Toward Open-Set Human Object Interaction Detection

Mingrui Wu\textsuperscript{1,2}*, Yuqi Liu\textsuperscript{1+}, Jiayi Ji\textsuperscript{1†}, Xiaoshuai Sun\textsuperscript{1,2}, Rongrong Ji\textsuperscript{1,2}

\textsuperscript{1}Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University
\textsuperscript{2}Institute of Artificial Intelligence, Xiamen University
mingrui0001@gmail.com, liuyuqi@stu.xmu.edu.cn, jjyxmu@gmail.com, \{xssun, rrji\}@xmu.edu.cn

Abstract

This work is oriented toward the task of open-set Human Object Interaction (HOI) detection. The challenge lies in identifying completely new, out-of-domain relationships, as opposed to in-domain ones which have seen improvements in zero-shot HOI detection. To address this challenge, we introduce a simple Disentangled HOI Detection (DHD) model for detecting novel relationships by integrating an open-set object detector with a Visual Language Model (VLM). We utilize a disentangled image-text contrastive learning metric for training and connect the bottom-up visual features to text embeddings through lightweight unary and pair-wise adapters. Our model can benefit from the open-set object detector and the VLM to detect novel action categories and combine actions with novel object categories. We further present the VG-HOI dataset, a comprehensive benchmark with over 17k HOI relationships for open-set scenarios. Experimental results show that our model can detect unknown action classes and combine unknown object classes. Furthermore, it can generalize to over 17k HOI classes while being trained on just 600 HOI classes.

Introduction

Human-Object Interaction (HOI) detection is a pivotal task in computer vision and artificial intelligence that facilitates the understanding of visual scenes. It involves pinpointing humans and objects within an image and understanding their interactions. While significant strides have been made in traditional HOI detection, these advancements are primarily in closed-set scenarios, with predefined and limited action and object categories, e.g., the COCO (Lin et al. 2014) 80 object categories and 117 action categories in the HICO-DET (Chao et al. 2018) dataset. However, open-set HOI detection, where novel object and action categories are not explicitly known beforehand, remains an underexplored domain. This necessitates a shift from the traditional closed-set mindset to a more adaptable and scalable paradigm.

Most current approaches (Hou et al. 2020, 2021b; Bansal et al. 2020; Liao et al. 2022; Ning et al. 2023) are tailored for zero-shot HOI detection, working on unseen but similar intra-domain relationships, especially within the HICO-DET dataset. This leaves a gap in models truly adaptable to real-world, open-set conditions. In response to the current gaps, we present VG-HOI, a comprehensive evaluation benchmark that features over 17k relationships, covering a broad spectrum of unfamiliar actions and objects. Our aim is to mimic real-world scenarios as closely as possible with a plethora of open-set interactions. The distinction between zero-shot and open-set HOI detection is shown in Figure 1.

Most of the existing methods cannot be directly evaluated on VG-HOI due to their limited focus either on specific actions or restricted object categories. Current mainstream zero-shot HOI detection paradigms (Liao et al. 2022; Ning et al. 2023) primarily function within closed-set object scenarios, involving the fine-tuning of object detectors and interaction classifiers in a tightly integrated, end-to-end manner. This often limits their scalability to new object categories. Conversely, emerging bottom-up HOI detection methods (Zhang, Campbell, and Gould 2022; Park, Park, and Lee 2023) highlight the significance of powerful bottom-up representations. Regrettably, these methods tend to concentrate on closed-set scenarios, either with bottom-up features focusing on object distinction or with the backbone fitting the long-tail pattern of the specified dataset, inevitably hampering their open-set detection and interactive classification capabilities.

The recent merger of image-text contrastive learning with the CLIP model (Radford et al. 2021) has significantly advanced open-set object detection (Liu et al. 2023). These new methods excel at identifying various new object categories with precise bounding boxes. Alongside this, the CLIP model, refined through training on extensive web-scale data, is skilled at capturing strong visual representations. This juncture offers a promising chance to merge these powerful models, propelling us toward open-set HOI detection. So, do we really need to repurpose these powerful object detectors or VLM backbone for HOI detection?

In light of this, we introduce the Disentangled HOI Detection (DHD) model, a straightforward bottom-up framework that includes an open-set object detector, a visual and language model, and an interaction head. We extract bottom-up features from VLM visual encoder with the bounding boxes from the open-set object detector. Then we attach lightweight unary and pair-wise adapters to establish connections be-
Figure 1: The comparison between the zero-shot HOI and open-set HOI detection. The zero-shot HOI detection (top) generalize to unseen classes in intra-domain. The open-set HOI detection (bottom) detect the unknown classes in out-of-domain. The seen actions or objects are shown in blue color, and the unseen or unknown classes are shown in red or green color.

The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)

Open-set Detection
Open-set detection aims to detect the categories which cannot be known in advance. Recently, great progress has been made in the field of open-set object detection (Li et al. 2022; Liu et al. 2023). GLIP (Li et al. 2022) implements region-level language-image pre-training by considering object detection as a phrase grounding task. Grounding DINO (Liu et al. 2023) pushes it by combining the end-to-end object detector and making visual language modalities fully interactive. RefCOD (Zhang et al. 2023) achieve open camouflaged object detection based on visual reference. There has been a lot of research focused on the zero-shot HOI detection (Hou et al. 2020, 2021b; Bansal et al. 2020; Liao et al. 2022; Ning et al. 2023; Wang et al. 2020, 2021, 2022; Wu et al. 2023). However, these methods either only recognize seen actions, or only can process limited objects. Benefiting from advances in open-set object detection, we combine the existing open-set object detectors e.g. (Liu et al. 2023) into a decoupled bottom-up structure to realize a more general open-set HOI detection.

Image-text Contrastive Learning
Inspired by CLIP’s breakthrough by contrastive learning with web-scale image-text pairs, a surge of research (Li et al. 2022; Liu et al. 2023; Wang et al. 2023) is now enhancing representations and open generalization through image-text contrastive learning. Meanwhile, in the field of HOI detection, UNIVRD (Zhao et al. 2023) begin to use contrastive learning for joint training on multi-source heterogeneously labeled HOI datasets. However, UNIVRD finetunes the VLM visual backbone to fit the object detector, which makes the bottom-up features focus on object discrimination and migrates the general representation of the VLM visual encoder. In contrast to UNIVRD, we make full use of the powerful generalization of the VLM visual encoder by freezing it and propose a disentangled image-text contrastive learning metric for towarding the open-set HOI detection.
**Preliminary**

**Problem Formulation:** Denote $C_a$ as a set of the known action categories, $C_o$ as a set of the known object categories, and $C_{rel}$ as a set of the known HOI relationship categories, in which the known HOI relationship categories consist of the known action and object categories.

Given an input image $I$, the human-object interaction detection aims to predict a set of $(b_h, b_o, c_a)$ triplets, in which $b_h$ denotes the bounding-box of human, $b_o$ is the bounding-box of object $c_o$, and the action $c_a$. For open-set human-object interaction detection, the model is tasked not only with detecting relationships within $C_{rel}$, but also with identifying relationships outside of this set during the inference stage. This includes novel actions beyond $C_a$ and novel objects beyond $C_o$.

**Method**

**Overall Architecture**

Our model is a simple bottom-up approach that classifies the bottom-up features by image-text matching, as shown in Figure 2. Given an image $I$, we first use a frozen open-set object detector to get a set of bounding boxes, and a frozen VLM visual encoder as the backbone to get the global-level features and the patch-level features of the image $I$. Further, the bottom-up visual features of humans and objects are extracted via ROI-Align. Then, these bottom-up visual features paired with global features are encoded by a learnable interaction head to get the result of the relationship classification. Specifically, this process can be divided into three decoupled steps, extracting bounding boxes, extracting bottom-up features, and classification with interaction head.

**Extracting Bottom-up Features.** This allows us to obtain a more general representation and retain the open-set image-text matching ability of the VLM. After passing the image $I$ into the ViT, we can get a CLS token embedding and a sequence of patch embeddings. We consider the CLS token embedding as the global-level feature $G$. Different from conventional bottom-up methods, we extract the instance embeddings $Z = \{z_i\}_{i=1}^N$ with 7x7 ROI-Align by applying the extracted bounding-boxes $B = \{b_i\}_{i=1}^N$ on the patch embeddings.

**Classifying with Interaction Head.** For general purposes, we do not specifically design the interaction head. We apply a simple interaction classifier like the one in UPT (Zhang, Campbell, and Gould 2022). We first flatten the $N$ instance embeddings $Z$ and convert them into a set of unary tokens by a feed-forward network (FFN). And the bounding box spatial positional encodings are injected into

---

**Extracting Bounding-boxes.** In this work, we apply a pre-trained GroundingDINO (Liu et al. 2023) for extracting bounding boxes. GroundingDINO gets the output bounding box in a grounding-liked method. Given $C_o$ object categories, we first combine them into a text query in the form of $'O_1, O_2, \cdots, O_{C_o} '$. Then passing the image $I$ and the text query to get a set of bounding boxes with corresponding labels and scores. After filtering by applying an instance thresholding on the score of instances, there are $N$ bounding boxes with corresponding object labels and scores retained.

---

**Extracting Bottom-up Features.** Unlike traditional bottom-up methods extracting bottom-up features directly from the object detector, we equip with VLM CLIP (Radford et al. 2021) visual encoder, a standard Vision Transformer (ViT), as the backbone for extracting bottom-up features. This allows us to obtain a more general representation and retain the open-set image-text matching ability of the VLM. After passing the image $I$ into the ViT, we can get a CLS token embedding and a sequence of patch embeddings. We consider the CLS token embedding as the global-level feature $G$. Different from conventional bottom-up methods, we extract the instance embeddings $Z = \{z_i\}_{i=1}^N$ with 7x7 ROI-Align by applying the extracted bounding-boxes $B = \{b_i\}_{i=1}^N$ on the patch embeddings.
unary tokens by a unary adapter layer. To this, we get the unary tokens of humans $H = \{h_i\}_{i=1}^K$ and the unary tokens of objects $O = \{o_i\}_{i=1}^N$ according to the corresponding box labels, note that objects can also be human because there exist the relation between the humans. Then the pairs can be formed by combining the human and objects. By fusing the unary tokens and positional encodings of the pairs, the pairwise tokens are obtained. For considering context outside the bottom-up features, the global-level feature $G$ is combined into pairwise tokens and then passing a pair-wise adapter to get the final pairwise tokens $P = \{p_i\}_{i=1}^{K \times N}$ for relationship classification.

### Training with Disentangled Contrastive Learning

Most of the close-set HOI detectors fit a linear classification layer to a specific dataset, which lacks generality in the open-set world. In order to achieve open-set HOI detection, we apply the image-text contrastive learning for training, which has been widely used in open-set object detection (Li et al. 2022; Liu et al. 2023; Wang et al. 2023).

We begin by applying pair-wise relationship contrastive learning, in which the contrastive happened between the normalized pair-wise tokens and the normalized relationship text embeddings. To do this, we have to first form relationship prompts by feeding a series of textual relationship descriptions, e.g. ‘person riding bicycle’, into a prompt ‘a picture of person (action)-ing (object)’$. Then these prompts are passed through the VLM text encoder to get a set of relationship text embeddings $T = \{t_i\}_{i=1}^{C_{rel}}$. We compute the image-text relationship contrastive loss between the pairwise tokens $P$ and the relationship text embeddings $T$, which can be formulated as,

$$
\mathcal{L}_{rel} = \sum_{i=1}^{K \times N} \mathcal{L}_{BCE}(e_i / \tau, y_i),
$$

where $e_i = [\text{sim}(p_1, t_1), \text{sim}(p_2, t_2), \ldots, \text{sim}(p_{C_{rel}}, t_{C_{rel}})]$, $y_i$ is matched multi-hot ground-truth relationship label, $\tau$ is a learnable temperature, and $\mathcal{L}_{BCE}$ is a sigmoid-based focal binary cross-entropy loss (Lin et al. 2017). By training with pair-wise relationship contrastive loss, the open-set capability of VLM is reawakened. However, we found that relationship-label semantic embeddings tend to be indistinguishable when the action is combined with the specific object category. As shown in Figure 3(a), the relationship ‘a picture of person riding bicycle’ has the cosine distances are close to the relationships with the same object ‘bicycle’, which means that there is a stronger discriminability of the object relative to the relation. This is very disadvantageous, after combining specific object categories, the classification of relations is fine-grained, which may have similar vi-

**Figure 3:** The PCA analysis of relationship label space and action label space. The middle column displays the visualization of the embedding space, where embeddings are represented with deeper colors for closer distances to the target relationship. The left column provides the zoomed view of the embeddings in box. The right column showcases the top closely related relationships, sorted by cosine distance to the target relationship.
tokens or pair-wise tokens) of the interaction head and the set of similarity scores split the triplets into related actions and objects, so the relationship triplets we want as queries, and then our experiments. During inference time, we need to first assign the relationship triplets we want as queries, and then split the triplets into related actions and objects, so the relationships outside the queries will be ignored. We compute a set of similarity scores \( s \) between the output tokens (human tokens or pair-wise tokens) of the interaction head and the text embeddings (verb embeddings or relationship embeddings) of the queries. Then we incorporate the object scores \( s_o \), verb similarity scores \( s_v \), and relationship scores \( s_r \) into the final scores of each human-object pair, which can be formulated as,

\[
s = s_o \cdot \sigma(s_v) \cdot \sigma(s_r),
\]

where \( \sigma \) is the sigmoid function.

### Experiments

#### Experimental Setup

**Datasets and Evaluation Metrics:** We perform our experiments on HICO-DET (Chao et al. 2018) and Visual Genome (VG) (Krishna et al. 2017). The HICO-DET dataset consists of 37,536 training images and 9,658 test images. It contains 600 HOI classes which are combined by 80 object classes as MS-COCO (Lin et al. 2014) and 117 action classes. VG dataset contains 108,077 images with a scene-generation graph (SGG) task, which has 100, 298 object classes and 36, 515 relation classes. We extract a subset from VG to form a VG-HOI dataset with 43118 images, which includes 17421 HOI categories (combined with 3542 action classes and 5385 object classes). Following the default setting on HOI detection, we report the mean average precision (mAP) on all datasets. A detected HOI triplet is considered matched with the grounding truth pair when both the predicted human and object bounding boxes have intersection-over-union (IoU) with a ground truth greater than 0.5. For the matched HOI triplets, the one is considered as a true positive if the predicted HOI category is correct.

**Open-set Setups:** We conduct open-set experiments on the HICO-DET and VG-HOI datasets. On HICO-Det, we split some HOIs as unseen setting following the previous works: including Rare-first unseen combination scenario (RF-UC) (Hou et al. 2020), Non-rare-first
UC (NF-UC) (Hou et al. 2020), unseen action scenario (UA) (Liu, Yuan, and Chen 2020) and unseen object scenario (UO) (Bansal et al. 2020). For the UC scenario, each of the action classes and object classes is seen at least once and the object-action pair during the training time. Note that there is an UC (Bansal et al. 2020) setting applying the 5 sets of 120 unseen HOI classes, but we found different results in different previous works. In order to avoid confusion, we did not use this setting because RF-UC and NF-UC settings play the same role. For the UO scenario, the HOIs including the same 12 objects as Functional (Bansal et al. 2020) are selected as unseen. For the UA scenario, the HOIs including the same 22 unseen actions as ConsNet (Liu, Yuan, and Chen 2020) selected as unseen, and there is a UV (Liao et al. 2022) scenario which same as UA, we follow the earliest UA (Liu, Yuan, and Chen 2020) setting. For the HICO-DET zero-shot scenario, we remove the images including the unseen HOIs, train the model on the remaining images and evaluate on the full test set.

On VG-HOI, we train the model with all HICO-DET training images and inference on all the VG-HOI data to simulate the open-set HOI detection. We consider the HOIs occurred in HICO-DET as known relationships and the others as novel HOIs. So there are about 17145 novel relationships, which is large enough to simulate the open-set HOI detection environment to some extent.

Implementations: We benchmark on the UPT (Zhang, Campbell, and Gould 2022) and apply the same settings for all models, unless explicitly specified. In the interaction head, the number of unary adapter layers is 2 and the pair-wise adapter layer is 1. For the open-set object detector, we use the GroundingDINO (Liu et al. 2023) with Swin-B backbone which is pre-trained on COCO, O365 (Shao et al. 2019), GoldG (Kamath et al. 2021), Cap4M, OpenImage (Kuznetsova et al. 2020), ODinW-35 and RefCOCO (Kazemzadeh et al. 2014). For VLM, we use the public pre-trained model CLIP\(^1\) with VIT-B/32 backbone, with an input size of 224 × 224. We feed prompt-engineered texts to the text encoder of CLIP with a prompt template a picture of person \{verb\} \{object\}. We apply the same data augmentation techniques as used in UPT, along with additional CLIP preprocessing to adapt the input image to the pre-trained CLIP visual encoder. The interaction head is trained for 20 epochs with about 7 hours on 2 NVIDIA GTX3090 GPUs, with a batch size of 4 per GPU.

Open-set HOI Detection

We present the results of open-set HOI detection by training with HICO-DET 600 relationship classes and testing on the VG-HOI dataset with 17421 relationships. And comparing the models under HICO-DET zero-shot settings.

From HICO-DET to VG-HOI. We first report the results of the models generalizing from the 37k HICO-DET dataset to the 43k VG-HOI dataset in Table 1. We clarify that GroundingDINO has two problems with this implementation. It cannot ground the excessively large vocabulary

\(^{1}\)https://github.com/openai/CLIP.

<table>
<thead>
<tr>
<th>Method</th>
<th>Type</th>
<th>Unseen</th>
<th>Seen</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>End-to-End methods</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GEN-VLKT</td>
<td>NF-UC</td>
<td>25.05</td>
<td>23.38</td>
<td>23.71</td>
</tr>
<tr>
<td>HOICLIP</td>
<td>NF-UC</td>
<td>25.71</td>
<td>27.18</td>
<td>26.88</td>
</tr>
<tr>
<td>HOICLIP(^*)</td>
<td>NF-UC</td>
<td>26.98</td>
<td>28.10</td>
<td>27.75</td>
</tr>
<tr>
<td>GEN-VLKT</td>
<td>RF-UC</td>
<td>21.36</td>
<td>32.91</td>
<td>30.56</td>
</tr>
<tr>
<td>HOICLIP</td>
<td>RF-UC</td>
<td>23.48</td>
<td>34.47</td>
<td>32.26</td>
</tr>
<tr>
<td>HOICLIP(^*)</td>
<td>RF-UC</td>
<td>25.53</td>
<td>34.85</td>
<td>32.99</td>
</tr>
<tr>
<td>GEN-VLKT</td>
<td>UV</td>
<td>20.96</td>
<td>30.23</td>
<td>28.74</td>
</tr>
<tr>
<td>HOICLIP</td>
<td>UV</td>
<td>23.37</td>
<td>31.65</td>
<td>30.49</td>
</tr>
<tr>
<td>HOICLIP(^*)</td>
<td>UV</td>
<td>24.30</td>
<td>32.19</td>
<td>31.09</td>
</tr>
<tr>
<td>GEN-VLKT</td>
<td>UO</td>
<td>10.51</td>
<td>28.92</td>
<td>25.63</td>
</tr>
<tr>
<td>HOICLIP</td>
<td>UO</td>
<td>9.36</td>
<td>30.32</td>
<td>26.82</td>
</tr>
<tr>
<td>HOICLIP(^*)</td>
<td>UO</td>
<td>16.20</td>
<td>30.99</td>
<td>28.53</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Type</th>
<th>Unseen</th>
<th>Seen</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bottom-up methods</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VCL</td>
<td>NF-UC</td>
<td>16.22</td>
<td>18.52</td>
<td>18.06</td>
</tr>
<tr>
<td>ATL</td>
<td>NF-UC</td>
<td>18.25</td>
<td>18.78</td>
<td>18.67</td>
</tr>
<tr>
<td>FCL</td>
<td>NF-UC</td>
<td>18.66</td>
<td>19.55</td>
<td>19.37</td>
</tr>
<tr>
<td>Ours</td>
<td>NF-UC</td>
<td>27.35</td>
<td>22.09</td>
<td>23.14</td>
</tr>
<tr>
<td>VCL</td>
<td>RF-UC</td>
<td>10.06</td>
<td>24.28</td>
<td>21.43</td>
</tr>
<tr>
<td>ATL</td>
<td>RF-UC</td>
<td>9.18</td>
<td>24.67</td>
<td>21.57</td>
</tr>
<tr>
<td>FCL</td>
<td>RF-UC</td>
<td>13.16</td>
<td>24.23</td>
<td>22.01</td>
</tr>
<tr>
<td>Ours</td>
<td>RF-UC</td>
<td>23.32</td>
<td>30.02</td>
<td>28.53</td>
</tr>
<tr>
<td>ConsNet</td>
<td>UA</td>
<td>14.12</td>
<td>20.02</td>
<td>19.04</td>
</tr>
<tr>
<td>Ours</td>
<td>UA</td>
<td>17.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Functional</td>
<td>UO</td>
<td>11.22</td>
<td>14.36</td>
<td>13.84</td>
</tr>
<tr>
<td>FCL</td>
<td>UO</td>
<td>15.54</td>
<td>20.74</td>
<td>19.87</td>
</tr>
<tr>
<td>ConsNet</td>
<td>UO</td>
<td>19.27</td>
<td>20.99</td>
<td>20.71</td>
</tr>
<tr>
<td>Ours</td>
<td>UO</td>
<td>27.05</td>
<td>27.87</td>
<td>27.23</td>
</tr>
</tbody>
</table>

Table 2: Zero-shot HOI Detection results on HICO-DET dataset. Denote UO and UA(UV) as unseen objects and unseen action scenarios, RF-UC and NF-UC as rare-first and non-rare-first unseen combination scenarios. The HOICLIP\(^*\) applies a training-free enhancement, which is orthogonal to our model. We primarily focus on presenting results for unseen HOIs, which are most related to open-set HOI detection. The seen and full results are for reference only, and colored with gray.

of categories. So we consider a known objects setting, we only feed the set of object classes present in the images into the query of GroundingDINO during inference on VG-HOI. This reduces the difficulty of the problem and better simulates the open-set detection because the class vocabulary is large enough, but the user usually needs to get the ones who want. In addition, we found that GroundingDINO fails to detect about 1455 object categories on VG-HOI even though pre-trained on 7 datasets due to bias toward common objects like the COCO 80 classes, which leads to 0 AP on some relations.

Despite these difficulties, our model still achieves 9.20 AP on VG-HOI with over 17k HOIs and 9.08 AP on HICO-DET zero-shot scenario, we remove the images including the same 12 objects as Functional (Bansal et al. 2020) scenario which same as UA, we follow the earliest UA (Liu, Yuan, and Chen 2020) setting. For the UA scenario, the HOIs in-
As far as we know, none of the existing methods have the ability to detect over 17k HOI relationships.

**Zero-shot HOI Detection on HICO-DET.** We compare the model with existing zero-shot HOI detection methods. The compared methods include the end-to-end methods and bottom-up methods. The end-to-end methods including GEN-VLKT (Liao et al. 2022) and HOICLIP (Ning et al. 2023), in which the GEN-VLKT distills the visual features from CLIP and the HOICLIP aggregates features from CLIP and DETR (Carion et al. 2020), they are all finetune the object detector on the seen pairs. The bottom-up methods including VCL (Hou et al. 2020), ATL (Hou et al. 2021a), FCL (Hou et al. 2021b), Functional (Bansal et al. 2020), and ConsNet (Liu, Yuan, and Chen 2020), in which the VCL, FCL, ATL, and Functional are all holding a novel idea that the actions can compositional to functional similar objects, but also limited to the seen actions.

We list the results in Table 2. Our model outperforms previous works on NF-UC and UO settings by a large margin, with 27.35 mAP and 27.05 mAP, which shows the generalize to combine the actions with seen or novel objects. And the model also shows the ability to detect novel actions compared to previous bottom-up methods.

**Ablation Study**

We perform ablation studies in this subsection. The experiments are based on a full-supervised setting, in which there are 138 low-frequency classes (Rare) and 462 classes (Non-rare) with more training instances. We first analyze the effect of the frozen VLM backbone and the frozen GroundingDINO, then we verify the validity of the proposed disentangled contrastive learning and neg-pair debias metric.

The results are shown in Table 3. We first extract the bottom-up features from the CLIP visual backbone, using only the bounding boxes from DETR. As shown in Table 3 line 3, this leads to better performance on rare classes, illustrating more general and balanced representations of the VLM. Then, we extract bounding boxes from a frozen GroundingDINO, which results in a balanced performance on both rare and nonrare classes, as shown in line 4. But the nonrare performance decreases by about 3% mAP, this is potential because the DETR finetuning on HICO-DET shows a dataset biased detection. Such as, the person can be detected inside the car although the person is completely invisible, as shown in Figure 4 (b). However, this bias will inevitably be hindered when encountering the open-set scene.

In order to achieve the open-set HOI detection, we replace the linear classifier with a pair-wise relation contrastive head. As shown in line 5, this makes the rare performance decrease by about 3 mAP due to the pair-wise relation contrastive learning fitting the long-tail pattern of HICO-DET to some extent. Further, we add proposed verb contrastive learning, which brings a significant performance improvement in both rare and non-rare classes, as shown in line 6. When the neg-pair debias is added, as shown in line 7, there is an improvement in rare classes, which shows the idea reduces the bias from significant head relationships.

**Limitations:** The performance of our model is currently constrained by the open-set capabilities of open-set object detectors and VLMs. We found that the pretrained VLM has difficult in some relationships with ambiguity or some inappropriate relationship prompts, such as “a photo of person jumping car” which actually means the car jumping away the road in the dataset. Using more diverse unstructured data to reconstruct the representation (Yuan et al. 2022; Zheng, Xu, and Jin 2023) of the basic model for focusing on relationship classification, and the fine-grained person-centered features (Li et al. 2020) could be beneficial in this regard. In addition, we did not explicitly design the interaction head, so a well-designed interaction head could significantly enhance the open set capability of model.

**Conclusions**

In this work, we present a simple disentangled bottom-up method for open-set human object interaction detection by leveraging the open-set object detector and the visual and language model. Our model can recognize novel actions and combine the actions with novel objects by training unary and pair-wise adapters only with a novel disentangled image-text contrastive learning method. Experiments show the effectiveness of the model on open-set human object interaction detection, and the general to over 17k relationships by training on 600 known relationships. We hope this work can bring new insight and inspire future works on the open-set human object interaction detection community.

**Acknowledgments**

This work was supported by National Key R&D Program of China (No.2022ZD0118201), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), the National Natural Science Fund for Young Scholars of China (No. 62302411), China Postdoctoral Science Foundation (No.2023M732948), and the Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001).
References


