Streamlined Open-Vocabulary Human-Object Interaction Detection

Chang Sun Dongliang Liao Changxing Ding*
South China University of Technology
eesunchang2024@mail.scut.edu.cn, {liaodl, chxding}@scut.edu.cn

Abstract

*Open-vocabulary human-object interaction (HOI) detection aims to localize and recognize all human-object interactions in an image, including those unseen during training. Existing approaches usually rely on the collaboration between a conventional HOI detector and a Vision-Language Model (VLM) to recognize unseen HOI categories. However, feature fusion in this paradigm is challenging due to significant gaps in cross-model representations. To address this issue, we introduce **SL-HOI**, a **S**tream**L**ined open-vocabulary **HOI** detection framework based solely on the powerful DINOv3 model. Our design leverages the complementary strengths of DINOv3's components: its backbone for fine-grained localization and its text-aligned vision head for open-vocabulary interaction classification. Moreover, to facilitate smooth cross-attention between the interaction queries and the vision head's output, we propose first feeding both the interaction queries and the backbone image tokens into the vision head, effectively bridging their representation gaps. All DINOv3 parameters in our approach are frozen, with only a small number of learnable parameters added, allowing a fast adaptation to the HOI detection task. Extensive experiments show that SL-HOI achieves state-of-the-art performance on both the SWiG-HOI and HICO-DET benchmarks, demonstrating the effectiveness of our streamlined model architecture. Code is available at https://github.com/MPI-Lab/SL-HOI.*

1. Introduction

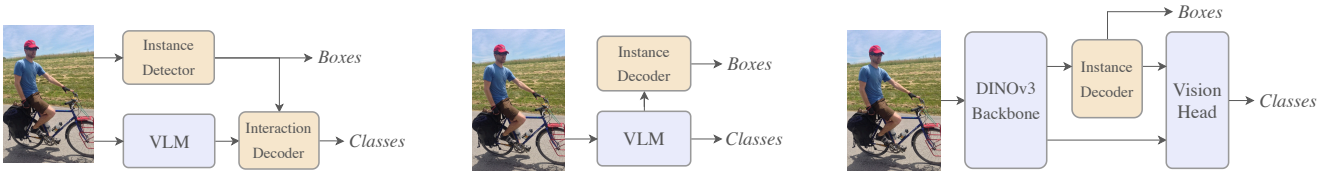
Human-Object Interaction (HOI) detection [9] is a fundamental vision task that involves not only localizing humans and objects in an image but also recognizing the interactions between each human-object pair. It is critical for applications such as video analysis [41], scene understanding [21], and robotics [31]. Compared with object detection, HOI detection is more dependent on the image context to infer the interaction categories. Moreover, in the open-vocabulary setting, HOI detectors face the additional challenge of category generalization, requiring classification of long-tailed or even unseen HOI categories during training.

Existing works rely on large-scale pre-trained Vision-Language Models (VLMs) to achieve open-vocabulary HOI detection. They can be categorized into two groups, as shown in Fig. 1(a) and Fig. 1(b). The first group of methods [2, 12, 24, 32] is based on collaboration between a VLM and a conventional HOI detector, primarily by extracting generalizable interaction representations from the VLM for the HOI detector. The second group of approaches [17, 19, 26, 40] directly transforms a VLM into an HOI detector for both interactive human-object detection and interaction classification.

Unfortunately, both categories of methods have limitations. Since the methods in the first group require two separately trained models, they tend to be complex in structure. Moreover, the fusion of features between the HOI detector and the VLM is challenging due to significant gaps in the cross-model representations. The methods in the second category are generally based on the CLIP model [34]. However, the CLIP model falls short in extracting fine-grained visual representations, since its training objective is to align holistic features between an image and its caption. The above analysis motivates us to develop the next-generation VLM-based HOI detection model that is both simple in model structure and superior in open-vocabulary HOI detection performance.

Accordingly, we propose a **S**tream**L**ined open-vocabulary **HOI** detector, namely **SL-HOI**, that streamlines interactive human-object detection and interaction classification. Specifically, we adopt the `dino.txt` variant [13] of the DINOv3 model [35] as the VLM. This variant consists of a DINOv3 backbone and a text-aligned vision head (hereafter "backbone" and "vision head", respectively). The backbone is pre-trained using large-scale self-supervised learning. It captures fine-grained visual features suitable for dense prediction, which we use for interactive human-object detection. To achieve this goal, we add a small detection decoder that uses the backbone's output patch tokens as the key and value. The vision

*Corresponding author



(a) VLM-collaborated method. (b) VLM-only method. (c) Our SL-HOI.

Figure 1. An illustration of the dominant architectural paradigms for open-vocabulary HOI detection. (a) VLM-collaborated methods that adopt both a VLM and a conventional HOI detector. (b) VLM-only methods that employ a single VLM for open-vocabulary HOI detection. (c) Our SL-HOI leverages the complementary strengths of DINOv3's backbone and vision head.

head aligns visual features with open-vocabulary captions, which is ideal for generalizable interaction classification. Similar to popular one-stage detectors HOI [24], the output embeddings of the detection decoder serve as interaction queries in this step.

However, directly performing cross-attention between the interaction queries and the vision head's output still suffers from a representation gap. To address this problem, we propose to force the interaction queries and the vision head's output tokens to share a common representation space. We achieve this by feeding both the interaction queries and the backbone's output image tokens into the vision head, rather than just the latter. Another advantage of this strategy is that it yields semantically enriched interaction queries. Then, we perform cross-attention between the refined interaction queries and the vision head's output tokens, and the output is used for open-vocabulary interaction classification.

In our approach, DINOv3 serves as the sole backbone for HOI detection, with all its parameters frozen. This streamlined design, as illustrated in Fig. 1(c), contains only a small number of trainable parameters for an end-to-end HOI detection framework, allowing efficient adaptation to HOI detection. Extensive experiments demonstrate the effectiveness of our design and show that SL-HOI achieves state-of-the-art performance on both the popular SWiG-HOI [39] and HICO-DET [4] benchmarks.

2. Related Work

2.1. HOI Detector Structures

Existing methods decompose HOI detection into two subtasks: object detection and interaction classification. Based on this division, HOI detection architectures are commonly grouped into two-stage and one-stage ones.

The two-stage models [8, 37, 46, 48, 50] typically employ an existing object detector to first localize humans and objects, and then perform human-object pairing and interaction classification in the second stage. Various features can support interaction classification, including visual features [50], spatial features [46], human pose [37, 48], and language features [8]. The two-stage methods have the advantage of a clear model structure: humans and objects are detected first, allowing the second stage to focus on interaction classification. Their main disadvantage is inefficiency in enumerating human-object pairs and potential error propagation from inaccurate detections in the first stage.

The one-stage methods perform interactive human-object detection and interaction classification in a single forward pass. Early one-stage designs represent an interacting human-object pair by an interaction region, e.g., a single point [23], a set of points [49], and the union box of a human-object pair [14]. Modern designs typically adopt the Detection Transformer (DETR) [3] as the backbone, and pre-train its parameters for the object detection task. Thanks to the powerful cross-attention mechanism in DETR, these methods can represent an interaction region more flexibly and incorporate more image-level context. Moreover, many variants of the HOI queries have been developed in DETR: some [33, 36] adopt a single query to predict the human, the object, and the interaction category in an HOI triplet simultaneously, while others [24, 45] employ independent queries for the three elements.

Our approach falls into the one-stage paradigm. Unlike most existing one-stage approaches, our objective is to design a simple and streamlined architecture that is strong for both open-vocabulary and closed-set HOI detection.

2.2. Open-Vocabulary HOI Detection

The annotation of HOI triples is time-consuming, which affects the diversity of HOI categories in the training data. Therefore, recent research has increasingly focused on open-vocabulary HOI detection, which aims to recognize HOI triplets that are even unseen during training. The two-stage HOI detection methods [15, 16, 18] are usually straightforward to extend to recognize unseen HOI categories. They typically adopt an off-the-shelf object detector to perform human and object detection in the first stage, and use a VLM to classify interactions within the detected human-object region in the second stage. The one-stage open-vocabulary HOI detection methods are more diverse and can be grouped into three categories. The first category of methods is based on compositional learning [10, 11], which encourages models to be generalizable by recombining seen <*human-verb-object*> triplets. The second cate-

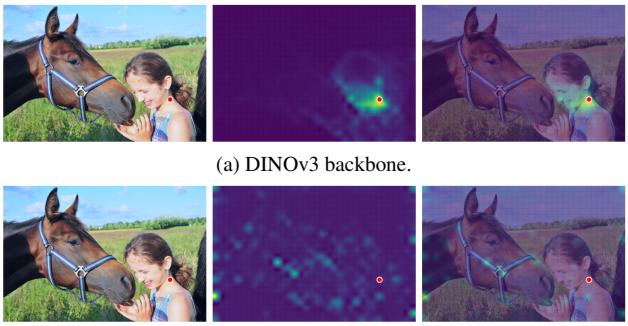
gory of methods [22, 43] employs large-scale retraining to enhance the generalization ability of HOI detectors. The third category of methods resorts to VLMs to obtain robust representations for unseen HOI categories. Moreover, VLM-based approaches can be divided into two types. The first type of approaches [2, 12, 24, 32] retains a conventional HOI detector for interactive human-object localization and leverages a VLM mainly for interaction classification. Since two separately trained models are adopted, these approaches tend to be more complex in their architectures and struggle to fuse features across the two models. To alleviate this problem, the second type of approaches [17, 19, 26, 40] employs a single VLM for both interactive human-object detection and interaction classification. Although they excel at open-vocabulary interaction classification, their detection performance is often weak due to a lack of fine-grained visual features. This is because most VLMs are pre-trained for image-level tasks, and their internal representations often lack the precise, region-specific details required for object localization.

In this paper, we also rely on a single VLM for open-vocabulary HOI detection. Unlike existing approaches, we adopt the latest DINOv3 model [35] as the backbone. We streamline it for HOI detection and carefully address the representation gaps between modules, achieving the state-of-the-art open-vocabulary HOI detection performance.

3. Preliminaries

DINOv3. DINOv3 [35] leverages self-distillation to learn rich feature representations and employs the Gram anchoring strategy to preserve dense spatial details during large-scale self-supervised training. It uses a Vision Transformer (ViT) [7] architecture with L self-attention layers. It splits an input image $I \in \mathbb{R}^{H\times W\times 3}$ into non-overlapping patches, projects them into patch tokens, and augments them with positional embeddings. These image patch tokens, a `[CLS]` token $\mathbf{x}_{\text{cls}}$, and a set of register tokens $\mathbf{x}_{\text{reg}}$ [6] form the input sequence of ViT. The VIT's output (denoted as final $\mathbf{Z}_L$) contains contextualized features for all image tokens.

`dino.txt`. `dino.txt` [13] extends the ViT-L/16 backbone of DINOv3 with a vision head and a text encoder for Locked-image Tuning (LiT) [44]. The vision head consists of two self-attention blocks that refine backbone features, jointly processing the class token, register tokens, and patch tokens to enhance inter-token dependencies and project visual features into the shared text embedding space. This increases the semantic richness of patch tokens while slightly reducing local details. During LiT, the DINOv3 backbone is frozen and only the text encoder and the vision head are trained. Text-image alignment is enforced by aligning both the class token and the mean-pooled patch features with the corresponding text embeddings.



(a) DINOv3 backbone.

(b) `dino.txt` vision head.

Figure 2. Visualization of attention maps from the last self-attention block of (a) DINOv3 backbone and (b) `dino.txt` vision head. The left column shows the original image of a person petting a horse, the middle column displays the attention map, and the right column overlays the attention on the original image. The red dot marks the queried patch located on the person. All other image patch tokens are as keys.

4. Method

4.1. Overview

Motivations. Our work begins with a fundamental question: can a single, unified model naturally provide the distinct feature types required for both precise localization and broad semantic classification? The attention maps in Fig. 2 provide a compelling affirmative. We observed an explicit functional specialization within the `dino.txt` model. The attention map of the DINOv3 backbone, shown in Fig. 2(a), is tightly focused, attending to small, specific areas in an image. This focus provides the fine-grained spatial detail essential for instance detection [5]. In contrast, the vision head in Fig. 2(b) exhibits holistic attention that aggregates the entire relational context. This allows it to form an ideal foundation for interaction classification [36]. This discovery of inherent, complementary roles becomes our architectural principle: we harness the backbone for detailed spatial representation, while the head provides semantic comprehension. This enables us to construct a streamlined one-stage framework.

Overall Architecture. We illustrate the overall architecture of SL-HOI in Fig. 3. It is a one-stage framework built on the DINOv3 model. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the frozen DINOv3 backbone produces image tokens $\mathbf{X}_b \in \mathbb{R}^{N\times D}$, where N and D denote the token number and the embedding dimension, respectively. These tokens are used for both interactive human-object instance detection and interaction classification tasks.

For the first task, we adopt the standard detection decoder in HOI detection works [24]. It has one set of learnable human queries $\mathbf{Q}_h \in \mathbb{R}^{N_q\times d}$ and one set of object queries $\mathbf{Q}_o \in \mathbb{R}^{N_q\times d}$. The obtained decoder embeddings $\mathbf{E}_h$ and $\mathbf{E}_o$ are used to predict the human and object bound-

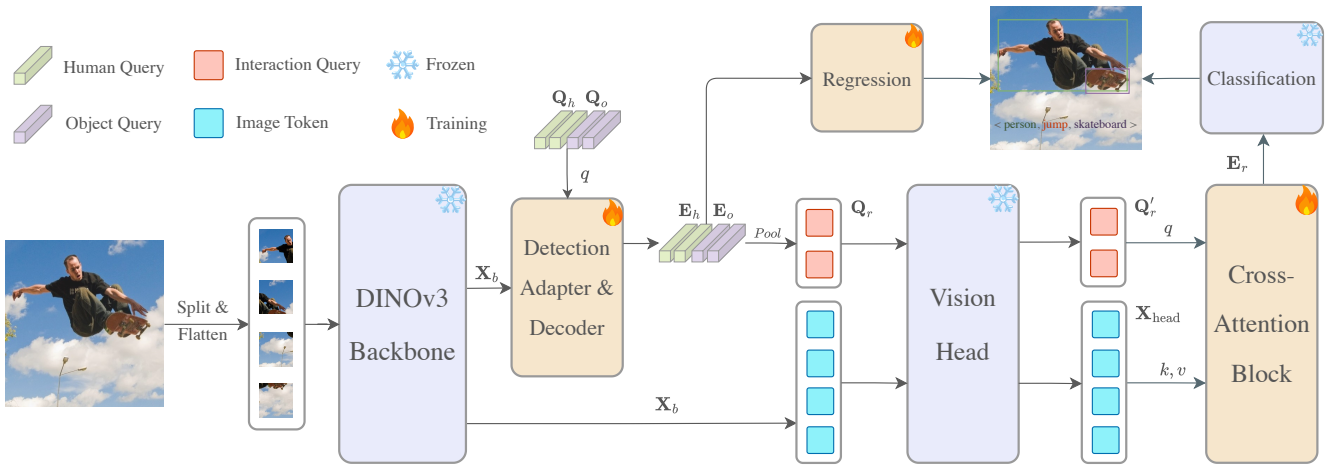


Figure 3. Overall architecture of our SL-HOI framework. A frozen DINOv3 ViT encoder (backbone) provides features for two branches. The first branch performs standard instance detection, localizing interactive human-object pairs. The second branch, our core contribution, refines interaction queries in a two-step process. We feed the initial interaction queries $\mathbf{Q}_r$ along with image tokens into the frozen vision head. This yields semantically enriched queries $\mathbf{Q}'_r$ and contextualized image tokens $\mathbf{X}_{\text{head}}$. Subsequently, we employ a single learnable cross-attention block that uses these enriched queries to re-attend to $\mathbf{X}_{\text{head}}$, producing higher-quality embeddings $\mathbf{E}_r$, which are used for open-vocabulary interaction classification.

ing boxes, respectively.

For the second task, we form initial interaction queries $\mathbf{Q}_r \in \mathbb{R}^{N_q \times D}$ by performing element-wise averaging on $\mathbf{E}_h$ and $\mathbf{E}_o$. We reduce the representation gap between $\mathbf{Q}_r$ and the output of the vision head by feeding $\mathbf{Q}_r$ and $\mathbf{X}_b$ together to the frozen vision head, resulting in mutually adapted interaction queries $\mathbf{Q}'_r$ and image tokens $\mathbf{X}_{\text{head}}$. Finally, we perform cross-attention between $\mathbf{Q}'_r$ and $\mathbf{X}_{\text{head}}$, and the output decoder embeddings $\mathbf{E}_r$ are used for open-vocabulary interaction classification.

4.2. Interactive Human-Object Detection

We first reduce the embedding dimension of $\mathbf{X}_b$ to d using a convolutional layer of 1×1. We then add positional encodings $\mathbf{E}_{pos}$ to the image patch tokens. The resulting features are further processed by a detection adapter consisting of L_E self-attention layers. The above process can be formulated as follows:

$$\mathbf{F} = \text{Adapter}(\text{Conv}(\mathbf{X}_b) + \mathbf{E}_{pos}). \tag{1}$$

Then, we adopt a transformer decoder that includes L_D cross-attention layers for interactive human-object detection. It has two independent sets of learnable queries: $\mathbf{Q}_h$ for humans and $\mathbf{Q}_o$ for objects. Both sets of queries adopt $\mathbf{F}$ as the key and value. This yields refined embeddings $\mathbf{E}_h$ and $\mathbf{E}_o$:

$$\mathbf{E}_h, \mathbf{E}_o = \text{Decoder}(\mathbf{Q}_h, \mathbf{Q}_o, \mathbf{F}). \tag{2}$$

Finally, we detect the human and object instances as follows:

$$\hat{b}_h = \text{MLP}_h(\mathbf{E}_h), \quad \hat{b}_o = \text{MLP}_o(\mathbf{E}_o), \tag{3}$$

where $\hat{b}_h$ and $\hat{b}_o$ denote the regressed human and object bounding boxes, respectively.

4.3. Interaction Classification

Although the pre-trained DINOv3 model offers rich features, these are not inherently optimized for classifying human-object interactions. To bridge this gap, we introduce a streamlined two-step process to adapt DINOv3's features for this task. We begin this by forming initial interaction queries, $\mathbf{Q}_r$, by projecting the mean of interactive human-object pair embeddings to the dimension of the frozen vision head:

$$\mathbf{Q}_r = \text{Proj}((\mathbf{E}_h + \mathbf{E}_o)/2). \tag{4}$$

Semantic Bootstrapping in the Frozen Vision Head. In the first step, we bootstrap the interaction queries, $\mathbf{Q}_r$, using high-level semantic context from the frozen vision head, denoted as $\mathcal{F}_{\text{head}}$.

The interaction queries Q_r are concatenated with the backbone's image tokens, $\mathbf{X}_b$, and passed through the vision head's self-attention layers at no additional training cost:

$$[\mathbf{Q}'_r; \mathbf{X}_{\text{head}}] = \mathcal{F}_{\text{head}}([\mathbf{Q}_r; \mathbf{X}_b]). \tag{5}$$

This operation yields two key outputs. The first is a set of semantically enriched interaction queries, $\mathbf{Q}'_r$, aligned with the head's text-semantic space. The second is a set of *query-influenced* image tokens, $\mathbf{X}_{\text{head}}$. These image tokens are now contextually modulated by task-specific interaction queries.

Hierarchical Refinement via a Cross-Attention Block. Although bootstrapped queries $\mathbf{Q}'_r$ alone improve classification performance, a streamlined architecture should take advantage of all available information. The query-influenced image tokens, $\mathbf{X}_{\text{head}}$, represent a valuable contextualized feature source that should not be ignored. In the second step, we leverage the contextualized information in $\mathbf{X}_{\text{head}}$. We introduce a lightweight, learnable decoder, $\mathcal{G}_{\text{decoder}}$, that refines the enriched queries $\mathbf{Q}'_r$ by conditioning on image tokens influenced by them:

$$\mathbf{E}_r = \mathcal{G}_{\text{decoder}}(\mathbf{Q}'_r, \mathbf{X}_{\text{head}}). \tag{6}$$

Composed of a single learnable cross-attention layer and an MLP layer, this decoder distills the most salient cues from contextualized tokens, producing the final specialized decoder embeddings $\mathbf{E}_r$ for open-vocabulary classification. This hierarchical process—a coarse semantic alignment followed by a focused, learnable refinement—is key to the performance of our method. The resulting decoder embeddings $\mathbf{E}_r$ are therefore maximally informed and specialized for accurate open-vocabulary classification of interactions.

Open-Vocabulary Predictions. For the final classification, the refined interaction decoder embeddings $\mathbf{E}_r = \{\mathbf{e}_r^{(i)}\}$ are first mapped to the text embedding space via a linear projection layer. Let the projected embeddings be $\mathbf{e}_r'^{(i)}$. We then compute class probabilities by measuring the cosine similarity between these projected embeddings and the text embeddings $\mathbf{E}_t = \{\mathbf{e}_t^{(j)}\}$, which are pre-computed for all interaction categories using a frozen text encoder. The probability is given by:

$$p_{ij} = \frac{\exp(\tau \cdot \cos(\mathbf{e}_r'^{(i)}, \mathbf{e}_t^{(j)}))}{\sum_{k \in \mathcal{R}} \exp(\tau \cdot \cos(\mathbf{e}_r'^{(i)}, \mathbf{e}_t^{(k)}))}, \tag{7}$$

where p_{ij} denotes the probability that the i-th interaction representation is classified into the j-th category, τ is a learnable temperature, and $\mathcal{R}$ is the set of all interaction categories. This yields the final open-vocabulary interaction classification results.

5. Experiments

5.1. Datasets and Metrics

Datasets. We conduct experiments on two widely used benchmarks, SWiG-HOI [39] and HICO-DET [4]. The SWiG-HOI dataset provides diverse human-object interactions across 406 action and 1,000 object categories. Its test set contains about 14,000 images and roughly 5,500 relation categories, among which over 1,000 relations are unseen during training, making it a suitable benchmark for open-vocabulary HOI detection. The HICO-DET dataset consists of 600 relation categories, formed by combining 117 action categories and 80 object categories, where the object categories are defined following COCO [25]. In the open-vocabulary setting, we follow [10, 40] to remove 120 rare interaction categories from the training set while retaining them in the test set.

Evaluation Metrics. We follow the settings of previous work [4, 24, 40] and use mean Average Precision (mAP) for the evaluation. We define a true positive when both human and object bounding boxes have an Intersection over Union (IoU) greater than 0.5 with the ground truth, and the predicted interaction label matches the ground truth.

5.2. Implementation Details

We adopt the ViT-L/16 variant of DINOv3 as our visual backbone. To facilitate a fair comparison, its parameter count is comparable to that of the CLIP model [34] with a ViT-L/14 backbone. The detection adapter consists of $L_E = 2$ self-attention layers, and the detection feature dimension is set to $d = 256$. The instance decoder is composed of $L_D = 3$ layers, and we use $N_q = 64$ learnable queries for human and object instances. The $\mathcal{G}_{\text{decoder}}$ before the interaction classification is a 1-layer transformer decoder, and its feature dimension is $D = 1024$. For training, we follow the settings of previous works: [40] for the SWiG-HOI dataset and [24] for the HICO-DET dataset. Specifically, SWiG-HOI adopts a contrastive objective in which in-batch negatives are used for training, whereas HICO-DET is trained to classify interactions across the entire category set. The model is optimized with AdamW [28] using a learning rate of 1×10^{-4}. All experiments are conducted on 8 NVIDIA RTX 4090 GPUs, with a batch size of 32 per GPU for SWiG-HOI and 2 per GPU for HICO-DET.

5.3. Comparison in the Open-Vocabulary Settings

We evaluate the performance of our model on both the SWiG-HOI and HICO-DET datasets. Following the experimental settings in [40] and [24], we conduct comparisons with existing methods from multiple perspectives.

SWiG-HOI. As presented in Tab. 1, SL-HOI establishes a new state-of-the-art across all metrics. We attribute this success not only to a stronger vision backbone but also to the intrinsic design of our model. Specifically, on the rare and non-rare categories, SL-HOI outperforms MP-HOI-L [42], the previous leading method in these categories by 6.10% and 4.86%, respectively. The generalization capability of our model is further highlighted by its performance on the unseen category, where it surpasses the second-best method, SGC-Net [26], by 6.58%. This advantage is maintained in the full category, where SL-HOI achieves a 7.47% improvement over SGC-Net. To isolate the contribution of our model's architecture, we note that simply equipping more powerful backbones does not yield equivalent performance. For instance, even when augmented with larger backbones

such as Swin-Large [27] and CLIP-ViT-L/14 along with additional pre-training, MP-HOI-L does not achieve commensurate gains. This result underscores that the streamlined architectural design of SL-HOI is uniquely effective in leveraging the rich features of DINOv3 for open-vocabulary HOI detection.

Table 1. Comparison on the SWiG-HOI dataset (mAP %).

Method	Unseen	Rare	Non-rare	Full
With object detection pre-training				
QPIC [36]	6.21	10.84	16.95	11.12
GEN-VLKT [24]	-	10.41	20.91	10.87
MP-HOI-S [42]	-	14.78	20.28	12.61
MP-HOI-L [42]	-	18.59	25.76	16.21
Without object detection pre-training				
THID [40]	10.04	12.82	17.67	13.26
CMD-SE [17]	10.70	14.64	21.46	15.26
SGC-Net [26]	12.46	16.55	23.67	17.20
INP-CC [19]	11.02	16.74	22.84	16.74
Ours	**19.04**	**24.69**	**30.62**	**24.67**

HICO-DET, Open-Vocabulary Setting. The results of our method on the HICO-DET dataset are presented in Tab. 2. In the HICO-DET open-vocabulary setting, object labels are still derived from COCO [25], so methods pre-trained on COCO object detection tend to perform better due to the overlapping label space [19, 26]. We therefore report two groups in Tab. 2 for fair comparison. Even under this biased condition, SL-HOI achieves a strong performance. Compared with methods that use object detection pre-training, SL-HOI achieves improvements of 2.16% and 1.50% in the seen and full categories, respectively. While BC-HOI [12] reports higher performance in the unseen category, our method remains overall competitive. Compared with approaches without object detection pre-training, SL-HOI achieves larger gains of 17.26%, 14.65%, and 15.27% in the unseen, seen, and full categories, respectively. These results clearly demonstrate the robustness of our method in different datasets and evaluation settings.

5.4. Comparison in the Closed Setting

We further evaluate our method in the closed setting of the HICO-DET dataset following [24], where all 600 interaction categories are present during training. We compare our results with two-stage and one-stage HOI detection methods, as summarized in Tab. 3. SL-HOI outperforms all other methods in this setting. Specifically, SL-HOI surpasses the previous state-of-the-art BC-HOI [12], with significant gains of +2.04% on the full set, +1.95% on rare, and +2.07% on non-rare categories. Notably, this robust performance in the closed setting is achieved without relying on object detection pre-training from datasets like COCO, demonstrating the powerful convergence capabilities and inherent strength of our framework.

Table 2. Comparison on the HICO-DET dataset in the open-vocabulary setting (mAP %).

Method	Backbone	Unseen	Seen	Full
With object detection pre-training				
GEN-VLKT [24]	ResNet50 + CLIP-ViT-B/32	21.36	32.91	30.56
HOICLIP [32]	ResNet50 + CLIP-ViT-B/32	25.53	34.85	32.99
CLIP4HOI [30]	ResNet50 + CLIP-ViT-B/16	28.47	35.48	34.08
LOGICHOI [20]	ResNet50 + CLIP-ViT-B/32	25.97	34.93	33.17
UniHOI [2]	ResNet50 + BLIP-2-ViT-G/14	28.68	33.16	32.27
BCOM [38]	ResNet50 + CLIP-ViT-L/14	28.52	35.04	33.74
CMMP [18]	ResNet50 + CLIP-ViT-L/14	35.98	37.42	37.13
EZ-HOI [15]	ResNet50 + CLIP-ViT-L/14	34.24	37.35	36.73
HOLa [16]	ResNet50 + CLIP-ViT-B/16	30.61	35.08	34.19
BC-HOI [12]	ResNet50 + BLIP-2-ViT-G/14	**42.31**	40.67	**40.99**
VRDiff [1]	ResNet50 + CLIP-ViT-L/14	38.92	**40.83**	40.45
Without object detection pre-training				
THID [40]	CLIP-ViT-B/16	15.53	24.32	22.96
CMD-SE [17]	CLIP-ViT-B/16	16.70	23.95	22.35
SGC-Net [26]	CLIP-ViT-B/16	23.27	28.34	27.22
INP-CC [19]	CLIP-ViT-B/16	17.38	24.74	23.13
Ours	DINOv3-ViT-L/16	**40.53**	**42.99**	**42.49**

Table 3. Comparison on the HICO-DET dataset in the closed setting (mAP %).

Method	Backbone	Rare	Non-rare	Full
Two-stage methods				
UPT [47]	ResNet50	25.94	33.36	31.66
CLIP4HOI [30]	ResNet50 + CLIP-ViT-B/16	33.95	35.74	35.33
CMMP [18]	ResNet50 + CLIP-ViT-L/14	37.75	38.25	38.14
BCOM [38]	ResNet50 + CLIP-ViT-L/14	39.90	39.17	39.34
EZ-HOI [15]	ResNet50 + CLIP-ViT-L/14	37.70	38.89	38.61
HOLa [16]	ResNet50 + CLIP-ViT-L/14	38.66	39.18	39.05
VRDiff [1]	ResNet50 + CLIP-ViT-L/14	41.69	41.31	41.40
One-stage methods				
QPIC [36]	ResNet50	21.85	31.23	29.07
CDN [45]	ResNet50	27.39	32.64	31.44
GEN-VLKT [24]	ResNet50 + CLIP-ViT-B/32	29.25	35.10	33.75
DOQ [33]	ResNet50 + CLIP-ViT-B/16	29.19	34.50	33.28
LOGICHOI [20]	ResNet50 + CLIP-ViT-B/32	32.03	36.22	35.47
HOICLIP [32]	ResNet50 + CLIP-ViT-B/32	31.12	35.74	34.69
FGAHOI [29]	Swin-Large	30.71	39.11	37.18
DP-HOI [22]	ResNet50 + CLIP-ViT-B/32	34.36	37.22	36.56
UniHOI [2]	ResNet50 + BLIP-2-ViT-G/14	39.91	40.11	40.06
BC-HOI [12]	ResNet50 + BLIP-2-ViT-G/14	45.76	42.18	43.01
Ours	DINOv3-ViT-L/16	**47.71**	**44.25**	**45.05**

5.5. Ablation Studies

We conduct comprehensive ablation studies on the SWiG-HOI dataset to validate the core design of our framework. Our analysis is threefold. First, we perform an additive analysis to quantify the contribution of each key architectural component, with results presented in Tab. 4. Second, we compare our final model against several plausible design variants to justify our specific architectural choices, as summarized in Tab. 5. Finally, we analyze the impact of varying the number of encoder layers in our detection adapter, as illustrated in Fig. 4.

Architecture Design. Tab. 4 presents ablation studies on the key architectural components of our method. The table begins with a strong baseline model, whose architecture is illustrated in the supplementary material. To construct this baseline, we replace our proposed interaction classification module with a more conventional design, such as that in HOICLIP [32]. Specifically, a 3-layer transformer decoder is used to perform cross-attention between interaction queries and the semantic features of the frozen vision head. We call this fusion strategy late-fusion hereafter. All other parameters remain identical to those of our final model. In particular, this baseline already achieves a high performance of 16.55%, 21.66%, 27.75%, and 21.82% on the unseen, rare, non-rare, and full categories of the SWiG-HOI dataset, respectively, which we attribute to the powerful representations provided by the DINOv3 backbone.

Next, we replace this late-fusion decoder with our **Semantic Bootstrapping**. In this step, the interaction queries are processed alongside the image tokens within the frozen vision head. This allows the queries to benefit from the pre-trained head parameters directly and to interact fully with the image tokens via the head's self-attention blocks. This single change yields substantial gains of +1.54%, +1.61%, +1.08%, and +1.46% across the unseen, rare, non-rare, and full categories. Finally, we introduce **Hierarchical Refinement**, which re-utilizes the query-influenced image tokens produced by the previous stage. The interaction queries re-attend to these contextualized tokens, forming our complete framework SL-HOI. This step further improves performance by +0.95%, +1.42%, +1.79%, and +1.39%.

The analysis reveals a clear division of benefits. Semantic Bootstrapping shares the rich semantic space of the frozen head, significantly boosting generalization on unseen and rare categories. Hierarchical Refinement, on the other hand, leverages image tokens cued by the HOI detection task, yielding larger gains across rare and non-rare categories. In total, SL-HOI achieves cumulative improvements of +2.49%, +3.03%, +2.87%, and +2.85% over the strong baseline. This clearly demonstrates the effectiveness of our design and its ability to successfully adapt the powerful DINOv3 model for the open-vocabulary HOI detection task.

Table 4. Ablation study of our model's architectural components on the SWiG-HOI dataset (mAP %).

Configuration	Unseen	Rare	Non-rare	Full
Baseline	16.55	21.66	27.75	21.82
+ Semantic Bootstrapping	18.09	23.27	28.83	23.28
+ Hierarchical Refinement	**19.04**	**24.69**	**30.62**	**24.67**

Variants of SL-HOI. Our method can be conceptually understood as a form of multi-scale feature fusion, combining features from before and after the vision head. However, our approach makes two critical distinctions from conventional multi-scale designs: 1) Rather than only fusion with the final output features, we leverage the head's internal, pre-trained computational forward pathway. 2) Task-specific interaction queries contextually modulate the image tokens processed by the vision head. To validate the importance of these two design choices, we introduce several variants in our ablation study, with results summarized in Tab. 5.

To investigate the first point, we compare our method against two alternatives that use a learnable decoder for fusion rather than our Semantic Bootstrapping. The first variant, labeled "Late Fusion (Head only)", serves as our baseline model, in which a decoder performs cross-attention solely over the head's output tokens. The second "Late Fusion (Multi-Scale)" extends this by attending to both the backbone and the head output tokens. As shown in Tab. 5, both of these learnable fusion strategies are suboptimal. A single standalone decoder alone struggles to match the performance of our approach. In contrast, our method effectively transfers the head's generalization capabilities by processing queries directly within its frozen, pre-trained self-attention blocks.

To address the second point, we return to our whole model and introduce a modification labeled "Ours w/ Attention Mask". In this variant, we mask the attention mechanism during Semantic Bootstrapping to prevent the image tokens from being influenced by the interaction queries. These "pure" image tokens, now lacking task-specific cues, lead to a drop in performance across all metrics. This result is significant: it demonstrates that the interaction queries function not only as information receivers but also as information givers, dynamically refining the image representations for the downstream HOI detection task.

Number of encoder layers. Fig. 4 presents the ablation studies on the number of encoder layers in our detection adapter. Adapting pre-trained DINOv3 features for downstream tasks, particularly for dense prediction, requires a delicate balance. Since DINOv3's representations are learned via self-supervision, its parameters are kept frozen to preserve their quality. However, a frozen backbone presents a challenge for DETR-based [3] architectures, which benefit from end-to-end optimization. This

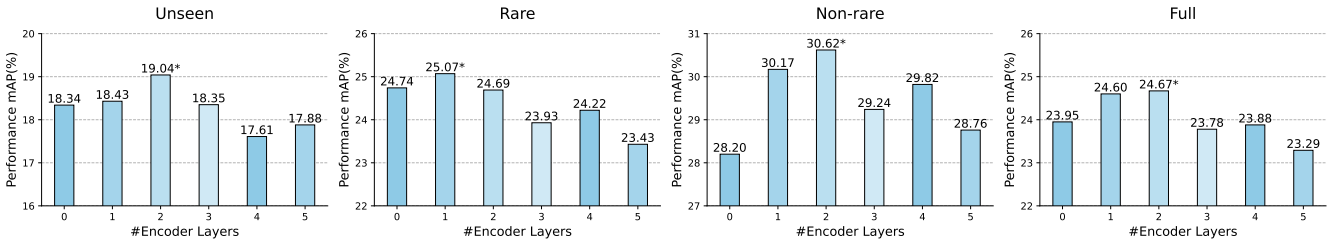


Figure 4. Ablation studies on the number of encoder layers in the detection adapter on the SWiG-HOI dataset (mAP %).

Table 5. Ablation study of variants of our proposed method on the SWiG-HOI dataset (mAP %).

Configuration	Unseen	Rare	Non-rare	Full
Late Fusion (Head only)	16.55	21.66	27.75	21.82
Late Fusion (Multi-Scale)	15.73	21.63	28.49	21.77
Semantic Bootstrapping	18.09	23.27	28.83	23.28
Ours w/ Attention Mask	17.28	24.64	29.81	24.01
Ours	**19.04**	**24.69**	**30.62**	**24.67**

makes the number of encoder layers in the detection adapter a critical hyperparameter. Using too many layers risks corrupting the rich DINOv3 features, while using too few may not adequately adapt them to the demands of the HOI detection task.

As illustrated in Fig. 4, simply increasing the number of encoder layers does not lead to better performance. This finding contrasts with the original DETR paper's conclusion [3], which found that deeper encoders generally yield monotonic performance gains. We observe that setting the number of encoder layers to 2 achieves the best trade-off, yielding the highest performance on the full category. Therefore, we use two encoder layers in all our experiments.

5.6. Qualitative Analysis

We also perform qualitative experiments to evaluate SL-HOI, focusing on analyzing the attention maps for our two-step interaction classification.

As shown in Fig. 5, the attention map during the semantic bootstrapping stage characteristically covers a broad area. This behavior stems mainly from the frozen vision head, which was pre-trained to align image patch tokens with textual captions. This objective encourages broader information exchange among tokens, leading to greater attention to capture rich contextual information. As the model transitions to the hierarchical refinement stage, the nature of the attention shifts. We identify two factors that drive this adaptation: (1) The preceding semantic bootstrapping stage aligns interaction queries and image tokens within a shared semantic space, which helps guide the attention toward potential interaction regions. (2) The now-unfrozen, learnable cross-attention block allows the mechanism to specialize for the HOI detection objective, refining its focus from the broader context. This two-stage process yields final attention maps that balance contextual understanding with a focus on salient interaction cues, a distinct characteristic of our model's decoding process.

(a) 1st Self-Attention (b) 2nd Self-Attention (c) Cross Attention

Figure 5. Visualization of attention maps across the interaction classification stage. The left two are in the self-attention blocks of the frozen head during Semantic Bootstrapping, and the right one is from the cross-attention block in Hierarchical Refinement, illustrating a Local-Global-Local interaction reasoning process.

6. Conclusion and Limitations

In this paper, we present SL-HOI, a streamlined one-stage framework for open-vocabulary HOI detection built upon the DINOv3 model. We leverage the complementary strengths of DINOv3's backbone and vision head to effectively address both interactive human-object detection and open-vocabulary interaction classification tasks. Our design includes a novel two-step interaction classification process that bridges representation gaps and enhances feature utilization. Extensive experiments on two popular benchmarks demonstrate that SL-HOI achieves state-of-the-art performance in open-vocabulary HOI detection while maintaining a simple architecture with few trainable parameters. Our work has certain limitations. For example, using the ViT backbone in the DINOv3 model may incur higher computational costs than traditional CNN-based HOI detectors.

Broader Impacts. By advancing HOI detection, our work can benefit many fields, such as robotics and assistive technologies. To the best of our knowledge, our method has no obvious negative social impacts.

Acknowledgement. This work was supported by the National Natural Science Foundation of China under Grant 62476099 and 62076101, Guangdong Basic and Applied Basic Research Foundation under Grant 2024B1515020082 and 2023A1515010007, the Guangdong Provincial Key Laboratory of Human Digital Twin under Grant 2022B1212010004, the TCL Young Scholars Program.

References

[1] Ping Cao, Yepeng Tang, Chunjie Zhang, Xiaolong Zheng, Chao Liang, Yunchao Wei, and Yao Zhao. Visual relation diffusion for human-object interaction detection. In *ICCV*, 2025. 6

[2] Yichao Cao, Qingfei Tang, Xiu Su, Song Chen, Shan You, Xiaobo Lu, and Chang Xu. Detecting any human-object interaction relationship: Universal HOI detector with spatial prompt learning on foundation models. In *NeurIPS*, 2023. 1, 3, 6

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2, 7, 8

[4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 2, 5

[5] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *ICLR*, 2023. 3

[6] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024. 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[8] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. DRG: dual relation graph for human-object interaction detection. In *ECCV*, 2020. 2

[9] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv*, abs/1505.04474, 2015. 1

[10] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, 2020. 2, 5

[11] Zhi Hou, Baosheng Yu, and Dacheng Tao. Discovering human-object interaction concepts via self-compositional learning. In *ECCV*, 2022. 2

[12] Yupeng Hu, Changxing Ding, Chang Sun, Shaoli Huang, and Xiangmin Xu. Bilateral collaboration with large vision-language models for open vocabulary human-object interaction detection. In *ICCV*, 2025. 1, 3, 6

[13] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, Oriane Siméoni, Huy V. Vo, Patrick Labatut, and Piotr Bojanowski. Dinov2 meets text: A unified framework for image- and pixel-level vision-language alignment. In *CVPR*, 2025. 1, 3

[14] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J. Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *ECCV*, 2020. 2

[15] Qinqian Lei, Bo Wang, and Robby T. Tan. EZ-HOI: VLM adaptation via guided prompt learning for zero-shot HOI detection. In *NeurIPS*, 2024. 2, 6

[16] Qinqian Lei, Bo Wang, and Robby T. Tan. Hola: Zero-shot hoi detection with low-rank decomposed vlm feature adaptation. In *ICCV*, 2025. 2, 6

[17] Ting Lei, Shaofeng Yin, and Yang Liu. Exploring the potential of large foundation models for open-vocabulary HOI detection. In *CVPR*, 2024. 1, 3, 6

[18] Ting Lei, Shaofeng Yin, Yuxin Peng, and Yang Liu. Exploring conditional multi-modal prompts for zero-shot HOI detection. In *ECCV*, 2024. 2, 6

[19] Ting Lei, Shaofeng Yin, Qingchao Chen, Yuxin Peng, and Yang Liu. Open-vocabulary hoi detection with interaction-aware prompt and concept calibration. In *ICCV*, 2025. 1, 3, 6

[20] Liulei Li, Jianan Wei, Wenguan Wang, and Yi Yang. Neural-logic human-object interaction detection. In *NeurIPS*, 2023. 6

[21] Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, and Xuming He. From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. In *CVPR*, 2024. 1

[22] Zhuolong Li, Xingao Li, Changxing Ding, and Xiangmin Xu. Disentangled pre-training for human-object interaction detection. In *CVPR*, 2024. 3, 6

[23] 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 *CVPR*, 2020. 2

[24] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. GEN-VLKT: simplify association and enhance interaction understanding for HOI detection. In *CVPR*, 2022. 1, 2, 3, 5, 6

[25] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 5, 6

[26] Xin Lin, Chong Shi, Zuopeng Yang, Haojin Tang, and Zhili Zhou. Sgc-net: Stratified granular comparison network for open-vocabulary HOI detection. In *CVPR*, 2025. 1, 3, 5, 6

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 6

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[29] Shuailei Ma, Yuefeng Wang, Shanze Wang, and Ying Wei. FGAHOI: fine-grained anchors for human-object interaction detection. *PAMI*, 2024. 6

[30] Yunyao Mao, Jiajun Deng, Wengang Zhou, Li Li, Yao Fang, and Houqiang Li. CLIP4HOI: towards adapting CLIP for practical zero-shot HOI detection. In *NeurIPS*, 2023. 6

[31] Esteve Valls Mascaro, Daniel Sliwowski, and Dongheui Lee. HOI4ABOT: human-object interaction anticipation for human intention reading collaborative robots. In *CoRL*, 2023. 1

[32] Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. HOICLIP: efficient knowledge transfer for HOI detection with vision-language models. In *CVPR*, 2023. 1, 3, 6, 7

[33] Xian Qu, Changxing Ding, Xingao Li, Xubin Zhong, and Dacheng Tao. Distillation using oracle queries for transformer-based human-object interaction detection. In *CVPR*, 2022. 2, 6

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 5

[35] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *arXiv*, abs/2508.10104, 2025. 1, 3

[36] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. QPIC: query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021. 2, 3, 6

[37] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *ICCV*, 2019. 2

[38] Guangzhi Wang, Yangyang Guo, Ziwei Xu, and Mohan S. Kankanhalli. Bilateral adaptation for human-object interaction detection with occlusion-robustness. In *CVPR*, 2024. 6

[39] Suchen Wang, Kim-Hui Yap, Henghui Ding, Jiyan Wu, Junsong Yuan, and Yap-Peng Tan. Discovering human interactions with large-vocabulary objects via query and multi-scale detection. In *ICCV*, 2021. 2, 5

[40] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *CVPR*, 2022. 1, 3, 5, 6

[41] Nan Xi, Jingjing Meng, and Junsong Yuan. Open set video HOI detection from action-centric chain-of-look prompting. In *ICCV*, 2023. 1

[42] Jie Yang, Bingliang Li, Ailing Zeng, Lei Zhang, and Ruimao Zhang. Open-world human-object interaction detection via multi-modal prompts. In *CVPR*, 2024. 5, 6

[43] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Samuel Albanie, Yining Pan, Tao Feng, Jianwen Jiang, Dong Ni, Yingya Zhang, and Deli Zhao. Rlipv2: Fast scaling of relational language-image pre-training. In *ICCV*, 2023. 3

[44] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022. 3

[45] 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. In *NeurIPS*, 2021. 2, 6

[46] Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *ICCV*, 2021. 2

[47] Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *CVPR*, 2022. 6

[48] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao. Polysemy deciphering network for robust human-object interaction detection. *IJCV*, 2021. 2

[49] Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *CVPR*, 2021. 2

[50] Tianfei Zhou, Wenguan Wang, Siyuan Qi, Haibin Ling, and Jianbing Shen. Cascaded human-object interaction recognition. In *CVPR*, 2020. 2