ZERO-SHOT HUMAN-OBJECT INTERACTION RECOGNITION BY BRIDGING GENERATIVE AND CONTRASTIVE IMAGE-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing studies in Human-Object Interaction (HOI) recognition rely heavily on costly human-annotated labels, limiting the application of HOI in real-world scenarios like retail and surveillance. To address this issue, this paper investigates a new zero-shot setup where no HOI labels are available for any image. We propose a novel *heterogenous teacher-student* framework that bridges two types of pre-trained models, namely contrastive (e.g., CLIP) and generative (e.g., GIT) image-language models. To bridge their gap, we introduce *pseudo-label distillation* to extract HOI probabilities from image captions to train the student classifier. Our method leverages the complementary strengths of both models. As a result, the student model has "the best of two worlds", e.g., the compact backbone of a contrastive model and the fine-grained discriminability of a generative (captioning) model. It achieves 49.6 mAP on the HICO dataset without any ground-truth labels, becoming a new state-of-the-art that outperforms previous supervised approaches. Code will be released upon acceptance.

1 INTRODUCTION

Human-Object Interaction (HOI) recognition is attracting growing interests (Li et al., 2020b; Tamura et al., 2021; Ma et al., 2022) due to its essential role in scene understanding. It retrieves all interactions that exist in the image, where each interaction class is a <*verb, object*> pair, e.g., <*ride, bicycle*>. HOI datasets have high annotation cost for two reasons: (1) large number of classes (e.g., 600 classes in HICO (Chao et al., 2015)) with fine-grained verb and object concepts, and (2) each class needs to be labeled separately as it is a multi-label problem. However, existing methods in HOI heavily rely on human annotations, which greatly undermines the applicability of these methods in real-world scenarios like retail, healthcare, and surveillance, where tailored sets of HOIs need to be labeled. This study aims to remove the strong dependency on such human annotation by introducing a zero-shot method for HOI recognition.

We introduce a new zero-shot HOI setting where only the list of class names and unlabeled images are provided during training, and no ground truth is available for any class. Compared with existing studies that are partially zero-shot (Shen et al., 2018; Ma et al., 2022), this setting is even more challenging, yet it resembles the real-world scenario where users can define a list of HOI classes and expect a tailored HOI model without data annotation efforts.

A natural solution to achieve zero-shot HOI classification is to leverage large image-text pre-trained models, such as image captioning models and contrastive models. However, neither of the models perform satisfactorily alone in our HOI experiments, because their training data is not HOI-specific, and they are not trained to perform classification tasks. Nonetheless, we found that their strengths are complementary: the captioning model can produce fine-grained verb and object concepts in its output, while the contrastive model provides modality-aligned representations with a relatively compact backbone. Can we deliver an HOI model that has "the best of two worlds"?

This paper demonstrate a novel *heterogenous teacher-student* framework that incorporates two types of image-text pre-trained models, including GIT (Wang et al., 2022), a large image captioning model as the teacher, and a CLIP-based (Radford et al., 2021) classifier as the student. To bridge the gap between the two models, we propose "pseudo-label distillation", which extracts HOI class

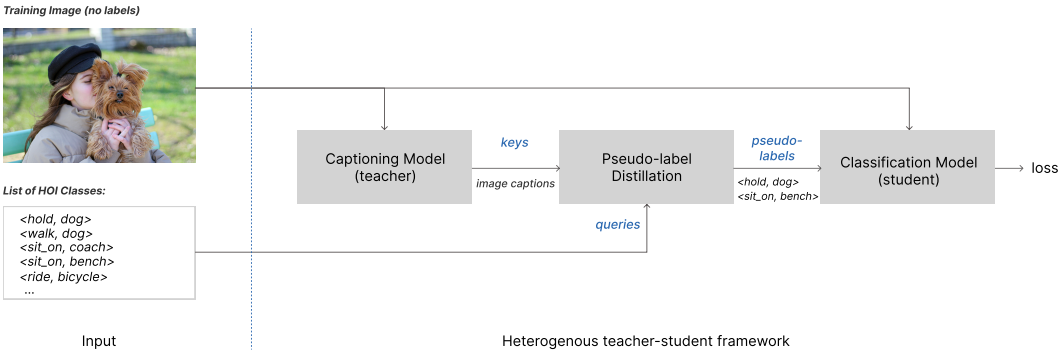

Figure 1: **Heterogenous teacher-student framework** for zero-shot HOI classification. Heterogenous means the teacher and student perform different tasks. The teacher generates captions on multiple regions in the image. Pseudo-label Distillation serves as a bridge which extracts HOI knowledge from captions into class probabilities to train the student classifier. The teacher and student models perform 35.6 mAP and 25.8 mAP, respectively. After learning, the student is improved to 41.1 mAP.

probabilities from image captions to train the student classifier. SimCSE (Gao et al., 2021) and WordNet (Miller, 1995) are incorporated in this step to better distill the HOI knowledge while filtering out non-HOI noise. The framework is different from conventional teacher-student methods in two significant ways. First, our teacher and student models are heterogenous as they perform different tasks. Second, the teacher is not supervised on the downstream HOI task.

Our zero-shot method achieves solid performance in multiple HOI classification benchmarks (49.6 mAP on HICO (Chao et al., 2015) and 35.2 mAP on MPII (Andriluka et al., 2014)), outperforming existing supervised methods. With only $1/7$ the size of the teacher model, the student learns to perform on par with the teacher. Extensive experiments show the indispensable role of each of our components and a satisfying transferability to the instance-level HOI detection task (11.6 mAP on HICO-DET (Chao et al., 2018) with an off-the-shelf object detector). The main contributions of this work are summarized as follows:

- We introduce a new zero-shot setting for HOI recognition.
- We demonstrate a novel heterogenous teacher-student framework that leverages image-text pre-trained models of diverse types.
- We provide an HOI recognition model that outperforms supervised baselines without any ground-truth data, removing the strong dependency on human annotations in the HOI task.

2 Related Work

2.1 HOI Recognition

HOI recognition includes an image-level classification task and an instance-level detection task. In the classification task, positions of the HOIs are unknown, therefore most existing methods (Gkioxari et al., 2015; Mallya & Lazebnik, 2016; Girdhar & Ramanan, 2017; Fang et al., 2018) depend on object detectors to extract human and object regions and conduct training with Multiple Instance Learning (MIL) (Maron & Lozano-Pérez, 1998). The state of the arts are PaStaNet (Li et al., 2020b) and HAKE (Li et al., 2019), which infer the HOI by reasoning from body part-level actions (extra labels required). HOI detection methods fall into three categories: (1) two-stage methods (Gao et al., 2018; Gkioxari et al., 2018; Gao et al., 2020; Liu et al., 2020a; Li et al., 2020b; Kim et al., 2020b; Zhang et al., 2021b) that detect objects first, then classify the HOI base on regional features, (2) one-stage methods (Liao et al., 2020; Kim et al., 2020a; Zhong et al., 2021; Hou et al., 2021a; Wang et al., 2020) that detect objects and HOI regions in parallel and then match them, and (3) end-to-end methods (Chen et al., 2021; Kim et al., 2021; Zou et al., 2021; Tamura et al., 2021; Zhang et al., 2021a; Dong et al., 2022; Qu et al., 2022; Zhou et al., 2022; Iftekhar et al., 2022; Kim et al., 2022), the recent trend which follows a DETR-based architecture (Carion et al., 2020).

A partially zero-shot setting was previously studied under the name of "zero-shot HOI recognition". The classes are split into a *seen* set (with labels) and an *unseen* set (without labels). The unseen classes can be novel combinations of known elements (verbs and object types). The model is trained on the seen set and is evaluated on the unseen set. This is different from our zero-shot setting as all the classes in our setting is unseen. For clarity, we call their setting "partially zero-shot" and ours "zero-shot". Ma et al. (2022) study partially zero-shot HOI classification based on auxiliary training objectives and EsViT (Li et al., 2021a). Methods on partially zero-shot HOI detection (Shen et al., 2018; Hou et al., 2020; 2021a) focus on learning features of individual verbs, objects, or object affordance to improve the generalization to new HOIs.

Besides these work, a list of methods (Li et al., 2021b; Peyre et al., 2019; Xu et al., 2019; Liao et al., 2022; Iftekhar et al., 2022; Qu et al., 2022) leverage text features or image-text pre-training as an additional feature, a weight initialization, or a distillation source.

2.2 Language models and image-text pre-training

As a language model, BERT (Devlin et al., 2018) established new baselines with unsupervised training and a transformer network. When adapted to text similarity or retrieval tasks, (Reimers & Gurevych, 2019; Li et al., 2020a; Su et al., 2021) adjusts the BERT embedding space to be more isotropic for improved performance, and SimCSE (Gao et al., 2021) shows that a contrastive fine-tuning on BERT with dropout as augmentation can achieve the state of the art performance.

Recent years also witness remarkable progress in image-text pre-training. Contrastive and generative models can be pre-trained on millions of image-text pairs. In CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021) and Florence (Yuan et al., 2021), the contrastive loss is applied to align image and text representations, which significantly boost the performance on various downstream tasks. DEFR (Jin et al., 2022) demonstrates a significant improvement in HOI recognition with proper initialization and fine-tuning. For generative models, GIT (Wang et al., 2022) achieves state-of-the-art in image captioning with the contrastive-based image representation. In the HOI scenario, the captioning models have fine-grained knowledge of verb and object concepts since their pre-training task enforces reconstructing each work in the whole sentence. In contrast, contrastive models are prone to ignore detailed features, e.g., visual relationships and small objects (Thrush et al., 2022).

3 Zero-shot HOI recognition defined

In this section, we define the zero-shot HOI recognition setting. Formally, the training set includes unlabeled images $X = \{I_1, I_2, \ldots, I_N\}$ and a list of HOI categories $Y = \{y_1, y_2, ...y_C\}$, where N and C are the number of images and classes, respectively. Each HOI category y_i is in terms of plain text like "hold dog", "ride bicycle". The test set has both images and labels for performance evaluation. This setup differs from previous studies (Shen et al., 2018; Ma et al., 2022) where training images have labels on a seen subset of classes (e.g., 480 seen classes for training and 120 unseen classes for testing). In other words, our zero-shot setting can be considered as an extreme case of previous studies where all classes are unseen.

The proposed zero-shot setup resembles real-world scenarios such as retail and surveillance more faithfully. Methods developed under this setting can significantly reduce human annotation cost. Users will be released from the cumbersome labeling process and only need to define a list of "interactions of interest" Y and record images X.

4 Heterogenous Teacher-Student Framework

We use a teacher-student framework that employs pre-trained image-text models to achieve zero-shot in HOI recognition. The framework is heterogenous because the teacher and student models perform different tasks. The goal of this framework is twofold. First, we want to leverage the complementary strengths of both models. Second, we aim to deliver a small model ($1/7$ of the teacher's size) with adequate performance.

Our heterogenous teacher-student framework is different from conventional knowledge distillation in three ways: (1) our teacher and student models are pre-trained for different tasks. The teacher is an

image captioner and the student is a classifier, (2) both the teacher and the student are not supervised on the downstream HOI task, and (3) knowledge is not transfered through feature distillation, but pseudo-label distillation which serves as the bridge.

In this section, we introduce the details of the framework, and discuss the rationale of our approach.

4.1 TEACHER AND STUDENT MODELS

Teacher: we use an image captioning model which predicts a text sequence τ to describe an image I. Although the model is trained for generic image captioning, it carries some HOI knowledge in the form of verb and object concepts in the output. We choose GIT (Wang et al., 2022), which has 0.7 billion parameters pre-trained on 0.8 billion image-text pairs from the web. Due to the large size, it is not feasible to be deployed.

Student: the student is an image classifier for HOI classes. The architecture follows Jin et al. (2022), where the network has a vision backbone and a linear classification head, both initialized by CLIP. We found that it crucial to use an image-text contrastive pre-trained model as initialization because the backbone and the classification head (initialized by text embeddings) is already aligned before fine-tuned for HOI.

The large pre-training data provides zero-shot capability for both GIT and CLIP. The teacher achieves 35.6 mAP on HICO when we convert the captioning output to class logits by evaluating text similarities. The student model achieves 25.8 mAP (the zero-shot performance of CLIP ViT-B/32). The performance of both models is not yet satisfactory. However, their strengths are complementary. The captioning model can recognize fine-grained verb and object concepts, because during the generative pre-training, it is required to reconstruct each word in the caption. The CLIP model, on the other hand, is prone to ignore details like small objects and visual relationships (Thrush et al., 2022). However, CLIP can initialize the student properly, giving it a good starting point to learn efficiently from the captioning model. Thus, we pursue to bridge them to have the best of two worlds. As a result, the student has only $1/7$ number of parameters of the teacher, but can perform on par after learning.

4.2 THE BRIDGE: PSEUDO-LABEL DISTILLATION

The caption is in the format of natural language description, which is different from the discrete HOI categories required to train the student. To mitigate this gap, we propose *pseudo-label distillation* to essentially query the probabilities of all HOI classes $Y = \{y_1, y_2, \dots, y_C\}$ from the captions. For example, given a caption "a man sitting on a couch holding a puppy", we want high pseudo-label values for `<sit_on, couch>` and `<hold, dog>`, and low values for other classes. A naive approach is to use a text-similarity model to evaluate the semantic similarities between the caption and the HOI classes, and use the similarities as pseudo-labels. A text similarity model can handle the vocabulary gap from captions to HOI classes, for example, to match "puppy" from the captions to "dog" in the HOI classes.

Formally, let τ denote an image caption, and t_i denote a sentence converted from the i^{th} HOI class (e.g., `<ride, bicycle>` is converted to "a person riding a bicycle"). Let $f(\cdot)$ denote a language model that encodes text sequences into embeddings. Here, we use SimCSE (Gao et al., 2021), which is a BERT (Devlin et al., 2018) model fine-tuned for text similarity tasks. The choice of SimCSE is due to its isotropic embedding space which promotes better performance. Firstly, we generate embeddings for caption τ and HOI class t_i as follows:

$$\boldsymbol{\phi} = f(\tau), \quad \boldsymbol{q}_i = f(t_i) \tag{1}$$

Then we use cosine similarity to roughly estimate the probability over HOI classes as:

$$P(y_i|\tau) \approx \frac{\boldsymbol{\phi}\, \boldsymbol{q}_i}{\|\boldsymbol{\phi}\|\|\boldsymbol{q}_i\|} \tag{2}$$

This estimation is rough for the HOI task, as the cosine distance of text embeddings reflects the semantic similarity of objects instead of the desired visual similarity. For example, "dog" and "elephant" are semantically similar but visually different. Visual similarity is usually sparser than

semantic similarity. To remove the noise from semantically similar but visually different false positives, we use WordNet (Miller, 1995) to sparsify the pseudo-labels as follows:

$$P(y_i|\tau) \approx \pi(o_i, o_\tau) \frac{\boldsymbol{\phi}\, \boldsymbol{q}_i}{\|\boldsymbol{\phi}\|\|\boldsymbol{q}_i\|} \tag{3}$$

where $\pi(o_i, o_\tau)$ represents the WordNet-based mask between the object in the i^{th} HOI class and the object in caption τ. It equals 1 if the objects are "synsets" or hypernyms of each other, and 0 otherwise. A visual example is provided in Appendix subsection A.1.

4.3 Multiple captions to pseudo-labels

There can be multiple HOIs in a single image. However, due to the limited length of the caption, the captioning model may not describe all expected interactions in the image. To solve this issue, we generate multiple regional crops R on varied positions of the image and generate one caption per region. Each of the regional captions provides one set of pseudo-labels $P(Y|\tau_{R_j})$, and we aggregate the results by max pooling as follows:

$$P(y_i) = max(P(y_i|r)), \forall r \in \Phi(R) \tag{4}$$

Where Φ denotes a sampling function that determines which regions r is included in the max pooling. Φ can be either a uniform sampling or giving more weights to regions that cover more objects detected in the scene. The performance difference is compared in subsection 6.3, and a visual example is provided in Appendix subsection A.2.

Finally, the pseudo-labels of all classes per image P is normalized to $[0, 1]$, given the prior knowledge that each image must contain both positive classes and negative classes. In sum, the whole process distills the HOI knowledge from generic image captions while filtering out non-HOI information, and thus is called *pseudo-label distillation*. It differs from conventional teacher-student frameworks that transfer knowledge through feature distillation or through soft labels that are right available.

5 Experiments

We introduce the implementation details in this section. Experiments on two HOI classification datasets are conducted. In the end, we demonstrate transfering the image-level model to the instance-level HOI detection task.

5.1 Datasets

HICO: the HICO dataset is the largest dataset for image level HOI classification, containing 600 HOI classes that are combinations of 80 object categories (same as COCO) and 117 verb categories. Each image contains one or multiple labels. It has 38,118 training images and 9,658 testing images. The rare HOI subset contains 162 classes that have no more than ten training images.

MPII: the MPII (Andriluka et al., 2014) dataset contains 15,205 training images and 5,708 testing images. There are a total of 393 interaction classes. Each image is labeled with only one interaction.

Zero-shot Variant: we remove the ground-truth labels from the training set of each dataset and only keep the images for the zero-shot HOI classification. The result is evaluated on the test set with ground-truth labels.

5.2 Training

Teacher model: We use GIT (Wang et al., 2022) as the image captioning model. We generate pseudo-labels on the training set, and use a threshold of 0.97 to convert them to binary labels.

Student model: The classification model has a ViT-B/32 backbone pre-trained by CLIP. Following (Jin et al., 2022), the classification head is initialized by normalized text embeddings of HOI class names, and was fine-tuned on pseudo-labels with the LSE-Sign loss for five epochs. We use the AdamW (Kingma & Ba, 2014) optimizer and a cosine learning rate scheduler with a base learning rate $1e^{-5}$. We use image resolution 672 and a batch size of 128 on eight V100 GPUs.

Table 1: **HOI classification on HICO**. We compare zero-shot methods with two existing settings include *Supervised* and *Partially zero-shot*. Methods depend on different extra supervisions or input, including bbox: object detection, pose: human pose, PaSta: part-level action labels and Image-text. † results obtained from (Ma et al., 2022). Methods may use different backbones.

Method	Extras				Ground Truth		mAP
	Bbox	Pose	PaSta	Image-text	None-rare	Rare	
Supervised methods							
R*CNN (Gkioxari et al., 2015)	✓				✓	✓	28.5
Girdhar & Ramanan (2017)		✓			✓	✓	34.6
Mallya & Lazebnik (2016)	✓				✓	✓	36.1
Pairwise-Part (Fang et al., 2018)	✓	✓			✓	✓	39.9
PastaNet (Li et al., 2020b)	✓	✓	✓		✓	✓	**46.3**
HAKE (Li et al., 2019)	✓	✓	✓		✓	✓	**47.1**
Partially zero-shot Methods							
VCL† (Hou et al., 2020)					✓		26.7
VCL† (Hou et al., 2020)						✓	21.8
RelViT (Ma et al., 2022)					✓		**37.2**
RelViT (Ma et al., 2022)						✓	**23.1**
Zero-shot Methods							
CLIP (ResNet101)				✓			25.8
CLIP (ViT-B/32)				✓			25.8
CLIP (ViT-L/14)				✓			33.8
HTS (ViT-B/32, ours)				✓			**41.1**
HTS (ViT-L/14, ours)				✓			**49.6**

5.3 HOI CLASSIFICATION RESULTS

On the HICO dataset, we compare with state of the arts under three settings: supervised, partially zero-shot and zero-shot, as is described in Table 1. Our zero-shot method HTS (Heterogenous Teacher-Student) achieves 41.1 mAP *without* dependencies on object detection or human pose. It outperforms existing partially zero-shot methods that can access ground-truth labels of 480 classes, and surpasses multiple supervised methods that require additional object and keypoint detections. With the same backbone, it outperforms the zero-shot CLIP baseline significantly. Table 2 compares our zero-shot method with supervised state-of-the-arts. Since we use no ground-truth labels, our method loses on classes with sufficient training samples, but wins on the few-shot subsets. The credit goes to the proper way of leveraging image-text pre-training.

On the MPII dataset, Table 3 shows that our method also outperforms multiple supervised baselines and the zero-shot CLIP. In this experiment, the gap between our method and zero-shot CLIP is not as significant as in the HICO dataset. The reason is that the interaction classes in MPII are not strictly mutually exclusive (e.g., "bicycling, general", "bicycling, mountain"; "jogging", "running"). Hence, the dataset is slightly biased, which the pseudo-label is harder to fit.

5.4 HOI DETECTION RESULTS

We transfer the student HOI classification model to HOI detection by connecting with an off-the-shelf object detector trained on COCO (Lin et al., 2014) by Gao et al. (2020). We pair detected humans and objects and use the classification model to recognize the regional HOI. To do so, we modify the last transformer layer in the classifier's backbone so that the `CLS` token only attends to the region specified by the boxes. Please see Appendix Appendix B for detailed implementations. The method achieves 11.6 mAP, showing a good transferability to HOI detection.

Table 2: **Few-shot performance** on HICO. Few@i stands for the set of classes with no more than i training images. Few@1, 5, 10 have 49, 125 and 162 classes, respectively. Please note that the methods may have different backbones and dependencies on object detectors, human pose detectors and etc.

Method	Labels	mAP	Few@1	Few@5	Few@10
Supervised methods					
Pairwise-Part (Fang et al., 2018)	HOI, Bbox, Pose	39.9	13.0	19.8	22.3
PastaNet (Li et al., 2020b)	HOI, Bbox, Pose, Pasta	46.3	24.7	31.8	33.1
HAKE (Li et al., 2019)	HOI, Bbox, Pose, Pasta	**47.1**	25.4	32.5	33.7
Zero-shot methods					
HTS (ViT-B/32, ours)	Image-text	41.4	**33.6**	**36.3**	**37.2**

Table 3: **HOI classification on the MPII dataset**. We report performance on the validation set that contains 6,987 images. The Zero-shot CLIP is their original model with an image encoder and a text encoder. HTS does not have a text encoder.

Method	Backbone	mAP
Supervised methods		
R*CNN (Gkioxari et al., 2015)	VGG16	21.7
Girdhar *et al.* (Girdhar & Ramanan, 2017)	ResNet101	30.6
Pairwise-Part (Fang et al., 2018)	ResNet101	32.0
Zero-shot methods		
CLIP (Radford et al., 2021)	ResNet101	33.7
HTS, ours	ResNet101	**35.2**

6 Ablation Studies

In this section, we conduct ablation studies on key decisions of the framework. We compare the teacher model of difference sizes, the language model that computes text similarities, and the design of the student model.

6.1 Image captioning models

To study the impact of different captioning models, we use UFO-B/32 and UFO-L/16 from (Wang et al., 2021) to replace GIT. The captioning performance on COCO (Lin et al., 2014) is 122.8, 131.2 and 144.8 in CIDEr (Vedantam et al., 2015) for UFO-B/32, UFO-L/16 and GIT, respectively.

Table 4 demonstrates the optimal result in the fine-tuned student model is dependent on the quality of the image captions. We evaluated both the pseudo-label performance and the eventual student performance. This comparison shows that our pseudo-label distillation provides consistent enhancements over naive text similarities (up to 5.3 mAP). The student model also consistently performs on par with or better than the teacher.

6.2 Language Models

A language model is used to generate text embeddings of the caption and HOI classes to compute the cosine similarity for pseudo-labels. We compare the difference in the pseudo-label performance when using language models of different types, including BERT, SimCSE and CLIP-T (text encoder from CLIP). Results in Table 5 show that SimCSE plays an important role in our framework. Please note that normalizing the cosine similarities to $[0, 1]$ range per image improves the result, especially for CLIP-T. This is because CLIP-T has aligned vision-text representation but is not fine-tuned with text-similarity as a training objective.

Table 4: **Comparison of image captioning models**. Performance is evaluated on the HICO test set using one caption per image. Cosine similarity: the performance of directly using the cosine similarity as the pseudo-label (e.g. object types not sparsified). Pseudo-label Distillation: the performance of the pseudo-label processed by our method. Student model: the performance of the classification model fine-tuned with pseudo-labels.

Captioning Model	**mAP**	**Few@1**	**Few@5**	**Few@10**
Pseudo-label: cosine similarity				
GIT	35.556	31.833	31.728	31.563
UFO-L/16	29.578	21.051	21.652	21.976
UFO-B/32	25.878	16.636	17.367	17.759
Pseudo-label: pseudo-label distillation output				
GIT	40.855	32.698	37.947	38.024
UFO-L/16	32.677	20.353	23.324	23.595
UFO-B/32	28.203	17.908	19.222	19.490
Student model: ViT-B/32				
GIT	40.43	31.90	35.76	35.85
UFO-L/16	38.39	31.67	33.82	33.85
UFO-B/32	35.24	26.46	28.01	28.68

Table 5: **Language models** in pseudo-label distillation. Three models has the same architecture while trained for different objectives. CLIP-T is the text encoder from CLIP, and SimCSE is BERT fine-tuned on text similarity tasks. Language models like SimCSE provide an isotropic embedding space making the cosine similarity more accurate in our pseudo-label generation. Normalization means scaling the pseudo-labels in each image to the full $[0, 1]$ range.

Language Model	**mAP**	**Few@1**	**Few@5**	**Few@10**
Without normalization				
BERT_{base}	1.018	0.212	0.661	0.551
CLIP-T_{base}	17.563	16.346	14.601	13.985
SimCSE_{base}	35.556	31.833	31.728	31.563
With normalization				
BERT_{base}	1.767	0.446	1.034	0.960
CLIP-T_{base}	34.784	27.972	32.320	32.388
SimCSE_{base}	38.438	31.343	34.460	35.172

6.3 Multi-region caption sampling strategy

When we generate multiple captions per image on different regional crops, a strategy is required to select the regions with clear HOI training signals and avoid regions like the sky or grass fields in the background. Table 6 compares three different sampling methods, and the results show that it influences the result critically. The best result achieved by *Object* × *10*, which selects the captions of regions that cover the most detected humans and objects.

6.4 Backbone of the classification model

A vital building block for the student model is the initialization method for both the backbone and the linear classification head. In Table 7, we compare a ViT-B/32 backbone initialized by image-only (ImageNet) and image-text (CLIP) pre-training, and the linear classification head initialized randomly (Glorot & Bengio, 2010; He et al., 2015), and initialized by language embeddings of the corresponding HOI class. Results show that the joint use of image-text pre-training in the backbone and language embedding initialized classifier is necessary. Besides, our method generalizes well to ResNet backbones.

Table 6: **Sampling strategy** of regional captions. *Center*: use the largest center crop and send to the captioning model. *Uniform*: uniformly sample 10 regional crops. *Object*: sample 10 regions that cover the most detected human and objects, which performs the best since it avoids regions not related to HOI. The pool of candidates contains regional crops of three sizes: $[1.0, 0.75, 0.5]$ of the shortest image edge. All regions are squared in shape for the best captioning performance.

Sampling Strategy	mAP	Few@1	Few@5	Few@10
Center × 1	40.43	31.90	35.76	35.85
Uniform × 10	38.74	30.57	33.47	33.81
Object × 10	41.42	33.56	36.30	37.24

Please note that the combined use of CLIP-initialized vision backbone and language embedding initialized classification head runs faster than CLIP at inference time since it does not carry a text encoder.

Table 7: **Pre-training and backbone architecture** for the classification model. Architectures include ResNets and vision transformers. The backbone initialization includes image-only (ImageNet-1K) and image-text contrastive learning (CLIP). Initializations for the classifier (classification head) include the conventional random (Kaiming) initialization (He et al., 2015) and language embedding initialization

Backbone	Pre-training	Classifier Init.	mAP	Few@1	Few@5	Few@10
ResNet-50	CLIP	Language	34.49	26.02	26.90	27.13
ResNet-101	CLIP	Language	36.49	28.91	30.20	31.13
ViT-B/32	ImageNet-1K	Random	27.53	5.61	9.69	11.46
ViT-B/32	CLIP	Random	18.91	2.71	4.87	5.52
ViT-B/32	CLIP	Language	41.42	33.56	36.30	37.24
ViT-B/16	CLIP	Language	45.36	35.49	40.26	41.75
ViT-L/14	CLIP	Language	49.58	41.57	47.09	47.41

7 Fine-tuning Iteratively

To inspire future work, we further experimented with switching the roles of teacher and student, so that in the second round, the classifier becomes the teacher to fine-tune GIT for the HOI task. In the third round, the HOI fine-tuned GIT teaches the classifier again. We observed an additional gain of 0.6 mAP on all classes and 2.8 mAP on the Few@1 subset. The fine-tuned GIT model predicts only HOI-related captions. We hope this new perspective will inspire future work in this area.

8 Discussion of limitations

As the first work that investigates fully zero-shot HOI recognition, we see additional room for improvements after this paper. For example, the process can be used to bring in additional training images from the web. Also, apart from using only the pseudo-labels, the visual feature from the captioning model can be a valuable training signal as well. We will leave these to the research community as potential future directions.

9 Conclusion

This paper introduces a new zero-shot setting for human-object interaction (HOI) recognition and presents a heterogenous teacher-student framework for this challenging setup. This method not only achieves solid performance by outperforming multiple supervised baselines, but also delivers a HOI classifier with a reasonable size (ViT-B) for deployment.

REFERENCES

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, pp. 3686–3693, 2014.

Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. 2020.

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

Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In Proceedings of the IEEE International Conference on Computer Vision, pp. 1017–1025, 2015.

Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In 2018 ieee winter conference on applications of computer vision (wacv), pp. 381–389. IEEE, 2018.

Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9004–9013, June 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

Leizhen Dong, Zhimin Li, Kunlun Xu, Zhijun Zhang, Luxin Yan, Sheng Zhong, and Xu Zou. Category-aware transformer network for better human-object interaction detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19538–19547, 2022.

Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In Proceedings of the European conference on computer vision (ECCV), pp. 51–67, 2018.

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

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

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Empirical Methods in Natural Language Processing (EMNLP), 2021.

Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. In Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 33–44, 2017.

Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with r* cnn. In Proceedings of the IEEE international conference on computer vision, pp. 1080–1088, 2015.

Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8359–8367, 2018.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision, pp. 1026–1034, 2015.

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

Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In CVPR, 2021a.

Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14646–14655, 2021b.

ASM Iftekhar, Hao Chen, Kaustav Kundu, Xinyu Li, Joseph Tighe, and Davide Modolo. What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5353–5363, 2022.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning, pp. 4904–4916. PMLR, 2021.

Ying Jin, Yinpeng Chen, Lijuan Wang, Jianfeng Wang, Pei Yu, Lin Liang, Jenq-Neng Hwang, and Zicheng Liu. The overlooked classifier in human-object interaction recognition. arXiv preprint arXiv:2203.05676, 2022.

Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In European Conference on Computer Vision, pp. 498–514. Springer, 2020a.

Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. Hotr: End-to-end human-object interaction detection with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 74–83, June 2021.

Bumsoo Kim, Jonghwan Mun, Kyoung-Woon On, Minchul Shin, Junhyun Lee, and Eun-Sol Kim. Mstr: Multi-scale transformer for end-to-end human-object interaction detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19578–19587, 2022.

Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In European Conference on Computer Vision, pp. 718–736. Springer, 2020b.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. arXiv preprint arXiv:2011.05864, 2020a.

Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. arXiv preprint arXiv:2106.09785, 2021a.

Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Mingyang Chen, Ze Ma, Shiyi Wang, Hao-Shu Fang, and Cewu Lu. Hake: Human activity knowledge engine. arXiv preprint arXiv:1904.06539, 2019.

Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 382–391, 2020b.

Zhimin Li, Cheng Zou, Yu Zhao, Boxun Li, and Sheng Zhong. Improving human-object interaction detection via phrase learning and label composition. arXiv preprint arXiv:2112.07383, 2021b.

Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 482–490, 2020.

Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. arXiv preprint arXiv:2203.13954, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pp. 740–755. Springer, 2014.

Yang Liu, Qingchao Chen, and Andrew Zisserman. Amplifying key cues for human-object-interaction detection. In European Conference on Computer Vision, pp. 248–265. Springer, 2020a.

Ye Liu, Junsong Yuan, and Chang Wen Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In Proceedings of the 28th ACM International Conference on Multimedia, pp. 4235–4243, 2020b.

Xiaojian Ma, Weili Nie, Zhiding Yu, Huaizu Jiang, Chaowei Xiao, Yuke Zhu, Song-Chun Zhu, and Anima Anandkumar. Relvit: Concept-guided vision transformer for visual relational reasoning. arXiv preprint arXiv:2204.11167, 2022.

Arun Mallya and Svetlana Lazebnik. Learning models for actions and person-object interactions with transfer to question answering. In European Conference on Computer Vision, pp. 414–428. Springer, 2016.

Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. Advances in neural information processing systems, pp. 570–576, 1998.

George A Miller. Wordnet: a lexical database for english. Communications of the ACM, 38(11): 39–41, 1995.

Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. 2019.

Xian Qu, Changxing Ding, Xingao Li, Xubin Zhong, and Dacheng Tao. Distillation using oracle queries for transformer-based human-object interaction detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19558–19567, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084, 2019.

Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. 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.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval. arXiv preprint arXiv:2103.15316, 2021.

Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. 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 (CVPR), pp. 10410–10419, June 2021.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5238–5248, 2022.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In CVPR, 2015.

Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. UFO: A unified transformer for vision-language representation learning. arXiv preprint arXiv:2111.10023, 2021.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. arXiv preprint arXiv:2205.14100, 2022.

Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In CVPR, 2020.

Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S. Kankanhalli. Learning to detect human-object interactions with knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432, 2021.

Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. arXiv preprint arXiv:2108.05077, 2021a.

Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13319–13327, October 2021b.

Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13234–13243, June 2021.

Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. Human-object interaction detection via disentangled transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19568–19577, 2022.

Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, and Jian Sun. End-to-end human object interaction detection with hoi transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11825–11834, June 2021.

A Appendix

A.1 Examples of image captions and object sparsification

Figure 2 provides examples of the image captions. It also shows that object sparsification using WordNet effectively reduces the noisy labels coming from semantically related objects, helping the ground-truth HOIs surface to the top.

To map the object types from the caption (open-vocabulary) to pre-defined object categories in the dataset, we use the following information from WordNet. Take *"bicycle"* for example:

- **Synsets** that provide synonyms, which includes *bike, cycle*
- **Hyponyms** that includes variations of the object, like *tandem, E-bike*
- **Derived Forms** for example *biker, cyclist, bicyclist*

Figure 2: **Examples of pseudo-labels** generated with and without object sparsification. Only top-ranking classes are shown. In both images, the ground-truth HOI classes (highlighted in grey) rank near the top of the 600 classes, showing the efficacy of the captioning model being a teacher. However, noisy labels are found which involve objects that are semantically related but visually distant to the ground-truth (highlighted in red). We effectively reduce such noise by sparsifying the objects.

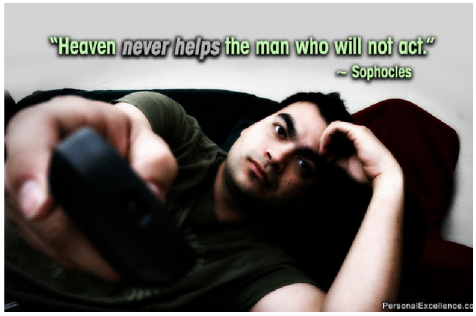

Image Caption:
a man laying on a couch holding a remote control.

Ground-truth HOI:
<hold, remote>, <point, remote>

Pseudo-HOI (w/o sparsification):		
×	1.00	lie_on, couch
×	0.97	none, couch
✓	0.88	hold, remote
×	0.88	sit_on, couch
×	0.87	none, remote
×	0.80	carry, couch
×	0.79	lie_on, chair
×	0.77	control, tv
×	0.76	none, tv
×	0.73	none, bed
×	0.72	swing, remote
×	0.69	none, chair
×	0.68	lie_on, bed
✓	0.67	point, remote

Pseudo-HOI (w/ sparsification):		
×	1.00	lie_on, couch
×	0.97	none, couch
✓	0.88	hold, remote
×	0.88	sit_on, couch
×	0.87	none, remote
×	0.80	carry, couch
×	0.72	swing, remote
✓	0.67	point, remote

(a)

Image Caption:
a man in a white hat and glasses with a dog in his lap.

Ground-truth HOI:
<hold, dog>, <hug, dog>

Pseudo-HOI (w/o sparsification):		
×	1.00	none, dog
✓	0.98	hold, dog
×	0.88	straddle, dog
×	0.84	pet, dog
×	0.83	carry, dog
✓	0.82	hug, dog
×	0.81	inspect, dog
×	0.76	kiss, dog
×	0.75	walk, dog
×	0.72	feed, dog
×	0.72	none, giraffe
×	0.72	hold, hot_dog
×	0.71	groom, dog
×	0.70	none, hot_dog

Pseudo-HOI (w/ sparsification):		
×	1.00	none, dog
✓	0.98	hold, dog
×	0.88	straddle, dog
×	0.84	pet, dog
×	0.83	carry, dog
✓	0.82	hug, dog
×	0.81	inspect, dog
×	0.76	kiss, dog
×	0.75	walk, dog
×	0.72	feed, dog
×	0.71	groom, dog

(b)

A.2 EXAMPLES OF REGIONAL IMAGE CAPTIONS

Figure 3 compares using a single caption for the whole image *v.s.* aggregating HOI from multiple regional image captions. When the interacting object is small, the captioning model tends to ignore it in the output. However, we can recover from this case by generating multiple regional captions. Then, we use object-guided sampling to select regions that cover more objects of interest, and max-pooling to aggregate regional pseudo-labels.

Figure 3: **Examples of regional captions**. **(a)** shows that the image caption for the whole image may not reveal the HOI information. The image caption for the left image doesn't mention the *book*, which is an object category in our HOI list. This can be fixed by using a regional image crop. **(b)** shows that when using regional captions, the sampling strategy becomes vital to find the HOI-related region (underlined) while filtering out the others.

Ground-truth HOI:
<hold, book>, <open, book>, <read, book>

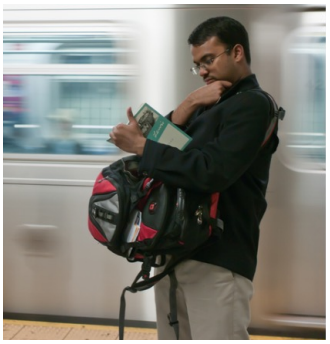

Image Caption:
a man and a woman standing in front of a subway.

Regional Image Caption:
a person holding a book and a backpack.

(a)

Ground-truth HOI:
<wear, tie>

Image Caption:
a woman and two children are standing next to each other.

Regional Image Caption:
a boy looking at a woman.

Regional Image Caption:
a boy wearing a tie and tie.

Regional Image Caption:
a person standing outside.

(b)

B ZERO-SHOT HOI DETECTION

We apply a binary self-attention mask Φ in the *Attention* (Mallya & Lazebnik, 2016) function in the last transformer layer:

$$Attention(Q, K, V) = softmax(\Phi + \frac{QK^T}{\sqrt{d_k}})V \tag{5}$$

Φ is a binary mask converted from the bounding boxes. $\Phi_{i,j}$ equals $-\infty$ if i is the `CLS` token and j a patch outside the given bounding boxes, and 0 otherwise. d_k is the dimension of Q, K and V.

Experiments are conduction on the HICO-DET (Chao et al., 2018) dataset. The whole pipeline uses no labels from the HICO-DET dataset. Table 8 compares our method with existing partially unsupervised methods. We achieve 11.62 mAP, surpassing multiple baselines on the unseen subset even we have no "seen" classes.

Table 8: **Comparison on HOI detection**. It includes four partially zero-shot settings, including **UC** (unseen classes are novel compositions of seen verbs and objects) and **UO** (unseen classes include novel objects). **RF** (rare first) selects rare HOIs for the unseen set, leaving more images for training, and **NF** (non-rare first) vise versa. Some results are from (Liao et al., 2022) where † indicates detected bounding box positions (no object categories) are used. Results are mAP.

Method	**Setting**	**Unseen**	**Seen**	**Full**
Shen et al. (Shen et al., 2018)	UC	5.62	-	6.26
FG (Bansal et al., 2020)		10.93	12.60	12.26
ConsNet (Liu et al., 2020b)		16.99	20.51	19.81
VCL (Hou et al., 2020)	UC (RF)	10.06	24.28	21.43
ATL (Hou et al., 2021a)		9.18	24.67	21.57
FCL (Hou et al., 2021b)		13.16	24.23	22.01
GEN-VLKT$_s$ (Liao et al., 2022)		21.36	32.91	30.56
VCL (Hou et al., 2020)	UC (NF)	16.22	18.52	18.06
ATL (Hou et al., 2021a)		18.25	18.78	18.67
FCL (Hou et al., 2021b)		18.66	19.55	19.37
GEN-VLKT$_s$ (Liao et al., 2022)		25.05	23.38	23.71
FCL† (Hou et al., 2021b)	UO	0.00	13.71	11.43
ATL† (Hou et al., 2021a)		5.05	14.69	13.08
GEN-VLKT$_s$ (Liao et al., 2022)		10.51	28.92	25.63
HTS (ViT-B/16, ours)	Zero-shot	-	-	**11.62**