

# End-to-End Zero-Shot HOI Detection via Vision and Language Knowledge Distillation

Mingrui Wu<sup>1†\*</sup>, Jiaxin Gu<sup>2†</sup>, Yunhang Shen<sup>2</sup>, Mingbao Lin<sup>2</sup>,  
Chao Chen<sup>2</sup>, Xiaoshuai Sun<sup>1,3,4†</sup>, Rongrong Ji<sup>1,3,4</sup>

<sup>1</sup>MAC Lab, School of Informatics, Xiamen University. <sup>2</sup>Youtu Lab, Tencent.

<sup>3</sup>Institute of Artificial Intelligence, Xiamen University.

<sup>4</sup>Fujian Engineering Research Center of Trusted Artificial Intelligence Analysis and Application, Xiamen University.

mingrui0001@gmail.com, {xssun, rrji}@xmu.edu.cn

{jiaxingu, odysseyshen, marlin, aaronccchen}@tencent.com

**Abstract.** Most existing Human-Object Interaction (HOI) Detection methods rely heavily on full annotations with predefined HOI categories, which is limited in diversity and costly to scale further. We aim at advancing zero-shot HOI detection to detect both seen and unseen HOIs simultaneously. The fundamental challenges are to discover potential human-object pairs and identify novel HOI categories. To overcome the above challenges, we propose a novel **End-to-end zero-shot HOI Detection (EoID)** framework via vision-language knowledge distillation. We first design an *Interactive Score* module combined with a *Two-stage Bipartite Matching* algorithm to achieve interaction distinguishment for human-object pairs in an action-agnostic manner. Then we transfer the distribution of action probability from the pretrained vision-language teacher as well as the seen ground truth to the HOI model to attain zero-shot HOI classification. Extensive experiments on HICO-Det dataset demonstrate that our model discovers potential interactive pairs and enables the recognition of unseen HOIs. Finally, our method outperforms the previous SOTA by 8.92% on unseen *mAP* and 10.18% on overall *mAP* under UA setting, by 6.02% on unseen *mAP* and 9.1% on overall *mAP* under UC setting. Moreover, our method is generalizable to large-scale object detection data to further scale up the action sets. The source code will be available at: <https://github.com/mrwu-mac/EoID>.

**Keywords:** HOI Detection; Zero-Shot; Vision and Language Model; Knowledge Distillation

## 1 Introduction

The task of Human-Object Interaction (HOI) detection aims to detect  $\langle human, verb, object \rangle$  triplets, which simultaneously localizes human-object pairs and identifies the corresponding interactive actions. HOI detection plays an important role in downstream visual understanding tasks, especially for human-centric scenes, such as Image Captioning [14] and Visual Question Answering [6].

---

\* Intern at Youtu Lab, Tencent. † Equal Contribution. ‡ Corresponding Author.



Fig. 1: CLIP shows promising zero-shot ability in detecting *unseen* actions (red) and *unseen* objects (e.g. *tiger*) with *seen* actions (blue). The *unseen* and *seen* here denote the divisions under the zero-shot setting.

Most of the existing works only focus on improving action classification performances for predefined HOI categories. However, they suffer from two major weaknesses: 1) excessive cost for new HOI dataset construction and 2) a lack of generalization for unseen actions. Previous works [1, 10, 18, 24] attempted to overcome the above drawbacks via zero-shot learning. Most of them are devoted to improving human-object visual representation [10] and introducing language model [1, 18], ignoring the implicit relations between vision and language.

As a recent technological breakthrough, CLIP [22] performs contrastive learning on 400 million image-text pairs collected from the Web and shows impressive zero-shot transferability on over 30 classification datasets. Some recent works also successfully transfer the pretrained CLIP model to various downstream tasks such as object detection [7], text-driven image manipulation [21] and semantic segmentation [28]. Compared to text embeddings simply extracted from pure language models, the text embeddings learned jointly with visual images can better encode the visual similarity between concepts [7]. Inspired by this, we attempt to transfer vision and language (V&L) knowledge of CLIP into the zero-shot HOI task. As shown in Fig. 1, CLIP has the ability to identify certain unseen actions (Fig. 1(a)), and also discover unseen objects with seen action interaction (Fig. 1(b)). Despite that, it remains an unexplored challenge to apply the vision and language pretrained model to discover both seen and unseen HOI pairs, without introducing any extra computation at the time of inference.

To this end, we propose EoID, an end-to-end zero-shot HOI detection framework to detect unseen HOI pairs by distilling the knowledge from CLIP. We first design a novel *Interactive Score (IS)* module with a *Two-stage Bipartite Matching* algorithm to discover potential action-agnostic interactive human-object pairs. Then we distill interactive knowledge from CLIP to teach the HOI model to identify unseen actions. Specifically, the probability distribution of the actions is obtained by CLIP given the cropped union regions of human-object pairs, with predefined HOI prompts. Finally, the HOI model learns from the distilled probability distribution as well as the ground truth actions. We evaluate the proposed method on HICO-Det [3] benchmark under unseen action (UA) and

unseen action-object combination (UC) settings. Extensive experiments validate that our framework can detect potential interactive human-object pairs. And the results show that our approach outperforms previous SOTA [18] by 8.92% on unseen  $mAP$  and 10.18% on overall  $mAP$  under UA setting, by 6.02% on unseen  $mAP$  and 9.1% on overall  $mAP$  under UC setting. In addition, our method can generalize to object detection datasets and obtain 47.15%  $mAP$  on unseen actions of V-COCO [9] only with the bounding boxes from MS-COCO [17].

To summarize, our contributions are:

- We propose an end-to-end zero-shot HOI detection framework which attains zero-shot HOI classification via V&L knowledge distillation, which is, as far as we know, the first work to distill interactive knowledge from CLIP.
- We succeed in detecting potential action-agnostic interactive human-object pairs by applying an *Interactive Score* module combined with a *Two-stage Bipartite Matching* algorithm, the effectiveness of which has been validated through extensive experiments.
- Experiments show that EoID is capable of detecting HOIs with unseen HOI categories and outperforms previous SOTA by a large margin under zero-shot settings. Moreover, our method is able to generalize to object detection datasets only with bounding boxes, which further scales up the action sets.

## 2 Related Works

### 2.1 Human-Object Interaction Detection

Most previous works can be categorized into two groups: two-stage and one-stage methods. Two-stage paradigm [3, 4, 15] first applies object detectors to generate human and object instances, and then combines the detected instances exhaustively, which are then fed to interaction classifiers. However, these methods are efficient but complex. In contrast, one-stage studies [5, 11, 12, 16, 25–27, 29] detect HOI triplets directly. Following transformer-based DETR [2] paradigm, some recent methods [12, 25, 27, 29] succeed in detecting HOI triplets in an end-to-end manner, which surpass the two-stage counterparts and previous one-stage methods in both efficiency and effectiveness. Following this paradigm, we build our framework on the transformer-based approach to achieve zero-shot HOI detection. We also replace the one-stage Hungarian matching algorithm [2, 13, 25, 27] with a novel two-stage matching algorithm for detecting potential interactive human-object pairs.

### 2.2 Knowledge Distillation from CLIP

CLIP [22] adopts contrastive learning to jointly train image-text embedding models on large-scale image-text pairs collected from the internet and has shown promising zero-shot transferability. It inspires subsequent studies to transfer the vision and language knowledge to various downstream tasks such as object detection [7], text-driven image manipulation [21], video clip retrieval [19] and semantic segmentation [23, 28]. ViLD [7] aligns the region embedding with visual

and semantic embeddings inferred from the CLIP and trains a student detector. DenseCLIP [28] modifies the image encoder of CLIP for a pixel-level dense prediction. However, the capability of CLIP on the HOI detection task remains unexplored. Different from previous attempts, we propose to distill action probability from CLIP without modifying model structures, which will not bring in extra inference computation cost.

### 2.3 Zero-Shot Learning on HOI detection

Zero-shot learning aims at classifying categories that are not seen during training. Previous works implement zero-shot learning on HOI detection task for three scenarios: unseen combination scenario (UC) [1, 10, 18, 24], unseen object cenario (UO) [1, 10, 18] and unseen action scenario (UA) [18]. ConsNet [18] performs zero-shot HOI detection for the three scenarios by learning from a consistency graph along with word embeddings. The consistency graph is built on existing HOI datasets and is hard to scale further. In contrast, we adopt the image-text pretrained CLIP as the teacher model to supervise the student HOI detector for both UC and UA scenarios, and the action set is scalable via adjusting the action prompts.

## 3 Preliminary

**Problem Formulation:** Denote  $\mathcal{A}_S = \{a_1, \dots, a_k\}$  as a set of the seen action categories, and  $\mathcal{A}_U = \{a_{k+1}, \dots, a_n\}$  as a set of unseen actions. Let  $I$  denote an input image, with corresponding labels  $\mathcal{T} = \{\mathcal{B}, \mathcal{Y}\}$  where  $\mathcal{B}$  is a set of bounding boxes including human boxes  $b_h$  and object boxes  $b_o$ , and  $\mathcal{Y}$  denote a set of known HOI triplets. Each  $y = \langle b_h, b_o, a \rangle$  in  $\mathcal{Y}$  is a HOI triplet, where  $b_h$  and  $b_o$  are the elements in the set  $\mathcal{B}$ , and  $a \in \mathcal{A}_S$ .

For our HOI model, the bounding boxes are trained in a paired manner. It requires to construct all possible human-object pairs one by one between human and objects in  $\mathcal{B}$ . The constructed pairs present in the annotated HOI triplets set of  $\mathcal{Y}$  denote *seen* pairs (or *known* pairs),  $y_s = \langle b_h, b_o, a \rangle$ , and the others absent in  $\mathcal{Y}$  denote *unknown* pairs,  $y_u = \langle b_h, b_o, \emptyset \rangle$ . The unknown pairs consist of the *unseen* pairs with potential interaction and the *non-interactive* pairs. Finally, our goal is to detect all interactive *seen* and *unseen* pairs, and also recognize their actions.

**Transformer-based HOI models:** Most of existing SOTA HOI models [25, 27, 29] are end-to-end transformer-based models. First, the input image  $I$  and learnable query vectors  $Q_e$  are fed to HOI model to predict human-object bounding boxes pairs and the corresponding actions. The paradigm is formulated as,  $\hat{y} = \text{Transformer}(I, Q_e)$ , where  $\hat{y}$  is the prediction. During training, a bipartite matching algorithm is adopted to match predictions with the best ground truth by the Hungarian algorithm, as follows,

$$\hat{\sigma} = \arg \min_{\sigma \in \Theta_N} \sum_{i=1}^N \mathcal{H}_{\text{match}}(y_i, \hat{y}_{\sigma(i)}), \quad (1)$$

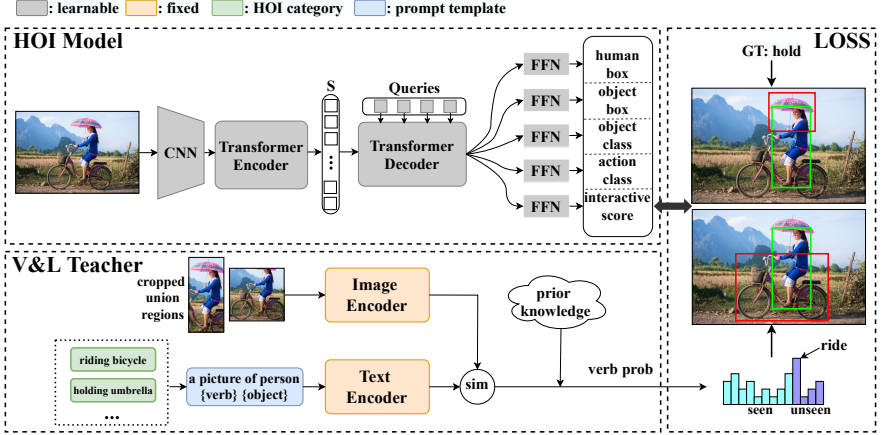


Fig. 2: The overview of the proposed EoID. After getting  $N$  predictions from the HOI model, we adopt the two-stage bipartite matching algorithm to select the predictions which best match the ground truth human-object pairs. Then we train the model with the selected predictions to learn the distribution of action probability from the pretrained V&L teacher as well as the seen ground truth to achieve zero-shot HOI classification.

where  $y_i \in \tilde{\mathcal{Y}}$ ,  $\tilde{\mathcal{Y}} = \{y_1, \dots, y_M, \emptyset_{M+1}, \dots, \emptyset_N\}$  denotes the  $M$  ground truth pairs padded with  $N - M$  no-pairs  $\emptyset$ ,  $\{\hat{y}_i\}_{i=1}^N$  denotes the set of  $N$  predictions, and  $\Theta_N$  is a search space for a permutation of  $N$  elements.  $\mathcal{H}_{match}$  is the matching cost [2] between ground truth  $y_i$  and a prediction with index  $\sigma(i)$ , which consists of four types of costs: the box-regression cost  $\mathcal{H}_b$ , intersection-over-union (IoU) cost  $\mathcal{H}_u$ , object-class cost  $\mathcal{H}_c$ , and action-class cost  $\mathcal{H}_a$ , as follows,

$$\mathcal{H}_{match} = \mathcal{H}_b + \mathcal{H}_u + \mathcal{H}_c + \mathcal{H}_a. \quad (2)$$

Finally, the losses of the matched pairs are optimized by a *Hungarian loss*, which can be formulated as:

$$\mathcal{L}_H = \sum_{i=1}^N \sum_{\mathcal{L} \in \Omega} \{\mathbb{1}_{\{y_i \neq \emptyset\}} \mathcal{L}(\hat{y}_{\sigma(i)}, y_i) + \mathbb{1}_{\{y_i = \emptyset\}} \mathcal{L}(\hat{y}_{\sigma(i)}, \emptyset)\}, \quad (3)$$

where  $\Omega = \{\mathcal{L}_b, \mathcal{L}_u, \mathcal{L}_c, \mathcal{L}_a\}$ ,  $\mathcal{L}_b = \mathcal{L}_b^{(h)} + \mathcal{L}_b^{(o)}$  is the box regression loss,  $\mathcal{L}_u = \mathcal{L}_u^{(h)} + \mathcal{L}_u^{(o)}$  is the intersection-over-union loss,  $\mathcal{L}_c$  is the object-class loss and  $\mathcal{L}_a$  the action-class loss, following QPIC [25] and CDN [27].

## 4 Method

### 4.1 Overall Architecture

Fig. 2 illustrates an overview of our EoID. We benchmark on transformer-based model CDN [27]. Given an image  $I$ , the CDN first encodes  $I$  into visual feature

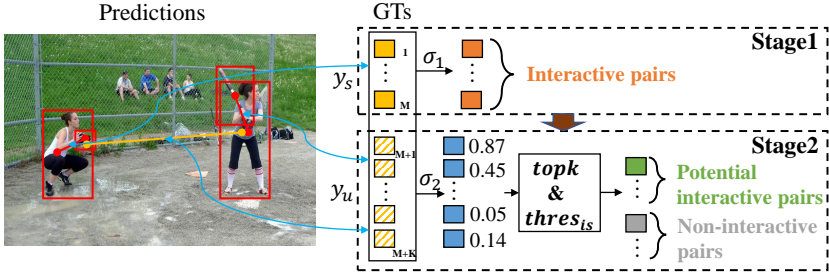


Fig. 3: The two-stage bipartite matching algorithm. Given  $N$  predictions, we first match all the predictions with  $M$  seen pairs ( $y_s$ ) as **interactive pairs** in the first stage. In the second stage, we use the remaining  $N - M$  predictions to match the  $K$  unknown pairs ( $y_u$ ). And we select  $topk$  predictions that have interactive score greater than interactive threshold  $thresh_{is}$  from the matched  $K$  predictions as **potential interactive pairs**, and the others as **non-interactive pairs**. The non-interactive pairs and the remaining  $N - M - K$  unmatched predictions will be regarded as **no-pairs**.

sequence  $S$ , then the cascaded human-object decoder and interaction decoder are assigned to decode the visual feature sequence into  $N$  predictions, including human-object bounding boxes pairs ( $b_h, b_o$ ) and action vectors  $\mathbf{a}$  with corresponding interactive scores  $s_{is}$  from a set of queries  $Q_e$ . For the vision and language teacher, the union regions of human-object pairs are cropped to extract the visual features, and prompt engineering is adopted before calculating the semantic features of the HOI texts. The cosine similarity between them combined with the prior knowledge represents the action probability for these predictions. Then a two-stage bipartite matching algorithm is applied to select the predictions that best match the ground truth human-object pairs for the seen pairs and the unknown pairs. These selected predictions from the two-stage bipartite matching algorithm are used to train bounding boxes regressive branches, object classification branch, interactive score branch, and action classification branch. Finally, we train the model to learn the distribution of action probability from the pretrained vision and language teacher as well as the seen ground truth to achieve zero-shot HOI classification.

## 4.2 Learning to Detect Potential Interactive Pairs

Exhaustively traversing all possible human-object pairs is computationally infeasible and might introduce excessive training noise. As a result, the first challenge in our work is to detect potential action-agnostic interactive pairs in both training and inference. We address this issue by introducing an interactive score module and a two-stage bipartite matching algorithm.

**Interactive Score Module:** To distinguish interactive and non-interactive pairs in the predictions, it is a natural idea to adopt the naive implementation of the interactive score module in CDN. However, it only considers the predictions

matching seen pairs  $y_s$  as interactive pairs and the others as non-interactive pairs, which suppresses the detection of unseen pairs.

Unlike CDN, we apply the interactive score head from the interaction decoder instead of the human-object decoder to detect unseen pairs. Specifically, given  $N$  predictions, we first apply a two-stage bipartite matching algorithm to select  $M + \text{topk}$  predictions which best match the  $M$  seen pairs and  $\text{topk}$  unknown pairs. Then we train the interactive score module by rewarding the  $M + \text{topk}$  matched predictions which have human and object IoUs with the matched ground truth pairs are greater than 0.5, and penalizing the others. For the rest  $N - M - \text{topk}$  predictions with any unknown pairs that satisfy the above condition, their loss is omitted during the optimization.

**Two-stage Bipartite Matching Algorithm:** Similar to OW-DETR [8], we apply a two-stage bipartite matching algorithm to match the seen pairs and the unknown pairs respectively, as shown in Fig. 3. We first match  $M$  predictions with the seen pairs as interactive pairs based on the box-regression cost  $H_b$ , IoU cost  $H_u$ , object-class cost  $H_c$ , and action-class cost  $H_a$ , as follows,

$$\hat{\sigma}_1 = \arg \min_{\sigma_1 \in \Theta_N} \sum_{i=1}^N \mathcal{H}_{\text{match}}^1(y_i, \hat{y}_{\sigma_1(i)}), \quad (4)$$

where  $y_i \in \tilde{\mathcal{Y}}_1 = \{y_1, \dots, y_M, \emptyset_{M+1}, \dots, \emptyset_N\}$ , and  $\mathcal{H}_{\text{match}}^1 = \mathcal{H}_b + \mathcal{H}_u + \mathcal{H}_c + \mathcal{H}_a$ . Then the  $N - M$  predictions not selected by the first stage matching will be used for the second stage matching. Since ground truth actions are not available for the unknown pairs, we implement the second stage matching only based on the box-regression cost  $H_b$ , IoU cost  $H_u$ , object-class cost  $H_c$ . The second stage matching process can be formulated as follows,

$$\hat{\sigma}_2 = \arg \min_{\sigma_2 \in \Theta_{N-M}} \sum_{i=M+1}^N \mathcal{H}_{\text{match}}^2(y_i, \hat{y}_{\sigma_2(i)}), \quad (5)$$

where  $y_i \in \tilde{\mathcal{Y}}_2 = \{y_{M+1}, \dots, y_{M+K}, \emptyset_{M+K+1}, \dots, \emptyset_N\}$ ,  $K$  is the number of unknown pairs, and  $\mathcal{H}_{\text{match}}^2 = \mathcal{H}_b + \mathcal{H}_u + \mathcal{H}_c$ . We select the  $\text{topk}$  predictions which have interactive scores  $s_{is}$  greater than interactive threshold  $\text{thres}_{is}$  as potential interactive pairs, the others as non-interactive pair. So we can get the  $M + \text{topk}$  predictions with  $\hat{\sigma} = \hat{\sigma}_1 \cup \text{topk}(\hat{\sigma}_2)$ . The non-interactive pairs and the remaining  $N - M - K$  unmatched predictions will be regarded as no-pairs. The no-pairs and  $M + \text{topk}$  predictions combined with matched ground truth will be used to train bounding boxes regressive branch, object classification branch, interactive score branch and action classification branch.

Such a strategy will help the model learn from the seen pairs to discriminate whether there exists interaction between human-object pairs at the early training stage, and gradually introduce potential interactive pairs for learning.



### 4.3 Knowledge Distillation from CLIP

After detecting potential interactive human-object pairs, we need to identify the corresponding action happening between the human and object. For this purpose, we transfer the interaction knowledge from the pretrained V&L model CLIP (teacher) into the HOI model (student). To make full use of the zero-shot transferability of CLIP and maintain the real-time inference performance of the HOI model, similar to ViLD [7], we train the HOI model to learn the interaction knowledge from CLIP via knowledge distillation, as shown in Fig. 2(bottom).

We first convert the HOI category texts, *e.g. riding bicycle*, into the prompts by feeding them into prompt template *a picture of person {verb} {object}*. Then we encode these prompts to generate the text embeddings  $t_e$  offline by the CLIP text encoder  $T$ . For the  $M + \text{top}k$  matched pairs of  $I$ , we crop the human-object union regions, and feed the preprocessed ones into the CLIP image encoder  $V$  to generate the image embeddings  $v_e$ . Then, we compute cosine similarities between the image and text embeddings, as  $s_i = v_e^T t_e^i / (\|v_e\| \cdot \|t_e^i\|)$ . According to the prior knowledge [3], we select the valid actions which is able to interact with the object for each union region. We apply a softmax activation on similarities of these HOI categories to get the probability distribution  $\mathbf{p}$  of the actions in  $\mathcal{A}_S + \mathcal{A}_U$  for each of union regions. The process can be formulated as follows,

$$\begin{aligned} p_i &= \frac{e^{\gamma s_i m_i}}{\sum_{j=1}^n e^{\gamma s_j m_j}}, \\ \text{s.t. } m_i &= \begin{cases} 1, & \text{if } \textit{valid} \\ -\infty, & \text{if } \textit{invalid} \end{cases}, \end{aligned} \quad (6)$$

where  $p_i$  is the probability of the action,  $\gamma$  is a scalar hyper-parameter and  $m_i$  is a correct coefficient to eliminate invalid HOI categories [3]. Finally, we train the model to learn from this probability distribution  $\mathbf{p}$  of the actions in  $\mathcal{A}_S + \mathcal{A}_U$  as well as the ground truth actions in  $\mathcal{A}_S$ .

### 4.4 Training and Inference

**Training:** We calculate the loss with extra interactive score loss and CLIP distillation loss, as follows,

$$\mathcal{L}_{total} = \mathcal{L}_H + \lambda_{is} \mathcal{L}_{is} + \lambda_{clip} \mathcal{L}_{clip}, \quad (7)$$

where  $\mathcal{L}_H$  is computed by Eq. 3,  $\mathcal{L}_{is}$  is the interactive score loss, and  $\mathcal{L}_{clip}$  is the CLIP distillation loss. The  $\mathcal{L}_{is}$  term adopts cross entropy loss, and the  $\mathcal{L}_{clip}$  term adopts binary cross entropy loss.  $\lambda_{is}$  and  $\lambda_{clip}$  are the hyper-parameters.

**Inference:** After distilling the action knowledge from CLIP, we only keep the learned CDN model for inference, avoiding extra computation cost. The post-process of our method remains unchanged as CDN.



Table 1: The recall on unseen pairs (U-R). We compare the models trained with different interactive score modules and supervision sources. Our method achieves comparable recall for unseen pairs on HICO-Det test set. Results show that our method can efficiently detect potential interactive pairs.

Method	Supervision	U-R@3	U-R@5	U-R@10
CDN	seen	61.47	66.68	71.70
EoID	seen	64.72	<b>71.45</b>	76.79
EoID	seen+ <i>topk</i>	<b>65.25</b>	71.16	<b>77.20</b>
CDN	full	67.03	73.41	78.64

## 5 Experiments

### 5.1 Experimental Setup

**Datasets and Evaluation Metrics:** We perform our experiments on two HOI detection benchmarks: HICO-DET [3] and V-COCO [9]. HICO-DET is a large-scale HOI detection dataset consisting of 47,776 images, 38,18 for training and 9,658 for test. It contains 80 object classes as MS-COCO [17] and 117 action classes, which result in 600 HOI triplets. V-COCO is a subset of the MS-COCO dataset consisting of 5,400 training images and 4,946 test images. It contains 29 action classes (25 HOIs and 4 body motions) and the same 80 object classes as MS-COCO. We follow the standard evaluation [3] to use the mean average precision (*mAP*) as the evaluation metric. A HOI triplet is considered as a true positive when (1) the predicted object and action categories are correct, and (2) both the predicted human and object bounding boxes have intersection-over-union (IoU) with a ground truth greater than 0.5.

**Zero-shot Setups:** We follow the common setups to conduct experiments on two scenarios on HICO-Det: unseen combination scenario (UC) and unseen action scenario (UA). For the UC scenario, we use the same 5 sets of 120 unseen HOI classes as Bansal et al. [1]. Each of the action classes and object classes is seen in at least one HOI action-object pair during the training procedure. For the UA scenario, we use the same 22 unseen actions as ConsNet [18], and remove all the training labels containing these actions. We train the model on the remaining labels and evaluate on the full test set.

**Implementations:** We benchmark on the CDN [27] and use the same settings for all models unless explicitly specified. The query number  $N$  is 64. The loss weights  $\lambda_{bbox}$ ,  $\lambda_{giou}$ ,  $\lambda_c$ ,  $\lambda_{is}$ ,  $\lambda_a$  and  $\lambda_{clip}$  are set to 2.5, 1, 1, 1, 1.6 and 700 respectively. For simplicity, we do not use the decoupling dynamic re-weighting in CDN. For CLIP, we use the publicly available pretrained model<sup>1</sup>, with an input size of  $224 \times 224$ , and  $\gamma=100$ . For cropped union regions, we preprocess them by square padding and resizing. We feed prompt engineered texts to the text

<sup>1</sup> <https://github.com/openai/CLIP>.

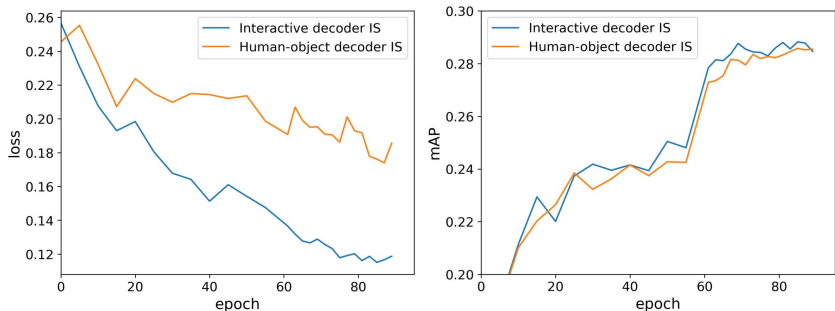


Fig. 4: The curves of the interactive score loss (left) and the  $mAP$  (right), compared between the models with IS branch from interactive decoder and human-object decoder. The model with IS branch from interactive decoder shows faster convergence and better performance.

encoder of CLIP with a prompt template *a picture of person {verb} {object}*. Experiments are conducted on 4 Tesla V100 GPUs, with a batch size of 16.

## 5.2 Learning to Detect Unseen Pairs

We first study whether our framework can detect potential interactive pairs. We compare the models trained with different interactive score (IS) modules and supervision sources. Table 1 shows the top-k recall (U-R@K) of unseen pairs on HICO-Det test set. Even though training only with seen pairs (seen), the recall on unseen pairs of our model outperforms the CDN with naive IS module by a large margin. After introducing the potential pairs (*topk*) for training, the recall is further improved and shows comparable results with the fully-supervised model. This experiment demonstrates that our method can detect more potential interactive pairs compared to the CDN with the naive IS module.

We also compare the curves of the interactive score loss  $\mathcal{L}_{is}$  and the overall  $mAP$  with IS branch from the interactive decoder ( $IS_{inter}$ ) and human-object decoder ( $IS_{ho}$ ), as shown in Fig. 4. Compared to the model with  $IS_{ho}$ , the interactive score loss of the model with  $IS_{inter}$  converges faster, and also shows a better overall performance. These results demonstrate the interactive decoder is more capable of extracting interactive information.

## 5.3 Zero-Shot Transferability of CLIP on HOI Classification

We replace the action classifier of the converged CDN model with CLIP to validate the zero-shot transferability of CLIP on HOI classification. The prior knowledge is also applied to avoid the dispersion of probability distribution of CLIP to some invalid actions. The CLIP model is RN50x16 and the CDN model is CDN-S. We report the  $mAP$  on the full set of HOI classes (Full), a rare set of the HOI classes (Rare) that have less than 10 training instances, and a non-rare set of the other HOI classes (Non-rare). As shown in Table 2, the HOI model with a fixed CLIP classifier only has a small degradation with the

Table 2: We replace the action classifier of the converged CDN with CLIP to validate the zero-shot transferability of CLIP on HOI classification, with or without prior knowledge. Using CLIP for HOI classification shows the competitive performance with the full-supervised model on rare HOI categories of HICO-Det.

Method	Full	Rare	Non-rare
CLIP	21.11	26.02	19.64
CLIP w/ prior	21.45	26.42	19.97
Full-Supervised(seen + unseen)	31.11	26.49	32.49

Table 3: Ablation of variants. We study the methods to cope with the problems occurred in Sec. 5.3.

$\mathcal{A}_U$ only	detach	Full	Seen	Unseen
		27.93	29.15	21.84
✓		28.83	30.20	21.98
	✓	28.68	29.88	22.71
✓	✓	<b>29.22</b>	<b>30.46</b>	<b>23.04</b>

fully supervised model on *Rare* categories. This gap is further reduced when prior knowledge is adopted, which shows the competitive performance with the full-supervised model. However, the performances of *Full* and *Non-rare* are not effective enough. Note that the CLIP classifies all the predicted pairs without considering whether it is interactive or not. There is still much improvement space for the overall performance and the inference speed.

## 5.4 Ablation Studies

We perform ablation experiments in Sec. 5.4. Unless otherwise specified, the CLIP model used here is RN50x16, the CDN model is CDN-S, the *topk* and *thres<sub>is</sub>* are set to 3 and 0.5 respectively. All ablation results are evaluated on the HICO-Det test set.

**Ablation of variants:** As shown in Sec. 5.3, the degradation on *Full* and *Non-rare* indicates the existence of noise from CLIP, which may lead to poor performance on the *Seen* category. In addition, extra loss terms are introduced into the framework which leads to the problem of convergence difficulties. We study the methods to overcome the above challenges: 1) we only distill CLIP to the HOI actions in  $\mathcal{A}_U$  under UA setting, short as  $\mathcal{A}_U$  only; 2) we use *detach technique* [20] to cut off the back-propagation of gradients between the human-object decoder and interaction decoder. As shown in Table 3, we can combine the  $\mathcal{A}_U$  only to alleviate the impact of the noise from CLIP, which improves the *Seen* and *Unseen* by 1.05% and 0.14% *mAP* respectively. With the *detach technique* adopted, the best *Full* performance is obtained to 29.22% *mAP*.

Table 4: Ablation studies. We perform ablation experiments to study the impact of  $topk$ ,  $thres_{is}$ , CLIP and CDN models.

(a) Impact of $topk$ .				(b) Impact of $thres_{is}$ .			
$topk$	Full	Seen	Unseen	$thres_{is}$	Full	Seen	Unseen
1	28.45	29.75	21.96	0.1	28.50	29.80	22.01
3	<b>28.83</b>	<b>30.20</b>	21.98	0.3	28.48	29.59	<b>22.93</b>
5	28.56	29.85	22.14	0.5	<b>28.83</b>	<b>30.20</b>	21.98
10	28.45	29.69	<b>22.29</b>	0.7	28.08	29.11	22.90
				0.9	28.44	29.99	20.75

(c) Impact of CLIP models.				(d) Impact of the CDN models.			
CLIP	Full	Seen	Unseen	CDN	Full	Seen	Unseen
ViT-B32	28.46	30.02	20.63	CDN-S	28.83	30.20	21.98
ViT-B16	28.32	29.71	21.40	CDN-B	28.70	30.17	21.39
RN50	28.75	30.13	21.84	CDN-L	<b>29.32</b>	<b>30.63</b>	<b>22.76</b>
RN50x16	<b>28.83</b>	<b>30.20</b>	<b>21.98</b>				

**Impact of  $topk$ :** We compare models with different  $topk \in \{1, 3, 5, 10\}$ , and results are shown in Table 4a. The model with  $topk=3$  obtains the best performance on *Full* and *Seen*. With the growth of the  $topk$ , the model obtains a better performance on *Unseen* while the overall *mAP* starts to drop for too much noise.

**Impact of  $thres_{is}$ :** We compare models with different  $thres_{is} \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ , and results are shown in Table 4b. The models with a larger  $thres_{is}$  will eliminate more non-interactive pairs, which may results in a larger precision but a lower recall of unseen pairs and vice versa. The best performance of *Full* and *Seen* are obtained when adopting a proper  $thres_{is}=0.5$ .

**Impact of CLIP models:** We compare models with different CLIP backbones, ViT-B32, ViT-B16, RN50, and RN50x16, and results are shown in Table 4c. The larger CLIP backbones provide a better performance boost but lead to more computational overhead in the training phase. However, the CLIP can be removed during inference, whose extra cost during training is acceptable.

**Impact of CDN models:** We compare variants of our method equipped with different CDN models, CDN-S, CDN-B, and CDN-L, and results are shown in Table 4d. The best performance is obtained by CDN-L, with more computational costs required during training and inference. It indicates that a larger HOI model leads to a better performance, which is consistent with CDN [27].

## 5.5 Zero-Shot HOI Detection

We compare our method with state-of-the-art models on the HICO-Det test set under UC and UA settings in Table 5. The compared models include: Shen et

Table 5: Zero-shot HOI Detection results on HICO-DET dataset. UC and UA denote unseen action-object combination and unseen action scenarios respectively. Our method outperforms all the other methods by a large margin. The \* indicates that the model training without using *A<sub>U</sub> only* and *detach technique*.

Method	Type	Full	Seen	Unseen
Shen <i>et al.</i> [24]	UC	6.26	-	5.62
Functional [1]	UC	12.45±0.16	12.74±0.34	11.31±1.03
FCL [10]	UC	19.37	19.55	18.66
ConsNet [18]	UC	19.81±0.32	20.51±0.62	16.99±1.67
EoID	UC	<b>28.91±0.33</b>	<b>30.39±0.4</b>	<b>23.01±1.98</b>
CDN+ConsNet	UA	26.53	28.77	15.30
EoID*	UA	<b>27.93</b>	<b>29.15</b>	<b>21.84</b>
ConsNet	UA	19.04	20.02	14.12
EoID	UA	<b>29.22</b>	<b>30.46</b>	<b>23.04</b>

Table 6: Transfer to object detection datasets. We study the performance of our method on V-COCO with the bounding box annotations from COCO. Our method can transfer to datasets with only bounding boxes annotated to further scale up existing HOI categories.

Training source	Method	Full	Seen	Unseen
HICO only	CDN	-	35.15	-
HICO+pseudo-V-COCO	CDN	-	38.13	-
HICO+pseudo-V-COCO	EoID	40.39	38.13	47.15
V-COCO(full)	CDN	56.43	54.56	62.05

al. [24], Functional [1], FCL [10] and ConsNet [18]. Besides, we introduce a new variant of CDN by applying the consistency graph of ConsNet (CDN+ConsNet, details in Appendix) to validate whether the better performance is obtained by the better backbone. In the top section, we show results under UC setting. In the middle section, we show results of the same backbone under UA setting. In the last section, we compare our model with the official ConsNet under UA setting. Our method outperforms the previous SOTA by 6.02% on unseen *mAP* and 9.1% on overall *mAP* for UC (row 5 v.s. row 6), by 8.92% on unseen *mAP* and 10.18% on overall *mAP* for UA (row 9 v.s. row 10). In addition, even though using the same CDN backbone, the performance on unseen actions is also significantly better than ConsNet without bells and whistles (row 7 v.s. row 8). Note that, our best *Unseen* performance is competitive with the fully-supervised method [29] (23.04% v.s. 23.46%) and these experiments indicate the effectiveness of our proposed zero-shot HOI detection framework.

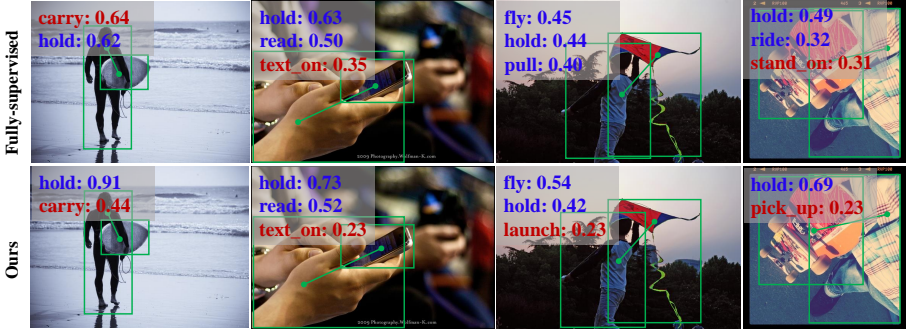


Fig. 5: Qualitative results on HICO-Det. Our method (second row) is able to identify the seen (blue) and unseen (red) actions similar to the full-supervised model (first row). In addition, our model can recognize some complex actions like *launch* (third column) and *pick\_up* (fourth column) compared to the full-supervised model.

## 5.6 Scaling up Action Sets via Object Detection Datasets

Current HOI detectors are limited in action size due to the small scale of HOI detection datasets. Meanwhile, large-scale object detection datasets contain potential unknown interaction pairs. As a result, we introduce a more difficult but practical experiment, in which object detection datasets are merged to further scale up the existing HOI categories. For this purpose, we study the performance of our method on V-COCO with the bounding box annotations from COCO. Specially, we compare the models trained on HICO-Det dataset (HICO only), HICO-Det with pseudo-V-COCO and full V-COCO dataset, where the pseudo-V-COCO consists of bounding boxes from COCO, pseudo-seen action labels from the predictions of the CDN trained on HICO-Det dataset. We test on V-COCO test set, and set the overlap actions between HICO-Det and V-COCO as *Seen* and the others as *Unseen*. More details are elaborated in Appendix. Results are shown in Table 6, and our method has an about 16% overall *mAP* gap compared to the fully-supervised method, while a smaller gap on *Unseen* categories. This experiment shows that our method can transfer to bounding boxes annotated datasets to further scale up the existing HOI categories.

## 5.7 Qualitative Results

We visualize the results of our method and the fully-supervised model in Fig. 5. Our method is able to identify the seen (blue) and unseen (red) actions similar to the fully-supervised model. In addition, our model can recognize some complex actions like *launch* and *pick\_up* compared to the fully-supervised model.

## 6 Conclusions

In this work, we present EoID, an end-to-end zero-shot HOI detection framework via knowledge distillation from multimodal vision-language embeddings. Our

method first detects potential action-agnostic interactive human-object pairs by applying a two-stage bipartite matching algorithm and an interactive score module. Then a zero-shot action classification is applied to identify novel HOIs. The experiments demonstrate that our detector is able to detect unseen pairs, which benefits the recognition of unseen HOIs. Our method outperforms the previous SOTA by a large margin under zero-shot setting and shows a promising generalization to utilize large-scale detection datasets to scale up the action sets.



## References

1. Bansal, A., Rambhatla, S.S., Shrivastava, A., Chellappa, R.: Detecting human-object interactions via functional generalization. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 34, pp. 10460–10469 (2020)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: *European Conference on Computer Vision*. pp. 213–229. Springer (2020)
3. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: *2018 IEEE winter conference on applications of computer vision (WACV)*. pp. 381–389. IEEE (2018)
4. Gao, C., Zou, Y., Huang, J.B.: ican: Instance-centric attention network for human-object interaction detection. In: *British Machine Vision Conference* (2018)
5. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 8359–8367 (2018)
6. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 6904–6913 (2017)
7. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921* **2** (2021)
8. Gupta, A., Narayan, S., Joseph, K., Khan, S., Khan, F.S., Shah, M.: Ow-detr: Open-world detection transformer. *arXiv preprint arXiv:2112.01513* (2021)
9. Gupta, S., Malik, J.: Visual semantic role labeling. *arXiv preprint arXiv:1505.04474* (2015)
10. Hou, Z., Yu, B., Qiao, Y., Peng, X., Tao, D.: Detecting human-object interaction via fabricated compositional learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14646–14655 (2021)
11. Kim, B., Choi, T., Kang, J., Kim, H.J.: Uniondet: Union-level detector towards real-time human-object interaction detection. In: *European Conference on Computer Vision*. pp. 498–514. Springer (2020)
12. Kim, B., Lee, J., Kang, J., Kim, E.S., Kim, H.J.: 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 (2021)
13. Kuhn, H.W.: The hungarian method for the assignment problem. *Naval research logistics quarterly* **2**(1-2), 83–97 (1955)
14. Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: *Proceedings of the IEEE international conference on computer vision*. pp. 1261–1270 (2017)
15. Li, Y.L., Zhou, S., Huang, X., Xu, L., Ma, Z., Fang, H.S., Wang, Y., Lu, C.: Transferable interactiveness knowledge for human-object interaction detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3585–3594 (2019)
16. Liao, Y., Liu, S., Wang, F., Chen, Y., Qian, C., Feng, J.: Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 482–490 (2020)
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: *European conference on computer vision*. pp. 740–755. Springer (2014)

18. Liu, Y., Yuan, J., Chen, C.W.: Consnet: Learning consistency graph for zero-shot human-object interaction detection. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 4235–4243 (2020)
19. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval. arXiv preprint arXiv:2104.08860 (2021)
20. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019)
21. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2085–2094 (2021)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
23. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. arXiv preprint arXiv:2112.01518 (2021)
24. Shen, L., Yeung, S., Hoffman, J., Mori, G., Fei-Fei, L.: Scaling human-object interaction recognition through zero-shot learning. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1568–1576. IEEE (2018)
25. Tamura, M., Ohashi, H., Yoshinaga, T.: 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 (2021)
26. Wang, T., Yang, T., Danelljan, M., Khan, F.S., Zhang, X., Sun, J.: Learning human-object interaction detection using interaction points. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4116–4125 (2020)
27. Zhang, A., Liao, Y., Liu, S., Lu, M., Wang, Y., Gao, C., Li, X.: Mining the benefits of two-stage and one-stage hoi detection. *Advances in Neural Information Processing Systems* **34** (2021)
28. Zhou, C., Loy, C.C., Dai, B.: Denseclip: Extract free dense labels from clip. arXiv preprint arXiv:2112.01071 (2021)
29. Zou, C., Wang, B., Hu, Y., Liu, J., Wu, Q., Zhao, Y., Li, B., Zhang, C., Zhang, C., Wei, Y., et al.: End-to-end human object interaction detection with hoi transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11825–11834 (2021)