dual Cross-Attention Decoder.

Method	HICO-DET (Default)			V-COCO	
	Full	Rare	Non-Rare	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
Base	35.71	33.19	36.46	64.72	69.81
+SIB	39.34	39.03	39.44	66.53	71.12
+MVVE	38.54	36.33	39.20	66.37	70.65
+SIB+MVVE	40.57	40.45	40.61	67.37	72.29
+SIB+MVVE+DCAD	**41.79**	**42.38**	**41.61**	**67.93**	**72.83**

Table 5. Performance comparison of different text prompts. **SP**: Simple Prompts. **AOP**: Action-oriented Posture. **SR**: Spatial Relationship. **SC**: Situational Condition.

Method	HICO-DET (Default)			V-COCO	
	Full	Rare	Non-Rare	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
Base+MVVE	38.54	36.33	39.20	66.37	70.65
+SP	39.66	38.14	40.11	66.90	71.75
+AOP	40.82	40.45	40.93	67.21	72.19
+SR	40.43	40.17	40.51	67.12	72.04
+SC	40.27	39.72	40.43	67.02	71.96
+AOP+SR+SC	**41.79**	**42.38**	**41.61**	**67.93**	**72.83**

performance can be further enhanced from 39.66 to 41.79 on HICO-DET and from 66.90 to 67.93 on V-COCO by integrating all the syntactic prompts from three levels, demonstrating the effectiveness of syntactic prompts.

4.3.2 Multi-View Visual Extractor

To evaluate the effectiveness of Multi-View Visual Extractor (MVVE), we create variants on "Base+SIB+DCAD" by using different visual features (i.e., instance, interaction and image level). As reported in Tab. 6, visual feature at each level works better than the baseline model "Base+SIB+DCAD". Among the three levels, the instance one performs the best, and it can improve "Base+SIB+DCAD" from 39.95 to 41.25 in the Full category of HICO-DET. The reason is that the instance segmentation implicitly contains information of spatial relation and human posture. Also, the background noise can be suppressed with INS. The result of "+INS+INT+IMG" indicates that it is essential to align the visual features with the text knowledge from the syntactic interaction bank.

4.3.3 Dual Cross-Attention Decoder

In SICHOI, we introduce the Dual Cross-Attention Decoder (DCAD) to facilitate context propagation between visual and textual features. To evaluate the effectiveness of DCAD, we also create variants on "Base+SIB+MVVE", by using different cross-attention strategies. E.g., V→L indicates that the visual feature and linguistic feature are treated as query and value in the cross-attention decoder, respectively, and vice versa. As summarized in Tab. 7, the dual cross-attention mode (+V→L+L→V) performs the best on the two datasets, which implies that the context propagation between the visual and contextual features is important to improve the HOI performance.

Table 6. Performance comparison of different visual features. **INS**, **INT**, and **IMG** indicate that the visual features are from Instance, Interaction, and Image levels, respectively.

Method	HICO-DET (Default)			V-COCO	
	Full	Rare	Non-Rare	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
Base+SIB+DCAD	39.95	39.67	40.04	66.87	71.56
+INS	41.25	41.72	41.11	67.45	72.08
+INT	41.14	41.67	40.98	67.36	72.16
+IMG	40.94	41.21	40.86	67.15	71.98
+INS+INT+IMG	**41.79**	**42.38**	**41.61**	**67.93**	**72.83**

Table 7. Effect of dual cross-attention decoder. **V→L** indicates that the vision feature and linguistic feature are treated as query and value, respectively, and vice versa.

Method	HICO-DET (Default)			V-COCO	
	Full	Rare	Non-Rare	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
Base+SIB+MVVE	40.57	40.45	40.61	67.37	72.29
+V→L	41.17	41.32	41.12	67.56	72.56
+L→V	41.35	41.57	41.29	67.69	72.80
+V→L+L→V	**41.79**	**42.38**	**41.61**	**67.93**	**72.83**

5. Conclusion

In this paper, we propose the SICHOI to discover the fine-grained prompts for HOI detection. Different from the existing VLM-based methods, the proposed model establishes a syntactic bank to explore text knowledge from three levels: *spatial relationship*, *action-oriented posture* and *situational condition*. To ensure alignment to the textual features, we adopt a multi-view extractor to aggregate visual features from instance, interaction and image levels accordingly. Also, a dual cross-attention decoder is designed to facilitate context propagation between visual feature and textual features. Experimental results indicate that our proposed SICHOI can achieve state-of-the-art results on HICO-DET and V-COCO datasets.

Acknowledgment. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB4703201, in part by the National Natural Science Foundation of China under Grants 62206075, 61733011, 62261160652, 62073205, U23A20343 and 61821005, in part by the Guang-Dong Basic and Applied Basic Research Foundation under Grant 2021A1515110438, in part by the Shenzhen Science and Technology Program under Grant RCBS20221008093220004, in part by CAS Project for Young Scientists in Basic Research under Grant YSBR-041, in part by Youth Innovation Promotion Association of the Chinese Academy of Sciences under Grant Y202051, and in part by the State Key Laboratory of Robotics under Grant 2023-O06.

References

[1] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. In *AAAI Conference on Artificial Intelligence*, pages 10460–10469, 2020. 2

[2] Yichao Cao, Qingfei Tang, Feng Yang, Xiu Su, Shan You, Xiaobo Lu, and Chang Xu. Re-mine, learn and reason: Exploring the cross-modal semantic correlations for language-guided hoi detection. In *IEEE International Conference on Computer Vision*, pages 23492–23503, 2023. 2, 6, 7

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1

[4] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *IEEE International Conference on Computer Vision*, pages 1017–1025, 2015. 4, 6

[5] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9004–9013, 2021. 6

[6] Leizhen Dong, Zhimin Li, Kunlun Xu, Zhijun Zhang, Luxin Yan, Sheng Zhong, and Xu Zou. Category-aware transformer network for better human-object interaction detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 19538–19547, 2022. 6

[7] Shuman Fang, Shuai Liu, Jie Li, Guannan Jiang, Xianming Lin, and Rongrong Ji. Improving human-object interaction detection via virtual image learning. In *ACM International Conference on Multimedia*, pages 5455–5463, 2023. 3

[8] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4768–4777, 2017. 1

[9] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *European Conference on Computer Vision*, pages 701–717. Springer, 2022. 2

[10] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018. 2

[11] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *European Conference on Computer Vision*, pages 696–712. Springer, 2020. 1, 2

[12] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018. 6

[13] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 6

[14] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *IEEE International Conference on Computer Vision*, pages 9677–9685, 2019. 1, 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[16] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *European Conference on Computer Vision*, pages 584–600. Springer, 2020. 7

[17] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 495–504, 2021. 7

[18] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14646–14655, 2021. 7

[19] ASM Iftekhar, Hao Chen, Kaustav Kundu, Xinyu Li, Joseph Tighe, and Davide Modolo. What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5353–5363, 2022. 2, 3

[20] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *European Conference on Computer Vision*, pages 498–514. Springer, 2020. 1, 2, 6

[21] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 74–83, 2021. 2

[22] Bumsoo Kim, Jonghwan Mun, Kyoung-Woon On, Minchul Shin, Junhyun Lee, and Eun-Sol Kim. Mstr: Multi-scale transformer for end-to-end human-object interaction detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 19578–19587, 2022. 2, 6

[23] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *European Conference on Computer Vision*, pages 718–736. Springer, 2020. 2, 6

[24] Sanghyun Kim, Deunsol Jung, and Minsu Cho. Relational context learning for human-object interaction detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2925–2934, 2023. 2, 3

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 5

[26] Ting Lei, Fabian Caba, Qingchao Chen, Hailin Jin, Yuxin Peng, and Yang Liu. Efficient adaptive human-object interaction detection with concept-guided memory. In *IEEE International Conference on Computer Vision*, pages 6480–6490, 2023. 3, 6, 7

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for uni-

fied vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2, 4, 6

[28] Zhimin Li, Cheng Zou, Yu Zhao, Boxun Li, and Sheng Zhong. Improving human-object interaction detection via phrase learning and label composition. In *AAAI Conference on Artificial Intelligence*, pages 1509–1517, 2022. 6

[29] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 482–490, 2020. 1, 2

[30] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 20123–20132, 2022. 3, 6, 7

[31] JunYi Lim, Vishnu Monn Baskaran, Joanne Mun-Yee Lim, KokSheik Wong, John See, and Massimo Tistarelli. Ernet: An efficient and reliable human-object interaction detection network. *IEEE Transactions on Image Processing*, 32:964–979, 2023. 2, 3, 6

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 6

[33] Xinpeng Liu, Yong-Lu Li, Xiaoqian Wu, Yu-Wing Tai, Cewu Lu, and Chi-Keung Tang. Interactiveness field in human-object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 20113–20122, 2022. 2

[34] Yang Liu, Qingchao Chen, and Andrew Zisserman. Amplifying key cues for human-object-interaction detection. In *European Conference on Computer Vision*, pages 248–265. Springer, 2020. 2

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[36] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14074–14083, 2022. 2

[37] Gyeongsik Moon, Heeseung Kwon, Kyoung Mu Lee, and Minsu Cho. Integralaction: Pose-driven feature integration for robust human action recognition in videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3339–3348, 2021. 1

[38] Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 23507–23517, 2023. 3, 6, 7

[39] Jeeseung Park, Jin-Woo Park, and Jong-Seok Lee. Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 17152–17162, 2023. 1, 3, 6, 7

[40] Xian Qu, Changxing Ding, Xingao Li, Xubin Zhong, and Dacheng Tao. Distillation using oracle queries for transformer-based human-object interaction detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 19558–19567, 2022. 3

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[42] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on Neural Networks*, 20(1): 61–80, 2008. 2

[43] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021. 1, 2, 6

[44] Danyang Tu, Wei Sun, Guangtao Zhai, and Wei Shen. Agglomerative transformer for human-object interaction detection. In *IEEE International Conference on Computer Vision*, pages 21614–21624, 2023. 3, 6

[45] Oytun Ulutan, A S M Iftekhar, and B. S. Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[47] Hai Wang, Wei-shi Zheng, and Ling Yingbiao. Contextual heterogeneous graph network for human-object interaction detection. In *European Conference on Computer Vision*, pages 248–264. Springer, 2020. 2

[48] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2022. 7

[49] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4116–4125, 2020. 1, 2, 6

[50] Hui Wu, Min Wang, Wengang Zhou, Houqiang Li, and Qi Tian. Contextual similarity distillation for asymmetric image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9489–9498, 2022. 1

[51] Mingrui Wu, Jiaxin Gu, Yunhang Shen, Mingbao Lin, Chao Chen, and Xiaoshuai Sun. End-to-end zero-shot hoi detection via vision and language knowledge distillation. In *AAAI Conference on Artificial Intelligence*, pages 2839–2846, 2023. 3, 7

[52] Xiaoqian Wu, Yong-Lu Li, Xinpeng Liu, Junyi Zhang, Yuzhe Wu, and Cewu Lu. Mining cross-person cues for body-part

interactiveness learning in hoi detection. In *European Conference on Computer Vision*, pages 121–136. Springer, 2022. 2, 6

[53] Chi Xie, Fangao Zeng, Yue Hu, Shuang Liang, and Yichen Wei. Category query learning for human-object interaction classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 15275–15284, 2023. 2, 6, 7

[54] Bingjie Xu, Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Interact as you intend: Intention-driven human-object interaction detection. *IEEE Transactions on Multimedia*, 22(6):1423–1432, 2019. 2

[55] Jie Yang, Bingliang Li, Fengyu Yang, Ailing Zeng, Lei Zhang, and Ruimao Zhang. Boosting human-object interaction detection with text-to-image diffusion model. *arXiv preprint arXiv:2305.12252*, 2023. 2, 6, 7

[56] Hangjie Yuan, Jianwen Jiang, Samuel Albanie, Tao Feng, Ziyuan Huang, Dong Ni, and Mingqian Tang. Rlip: Relational language-image pre-training for human-object interaction detection. *Advances in Neural Information Processing Systems*, 35:37416–37431, 2022. 3, 7

[57] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Samuel Albanie, Yining Pan, Tao Feng, Jianwen Jiang, Dong Ni, Yingya Zhang, and Deli Zhao. Rlipv2: Fast scaling of relational language-image pre-training. In *IEEE International Conference on Computer Vision*, pages 21649–21661, 2023. 2, 3, 6, 7

[58] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. *Advances in Neural Information Processing Systems*, 34:17209–17220, 2021. 6

[59] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 20104–20112, 2022. 2, 6

[60] Frederic Z Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring predicate visual context in detecting of human-object interactions. In *IEEE International Conference on Computer Vision*, pages 10411–10421, 2023. 6

[61] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 19548–19557, 2022. 1, 3, 6

[62] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *European Conference on Computer Vision*, pages 159–175. Springer, 2022. 2

[63] Sipeng Zheng, Boshen Xu, and Qin Jin. Open-category human-object interaction pre-training via language modeling framework. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 19392–19402, 2023. 2, 3, 6

[64] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao. Polysemy deciphering network for robust human–object interaction detection. *International Journal of Computer Vision*, 129:1910–1929, 2021. 2

[65] Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. Human-object interaction detection via disentangled transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 19568–19577, 2022. 1, 6

[66] Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. Human-object interaction detection via disentangled transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 19568–19577, 2022. 2

[67] Tianfei Zhou, Siyuan Qi, Wenguan Wang, Jianbing Shen, and Song-Chun Zhu. Cascaded parsing of human-object interaction recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2827–2840, 2021. 6

[68] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, and Jian Sun. End-to-end human object interaction detection with hoi transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11825–11834, 2021. 2

[69] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11825–11834, 2021. 1, 2, 6