Focusing on what to decode and what to train: SOV Decoding with Specific Target Guided DeNoising and Vision Language Advisor

Junwen Chen Yingcheng Wang Keiji Yanai
Department of Informatics, The University of Electro-Communications, Tokyo, Japan
{chen-j, wang-y, yanai}@mm.inf.uec.ac.jp

Abstract

Recent transformer-based methods achieve notable gains in the Human-object Interaction Detection (HOID) task by leveraging the detection of DETR and the prior knowledge of Vision-Language Model (VLM). However, these methods suffer from extended training times and complex optimization due to the entanglement of object detection and HOI recognition during the decoding process. Especially, the query embeddings used to predict both labels and boxes suffer from ambiguous representations, and the gap between the prediction of HOI labels and verb labels is not considered. To address these challenges, we introduce SOV-STG-VLA with three key components: Subject-Object-Verb (SOV) decoding, Specific Target Guided (STG) denoising, and a Vision-Language Advisor (VLA). Our SOV decoders disentangle object detection and verb recognition with a novel interaction region representation. The STG denoising strategy learns label embeddings with ground-truth information to guide the training and inference. Our SOV-STG achieves a fast convergence speed and high accuracy and builds a foundation for the VLA to incorporate the prior knowledge of the VLM. We introduce a vision advisor decoder to fuse both the interaction region information and the VLM's vision knowledge and a Verb-HOI prediction bridge to promote interaction representation learning. Our VLA notably improves our SOV-STG and achieves SOTA performance with one-sixth of training epochs compared to recent SOTA. Code and models are available at https://github.com/cjw2021/SOV-STG-VLA.

1. Introduction

Recent Human-Object Interaction (HOI) detection studies are mainly built on the object detection framework. The most widely used datasets, HICO-DET [3] and V-COCO [11], share the same object categories as the MS-COCO dataset [25]. Following the definition of the HOI instance $\{B_s, (B_o, C_o), C_v\}$, which is a tuple of the subject (human) box B_s, the object box B_o with class C_o, and the verb class C_v. The HOI class C_{HOI} is defined as the possible combination of the object class C_o and the verb class C_v. In the beginning, a multi-stream architecture built on top of a CNN-based object detector is commonly adopted in the two-stage methods [3, 9, 10, 34]. By introducing the human pose information [17, 22, 51], the language priors [8, 51], or graph structure [8, 38, 46], CNN-based methods achieve considerable accuracy. On the other hand, CNN-based one-stage methods [23, 39, 52] leverage interaction points to detect possible interaction between the subject and object and achieve promising performance.

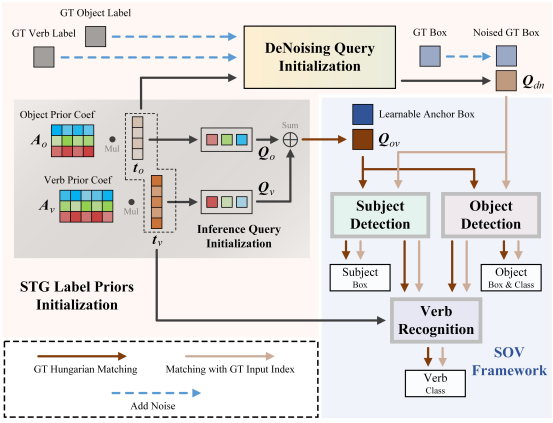

Figure 1. **End-to-end training pipeline of our SOV-STG.** Our SOV framework splits the decoding process into three parts for each element of the HOI instance. Our STG training strategy efficiently transfers the ground-truth information to label embeddings through additional denoising queries.

The attention mechanism of the transformer is more flexible than the CNN architecture in handling the relationships of features at different locations in the feature map and extracting global context information [7]. In the HOID task,

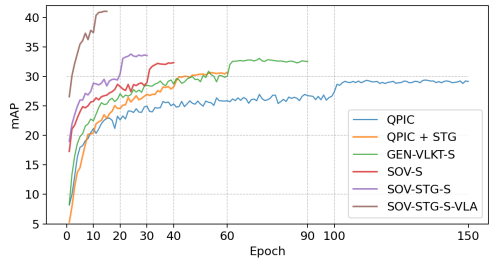

Figure 2. Comparison of the training convergence curves of the state-of-the-art methods on the HICO-DET dataset.

transformer-based methods [5, 15, 37, 55] show the advantage of the attention mechanism by adopting DETR [2]. Without the matching process in one-stage and two-stage CNN-based methods, QPIC [37] and HOITrans [55] adopt a compact encoder-decoder architecture to predict the HOI instances directly. However, the compact architecture with a single decoder binds the feature of the subject and object localization and verb recognition together. As a result, the finetuning for QPIC and HOITrans needs 150 and 250 epochs to converge, respectively. Current one-stage methods [13, 18, 24, 40, 42, 43, 45, 53] improve the single decoder design by disentangling the object localization and the verb recognition in a cascade manner. Specifically, GEN-VLKT [24] improves the cascade decoder design of CDN [45] by introducing two isolated queries of humans and objects in an instance decoder and fusing the human and object features in an interaction decoder. However, the subject and object detection are still tangled in the instance decoder, and the spatial information is implicitly represented by the query embeddings. Consequently, the training of GEN-VLKT is still hard, and it needs 90 epochs. Recently, vision language model (VLM) based methods [1, 31, 32] incorporates the pretrained VLM, CLIP [36] or BLIP2 [21] into the interaction decoding, which boosts the interaction representation learning. However, the interaction decoder for the visual encoder of the VLM lacks the guidance of the spatial information of the interaction region to focus on the interaction region. Moreover, the prediction of HOI labels and verb labels with the HOI embeddings is not optimized.

To improve the training convergence and performance, our motivation can be summarized in three aspects: 1) how to focus on decoding specific targets, 2) how to guide the training with specific priors, and 3) how to align the pretrained VLM with the HOI and verb recognition. For the first aspect, we revisit the decoding pipeline of the transformer-based method. Recent one-stage methods [24, 37, 45] redirect the decoding target of the decoder pretrained from the object detection task, which leads to slow training convergence. To this end, as shown in Fig. 1, according to the definition of the HOI instance, we propose a new **S**ubject-**O**bject-**V**erb decoding framework, which splits the decoding process into three parts: subject detection, object detection, and verb recognition. The spatial information (anchor boxes) and label information (label queries) are explicitly disentangled and fed into the decoders to guide the feature extraction. focus on specific targets and share the training burden In Fig. 2, we compare the training convergence with recent SOTA methods. From the results, SOV takes advantage of the balanced decoding pipeline and the training converges faster. The object detection part of SOV is the same as the pretrained detection model, which makes the training more stable and achieves a notable high accuracy at the early stage of the training.

For the second aspect, we focus on how to obtain specific label priors to initialize the label queries for HOI detection with an effective training strategy. As shown in Fig. 1, we introduce a novel **S**pcific **T**arget **G**uided (STG) denoising training strategy for HOI detection, which constructs a connection between the ground-truth label information and predefined label priors (embeddings) to guide the training. Moreover, we leverage the verb label embeddings to guide the verb recognition in the verb recognition part to improve the verb representation learning capabilities. In Fig. 2, we illustrate the training convergence of SOV and QPIC [37] with STG, and the results show that our STG strategy effectively accelerates the training convergence before the learning rate drops and finally improves the performance.

With our SOV framework and STG training strategy, we introduce **V**ision **L**anguage **A**dvisor (VLA) to leverage the pretrained vision language model to further improve the interaction recognition and training convergence. Our VLA consists of two parts: a vision advisor and a **V**erb-HOI (V-HOI) Bridge (language advisor). VLA incorporates additional image-level information into the verb embeddings and considers the spatial information of the interaction region with our verb box representation. V-HOI Bridge aims to fill the gap between the prediction of HOI labels and verb labels and aligns the pretrained VLM with the HOI recognition. With the above advancements, our SOV-STG-VLA notably improves the performance and only needs 15 epochs to achieve the SOTA performance.

2. Related Work

Predicting interactions with specific priors. For one-stage transformer-based methods, how to extract the interaction information under a predefined representation of the interaction region is a key issue. Recent studies [4, 16, 30] attempt to leverage the deformable attention mechanism [54] to guide the decoding by reference points. QAHOI [4] and FGAHOI [30] view the deformable transformer decoder's reference point as the HOI instance's anchor and use the anchor to guide the subject and object detection. MSTR [16] proposes to use the subject, object, and context

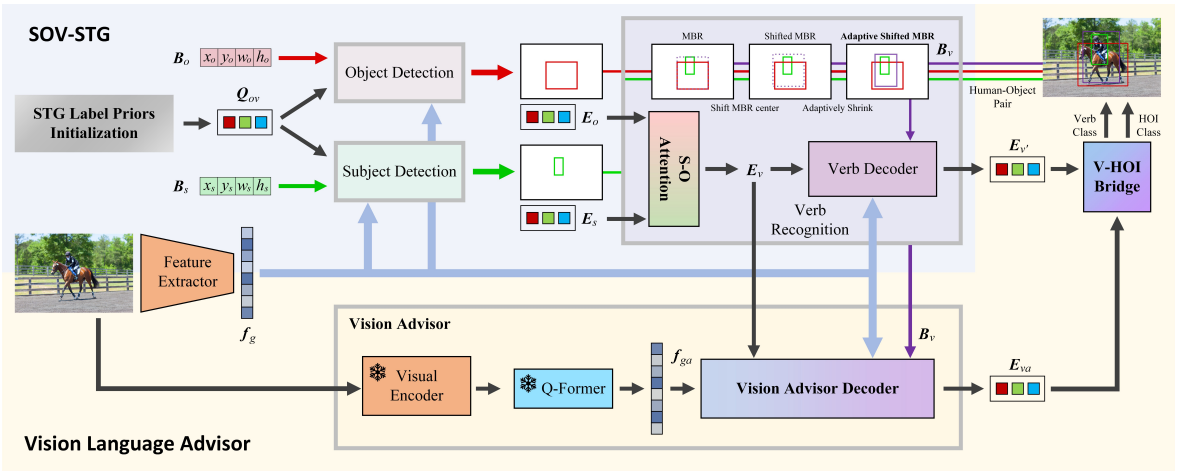

Figure 3. **The inference pipeline of SOV-STG-VLA.** SOV-STG consists of three parts: the STG label priors initialization, the subject and object detection, and the verb recognition. The label embeddings learned by our STG training strategy are used to initialize the label queries $\boldsymbol{Q}_{ov}$. The subject and object decoders update the learnable anchor boxes B_s and B_v to predict the subject and object, and the verb boxes B_v are generated by our adaptive shifted MBR. Our SOV-STG-VLA is built on the SOV-STG framework. VLA enriches the expression of the verb embeddings E_v by Vision Advisor with the global context information from the feature extractor and the pretrained VLM and the spatial information from the verb box. Then, V-HOI Bridge connects the prediction of HOI labels and verb labels.

reference points to represent the HOI instance and predict the subject, object, and verb based on the reference points. However, QAHOI, FGAHOI, and MSTR use x-y coordinates as the spatial priors to guide the decoding, the box size priors are not considered, and the query embeddings used to predict both labels and boxes suffer from ambiguous representations. In contrast, our SOV explicitly defines the subject and object anchor boxes as the spatial priors and refines the anchor boxes layer by layer. Moreover, for the verb recognition part of our SOV, we introduce a novel verb box to guide the verb feature extraction.

Effective learning with ground-truth guided. For the object detection methods of the DETR family [2, 26, 54], DN-DETR [20] shows that using the ground-truth information to guide the training can accelerate the training convergence and improve the performance. In the HOID task, HQM [50] encodes the shifted ground-truth boxes as hard-positive queries to guide the training. However, the ground-truth label information is not considered in HQM. DOQ [35] introduces the oracle queries to implicitly encode the ground-truth boxes of human-object pairs and the object labels, and guide However, DOQ implicitly encodes the same number of oracle queries as the ground-truth with learnable weights and only uses the learned weights during training. Without a complete and clear usage of ground-truth information, both HQM and DOQ still need 80 epochs to converge. Different from DOQ and HQM, we introduce denoising queries to encode the ground-truth information and guide the training. Moreover, our STG is used to learn the label priors for our model, and we intuitively use a "select" and a "weighted sum" approach to transfer the ground-truth label information to the denoising queries and inference queries, respectively.

VLM based HOID approaches. Recent studies [14, 24, 31, 32] leverage the prior knowledge of the vision language model and achieve notable gains. HOICLIP [32] and CLIP4HOI [31] implement the pretrained CLIP model into the interaction decoding process. UniHOI [1] introduces an HO Prompt Guideed Decoder to refine the HOI query embeddings with the visual features extracted by BLIP2 [21]. However, these three methods only predict the HOI labels or use the same decoder to predict the verb labels and HOI labels, which limits the representation of the HOI recognition prior knowledge of the VLM. Instead of directly assigning the verb recognition task to the pretrained VLM, based on the spatial and refined label information provided by our SOV-STG, we introduce VLA to gradually incorporate the prior knowledge of the pretrained VLM into our framework and improve the interaction recognition.

3. Methodology

Fig. 3 shows the inference pipeline of SOV-STG-VLA. First, a feature extractor is used to extract the multi-scale global context features. Then, the global features are fed into the object and subject decoder with learnable anchor boxes and label queries to predict pairs of subjects and objects. The label queries are initialized by the label embeddings and learnable coefficient matrices as shown in Fig. 1. The STG denoising training strategy (In Sec. 3.2) is used to learn the label embeddings with the ground-truth information. The **S**ubject-**O**bject (S-O) attention module and

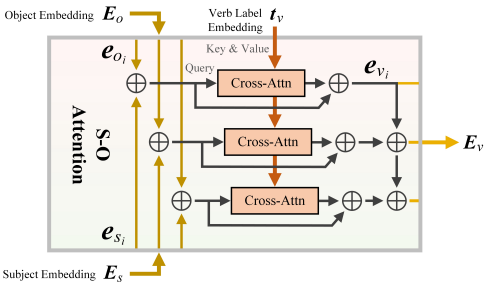

Figure 4. The illustration of the S-O attention module.

Adaptive **S**hifted **M**inimum **B**ounding **R**ectangle (ASMBR) used to fuse the subject and object embeddings and generate the verb box for the verb decoder are introduced in Sec. 3.1. Vision Advisor and V-HOI Bridge used to give additional guidance to our SOV-STG are introduced in Sec. 3.3.

3.1. SOV Decoding

To clarify the decoding target, the design of the split decoders is crucial for our framework. Different from recent one-stage transformer-based methods [24,37,45], which use a single decoder to detect objects and subjects, we split the detection part into two decoders, the subject decoder, and the object decoder, and share the prediction burden of the verb decoder. Moreover, we explore the design of the multi-branch feature fusion module and a new way to represent the interaction region for the verb decoder.

Subject and Object Detection. First, we leverage a hierarchical backbone and deformable transformer encoder [54] as the feature extractor to extract the multi-scale global features $\boldsymbol{f}_g$. Then, we adopt an improved deformable transformer decoder proposed in the recent object detection method [26] as the object and subject decoder, which can process the label queries with the constraint of anchor boxes. We clone the object decoder to initialize the subject decoder and alleviate the learning burden of the subject decoder. As shown in Fig. 3, the subject and object decoder both use the label queries $\boldsymbol{Q}_{ov} \in \mathbb{R}^{N_q \times D}$ as the input queries, where N_q is the number of queries and D is the hidden dimension of the embeddings. With the same query input, the pair of subject embeddings $\boldsymbol{E}_s \in \mathbb{R}^{N_l \times N_q \times D}$ and object embeddings $\boldsymbol{E}_o \in \mathbb{R}^{N_l \times N_q \times D}$ can share the same prior label information (N_l indicates the number of the decoder layers). The subject and object embeddings with corresponding learnable anchor boxes $\boldsymbol{B}_s$ and $\boldsymbol{B}_o$ are updated layer by layer during decoding. Then, the object embeddings from the object decoder are used to predict the object classes.

Verb Decoder with S-O attention module. As shown in Fig. 3, For the label queries of the verb decoder, we introduce S-O attention to fuse the subject and object embeddings in a multi-layer manner. In Fig. 4, we illustrate the fusion process of our S-O attention module. Given the subject embedding $\boldsymbol{e}_{s_i}$ and object embedding $\boldsymbol{e}_{o_i}$ from the i-th layer ($i > 1$), first, we sum the subject and object embeddings to fuse an intermediate embedding $\boldsymbol{e}_{so_i} = (\boldsymbol{e}_{o_i} + \boldsymbol{e}_{s_i})/2$. Then, to guide the verb recognition with our predefined label priors, the intermediate embeddings $\boldsymbol{e}_{so_i}$ are used to absorb the prior knowledge from the verb label embeddings $\boldsymbol{t}_v \in \mathbb{R}^{N_q \times D}$ learned by our STG training strategy (in Sec. 3.2) through a cross-attention module. Furthermore, we introduce a bottom-up path to amplify the information from the bottom to the top layer. Finally, the verb embedding $\boldsymbol{e}_{v_i}$ after the bottom-up path can be defined as:

$$\begin{aligned} \boldsymbol{e}_{v_i} =& ((\text{CrossAttn}(\boldsymbol{e}_{so_{i-1}}, \boldsymbol{t}_v) + \boldsymbol{e}_{so_{i-1}}) \\ &+ (\text{CrossAttn}(\boldsymbol{e}_{so_i}, \boldsymbol{t}_v) + \boldsymbol{e}_{so_i}))/2 \end{aligned} \tag{1}$$

Verb box represented by ASMBR To constrain the verb feature extraction with positional information in the verb decoder, as shown in Fig. 3, we introduce a novel verb box, Adaptive Shifted Minimum Bounding Rectangle (ASMBR) as the representation of the interaction region. The verb box is directly initialized from the last layer of subject and object boxes from the subject and object decoder. To balance the attention between the subject and object, we shift the center of the MBR to the center of the subject and object boxes. Considering the boxes will overlap with each other, we shrink the width and height of the MBR according to the spatial relationship between the two boxes. With the shift and adapt operation, the verb box can constrain the interaction region for sampling points of the deformable attention and extract interaction information from specific subject and object pairs. Finally, given the last layer subject box $\boldsymbol{B}_s = (x_s, y_s, w_s, h_s)$ and object box $\boldsymbol{B}_o = (x_o, y_o, w_o, h_o)$, where (x, y) indicates the box center, the verb box is defined as:

$$\boldsymbol{B}_v = \left(\frac{x_s + x_o}{2}, \frac{y_s + y_o}{2}, w_v, h_v \right) \tag{2}$$

$$w_v = \frac{w_s + w_o}{2} + |x_s - x_o|, h_v = \frac{h_s + h_o}{2} + |y_s - y_o| \tag{3}$$

3.2. Specific Target Guided DeNoising Training

For our SOV framework, we generate inference and denoising label queries during training. Since the two kinds of queries are generated from the specific target priors and learned during the denoising training, we call our training strategy as **S**pecific **T**arget **G**uided (STG) denoising.

Label-specific Priors To explicitly equip the prior label knowledge into the decoders and disentangle the training and decoding target, as shown in Fig. 1, two kinds of learnable label embeddings are used to initialize the query embeddings. Specifically, we define the object label embeddings $\boldsymbol{t}_o \in \mathbb{R}^{C_o \times D}$ as the object label priors, which consist

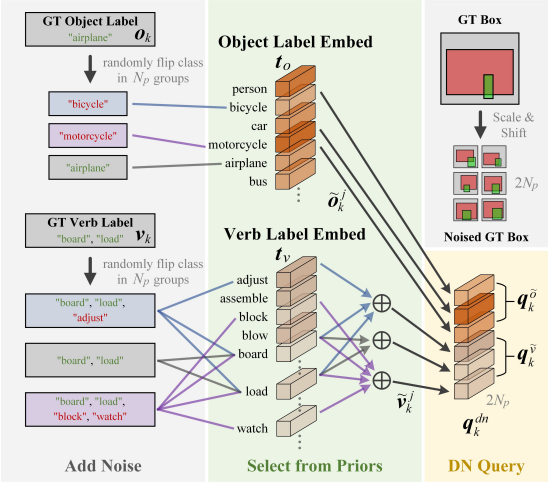

Figure 5. **Illustration of adding noise to a ground-truth HOI instance.** The initialization consists of two parts, the object label and the verb label DN queries initialization. The final DN query embeddings $\boldsymbol{q}_k^{dn}$ are concatenated with the object label DN queries $\boldsymbol{q}_k^{\tilde{o}}$ and the verb label DN queries $\boldsymbol{q}_k^{\tilde{v}}$.

of C_o vectors with D dimensions, where C_o is the number of object classes. Similarly, the verb label embeddings $\boldsymbol{t}_v \in \mathbb{R}^{C_v \times D}$ are defined as the verb label priors. With the object label and verb label priors, we first initialize the query embeddings of object label $\boldsymbol{q}_o \in \mathbb{R}^{N_q \times D}$ and verb label $\boldsymbol{q}_v \in \mathbb{R}^{N_q \times D}$ by linear combining the object label and verb label embeddings with two learnable coefficient matrices $\boldsymbol{A}_o \in \mathbb{R}^{N_q \times C_o}$ and $\boldsymbol{A}_v \in \mathbb{R}^{N_q \times C_v}$, respectively. Then, we add the object and verb label embeddings to obtain the inference query embeddings $\boldsymbol{q}_{ov} \in \mathbb{R}^{N_q \times D}$. The initialization of $\boldsymbol{q}_o$, $\boldsymbol{q}_v$, and $\boldsymbol{q}_{ov}$ is defined as follows:

$$\boldsymbol{q}_o = \boldsymbol{A}_o \boldsymbol{t}_o, \quad \boldsymbol{q}_v = \boldsymbol{A}_v \boldsymbol{t}_v \tag{4}$$

$$\boldsymbol{q}_{ov} = \boldsymbol{q}_o + \boldsymbol{q}_v \tag{5}$$

Different from DN-DETR [20] and DOQ [35], which learn an encoding weight to generate queries only used in training, we use the label embeddings both in the denoising and inference parts and enable the inference part to obtain the input query with label-specific information from the beginning.

Learning Priors with DeNoising Training In Fig. 5, we show the initialization of the DN (DeNoising) query embeddings and visualize the process of adding noise to a ground-truth HOI instance. Given the ground-truth object label set $\boldsymbol{O}_{gt} = \{\boldsymbol{o}_i\}_{i=1}^{K}$ and verb label set $\boldsymbol{V}_{gt} = \{\boldsymbol{v}_i\}_{i=1}^{K}$ of an image, where $\boldsymbol{o}_i$ and $\boldsymbol{v}_i$ are the labels of the object and verb classes, K is the number of ground-truth HOI instances, we generate N_p groups of noised labels for each of the ground-truth HOI instances. For the k-th ground-truth HOI instance, the noised object labels are obtained by randomly flipping the ground-truth index of the object label $\boldsymbol{o}_k$ to another object class index. Because the verb label $\boldsymbol{v}_k$ consists of co-occurrence ground-truth classes, to keep the co-occurrence ground-truth indices appearing in the noised verb labels, we randomly flip the other indices of the ground-truth verb label to generate the noised verb labels. Two flipping rate hyper-parameters $\eta_o \in (0, 1)$ and $\eta_v \in (0, 1)$ are used to control the percentage of the noised object labels and verb labels, respectively. Besides, a verb class flipping rate hyper-parameter $\lambda_v \in (0, 1)$ is used to control the class-wise flipping rate in the verb labels. Next, we introduce a "select" approach to "encode" the noised labels to DN query embeddings. Specifically, we directly compose the object DN query embeddings $\boldsymbol{q}_k^{\tilde{o}} \in \mathbb{R}^{N_p \times D}$ by selecting class-specific vectors $\{\tilde{\boldsymbol{o}}_k^j\}_{j=1}^{N_p}$ from the object label embeddings $\boldsymbol{t}_o$ according to the indices of the noised object labels. For encoding of the noised verb labels, we select and sum the class-specific vectors to construct multi-class vectors $\{\tilde{\boldsymbol{v}}_k^j\}_{j=1}^{N_p}$, and compose the verb DN query embeddings $\boldsymbol{q}_k^{\tilde{v}} \in \mathbb{R}^{N_p \times D}$. Finally, we concatenate the object DN query embeddings $\boldsymbol{q}_k^{\tilde{o}}$ and verb DN query embeddings $\boldsymbol{q}_k^{\tilde{v}}$ to form the DN query embeddings $\boldsymbol{q}_k^{dn} \in \mathbb{R}^{2N_p \times D}$ for the denoising training. Since the specific target priors learned by the denoising training are also used to guide the inference during end-to-end training, our STG can accelerate the training convergence and improve the inference performance at the same time. In addition, motivated by the box denoising strategy of DN-DETR [20], we scale and shift pairs of ground-truth subject and object boxes to generate $2N_p$ groups of noised anchor boxes for corresponding DN query embeddings.

3.3. Vision Language Advisor

Our Vision Language Advisor (VLA) is designed to release the burden of aligning the pretrained VLM knowledge with the verb recognition part of SOV-STG and fill the gap between the verb and HOI label prediction.

Vision Advisor. We leverage the visual encoder and Q-Former of BLIP2 [21] to extract the image-level features $\boldsymbol{f}_{ga} \in \mathbb{R}^{N_{ga} \times D_a}$. We introduce the vision advisor decoder to incorporate the global visual features $\boldsymbol{f}_g$ and $\boldsymbol{f}_{ga}$ from our SOV model and the vision advisor into our verb embeddings $\boldsymbol{E}_v$ while considering the spatial information of interaction region. To better connect the visual knowledge to each HOI query embedding, as shown in Fig. 6, we design a cascade attention mechanism for our vision advisor decoder. Different from recent VLM-based HOID methods [1, 31, 32] using the self-attention without the positional embedding, we convert the verb box to a positional embedding through a positional encoding method [26] for the self-attention module. Then, the same as the original transformer decoder, a cross-attention is used to extract the global context information from the visual encoder of the

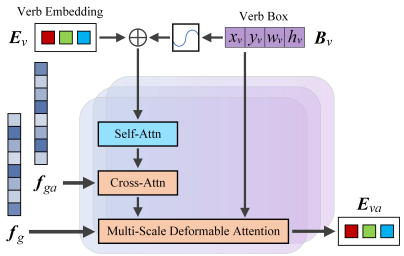

Figure 6. The illustration of the vision advisor decoder.

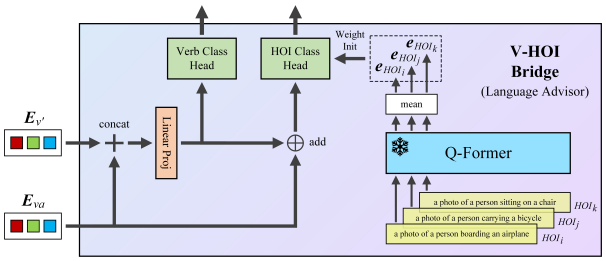

Figure 7. The illustration of V-HOI Bridge.

VLM. In addition to aligning the prior knowledge with the verb recognition part, we add a multi-scale deformable attention module [54] after the cross-attention to focus on the region of the verb box.

For each layer of the vision advisor decoder, the refined verb embeddings e_{va} can be calculated as follows:

$$\boldsymbol{e}_{va} = \text{MSACrossAttn}(\boldsymbol{e}_t, \boldsymbol{f}_g, \boldsymbol{B}_v) \quad (6)$$

$$\boldsymbol{e}_t = \text{MHACrossAttn}(\text{SelfAttn}(\boldsymbol{e}_v, \text{PE}(\boldsymbol{B}_v)), \boldsymbol{f}_{ga}) \quad (7)$$

where MSACrossAttn, MHACrossAttn, and PE are multi-scale, multi-head cross-attention mechanisms, and positional encoding, respectively. With the MHACrossAttn, the vision advisor is able to focus on the region of the verb box, which is related to each HOI query embedding.

Verb-HOI Bridge. For our Verb-HOI Bridge (also language advisor) in Fig. 7, we predict the HOI classes in a two-step manner to fill the gap between the verb recognition and HOI recognition. In the first step, we concatenate $\boldsymbol{E}_{v'}$ after our verb decoder and $\boldsymbol{E}_{va}$ from the vision advisor decoder and feed into a linear projection layer to form E_{vt} and predict the verb classes with a linear verb prediction head in original SOV-STG framework. As the first step follows the original training pipeline of SOV-STG, the vision advisor can align VLM's visual knowledge in a straight feed and return style. In the second step, we add $\boldsymbol{E}_{vt}$ to $\boldsymbol{E}_{va}$ and predict the HOI classes with the HOI prediction head. Similar to previous VLM-based HOID methods [14,24,32], we generate text embeddings to initialize the weights of the HOI prediction head. However, different from previous works that use CLIP text encoder [24, 32] to encode the HOI text prompt into a single vector, we keep the same instruction as the visual feature extraction, we use Q-Former to encode the HOI text prompt into a set of vectors and use the average of the set as the weight of the HOI classifier. Since BLIP2 introduced Q-Former as the bridge between the visual and language information and complete the relation between the vision advisor and the language advisor, thus, we call this part as V-HOI Bridge.

3.4. Training and Inference

As shown in Fig. 1, our proposed method SOV-STG is trained in an end-to-end manner. For the inference queries $\boldsymbol{Q}_{ov}$, the Hungarian algorithm [19] is used to match the ground-truth HOI instances with the predicted HOI instances, and the matching cost and the training loss of predicted HOI instances follow the previous deformable transformer based method [4]. For the DN queries $\boldsymbol{Q}_{dn}$, the ground-truth indices used in query initialization are used to match the predicted HOI instances, and the loss function is the same as the inference queries. As the label embeddings used in the denoising training part are also the specific target priors of the inference part, SOV-STG uses all of the parameters in training and inference.

3.5. Experiments

We evaluate our proposed SOV-STG on the HICO-DET [3] and V-COCO [11] datasets to compare with other SOTA methods and conduct extensive ablation studies to analyze the contributions of each component and show the effectiveness of our proposed method.

3.6. Experimental Settings

Dataset and Metric. The HICO-DET [3] dataset contains 38,118 images for training and 9,658 images for the test. The 117 verb classes and 80 object classes in HICO-DET form 600 HOI classes. According to the number of HOI instances appearing in the dataset, the HOI classes are divided into three categories: *Full*, *Rare*, and *Non-Rare*. Moreover, considering HOI instances including or not including the unknown objects, the evaluation of HICO-DET is divided into two settings: Default and Known Object. The V-COCO [11] dataset contains 5,400 images for training and 4,946 images for the test. In V-COCO, 80 object classes and 29 verb classes are annotated, and two scenarios are considered: scenario 1 with 29 verb classes and scenario 2 with 25 verb classes. We follow the standard evaluation [3] and report the mAP scores.

Implementation Details. We use the DAB-Deformable-DETR trained on COCO [25] to initialize the weight of the feature extractor, the subject decoder, and the object decoder. The feature extractor consists of a ResNet-50 [12] backbone and a 6-layer deformable transformer encoder. We implement three variants of SOV-STG, which are denoted as **SOV-STG-S** with ResNet-50 and 3-layer decoders,

Method	Epoch	Backbone	Default			Known Object		
			Full	*Rare*	*Non-Rare*	*Full*	*Rare*	*Non-Rare*
Two-stage								
CATN [6]	12	ResNet-50	31.86	25.15	33.84	34.44	27.69	36.45
STIP [49]	30	ResNet-50	32.22	28.15	33.43	35.29	31.43	36.45
UPT [47]	20	ResNet-101-DC5	32.62	28.62	33.81	36.08	31.41	37.47
Liu *et al.* [27]	129	ResNet-50	33.51	30.30	34.46	36.28	33.16	37.21
ViPLO [33]	8	ViT-B/32	34.95	33.83	35.28	38.15	36.77	38.56
One-stage								
QAHOI [4]	150	ResNet-50	26.18	18.06	28.61	-	-	-
QPIC [37]	150	ResNet-50	29.07	21.85	31.23	31.68	24.14	33.93
MSTR [16]	50	ResNet-50	31.17	25.31	32.92	34.02	28.83	35.57
CDN-L [45]	100	ResNet-101	32.07	27.19	33.53	34.79	29.48	36.38
HQM (CDN-S) [50]	80	ResNet-50	32.47	28.15	33.76	35.17	30.73	36.50
RLIP-ParSe [42]	90	ResNet-50	32.84	34.63	26.85	-	-	-
DOQ (CDN-S) [35]	80	ResNet-50	33.28	29.19	34.50	-	-	-
GEN-VLKT-S [24]	90	ResNet-50	33.75	29.25	35.10	36.78	32.75	37.99
HOICLIP [32]	90	ResNet-50	34.69	31.12	35.74	37.61	34.47	38.54
GEN-VLKT-M [24]	90	ResNet-101	34.78	31.50	35.77	38.07	34.94	39.01
GEN-VLKT-L [24]	90	ResNet-101	34.95	31.18	36.08	38.22	34.36	39.37
CLIP4HOI [31]	100	ResNet-50	35.33	33.95	35.74	37.19	35.27	37.77
CQL(+GEN-VLKT-S) [40]	90	ResNet-50	35.36	32.97	36.07	38.43	34.85	39.50
UniHOI-S [1]	90	ResNet-50	40.06	**39.91**	40.11	42.20	42.60	42.08
QAHOI-Swin-L [4]	150	Swin-Large-22K	35.78	29.80	37.56	37.59	31.36	39.36
FGAHOI-Swin-L [30]	190	Swin-Large-22K	37.18	30.71	39.11	38.93	31.93	41.02
DiffHOI-Swin-L [41]	90	Swin-Large-22K	41.50	39.96	41.96	43.62	41.41	44.28
PViC-Swin-L [48]	30	Swin-Large	44.32	**44.61**	44.24	47.81	**48.38**	47.64
RLIPv2-ParSeDA-Swin-L [44]	20	Swin-Large	45.09	43.23	45.64	-	-	-
SOV-STG-S	30	ResNet-50	33.80	29.28	35.15	36.22	30.99	37.78
SOV-STG-M	30	ResNet-101	34.87	30.41	36.20	37.35	32.46	38.81
SOV-STG-L	30	ResNet-101	35.01	30.63	36.32	37.60	32.77	39.05
SOV-STG-VLA-S	15	ResNet-50	**41.16**	39.48	**41.67**	**43.81**	**42.63**	**44.17**
SOV-STG-Swin-L	30	Swin-Large-22K	43.35	42.25	43.69	45.53	43.62	46.11
SOV-STG-VLA-Swin-L	15	Swin-Large-22K	**45.64**	44.35	**46.03**	**48.22**	47.12	**48.55**

Table 1. Comparison to state-of-the-arts on the HICO-DET.

#	oDec	sDec	vDec	S-O Attn	STG	VLA	Default		
							Full	*Rare*	*Non-Rare*
(1)	✓		✓		✓		32.68	28.21	34.02
(2)	✓	✓			✓		32.35	27.64	33.63
(3)	✓						30.14	22.82	32.32
(4)	✓	✓					30.62	24.60	32.42
(5)	✓	✓	✓				31.90	25.92	33.69
(6)	✓	✓	✓		✓		33.01	27.83	34.55
(7)	✓	✓	✓	✓	✓		33.80	29.28	35.15
(8)	✓	✓		✓	✓	✓	40.58	38.65	41.16
(9)	✓	✓	✓	✓	✓	✓	**41.16**	**39.48**	**41.67**

Table 2. Contributions of each module. The "oDec", "sDec", and "vDec" denote the object, subject, and verb decoder, respectively.

Method	Backbone	AP^{S1}_{role}	AP^{S2}_{role}
FGAHOI [30]	ResNet-50	59.0	59.3
RLIP-ParSe [42]	ResNet-50	61.9	64.2
MSTR [16]	ResNet-50	62.0	65.2
GEN-VLKT-S [24]	ResNet-50	62.4	64.5
GEN-VLKT-M [24]	ResNet-101	63.3	65.6
HOICLIP [32]	ResNet-50	63.5	64.8
GEN-VLKT-L [24]	ResNet-101	63.6	65.9
CDN-L [45]	ResNet-101	63.9	65.9
UniHOI-S [1]	ResNet-50	65.6	68.3
SOV-STG-S	ResNet-50	61.2	62.5
SOV-STG-M	ResNet-101	63.7	65.2
SOV-STG-L	ResNet-101	63.9	65.4
SOV-STG-VLA-S	ResNet-50	63.8	65.7

Table 3. Comparison on V-COCO.

#	Verb Prediction	Q-Former Text Embed	PE($\boldsymbol{B}_v$)	Default		
				Full	*Rare*	*Non-Rare*
(1)	✓	✓	✓	41.16	39.48	41.67
(2)	✓	✓		40.47	37.85	41.25
(3)	✓		✓	40.17	37.25	41.04
(4)		✓	✓	40.21	38.01	40.86
(5)			✓	39.78	35.51	41.05

Table 4. Ablation studies of VLA.

#	Verb Box	Default		
		Full	*Rare*	*Non-Rare*
(1)	Object Box	33.16	27.21	34.94
(2)	Subject Box	32.78	28.01	34.21
(3)	MBR	33.44	27.84	35.11
(4)	SMBR	33.41	28.22	34.97
(5)	ASMBR	**33.80**	**29.28**	**35.15**

Table 5. Different designs for the verb box. The "SMBR" indicates the Shifted MBR.

#	Denoising Strategies			Default		
	Box	Obj	Verb	*Full*	*Rare*	*Non-Rare*
(1)				32.99	28.28	34.40
(2)	✓			33.27	29.07	34.53
(3)	✓		✓	33.28	28.57	34.69
(4)		✓	✓	33.39	28.82	34.76
(5)	✓	✓		33.51	29.05	34.84
(6)	✓	✓	✓	**33.80**	**29.28**	**35.15**

Table 6. Ablation studies of denoising strategies. The symbol of ✓ means adding noise to the ground-truth.

SOV-STG-M with ResNet-101 and 3-layer decoders, and **SOV-STG-L** with ResNet-101 and 6-layer decoders. The hidden dimension of the transformer is $D = 256$, and the number of the query N_q is set to 64 for HICO-DET and 100 for V-COCO. For the DN part, $2N_p = 6$ groups of noised labels are generated for each ground-truth HOI instance. The same as previous works [1, 31, 32], we use ResNet-50 as the backbone and add VLA to our SOV-STG-S to form **SOV-STG-VLA-S**. In addition, we also use Swin-Transformer [28] as the backbone to achieve the best performance. The ViT-32/B variant is used as the visual encoder in our VLA. The vision advisor decoder has the same number of layers as SOV decoders. We train SOV-STG with the AdamW optimizer [29] with a learning rate of 2e-4 (except for the backbone, which is 1e-5 for HICO-DET, 2e-6 for V-COCO) and a weight decay of 1e-4. The batch size is set to 32 (8 NVIDIA A6000 GPUs, 4 images per GPU), and the training epochs are 30 (the learning rate drops at the 20th epoch). We train SOV-STG-VLA with half of the learning rate, batch size, and training epochs.

3.7. Comparison to State-of-the-Arts

In Tab. 1, we compare SOV-STG and SOV-STG-VLA with the recent SOTA methods on the HICO-DET dataset. Our SOV-STG-S with ResNet-50 backbone achieves 33.80 mAP on the *Full* category of the Default setting. Compared with the transformer-based one-stage methods, QAHOI and MSTR, which are based on the reference point, SOV-STG benefits from the anchor box priors and label priors and achieves 7.62 (29.11%) and 2.63 (8.44%) mAP improvements, respectively. Note that, without any extra language prior knowledge [36], SOV-STG-M outperforms GEN-VLKT-M by 0.26% in one-third of the training epochs. Since our SOV-STG explicitly makes full use of the ground-truth information, compared with DOQ, which also uses ground-truth to guide the training, SOV-STG-S achieves 1.05% mAP improvement with less than half of the training epochs of DOQ. Especially, compared with UniHOI-S and GEN-VLKT-S, UniHOI gains 18.70% mAP with the use of VLM and additional modules. Our VLA ef-

fectively connects the verb prediction part of SOV-STG and VLM, as a result, SOV-STG-VLA-S achieves 41.16 mAP and improves the performance by 21.8% compared with the SOV-STG-S, which is higher than the improvement of UniHOI-S, and our training epochs is one-sixth of UniHOI-S. Moreover, our SOV-STG-VLA-Swin-L achieves 45.64 mAP also with only 15 training epochs. In Tab. 3, on the V-COCO dataset, our SOV-STG-VLA-S improves SOV-STG-S by 4.25% and 5.12% on the scenario 1 and 2, respectively.

3.8. Ablation Study

We conduct all the ablation experiments on the HICO-DET dataset with the SOV-STG-S model.

Contributions of proposed modules. SOV-STG is composed of flexible decoding architecture and training strategies. To clarify the contributions of each proposed module, in Tab. 2, we remove the proposed modules one by one and conduct ablation studies on the HICO-DET dataset. The row of (5) indicates the experiment removing the STG strategy and the S-O attention module is degraded to a sum fusion module. From the result, compared with SOV-STG in (7), the STG strategy and S-O attention improve the performance by 5.96% on the *Full* category. Next, in (4), we remove the verb decoder in (5). As a result, comparing (4) with (5), without the verb decoder, the performance drops by 4.01%. Then, in (3), we remove the subject decoder and the sum fusion module and update both the subject and object boxes by the object decoder. Without balancing the decoding burden of the detection, compared with (4), the performance drops by 1.57%. Furthermore, in (1) and (2), we conduct drop-one-out experiments on the subject and verb decoder, respectively. Compared with (1) and (2), the model without the verb decoder is worse than the model without the subject decoder, which indicates that the verb decoder plays a more critical role. With the MSACrossAttn in our vision advisor decoder, our vision advisor decoder can extract specific region features as the verb decoder, therefore, in (8), we remove the verb decoder of SOV-STG-VLA-S, and the model still performs well.

Vision Language Advisor. Our VLA connects the verb and HOI label prediction both in the visual and language space. To verify the effectiveness of the VLA, we conduct ablation studies in Tab. 4. In row (2), we remove the positional embeddings from the vision advisor decoder, and from the result, the performance drops by 4.13% on the *Full* category. In row (3), we train the HOI class head from scratch without initializing the weight from the text embeddings. From the result, without the guidance of Q-Former in language, the performance on the *Rare* category drops by 5.65%. In row (4), we investigate the effect of the connection of verb and HOI prediction by removing the verb prediction part from VLA. Specifically, the verb embeddings $E_{v'}$ from the verb decoder are directly added to the refined verb embeddings E_{va} from the vision advisor decoder and fed into the HOI class head. As a result, the performance drops by 2.31% on the *Full* category and 3.72% on the *Rare* category compared with the full VLA in (1). Then, in row (5), we remove the verb prediction and train the HOI class head from scratch, and the performance on the *Rare* category drops by 10.06% compared with row (1).

Denoising Strategy. In Tab. 6, we investigate the denoising strategies of three parts of the targets, i.e., the box coordinates, the object labels, and the verb labels. The result of (6) indicates the result of SOV-STG-S. In (1), we set the noise rate of box coordinates to $\delta_b = 0$, the object label flipping rate to $\eta_o = 0$, and the verb label flipping rate to $\eta_v = 0$, thus, the ground-truth box coordinates, object labels, and verb labels are directly fed into the model without any noise. From the result, the accuracy drops by 2.40% compared with the full denoising training in (6). In (3), (4), and (5), we conduct drop-one-out experiments, and the results show that each part of the denoising strategy is effective. For the results between (2) and (3), and (4) and (3), the verb denoising increases the performance when it is used with the object denoising.

Formulations of the verb box. To verify the effectiveness of ASMBR, we use the verb box degraded from the ASMBR to conduct ablation studies, and the results are shown in Tab. 5. From the results of (3) to (5), the adaptive and shift operations for the MBR promote the performance of the verb box, by 1.08% on the *Full* category and 5.17% on the *Rare* category. Furthermore, in (1) and (2), we directly use the object or subject box as the verb box, and the results show that the object box is better for non-rare class recognition, while the subject box is better for rare class recognition.

4. Conclusion

In this paper, we propose a novel one-stage HOID framework, SOV for target-specific decoding, and a specific target guided denoising strategy, STG, for efficient training. We introduce a new format to represent the interaction region in a verb box and improve the verb prediction. Our SOV-STG shows the effectiveness of splitting the HOI decoding with each element of the HOI triplet. Based on SOV-STG, we further propose VLA, which integrates VLM to enhance the interaction representation. Our vision advisor focuses on the alignment between the VLM's prior knowledge and the verb embeddings with the guidance of spatial information. Our V-HOI Bridge connects the visual and language information for verb and HOI label prediction. With the well-designed architecture and training strategy, our framework achieves state-of-the-art performance with fast convergence. In the future, we will explore the potential of the proposed framework in other vision-language tasks and further improve the performance of the model.

References

[1] 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*, 2024. 2, 3, 5, 7

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

[3] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 1, 6

[4] Junwen Chen and Keiji Yanai. Qahoi: Query-based anchors for human-object interaction detection. In *2023 18th International Conference on Machine Vision and Applications (MVA)*, pages 1–5. IEEE, 2023. 2, 6, 7

[5] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *CVPR*, 2021. 2

[6] 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 *CVPR*, 2022. 7

[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. 1

[8] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. DRG: Dual relation graph for human-object interaction detection. In *ECCV*, 2020. 1

[9] Chen Gao, Yuliang Zou, and Jia-Bin Huang. iCAN: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018. 1

[10] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018. 1

[11] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 1, 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[13] ASM Iftekhar, Hao Chen, Kaustav Kundu, Xinyu Li, Joseph Tighe, and Davide Modolo. What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions. In *CVPR*, 2022. 2

[14] Ying Jin, Yinpeng Chen, Lijuan Wang, Jianfeng Wang, Pei Yu, Lin Liang, Jenq-Neng Hwang, and Zicheng Liu. The overlooked classifier in human-object interaction recognition. *arXiv preprint arXiv:2203.05676*, 2022. 3, 6

[15] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. HOTR: End-to-end human-object interaction detection with transformers. In *CVPR*, 2021. 2

[16] 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 *CVPR*, 2022. 2, 7

[17] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *ECCV*, 2020. 1

[18] Sanghyun Kim, Deunsol Jung, and Minsu Cho. Relational context learning for human-object interaction detection. In *CVPR*, 2023. 2

[19] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Res. Logist. Quart*, pages 83–97, 1955. 6

[20] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *CVPR*, 2022. 3, 5

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2, 3, 5

[22] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *CVPR*, 2020. 1

[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. 1

[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. 2, 3, 4, 6, 7

[25] 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 *ECCV*, 2014. 1, 6

[26] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *ICLR*, 2022. 3, 4, 5

[27] Xinpeng Liu, Yong-Lu Li, Xiaoqian Wu, Yu-Wing Tai, Cewu Lu, and Chi-Keung Tang. Interactiveness field in human-object interactions. In *CVPR*, 2022. 7

[28] 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. 7

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 7

[30] Shuailei Ma, Yuefeng Wang, Shanze Wang, and Ying Wei. Fgahoi: Fine-grained anchors for human-object interaction detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 7

[31] 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*, 2024. 2, 3, 5, 7

[32] Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In *CVPR*, 2023. 2, 3, 5, 6, 7

[33] Jeeseung Park, Jin-Woo Park, and Jong-Seok Lee. Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In *CVPR*, 2023. 7

[34] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018. 1

[35] 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. 3, 5, 7

[36] 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 *ICML*, 2021. 2, 7

[37] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021. 2, 4, 7

[38] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. VSGNet: Spatial attention network for detecting human object interactions using graph convolutions. In *CVPR*, 2020. 1

[39] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *CVPR*, 2020. 1

[40] Chi Xie, Fangao Zeng, Yue Hu, Shuang Liang, and Yichen Wei. Category query learning for human-object interaction classification. In *CVPR*, 2023. 2, 7

[41] Jie Yang, Bingliang Li, Fengyu Yang, Ailing Zeng, Lei Zhang, and Ruimao Zhang. Boosting human-object interaction detection with text-to-image diffusion model. *arXiv preprint arXiv:2305.12252*, 2023. 7

[42] Hangjie Yuan, Jianwen Jiang, Samuel Albanie, Tao Feng, Ziyuan Huang, Dong Ni, and Mingqian Tang. Rlip: Relational language-image pre-training for human-object interaction detection. In *NeurIPS*, 2022. 2, 7

[43] Hangjie Yuan, Mang Wang, Dong Ni, and Liangpeng Xu. Detecting human-object interactions with object-guided cross-modal calibrated semantics. In *AAAI*, 2022. 2

[44] 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. 7

[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, 4, 7

[46] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *ICCV*, 2021. 1

[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. 7

[48] Frederic Z. Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring predicate visual context in detecting human-object interactions. In *ICCV*, 2023. 7

[49] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *CVPR*, 2022. 7

[50] Xubin Zhong, Changxing Ding, Zijian Li, and Shaoli Huang. Towards hard-positive query mining for detr-based human-object interaction detection. In *ECCV*, 2022. 3, 7

[51] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao. Polysemy deciphering network for robust human-object interaction detection. In *IJCV*, 2021. 1

[52] 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. 1

[53] Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. Human-object interaction detection via disentangled transformer. In *CVPR*, 2022. 2

[54] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2020. 2, 3, 4, 6

[55] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *CVPR*, 2021. 2