Toward Open-Set Human Object Interaction Detection

Mingrui Wu[1,2*], Yuqi Liu[1*], Jiayi Ji[1†], Xiaoshuai Sun[1,2], Rongrong Ji[1,2]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University
[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.

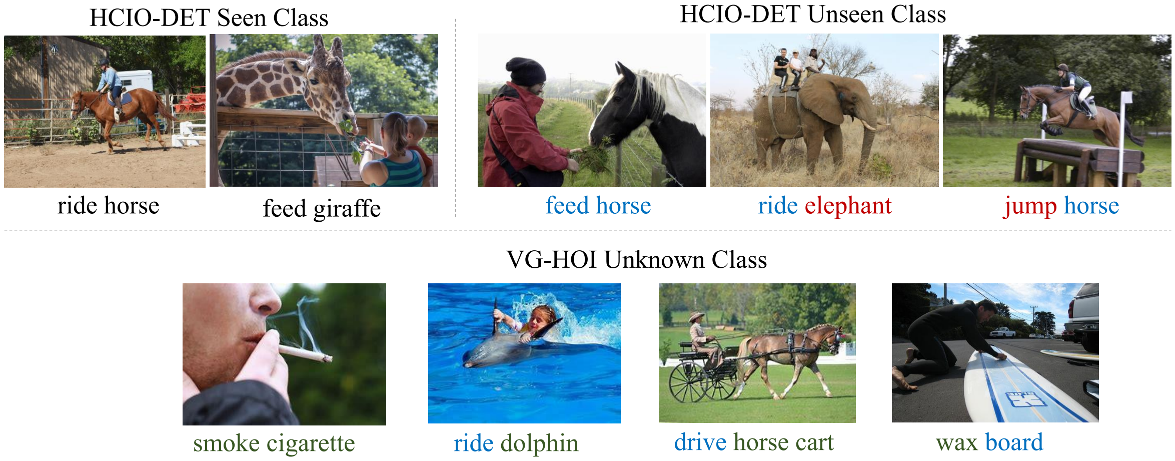
Most of the existing methods cannot be directly evaluated on VG-HOI due to their limited focus either on specific actions or restricted object categories. Current mainstream zero-shot HOI detection paradigms (Liao et al. 2022; Ning et al. 2023) primarily function within closed-set object scenarios, involving 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 openset object detector. Then we attach lightweight unary and pair-wise adapters to establish connections be-

[*] Equal Contribution.

[†] Corresponding Author.



HCIO-DET Seen Class

HCIO-DET Unseen Class

VG-HOI Unknown Class

Figure 1: The comparision 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.

tween the bottom-up visual features and the text embeddings of labels. By exclusively training the adapters using a disentangled image-text contrastive learning metric, we can retain the learned knowledge of pretrained model. This approach enables our model to leverage the capabilities of the open-set object detector and the VLM, aiding in the detection of new action categories and the combination of actions with new object categories.

To summarize, our contributions are:

- We introduce the VG-HOI dataset, a comprehensive evaluation benchmark comprising over 17k relationships designed for open-set scenarios.
- We present a simple disentangled bottom-up paradigm, by training the unary and pair-wise adapters with disentagled image-text contrastive learning, to achieve the open-set HOI detection.
- Our model demonstrates its success in evaluating over 17k HOI classes on the VG-HOI dataset, despite being trained on a modest 600 classes. This underscores its potential to detect a wide spectrum of HOI classes.

Related Works

Human-Object Interaction Detection

Current mainstream HOI detection method can be categorized into two main frameworks, bottom-up models (Chao et al. 2018; Zhang, Campbell, and Gould 2022; Park, Park, and Lee 2023) and end-to-end models (Tamura, Ohashi, and Yoshinaga 2021; Liao et al. 2022; Ning et al. 2023; Kim, Jung, and Cho 2023). Bottom-up approaches typically use a detector to detect all people and objects in the image, and then an interaction classifier to implement HOI classification. And the end-to-end models perform object detection and interaction classification simultaneously. However, the end-to-end models are more suitable for closed-set scenarios. Thus, in this work, we construct our model based on the bottom-up paradigm.

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.

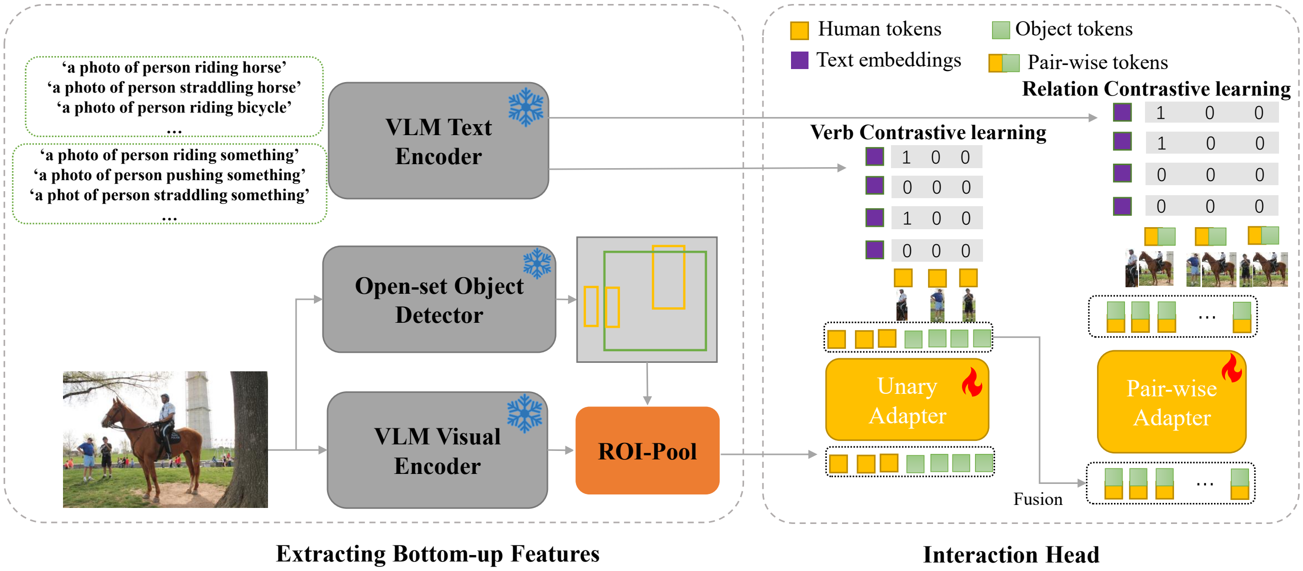

'a photo of person riding horse'
'a photo of person straddling horse'
'a photo of person riding bicycle'
...
'a photo of person riding something'
'a photo of person pushing something'
'a phot of person straddling something'
...
VLM Text Encoder
Open-set Object Detector
VLM Visual Encoder
ROI-Pool
Extracting Bottom-up Features
Human tokens
Object tokens
Text embeddings
Pair-wise tokens
Relation Contrastive learning
Verb Contrastive learning
1 0 0
0 0 0
1 0 0
0 0 0
1 0 0
1 0 0
0 0 0
0 0 0
Unary Adapter
Pair-wise Adapter
...
Fusion
Interaction Head

Figure 2: Overview of our method. We begin by utilizing a pre-trained open-set object detector and a frozen VLM visual backbone to extract the bottom-up features. Subsequently, we engage in relationship contrastive learning, where the encoded pair-wise tokens are contrasted with relationship text embeddings. Additionally, we conduct verb contrastive learning between the encoded human tokens with action text embeddings.

Preliminary

Problem Formulation: Denote $\mathcal{C}_a$ as a set of the known action categories, $\mathcal{C}_o$ as a set of the known object categories, and $\mathcal{C}_{rel}$ 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 $\langle \boldsymbol{b}_h, \boldsymbol{b}_o, c_a \rangle$ triplets, in which $\boldsymbol{b}_h$ denotes the bounding-box of human, $\boldsymbol{b}_o$ 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 $\mathcal{C}_{rel}$, but also with identifying relationships outside of this set during the inference stage. This includes novel actions beyond $\mathcal{C}_a$ and novel objects beyond $\mathcal{C}_o$.

Method

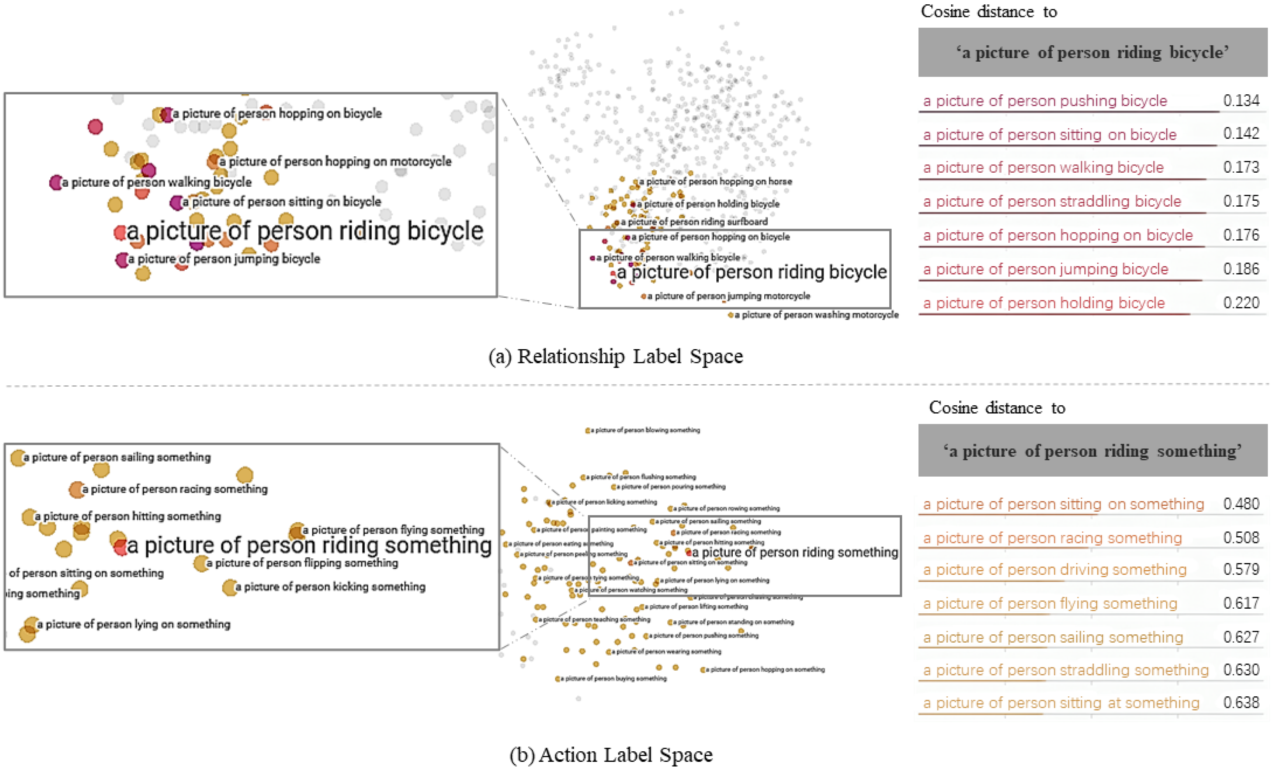
Overall Architecture

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 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 $\mathcal{C}_o$ object categories, we first combine them into a text query in the form of '$O_1.O_2.\cdots.O_{\mathcal{C}_o}$'. 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 $\mathcal{Z} = \{\boldsymbol{z}_i\}_{i=1}^{N}$ with 7x7 ROI-Align by applying the extracted bounding-boxes $\mathcal{B} = \{\boldsymbol{b}_i\}_{i=1}^{N}$ on the patch embeddings.

Classification 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 $\mathcal{Z}$ 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



(a) Relationship Label Space

(b) Action Label Space

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.

unary tokens by a unary adapter layer. To this, we get the unary tokens of humans $\mathcal{H} = \{\boldsymbol{h}_i\}_{i=1}^{K}$ and the unary tokens of objects $\mathcal{O} = \{\boldsymbol{o}_i\}_{i=1}^{N}$ 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 $\mathcal{P} = \{\boldsymbol{p}_i\}_{i=1}^{K\times N}$ 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 $\langle action \rangle$-ing $\langle object \rangle$'. Then these prompts are passed through the VLM text encoder to get a set of relationship text embeddings $\mathcal{T} = \{\boldsymbol{t}_i\}_{i=1}^{\mathcal{C}_{rel}}$. We compute the image-text relationship contrastive loss between the pairwise tokens $\mathcal{P}$ and the relationship text embeddings $\mathcal{T}$, which can be formulated as,

$$\mathcal{L}_{\text{rel}} = \sum_{i=1}^{K\times N} \mathcal{L}_{\text{BCE}}(\boldsymbol{e}_i/\tau, \boldsymbol{y}_i), \tag{1}$$

where $\boldsymbol{e}_i = [sim(\boldsymbol{p}_i, \boldsymbol{t}_1), sim(\boldsymbol{p}_i, \boldsymbol{t}_2), \cdots, sim(\boldsymbol{p}_i, \boldsymbol{t}_{\mathcal{C}_{rel}})]$, $\boldsymbol{y}_i$ is matched multi-hot ground-truth relationship label, τ is a learnable temperature, and $\mathcal{L}_{\text{BCE}}$ is a sigmoid-based focal binary cross-entropy vloss (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 very similar vi-

Methods	Source Dataset	Target Dataset	Full	Known	Novel
Ours	HICO-DET (class=600)	VG-HOI (class=17421)	9.20	16.74	9.08
w/o Neg-pair Debias	HICO-DET (class=600)	VG-HOI (class=17421)	9.11	16.38	9.00
w/o Verb Contrastive	HICO-DET (class=600)	VG-HOI (class=17421)	8.85	16.41	8.73

Table 1: Open-set test on VG-HOI dataset with the models training on HICO-DET dataset. The Known and Novel refer to whether the relationships occurred in the HICO-DET dataset or not.

sual representations between the different relations.

So what happens if the actions are not combined with the specific object categories? We try to combine the actions with ‘something’ instead of the specific objects. As shown in Figure 3(b), in the action label space, there is a significant discriminability between action embeddings. So, performing verb contrastive learning is on the agenda. However, it is not feasible to directly replace pair-wise contrastive learning with verb contrastive learning, because the pair-wise features of the same action combined with different objects will tend to be consistent, which will excessively change the visual features of VLM and affect its open-set general ability. At the same time, we note that person-centric features tend to be consistent when encountering functionally similar objects (Shen et al. 2018; Bansal et al. 2020). This inspires us to perform a verb contrastive learning between the human tokens and the actions, which can be formulated as,

$$\mathcal{L}_{\text{verb}} = \sum_{i=1}^{K} \mathcal{L}_{\text{BCE}}(\boldsymbol{e}_i/\tau, \boldsymbol{y}_i), \tag{2}$$

where $\boldsymbol{e}_i = [sim(\boldsymbol{h}_i, \boldsymbol{t}_1), sim(\boldsymbol{h}_i, \boldsymbol{t}_2), \cdots, sim(\boldsymbol{h}_i, \boldsymbol{t}_{\mathcal{C}_a})]$, $\boldsymbol{y}_i$ is matched multi-hot ground-truth action label. Such a verb contrastive learning can better focus on person-centric action features, without having to consider the specific objects. This helps the model generalize to unseen object scenes by learning from the seen objects.

In addition, we revisit the currently widely used HOI detection dataset, such as HICO-DET (Chao et al. 2018). We found that there are subjective bias inherent in annotators in the dataset, *e.g.*, the annotations bias toward the the significant relationship in the picture, and the non-significant relationship is left out, as shown in Figure 4 (a). This leads to a large number of false negative samples in the data, which in turn affects the open set performance of the model. We deal with this problem through negative pair debias, which simply sets a smaller loss weight for negative samples that do not match the ground-truth pairs.

Inference

We can use two forms for the model inference, including multi-label contrastive classification and one-shot query. The multi-label contrastive classification is mainly used in 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_a, and relationship scores s_r into the final scores of each human-object pair, which can be formulated as,

$$s = s_o \cdot \sigma(s_a) \cdot \sigma(s_r), \tag{3}$$

where the σ is the sigmoid function.

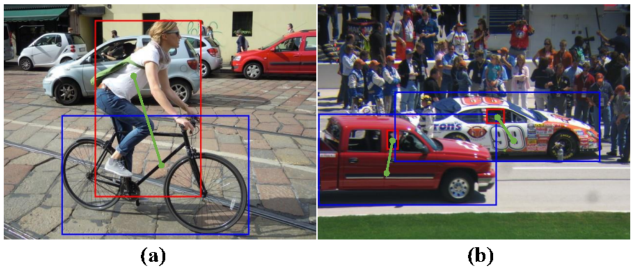


Figure 4: The visualization examples of annotation bias.

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-graph generation (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 in at least one HOI action-object pair during the training time. Note that there is a 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 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 relationships. Despite these difficulties, our model still achieves 9.20 mAP on VG-HOI with over 17k HOIs and 9.08 mAP on the unknown relationships which has never been seen before. In addition, we found that reducing the impact of the false negative on HICO-DET improves the 0.09 mAP, which is significant because a total of over $0.26 \times 17000 = 4420$ AP has been improved. And the proposed verb contrastive learning improves the model by 0.26 mAP. The results illustrate the ability of our model to generalize to open-set HOIs.

Method	Type	Unseen	Seen	Full
End-to-End methods				
GEN-VLKT	NF-UC	25.05	23.38	23.71
HOICLIP	NF-UC	25.71	27.18	26.88
HOICLIP*	NF-UC	26.39	28.10	27.75
GEN-VLKT	RF-UC	21.36	32.91	30.56
HOICLIP	RF-UC	**23.48**	34.47	32.26
HOICLIP*	RF-UC	25.53	34.85	32.99
GEN-VLKT	UV	20.96	30.23	28.74
HOICLIP	UV	**23.37**	31.65	30.49
HOICLIP*	UV	24.30	32.19	31.09
GEN-VLKT	UO	10.51	28.92	25.63
HOICLIP	UO	9.36	30.32	26.82
HOICLIP*	UO	16.20	30.99	28.53
Bottom-up methods				
VCL	NF-UC	16.22	18.52	18.06
ATL	NF-UC	18.25	18.78	18.67
FCL	NF-UC	18.66	19.55	19.37
Ours	NF-UC	**27.35**	22.09	23.14
VCL	RF-UC	10.06	24.28	21.43
ATL	RF-UC	9.18	24.67	21.57
FCL	RF-UC	13.16	24.23	22.01
Ours	RF-UC	**23.32**	30.09	28.53
ConsNet	UA	14.12	20.02	19.04
Ours	UA	**17.92**	28.13	26.43
Functional	UO	11.22	14.36	13.84
FCL	UO	15.54	20.74	19.87
ConsNet	UO	19.27	20.99	20.71
Ours	UO	**27.05**	27.87	27.73

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.

[1] https://github.com/openai/CLIP.

Ablation	Full	Rare	Non-rare
UPT	31.65	26.52	33.18
Frozen VLM Backbone	31.49	28.23	32.46
+Frozen GroundingDINO	29.30	29.29	29.30
Pair-wise Contrastive	28.98	26.19	29.82
+Verb Contrastive	29.85	27.38	30.59
+Neg-pair Debias	29.91	28.42	30.35

Table 3: Ablation study on HICO-DET dataset under full-supervised setting. The Rare and Non-rare denote the low-frequency and high-frequency relationships respectively.

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 adapaters 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

Bansal, A.; Rambhatla, S. S.; Shrivastava, A.; and Chellappa, R. 2020. Detecting human-object interactions via functional generalization. In *AAAI*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*.

Chao, Y.-W.; Liu, Y.; Liu, X.; Zeng, H.; and Deng, J. 2018. Learning to detect human-object interactions. In *WACV*.

Hou, Z.; Peng, X.; Qiao, Y.; and Tao, D. 2020. Visual compositional learning for human-object interaction detection. In *ECCV*.

Hou, Z.; Yu, B.; Qiao, Y.; Peng, X.; and Tao, D. 2021a. Affordance transfer learning for human-object interaction detection. In *CVPR*.

Hou, Z.; Yu, B.; Qiao, Y.; Peng, X.; and Tao, D. 2021b. Detecting human-object interaction via fabricated compositional learning. In *CVPR*.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*.

Kim, S.; Jung, D.; and Cho, M. 2023. Relational Context Learning for Human-Object Interaction Detection. In *CVPR*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022. Grounded language-image pre-training. In *CVPR*.

Li, Y.-L.; Xu, L.; Liu, X.; Huang, X.; Xu, Y.; Wang, S.; Fang, H.-S.; Ma, Z.; Chen, M.; and Lu, C. 2020. Pastanet: Toward human activity knowledge engine. In *CVPR*.

Liao, Y.; Zhang, A.; Lu, M.; Wang, Y.; Li, X.; and Liu, S. 2022. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *CVPR*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *ICCV*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Liu, Y.; Yuan, J.; and Chen, C. W. 2020. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *ACM Multimedia*.

Ning, S.; Qiu, L.; Liu, Y.; and He, X. 2023. HOICLIP: Efficient Knowledge Transfer for HOI Detection with Vision-Language Models. In *CVPR*.

Park, J.; Park, J.-W.; and Lee, J.-S. 2023. ViPLO: Vision Transformer based Pose-Conditioned Self-Loop Graph for Human-Object Interaction Detection. In *CVPR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*.

Shen, L.; Yeung, S.; Hoffman, J.; Mori, G.; and Fei-Fei, L. 2018. Scaling human-object interaction recognition through zero-shot learning. In *WACV*.

Tamura, M.; Ohashi, H.; and Yoshinaga, T. 2021. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*.

Wang, S.; Duan, Y.; Ding, H.; Tan, Y.-P.; Yap, K.-H.; and Yuan, J. 2022. Learning transferable human-object interaction detector with natural language supervision. In *CVPR*.

Wang, S.; Yap, K.-H.; Ding, H.; Wu, J.; Yuan, J.; and Tan, Y.-P. 2021. Discovering human interactions with large-vocabulary objects via query and multi-scale detection. In *ICCV*.

Wang, S.; Yap, K.-H.; Yuan, J.; and Tan, Y.-P. 2020. Discovering human interactions with novel objects via zero-shot learning. In *CVPR*.

Wang, Z.; Li, Y.; Chen, X.; Lim, S.-N.; Torralba, A.; Zhao, H.; and Wang, S. 2023. Detecting everything in the open world: Towards universal object detection. In *CVPR*.

Wu, M.; Gu, J.; Shen, Y.; Lin, M.; Chen, C.; and Sun, X. 2023. End-to-end zero-shot hoi detection via vision and language knowledge distillation. In *AAAI*.

Yuan, H.; Jiang, J.; Albanie, S.; Feng, T.; Huang, Z.; Ni, D.; and Tang, M. 2022. Rlip: Relational language-image pre-training for human-object interaction detection. *NeurIPS*.

Zhang, F. Z.; Campbell, D.; and Gould, S. 2022. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *CVPR*.

Zhang, X.; Yin, B.; Lin, Z.; Hou, Q.; Fan, D.-P.; and Cheng, M.-M. 2023. Referring Camouflaged Object Detection. *arXiv preprint arXiv:2306.07532*.

Zhao, L.; Yuan, L.; Gong, B.; Cui, Y.; Schroff, F.; Yang, M.-H.; Adam, H.; and Liu, T. 2023. Unified Visual Relationship Detection with Vision and Language Models. *arXiv preprint arXiv:2303.08998*.

Zheng, S.; Xu, B.; and Jin, Q. 2023. Open-Category Human-Object Interaction Pre-Training via Language Modeling Framework. In *CVPR*.