

# Weakly-Supervised HOI Detection from Interaction Labels Only and Language/Vision-Language Priors

Mesut Erhan Unal

Department of Computer Science  
University of Pittsburgh

meu6@pitt.edu

Adriana Kovashka

Department of Computer Science  
University of Pittsburgh

kovashka@cs.pitt.edu

## Abstract

*Human-object interaction (HOI) detection aims to extract interacting human-object pairs and their interaction categories from a given natural image. Even though the labeling effort required for building HOI detection datasets is inherently more extensive than for many other computer vision tasks, weakly-supervised directions in this area have not been sufficiently explored due to the difficulty of learning human-object interactions with weak supervision, rooted in the combinatorial nature of interactions over the object and predicate space. In this paper, we tackle HOI detection with the weakest supervision setting in the literature, using only image-level interaction labels, with the help of a pretrained vision-language model (VLM) and a large language model (LLM). We first propose an approach to prune non-interacting human and object proposals to increase the quality of positive pairs within the bag, exploiting the grounding capability of the vision-language model. Second, we use a large language model to query which interactions are possible between a human and a given object category, in order to force the model not to put emphasis on unlikely interactions. Lastly, we use an auxiliary weakly-supervised preposition prediction task to make our model explicitly reason about space. Extensive experiments and ablations show that all of our contributions increase HOI detection performance.*

## 1. Introduction

Human-object interaction (HOI) detection is formally defined as correctly localizing interacting human-object pairs and classifying their interaction in a given natural image. The problem has been formulated in different ways, either end-to-end, or more commonly as a two-step procedure wherein all human and object instances get detected first, then interacting human-object pairs are identified. Regardless of its formulation, researchers rely on strong supervision to tackle HOI detection. This strong supervision is in the form of bounding box annotations for interacting human-object pairs as well as semantic labels for their interactions, which are costly to acquire and cognitively demanding for annotators as they require one to fully understand the image content<sup>1</sup>. Despite the excessive cost of gathering annotations for HOI detection, weakly-supervised directions to relax this strong supervision need have not

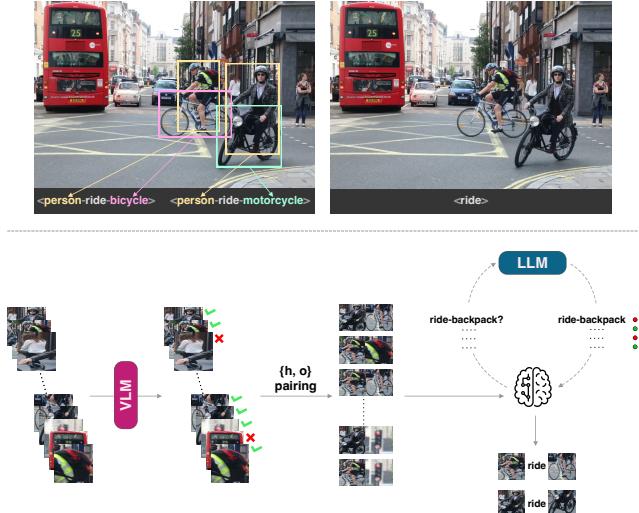


Figure 1. **Top-left:** Existing HOI detection methods need costly annotations which contain bounding boxes for interacting human-object pairs as well as their interaction categories. **Top-right:** Our method relies on image-level interaction labels, without any information on where, between whom and how many times those interactions occur. **Bottom:** During training, our approach utilizes image captions to prune non-interacting human/object proposals with the help of a vision-language model. Remaining human and object proposals will be paired for classification and a large language model will verify if predicted interactions are plausible. Best viewed in color with zoom.

<sup>1</sup>[http://www-personal.umich.edu/~ywchao/hico-hoi-det-ui/demo\\_20171121.html](http://www-personal.umich.edu/~ywchao/hico-hoi-det-ui/demo_20171121.html)

been fully explored, due to the combinatorial complexity of object interactions over object and predicate space.

In this paper, we tackle weakly-supervised HOI detection using the weakest supervision in the literature, namely image-level interaction labels (e.g. ride). This supervision level is less costly and more natural to acquire than ones required in existing efforts, as annotators would be required to answer a simple question: “What are the individuals doing in this picture?”. To make learning possible, our approach utilizes free-form captions paired with images to weakly-supervise an auxiliary task and to prune non-interacting humans and objects. We query a large language model (LLM) to eliminate unlikely human-object interactions (e.g. riding toothbrush). To increase the spatial reasoning capability of our model, we further formulate an auxiliary preposition prediction task. In this task, our model learns to assign one of the predefined prepositions to each human-object pair during training via weak supervision. Having free-form captions in hand also gives us the ability of extracting image-level interaction labels using a language parser, and hence further relax the level of supervision. Our code will be released upon publication. To summarize, our main contributions are as follows:

- We formulate a weakly-supervised HOI detection setting where supervision comes from image-level interaction labels (e.g. ride, eat). This weak supervision has not previously been used in the literature.
- We utilize free-form captions paired with images to exploit the implicit grounding capability of a vision language model (VLM) in order to prune non-interacting human and object proposals.
- We make use of a large language model (LLM) to verify if a given  $\langle$ interaction, object $\rangle$  pair is plausible.
- To further increase our model’s spatial reasoning capability, we formulate a weakly-supervised preposition prediction task.
- For the first time in the literature, we train an HOI detection model using image-caption pairs which are abundant on the web.

## 2. Related Work

### 2.1. Human-object interaction detection

The problem of detecting interactions between humans and objects was originally introduced in [11] and has drawn immense attention in the computer vision community since then. Most of the research efforts on this topic [9, 23, 38, 20, 22, 41] use a two-stage solution in which human/object locations are extracted along with their semantic labels by

an off-the-shelf object detector first, and an interaction classification model is learnt on pairwise human-object features. Apart from human/object appearances, there exist models that make use of contextual features [9, 38], spatial layouts [23, 38, 41] and human pose estimations [23]. Inspired by one-stage object detection efforts, researchers lately try to formulate end-to-end HOI detection approaches where human/instance localization and interaction classification are performed in parallel [24, 17, 18, 19]. These methods are analogous to CNN-based (e.g. YOLO[33]) and Transformer-based (e.g. DETR[3]) end-to-end object detectors. PPDM[24] takes a step forward and drops the need for heuristically created “anchors”, formulating HOI detection as a point matching problem between human and object locations.

Regardless of being one-stage or two-stage, these methods rely on strong supervision which is costly to acquire. This supervision is in the form of quadruplets that contain interacting human-object locations, object category and interaction category. Even though HOI detection is extremely costly to supervise, there exists a lack of weakly-supervised efforts in the literature. Among the existing weakly-supervised methods, MX-HOI [21] proposes a momentum-independent learning framework where they utilize both weak and strong supervision. Additionally, AlignFormer [16] formulates an alignment layer in transformer framework, that generates pseudo-aligned human-object pairs from weak annotations, conditioning on geometric and visual priors. Both of these methods utilize image-level  $\langle$ interaction, object $\rangle$  annotations (e.g. {eat-banana}) as weak supervision as opposed to the much weaker supervision we use in our work, namely image-level interaction labels (e.g. {eat}).

### 2.2. Using cues from vision-language models

Following vision-language models’ breakthrough, researchers have explored their usage in aiding diverse computer vision tasks. For example, one of the most popular VLMs, namely CLIP [32], has been researched extensively in the context of image generation [31], cross-modal retrieval [1, 8], image classification [1], object detection [10], HOI detection [7, 25] and image captioning [4], thanks to its robust image-text joint space learned on a massive dataset.

Even though [7, 25] also utilize CLIP in the context of HOI detection, how CLIP is employed within our approach is quite different. [7] uses CLIP’s text encoder to initialize context-aware HOI queries within a fully-supervised Transformer-based HOI detector. [25] utilize CLIP as a teacher within their model and distill knowledge for both visual and textual understanding of interactions. The most similar work to ours in terms of how CLIP is employed is ProposalCLIP [36], where authors prune low-quality object proposals produced by a static algorithm (e.g. Edge-

Boxes [42]). Their method runs cropped proposal regions along with produced captions for object categories (i.e. {“a photo of a  $c_i^{(obj)}$ ,”} $_{i=1}^{|C^{(obj)}|}$ ) through CLIP and removes proposals based on the alignment entropy over caption set. In our work, on the other hand, we need to quantify if a given proposal is a part of an interaction or not. Unlike running CLIP on large number of proposals, we run whole image through CLIP once and create grounding maps to calculate an interaction score on each proposal.

### 3. Method

#### 3.1. Formulating HOI detection with weak supervision

Assume an object detector that outputs a set of human and object predictions for a given image,  $\mathcal{H} = \{h_i\}_{i=1}^N$  and  $\mathcal{O} = \{o_j\}_{j=1}^M$  respectively. Each of these predictions is in the form of  $\{x^{(1)}, y^{(1)}, x^{(2)}, y^{(2)}, c^o, s^o\}$  where  $x^{(1)}, y^{(1)}$  and  $x^{(2)}, y^{(2)}$  denote the top-left and bottom-right corner coordinates of the proposal bounding box respectively,  $c^o$  is the semantic category assigned by the object detector (“person” for each proposal in  $\mathcal{H}$ ) and  $s^o \in [0, 1]$  is the confidence score.

Given (1) the above set of human and object proposals, and (2) provided image-level interaction labels (e.g. ride), our goal is to learn an HOI detection model,  $F(\cdot)$ , that can map each human-object pair  $\{h, o\} \in \mathcal{H} \times \mathcal{O}$  to an interaction class  $c^v$  that belongs to a predefined set of classes  $C^v = \{c^{v,(k)}\}_{k=1}^K$  and a confidence score  $s^v \in [0, 1]$ , yielding  $\mathcal{H} \times \mathcal{O} \xrightarrow{F(\cdot)} \{c_{i,j}^v, s_{i,j}^v\}_{i=1,j=1}^{N,M}$ . Here the  $v$  superscript denotes “verb” (interaction), while  $k$  denotes the specific verb/interaction class. Please note that this definition of weakly-supervised HOI detection is slightly different than the one given in §Sec. 1 as we offload localization and semantic labeling of humans and objects to an object detector as in every two-stage HOI detection work. We can further rewrite  $F$  as composition of two separate functions,  $F_1$  and  $F_2$ , where  $F_1$  is responsible for extracting pairwise features and modeling interactions while  $F_2$  performs classification, yielding  $F = F_2 \circ F_1$ .

In the fully-supervised case, learning is facilitated by giving the model access to the correct HOI targets  $Y = \{\text{bbox}_i^{\text{human}}, \text{bbox}_i^{\text{object}}, c_i^o, c_i^v\}_{i=1}^L$  which contain ground-truth human and object locations as bounding box coordinates, as well as the semantic categories for the object and the interaction. When one has ground-truth targets at hand, an HOI detection model can be trained to increase the likelihood of  $\{h, o\}$  pairs that spatially and semantically overlap with a HOI target to have the same interaction class as that target. In our case, however, the model can only access a set of ground-truth interaction classes for a given image, without even knowing if a certain interaction happens once or

multiple times in the image (see Figure 1).

Inspired by existing weakly-supervised object detection (WSOD) literature, we formulate weakly-supervised HOI detection as a multiple instance learning (MIL) problem. In this formulation, each image is considered as a bag of human-object pairs (i.e.  $\mathcal{H} \times \mathcal{O}$ ). If a bag (i.e. image) is labeled positive for a certain interaction, it has to contain at least one  $\{h, o\}$  pair of that interaction. Similar to WSDDN [2], we split the final classification layer,  $F_2$ , into a two-stream head (i.e.  $F_2^{(1)}$  and  $F_2^{(2)}$ ) where one models “what is the most probable interaction class for a given human-object pair?” (i.e.  $P(C^v | \{h, o\}, F_2^{(1)})$ ) while the other models “what is the most probable human-object pair for a given interaction class?” (i.e.  $P(\{h_i, o_j\}_{i=1,j=1}^{N,M} | c^v, F_2^{(2)})$ ). Assuming we get  $d$ -dimensional feature for each pair through  $F_1$  as row vectors in a feature matrix  $\mathbf{Z}$ , such as  $F_1(\mathcal{H} \times \mathcal{O}) = \mathbf{Z} \in \mathbb{R}^{NM \times d}$ , the aforementioned probabilities can be calculated by mapping  $\mathbf{Z}$  to a  $|C^v|$ -dimensional space first and then applying softmax on different dimensions. Hence, we can formulate  $F_2^{(1)}$  as a mapping  $\mathbb{R}^d \rightarrow \mathbb{R}^{|C^v|}$  followed by a row-wise softmax on  $\mathbf{Z}$  while  $F_2^{(2)}$  can be formulated as a mapping  $\mathbb{R}^d \rightarrow \mathbb{R}^{|C^v|}$  followed by a column-wise softmax on  $\mathbf{Z}$ . Then, we can write  $F_2$ ’s output  $\mathbf{Z}^{\text{HOI}}$  as:

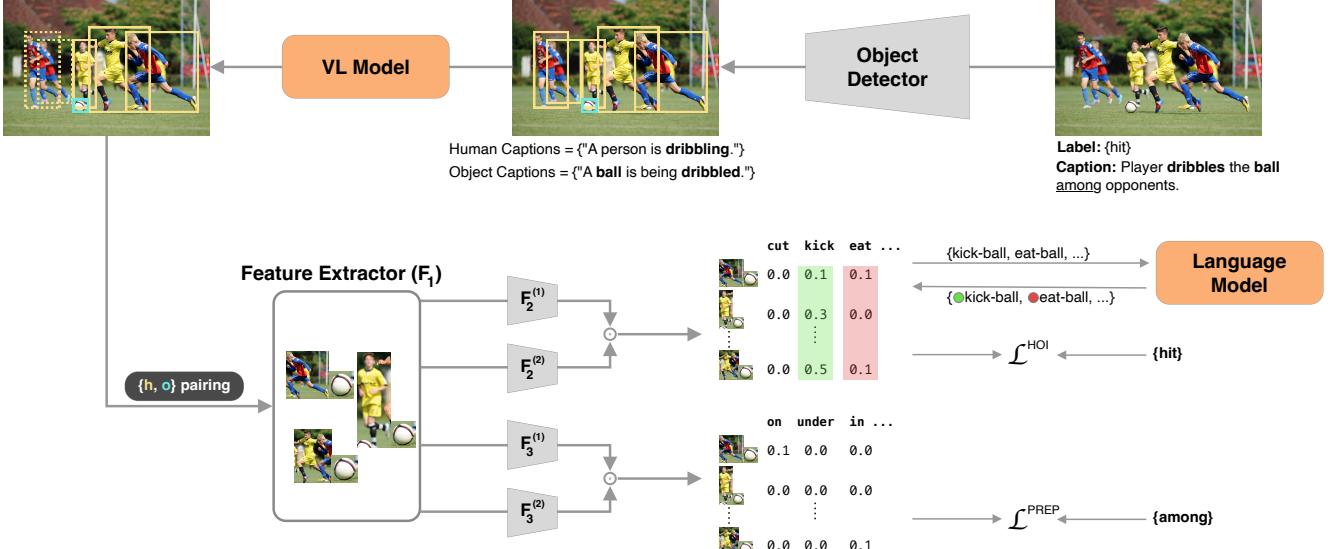
$$\begin{aligned} \mathbf{Z}^{\text{HOI}} &= F_2^{(1)}(\mathbf{Z}) \odot F_2^{(2)}(\mathbf{Z}) \in \mathbb{R}^{NM \times |C^v|} \quad \text{where} \\ F_2^{(1)}(\mathbf{Z}) &= \sigma^\rightarrow(\mathbf{Z} * W_{F_2^{(1)}}) \\ F_2^{(2)}(\mathbf{Z}) &= \sigma^\downarrow(\mathbf{Z} * W_{F_2^{(2)}}) \end{aligned} \quad (1)$$

where  $\odot$  represents Hadamard product,  $N$  is the number of human proposals,  $M$  is the number of object proposals,  $|C^v|$  is the number of interaction classes,  $W_{F_2^{(1)}}, W_{F_2^{(2)}} \in \mathbb{R}^{d \times |C^v|}$  are weight matrices, and  $\sigma^\rightarrow$  and  $\sigma^\downarrow$  are row-wise and column-wise softmax operations, respectively.

Finally, we can formulate our learning objective as minimizing  $|C^v|$  binary classification losses, one for each interaction class  $c^v \in C^v$ :

$$\begin{aligned} \mathcal{L}^{\text{HOI}}(\hat{Y}^v, Y^v) &= \frac{1}{|C^v|} \sum_{k=1}^{|C^v|} \ell(\hat{y}^{v,(k)}, y^{v,(k)}) \\ \hat{Y}^v &= \sum_{i=1}^{NM} \mathbf{Z}^{\text{HOI}} \end{aligned} \quad (2)$$

where  $Y^v$  is the binary image-level interaction labels and  $y^{v,(k)} = 1$  iff  $c^{v,(k)}$  is apparent in the image, 0 otherwise. Let  $\mathbf{Z}_{[i \times j]}^{\text{HOI}}$  denote the  $|C^v|$ -dimensional row vector in  $\mathbf{Z}^{\text{HOI}}$  that corresponds to the pair  $\{h_i, o_j\}$ . Then one can obtain the pair’s interaction class  $c_{i,j}^v$  and confidence score  $s_{i,j}^v$  during inference as:



**Figure 2. Overview of our method during training.** Retrieving human and object proposals from an object detector, our method first prunes non-interacting human/object proposals with the help of a vision-language model, calculating an interaction score for each proposal. Next, we pair remaining human-object proposals and run those pairs through a two-stream feed-forward neural net ( $F_2$ ) that operates on  $F_1$ 's output space. Finally, image-level predictions are calculated by summing  $F_2$ 's output over region pairs. We query a large-language model to restrict our model's output space only to meaningful interactions. In order to improve our model's spatial reasoning capability, we formulate a weakly-supervised preposition prediction task wherein supervision comes from preposition extracted from captions. During inference, we drop proposal pruning and preposition prediction modules, requiring only an image to detect HOI instances.

$$\begin{aligned} c_{i,j}^v &= c^{v, (\arg \max_k \mathbf{Z}_{[i \times j]}^{\text{HOI},(k)})} \\ s_{i,j}^v &= \max(\mathbf{Z}_{[i \times j]}^{\text{HOI}}) \cdot s_{h_i}^o \cdot s_{o_j}^o \end{aligned} \quad (3)$$

where  $s_{h_i}^o$  and  $s_{o_j}^o$  are confidence scores assigned to  $h_i$  and  $o_j$  by the object detector, respectively.

### 3.2. Extracting interaction labels from captions

Our weakly-supervised HOI detector learning procedure requires image-level interaction labels for supervision. However, one can utilize captions to extract those annotations to further relax the level of annotation required. In this work, we demonstrate how one can train an HOI detector on a dataset scraped from the web and contains noisy captions, using a simple technique.

We start with extracting nouns and verbs from captions using a POS tagger [14]. Consider we have a set of predefined interaction categories  $C^v$ , and verb and noun sets for a particular image  $\mathcal{V} = \{v_i\}_{i=1}^A$  and  $\mathcal{N} = \{n_i\}_{i=1}^B$ , respectively. For each image where “person”  $\in \mathcal{N}$ , we construct its label as  $Y^v = \{v \mid v \in \mathcal{V} \text{ and } v \in C^v\}$ . We use a synonym list to match “person”, as given in [30].

### 3.3. Pruning non-interacting proposals

Learning an HOI detector using only image-level interaction labels is inherently a difficult task. A model needs to

learn how to identify interacting human-object pairs among a large candidate pool and classify their interactions correctly. Without bounding box supervision, the model is left by itself to learn how an interacting human or object should look like, and what combination of those maps to a certain interaction class. For instance, consider learning the interaction “kick” with object “ball”. We can expect that most of the images containing this particular interaction and object would portray a game field where more than one person is apparent. **To a weakly-supervised model, each person would be equally likely to be the subject of the “kick” interaction.** One can try to build coarse heuristic rules (e.g. interacting human-object pairs should be close in space) or more fine-grained ones (e.g. human-object pairs for kick interaction should be close in space, but they can be further for another interaction) to reduce the search space but it is impossible to precisely develop rules for every natural interaction. To this end, we propose to exploit the implicit grounding capability of a vision-language model to prune non-interacting human and object proposals. In this work, we employ CLIP [32] and produce visual grounding maps for image-text pairs using [6].

Consider that we have access to free-form captions for the images in our training data; we do *not* require captions at inference time. Given an image and its caption, we first extract all verbs  $\mathcal{V} = \{v_i\}_{i=1}^A$  and nouns  $\mathcal{N} = \{n_i\}_{i=1}^B$  out of the caption using a POS tagger [14]. We then create hu-

man captions as  $HC = \{\text{“a person is } v_i\text{-ing”}\}_{i=1}^A$  and object captions as  $OC = \{\text{“a } n_i \text{ is being } v_j\text{-ed”}\}_{i=1,j=1}^{B,A}$ . We run the image and created captions through CLIP to produce a grounding map per caption and resize them into original image dimensions via bilinear interpolation. Finally, grounding maps are min-max normalized to map their values into  $[0, 1]$  range.

Retrieving grounding maps  $GH$  and  $GO$  for human and object captions respectively, one can calculate a grounding score,  $g$ , for each human and object proposal,  $h \in \mathcal{H}, o \in \mathcal{O}$ . Intuitively,  $g$  should measure how likely a certain proposal is to engage in an interaction. We calculate  $g$  for each proposal as follows:

$$g_h = \frac{1}{(x_h^{(2)} - x_h^{(1)})(y_h^{(2)} - y_h^{(1)})} \frac{1}{|GH|} \sum_{k=1}^{|GH|} \sum_{i=x_h^{(1)}, j=y_h^{(1)}}^{x_h^{(2)}, y_h^{(2)}} GH_{i,j}^{(k)}$$

$$g_o = \frac{1}{(x_o^{(2)} - x_o^{(1)})(y_o^{(2)} - y_o^{(1)})} \frac{1}{|GO|} \sum_{k=1}^{|GO|} \sum_{i=x_o^{(1)}, j=y_o^{(1)}}^{x_o^{(2)}, y_o^{(2)}} GO_{i,j}^{(k)} \quad (4)$$

The above equations simply calculate the average grounding score that falls into each human/object proposal region using the corresponding grounding maps. Finally, the interaction score for a human proposal  $h_i$  or an object proposal  $o_j$  is calculated as the multiplication of its grounding score  $g$  and confidence score given by the object detector  $s^o$ :

$$I_{h_i} = g_{h_i} \cdot s_{h_i}^o, I_{o_j} = g_{o_j} \cdot s_{o_j}^o \quad (5)$$

The reason behind multiplying  $g$  and  $s^o$  for interaction score calculation is quite simple. In our experiments, we have seen that the generated grounding maps usually focus on the most distinct parts of the interacting human/object, which would result in proposals covering only those distinct areas to get the highest interaction scores if only  $g$  was used.

Lastly, we sort human/object proposals in descending order of their interaction scores  $I_h/I_o$  and keep top 50% as is while assigning a special “background” class to others. These “background” proposals still get paired with human proposals within the model and will serve as negatives.

### 3.4. Suppressing implausible interactions

Previous work [12, 41] has shown that it can be beneficial to restrict a model’s output space only to meaningful interactions, conditioning on some type of lookup table in which plausible interactions are encoded. While [12] proposed to learn these conditions within the model optimizing an indicator function over possible interactions given human and object proposals, [41] compute them directly on data, iterating over ground-truth HOI targets. There also

exist works that learn such conditions by modeling interactions as phrases (e.g. “person eat banana”) in a textual [29] or multi-modal space [40]. Unlike these methods, ours does not require subject-predicate-object annotations nor multi-modal training.

In this work, we propose to use a large language model (LLM) to query which interactions are plausible for a given object category. Our hypothesis is that these models would have learnt natural co-occurrences throughout their training on massive text, and this information would also be applicable to the visual domain. We consider two natural approaches for how an LLM can be used for this purpose: (1) inputting “A person is [MASK]  $c^o$ ” caption to the model (where  $c^o$  denotes a particular object category) and calculating a probability distribution over possible interaction categories  $C^v = \{c^{v,(k)}\}_{k=1}^K$  at the masked-language modeling (MLM) head to obtain the [MASK] token, ignoring the rest of the vocabulary, and (2) plugging “What a person do with  $c^o$ ?” as a question and interaction classes  $C^v = \{c^{v,(k)}\}_{k=1}^K$  as an answer set, then retrieving the language model’s output distribution over  $C^v$  at the multiple choice question answering (MCQA) head. After obtaining a probability distribution over interaction classes given an object category i.e.  $P(C^v | c^o)$ , we create a binary lookup table for each object category, wherein interaction categories are encoded as plausible (if their probability is larger than average) or otherwise implausible.

$$\Phi_{c^o} = \{\phi(c^o, c^{v,(k)})\}_{k=1}^{|C^v|}$$

$$\phi(c^o, c^{v,(k)}) = \begin{cases} 1 & P(c^{v,(k)} | c^o) > \frac{1}{|C^v|} \sum^{|C^v|} P(c^v | c^o) \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Lastly, we double the confidence score of a human-object pair  $\{h_i, o_j\}$  if its predicted label is plausible given the object category of  $o_j$ :

$$s_{i,j}^{v'} = s_{i,j}^v \cdot (1 + \phi(c_{o_j}^o, c_{i,j}^v)) \quad (7)$$

We use RoBERTa [28] to query if a given  $\langle$ interaction, object $\rangle$  pair is plausible. We build  $\Phi_{c^o}$  for each dataset before training, instead of querying LLM constantly.

### 3.5. Formulating weakly-supervised preposition prediction

Prior work [15] demonstrates that encoding pairwise spatial relations as discrete labels (e.g. inside of, contains) within a model improves performance on tasks that require explicit spatial understanding, such as TextVQA [37]. Inspired from but different from their work, we formulate a preposition prediction task in which the model is forced to learn a mapping from pairwise features to discrete spatial

labels in weakly-supervised manner, in the unique context of human-object interaction.

Similar to our weakly-supervised HOI detection formulation given in §Sec. 3.1, we employ a two-stream head,  $F_3$ , that operates on  $F_1$ 's output space. Assuming our pre-defined preposition set is  $C^p = \{c^{p,(k)}\}_{k=1}^K$  and we get a  $d$ -dimensional feature for each human-object pair through  $F_1$  as row vectors in feature matrix  $\mathbf{Z}$ , such as  $F_1(\mathcal{H} \times \mathcal{O}) = \mathbf{Z} \in \mathbb{R}^{NM \times d}$ , we formulate  $F_3^{(1)}$  as a mapping  $\mathbb{R}^d \rightarrow \mathbb{R}^{|C^p|}$  followed by a row-wise softmax on  $\mathbf{Z}$  while  $F_3^{(2)}$  is formulated as a mapping  $\mathbb{R}^d \rightarrow \mathbb{R}^{|C^p|}$  followed by a column-wise softmax on  $\mathbf{Z}$ . Then, we can write  $F_3$ 's output  $\mathbf{Z}^{\text{PREP}}$  as:

$$\begin{aligned} \mathbf{Z}^{\text{PREP}} &= F_3^{(1)}(\mathbf{Z}) \odot F_3^{(2)}(\mathbf{Z}) \in \mathbb{R}^{NM \times |C^p|} \quad \text{where} \\ F_3^{(1)}(\mathbf{Z}) &= \sigma^\rightarrow(\mathbf{Z} * W_{F_3^{(1)}}) \\ F_3^{(2)}(\mathbf{Z}) &= \sigma^\downarrow(\mathbf{Z} * W_{F_3^{(2)}}) \end{aligned} \quad (8)$$

where  $\odot$  represents Hadamard product,  $N$  is the number of human proposals,  $M$  is the number of object proposals,  $|C^p|$  is the number of preposition classes,  $W_{F_3^{(1)}}, W_{F_3^{(2)}} \in \mathbb{R}^{d \times |C^p|}$  are weight matrices, and  $\sigma^\rightarrow$  and  $\sigma^\downarrow$  are row-wise and column-wise softmax operations, respectively.

Finally, we formulate our learning objective as minimizing  $|C^p|$  binary classification losses, one for each preposition class  $c^p \in C^p$ :

$$\begin{aligned} \mathcal{L}^{\text{PREP}}(\hat{Y}^p, Y^p) &= \frac{1}{|C^p|} \sum_{k=1}^{|C^p|} \ell(\hat{y}^{p,(k)}, y^{p,(k)}) \\ \hat{Y}^p &= \sum_{NM} \mathbf{Z}^{\text{PREP}} \end{aligned} \quad (9)$$

where  $Y^p$  is the binary image-level preposition labels and  $y^{p,(k)} = 1$  iff  $c^{p,(k)}$  is apparent in the image, 0 otherwise. Adding a new task in the model, our overall training objective now becomes minimizing both  $\mathcal{L}^{\text{HOI}}$  (Eq. 2) and  $\mathcal{L}^{\text{PREP}}$ :

$$\mathcal{L} = \mathcal{L}^{\text{HOI}} + \lambda \mathcal{L}^{\text{PREP}} \quad (10)$$

Since none of the datasets we use in our experiments comes with such preposition annotations, we utilize captions to extract them. Specifically, we run captions through a scene graph parser (e.g. Stanford Scene Graph Parser [35]) to extract  $\langle \text{subject}, \text{predicate}, \text{object} \rangle$  triplets. We filter out triplets whose subject is not “person” or predicate is not in  $C^p$ , which we curated by hand collecting 32 most common prepositions. We use the same synonym list for “person”, mentioned in §Sec. 3.2. After the filtering process, the unique predicates from the remaining triplets are used as image-level preposition labels for their corresponding images.

## 4. Experiments

### 4.1. Setup

**Datasets and metrics.** We use the well-established HOI detection benchmark datasets, HICO-DET [5] and V-COCO [11], in our experiments. HICO-DET contains 37,633 training and 9,546 test images with bounding box annotations for interacting human-object pairs and their interaction labels. There are 80 object (same as in MS COCO [27]) and 117 interaction categories in HICO-DET with 600 unique  $\langle \text{interaction}, \text{object} \rangle$  pairs. As HICO-DET instances do not come with a paired caption, we use a state-of-the-art image captioning model, OFA [39], to generate one for each image in the training split. On the other hand, V-COCO is a relatively smaller dataset with 5,400 images in trainval and 4,946 images in test split. There are 80 object and 26 interaction categories. As V-COCO is a subset of MS COCO, each image is paired with 5 captions. We use standard metrics for each dataset which are Agent AP and Role AP for V-COCO, and Full mAP for HICO-DET. Please note that HICO-DET’s Full mAP is analogous to V-COCO’s Role AP, which requires predicted human and object bounding boxes to have at least 0.5 IoU with corresponding HOI target, and predicted interaction category should be the same as the target interaction label. V-COCO’s Agent AP, on the other hand, requires correct localization of humans (i.e.  $\text{IoU} > 0.5$ ) engaging in a particular interaction.

Furthermore, for the first time in the literature we learn an HOI detection model on a small subset of Conceptual Captions, which consists of roughly 18,000 image-caption pairs, without any HOI-related annotation. We extract image-level interaction labels from captions as explained in §Sec. 3.2. We use the V-COCO test split to evaluate models trained on this new dataset.

**Baseline and training procedure.** Please note that our proposed approach as a whole is applicable to any existing two-stage HOI detection method. In our experiments, we use SCG [41] as our baseline and implement our main contributions on top of it. We choose SCG because it is one of the best performing fully-supervised two-stage HOI detector with a publicly-available implementation. Unless noted otherwise, we use the same hyperparameter settings as [41]. We use Faster R-CNN[34] with ResNet50-FPN[13] pretrained on MS COCO to generate detections. We train all models on  $4 \times$  NVIDIA Quadro RTX 5000 GPUs with an initial learning rate of  $1e-4$  and batch size of 16 (4 images per GPU). On V-COCO and HICO-DET, all models are trained for 8 epochs, reducing learning rate to  $1e-5$  after 6<sup>th</sup> epoch. On the other hand, we train models on Conceptual Captions subset for 5 epochs, without applying any decay strategy on learning rate. For weakly-supervised HOI detection task, we use binary adaptation of focal loss [26]

Method	Sup.	Backbone	Role AP
iCAN [9]	Full	RN50	52.04
VSGNet [38]	Full	RN152	57.00
SCG [41]	Full	RN50 FPN	58.02
IDN [22]	Full	RN50	60.30
HOTR [18]	Full	RN50+Transformer	64.40
MSTR [19]	Full	RN50+Transformer	<b>65.20</b>
MX-HOI [21]	Weak+	RN101	-
AlignFormer [16]	Weak+	RN50	<b>14.15</b>
Baseline [41] (\$3.1)	Weak	RN50 FPN	20.05
Ours	Weak	RN50 FPN	<b>29.59</b>
Ours-CC	Weak-	RN50 FPN	<b>17.71</b>

Table 1. V-COCO test Role AP performance among methods trained on V-COCO trainval split (except OursCC). Ours outperforms AlignFormer by a large margin (absolute 15.54%), even though its supervision comes from image-level  $\langle$ interaction, object $\rangle$  labels (**Weak+**) rather than image-level  $\langle$ interaction $\rangle$  only labels (**Weak**) we use. It also greatly improves (absolute 9.54%) over Baseline, which is close to SOTA when trained fully-supervised, verifying effectiveness of our contributions. Trained on a dataset scraped from the web, extracting image-level  $\langle$ interaction $\rangle$  only labels from captions (**Weak-**), our method (OursCC) still outperforms AlignFormer by absolute 3.56%. RN denotes ResNet. MX-HOI did not report V-COCO results. **Bolding** shows the best method within each supervision level.

( $\ell$  in Eq. 2) following the baseline, and binary cross entropy loss for weakly-supervised preposition prediction task ( $\ell$  in Eq. 9). We set the weight for weakly-supervised preposition prediction task as 0.1 ( $\lambda$  in Eq. 10). Interested readers may consult the original paper for additional details on model implementation and training procedure.

## 4.2. Comparison with the SOTA

In this subsection, we compare our model against the state-of-the-art HOI detection efforts. We also include fully-supervised approaches to inform readers on the performance gap between fully- and weakly-supervised HOI detection literature, and to show that our baseline SCG [41] is comparable with SOTA when trained with strong supervision. We would like to stress that there are not many weakly-supervised work in the HOI detection literature and existing approaches (e.g. MX-HOI [21] and AlignFormer [16]) use image-level  $\langle$ interaction, object $\rangle$  annotations as weak supervision. Careful readers would have already noticed that this specific definition of “weak supervision” is considerably stronger than is in our formulation as it reduces the search space on object proposals to match with a specific interaction category in a given image. Consider a natural image that portrays a person riding a bike while another riding a motorcycle in an urban setting. For that particular image, image-level  $\langle$ interaction, object $\rangle$  annotation will yield to {ride-bike, ride-motorcycle} while image-level interaction annotation (our supervision) will be {ride}

Method	Sup.	Backbone	mAP
iCAN [9]	Full	RN50	14.84
VSGNet [38]	Full	RN152	19.80
SCG [41]	Full	RN50 FPN	21.85
IDN [22]	Full	RN50	23.36
HOTR [18]	Full	RN50+Transformer	23.46
MSTR [19] †	Full	RN50+Transformer	<b>31.17</b>
MX-HOI [21] †	Weak+	RN101	16.14
AlignFormer [16] †	Weak+	RN50	<b>19.26</b>
Baseline [41] (\$3.1)	Weak	RN50 FPN	7.05
Ours	Weak	RN50 FPN	<b>8.38</b>

Table 2. Full mAP (Default) comparison on HICO-DET. Unsurprisingly, AlignFormer and MX-HOI benefit from having much stronger supervision, namely image-level  $\langle$ interaction, object $\rangle$  labels (**Weak+**), when combinatorial complexity over interaction space is increased moving from V-COCO to HICO-DET. However, Ours still improves over Baseline, which uses the same image-level  $\langle$ interaction $\rangle$  only labels (**Weak**) verifying effectiveness of our contributions. RN denotes ResNet. † denotes using an object detector fine-tuned on HICO-DET.

only. If  $\langle$ interaction, object $\rangle$  annotation are exploited, object proposals that can be matched with “ride” will be narrowed down to bike and motorcycle. In our case, however, the object space is excessively larger, including other objects that may be apparent in the image (e.g. bus, car, etc.).

In Table 1, we compare HOI detection performance on V-COCO test split among the models trained on the V-COCO trainval set (except Ours-CC). Our method improves absolute 9.54% over weakly-supervised variant of SCG, which we build our contributions upon, and absolute 15.54% over AlignFormer, which uses stronger supervision in the form of image-level  $\langle$ interaction, object $\rangle$  labels. Our method trained on the Conceptual Captions subset (Ours-CC) also surpasses AlignFormer and achieves a comparable performance to weakly-supervised SCG, even though we extract image-level interaction labels from captions to supervise its learning (§Sec. 3.2) and use MS COCO trained object detector to produce human/object proposals.

Similarly in Table 2, we compare HOI detection performance on the HICO-DET test split among the models trained on the HICO-DET training set. Results show that both weakly-supervised SCG and Ours fail to sustain their improvement over AlignFormer, due to the increased number of interaction categories over V-COCO (26 vs 117). Unsurprisingly, AlignFormer was not affected heavily by increased combinatorial complexity over the  $\langle$ interaction, object $\rangle$  joint space thanks to its stronger supervision than ours. It is also worth noting that both AlignFormer and MX-HOI use an object detector fine-tuned on HICO-DET (denoted by †) while we do not. Regardless of its unsustained performance on HICO-DET, our method still improves over the baseline weakly-supervised SCG by abso-

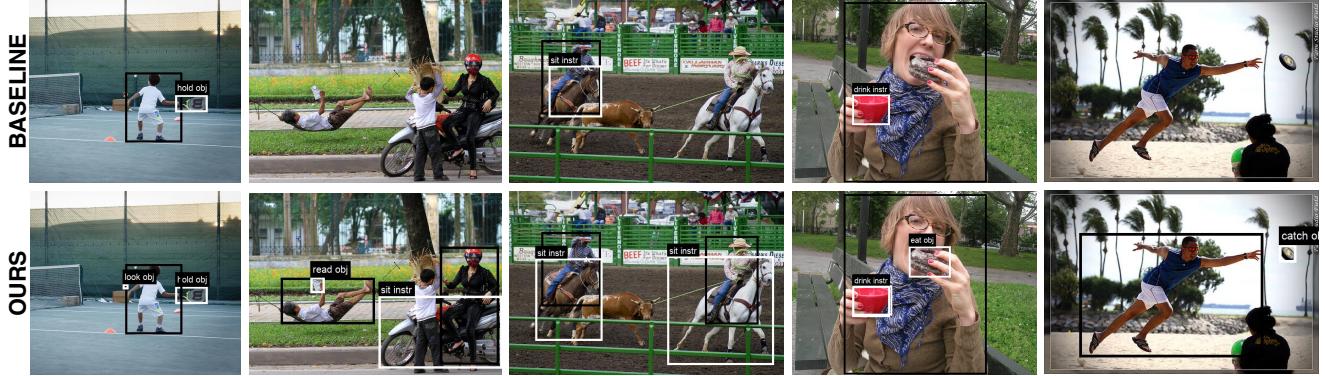


Figure 3. **Qualitative examples** sampled from V-COCO test split. Black and white boxes show interacting humans and objects, respectively. Our method successfully detects more interactions than Baseline, especially when the same human is subject to more than one interactions. Moreover, the 3<sup>rd</sup> example shows that it selects the better object proposal for “sit” interaction (horse). Interaction label explanations can be found in original V-COCO paper [11], Table 1.

Method	Sup.	Agent AP ( $\Delta$ )	Role AP ( $\Delta$ )
Baseline [41] (§3.1)	Weak	32.41	20.05
+Pruning (§3.3)	Weak	33.88 (+1.47)	21.80 (+1.75)
+Suppressing (§3.4)	Weak	37.04 (+4.63)	28.28 (+8.23)
+Preposition (§3.5)	Weak	40.53 (+8.12)	29.59 (+9.54)

Table 3. Incremental ablations on V-COCO.  $\Delta$  denotes performance difference over Baseline. All three of our contributions help improving HOI detection performance.

Method	Sup.	mAP ( $\Delta$ )
Baseline [41] (§3.1)	Weak	7.05
+Pruning (§3.3)	Weak	7.55 (+0.50)
+Suppressing (§3.4)	Weak	7.81 (+0.76)
+Preposition (§3.5)	Weak	8.38 (+1.33)

Table 4. Incremental ablations on HICO-DET.  $\Delta$  denotes performance difference over Baseline. All three of our contributions help improving HOI detection performance.

lute 1.33% (relative 18.87%), which has been trained with the same level of supervision as Ours.

### 4.3. Ablation study

To demonstrate effectiveness of our contributions, we incrementally ablate them over the baseline weakly-supervised SCG on V-COCO, HICO-DET and Conceptual Captions. The results are shown in Tables 3, 4 and 5. While all of our contributions clearly improve the performance over the baseline, results also show that caption-dependent parts of our method (§Sec. 3.3 & §Sec. 3.5) are not affected heavily from the caption source. Independent of whether captions are collected in a controlled setting (V-COCO), scraped from the web (Conceptual Captions) or generated by a captioning model (HICO-DET), our model can utilize them to boost model performance.

Method	Sup.	Agent AP ( $\Delta$ )	Role AP ( $\Delta$ )
Baseline [41] (§3.1)	Weak	17.71	14.33
+Pruning (§3.3)	Weak	19.44 (+1.73)	15.95 (+1.62)
+Suppressing (§3.4)	Weak	20.00 (+2.29)	18.23 (+3.90)
+Preposition (§3.5)	Weak	20.75 (+3.04)	17.71 (+3.38)

Table 5. Incremental ablations on Conceptual Captions.  $\Delta$  denotes performance difference over Baseline. While all three contributions help improve performance over Baseline, the preposition prediction task slightly decreases Role AP when added on top of implausible interaction suppression (but still boosts Agent AP).

## 5. Conclusion

In this work, we tackle HOI detection problem with the weakest supervision setting in the literature, using image-level interaction labels only (e.g. “ride”). We exploit the implicit grounding capability of a vision-language model, in order to prune non-interacting human and object proposals. We restrict our model’s output space to natural interactions only, querying a large language model if a given  $\langle$ interactions, object $\rangle$  is plausible. We lastly formulate a weakly-supervised preposition prediction task to improve spatial reasoning capability of our model explicitly. For the first time in the literature, we learn an HOI detector on image-caption pairs, extracting image-level interaction labels out of captions.

**Ethical concerns.** VLMs and LLMs can contain implicit biases inherited from their training data. Even though their usage within this work’s context did not pose any explicit harm during our experimentation, we would like to warn users that their usage in a different context may expose people to potentially unethical content.

## References

- [1] Shuai Bai, Zhen Dong Zheng, Xiaohan Wang, Junyang Lin, Zhu Zhang, Chang Zhou, Hongxia Yang, and Yi Yang. Connecting language and vision for natural language-based vehicle retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4034–4043, 2021.
- [2] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2846–2854, 2016.
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.
- [4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3558–3568, 2021.
- [5] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 381–389. IEEE, 2018.
- [6] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406, 2021.
- [7] 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 (CVPR)*, pages 19538–19547, 2022.
- [8] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. Mdmmmt: Multidomain multi-modal transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3363, 2021.
- [9] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *British Machine Vision Conference*, 2018.
- [10] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022.
- [11] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.
- [12] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9677–9685, 2019.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [14] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020.
- [15] Yash Kant, Dhruv Batra, Peter Anderson, Alexander Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. Spatially aware multimodal transformers for textvqa. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 715–732. Springer, 2020.
- [16] Mert Kilickaya and Arnold WM Smeulders. Human-object interaction detection without alignment supervision. In *British Machine Vision Conference*, 2021.
- [17] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 498–514. Springer, 2020.
- [18] 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)*, pages 74–83, 2021.
- [19] 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 (CVPR)*, pages 19578–19587, 2022.
- [20] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 718–736. Springer, 2020.
- [21] Suresh Kirthi Kumaraswamy, Miaojing Shi, and Ewa Kijak. Detecting human-object interaction with mixed supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1228–1237, 2021.
- [22] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. Hoi analysis: Integrating and decomposing human-object interaction. *Advances in Neural Information Processing Systems*, 33:5011–5022, 2020.
- [23] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3585–3594, 2019.
- [24] 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 *Proceed-*

- ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 482–490, 2020.
- [25] Y. Liao, A. Zhang, M. Lu, Y. Wang, X. Li, and S. Liu. Genvlkt: Simplify association and enhance interaction understanding for hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
  - [26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.
  - [27] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
  - [28] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.
  - [29] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 852–869. Springer, 2016.
  - [30] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7219–7228, 2018.
  - [31] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, 2021.
  - [32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
  - [33] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.
  - [34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015.
  - [35] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the Fourth Workshop on Vision and Language*. Association for Computational Linguistics, 2015.
  - [36] Hengcan Shi, Munawar Hayat, Yicheng Wu, and Jianfei Cai. Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9611–9620, 2022.
  - [37] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8317–8326, 2019.
  - [38] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13617–13626, 2020.
  - [39] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.
  - [40] Keren Ye and Adriana Kovashka. Linguistic structures as weak supervision for visual scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8289–8299, June 2021.
  - [41] 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)*, pages 13319–13327, 2021.
  - [42] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 391–405. Springer, 2014.