FOCUSING ON WHAT TO DECODE AND WHAT TO TRAIN: EFFICIENT TRAINING WITH HOI SPLIT DECODERS AND SPLIT TARGET GUIDED DENOISING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent one-stage transformer-based methods achieve notable gains in the Human-object Interaction Detection (HOI) task by leveraging the detection of DETR. However, the current methods redirect the detection target of the object decoder, and the box target is not explicitly separated from the query embeddings, which leads to long and hard training. Furthermore, matching the predicted HOI instances with the ground-truth is more challenging than object detection, simply adapting training strategies from the object detection makes the training more difficult. To clear the ambiguity between human and object detection, we propose a novel one-stage framework (SOV), which consists of a subject decoder, an object decoder, and a well-designed verb decoder. Three split decoders with explicitly defined box queries share the prediction burden and accelerate the training convergence. To further improve the training efficiency, we propose a novel Split Target Guided (STG) DeNoising strategy, which leverages learnable object label embeddings and verb label embeddings to guide the training. In addition, for the prediction part, the label-specific information is directly fed into the decoders by initializing the query embeddings from the learnable label embeddings. Extensive experiments show that our method (SOV-STG) achieves 3$\times$ fewer training epochs and 4.68% higher accuracy than the state-of-the-art method.

1 INTRODUCTION

Recent Human-Object Interaction (HOI) detection studies are mainly built on the object detection framework. The most widely used datasets, HICO-DET (Chao et al., 2018) and V-COCO (Gupta & Malik, 2015), share the same object categories as the MS-COCO dataset (Lin et al., 2014). Following the definition of the HOI instance, which is a tuple of the subject (human), the object, and the verb, detecting methods are split into one-stage and two-stage methods. In the beginning, a multi-stream architecture built on top of a CNN-based object detector is commonly adopted in the two-stage methods (Chao et al., 2018; Gkioxari et al., 2018; Qi et al., 2018; Gao et al., 2018). Multi-stream methods resolve the HOI detection problem in split parts and have a good potential to improve. By introducing the human pose information (Kim et al., 2020b; Li et al., 2020; Zhong et al., 2021a), the language priors (Gao et al., 2020; Zhong et al., 2021a), or graph structure (Gao et al., 2020; Ulutan et al., 2020; Zhang et al., 2021b), CNN-based two-stage methods achieve considerable accuracy. On the other hand, CNN-based one-stage methods (Liao et al., 2020; Zhong et al., 2021b; Wang et al., 2020) leverage interaction points to detect possible interaction between the subject and object and achieve promising performance.

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 (Dosovitskiy et al., 2021). At first, the transformer-based methods (Tamura et al., 2021; Zou et al., 2021; Chen et al., 2021; Kim et al., 2021) show the advantage of the attention mechanism by adopting DETR (Carion et al., 2020) in the HOI detection task. QPIC (Tamura et al., 2021) and HOITrans (Zou et al., 2021) follow the same training pipeline as the DETR by viewing the HOI detection problem as a set prediction problem. Without the matching process in one-stage and two-stage CNN-based methods, QPIC and HOITrans 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 interaction recognition together. Even though the following one-stage methods (Zhang et al., 2021a; Liao et al., 2022; Yuan et al., 2022; Iftekhar et al., 2022; Zhou et al., 2022) improve the single decoder design by disentangling the object localization and the interaction recognition in a cascade manner, the subject detection and object detection are still tangled in an instance decoder. By construct, the two-stage transformer-based methods (Zhang et al., 2022; Liu et al., 2022b) stack additional interaction pair detection modules on top of the object decoder without modifying the subject and object detection part. Thus, two-stage methods can focus on filtering the interaction pairs and achieve higher accuracy than the one-stage transformer-based methods.

Both the studies of one-stage (Zhong et al., 2021b) and two-stage (Zhang et al., 2021b) methods show that better detection results promote the final performance a lot. Different from previous one-stage methods (Zhang et al., 2021a; Zhou et al., 2022), which focus on how to disentangle the detection and recognition. In this paper, we start by reviewing the definition of the HOI instance and split the decoding process into three parts, the subject decoding, the object decoding, and the verb decoding, according to the composition of the HOI instance. By doing so, the object decoder maintains the object detection capability from the beginning of the training and accelerates the training convergence. Besides, we introduce subject-object (S-O) attention in the verb decoder to fuse the subject and object information and improve the verb representation learning capabilities.

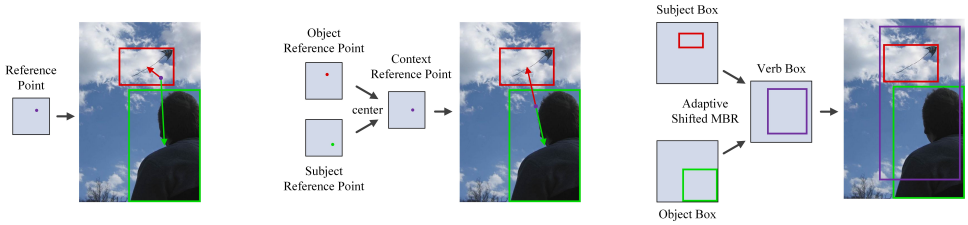
Recently, the variant (Liu et al., 2022a) of DETR formulates explicit learnable anchor boxes as the box queries to improve the connection between query and feature and accelerate the training convergence. Profiting from the explicitly formulated anchor boxes, DN-DETR (Li et al., 2022) introduces the denoising strategy to improve training efficiency and detection performance. DN-DETR uses the coordinates of ground-truth boxes with noise as the anchor boxes and generates object class label queries by encoding the randomly flipped ground-truth object class labels. This work extends the anchor box formulation and denoising strategy to the HOI detection task. Firstly, we split the subject anchor box and object anchor box from the HOI query and introduce an adaptive shifted minimum bounding rectangle (MBR) as the verb anchor box to represent the interaction region between the subject and object. Then, we propose a novel Split Target Guided (STG) DeNoising strategy, which leverages learnable object label embeddings and verb label embeddings to initialize the label queries, both the denoising part and the inference part. With the STG denoising strategy, the matching process between the predicted HOI instances and the ground-truth HOI instances is guided by the denoising queries, and the training is more efficient.

In summary, our contributions are mainly in two aspects: (1) we propose a novel one-stage framework (SOV) to enable the model to concentrate on what to detect and what to recognize during decoding; (2) we propose a novel training strategy (STG) to allow the model to learn the label-specific information between the queries and the results during training. After combining the decoding optimization design and the HOI-specific training strategy, we achieve a new state-of-the-art performance on the HOI detection benchmark with $3\times$ fewer training epochs than the current state-of-the-art method.

2 Representing an HOI instance in anchor boxes

Learning to localize the interaction region. Before the transformer-based methods represent the HOI detection as a set prediction problem, the difficulties for one-stage methods (Liao et al., 2020; Wang et al., 2020; Kim et al., 2020a) lie in how to aggregate the interaction information from a proper region and allocate it to a pair of subject and object. PPDM (Liao et al., 2020) and IP-Net (Wang et al., 2020) use the interaction points and vectors from heatmaps to represent the interaction and require a post-process to match the interaction and the pair of subject and object. UnionDet (Kim et al., 2020a) predicts the union box to represent the interaction and matches the union box with the subject and object pair.

Predicting interactions based on point priors. Transformer-based methods (Tamura et al., 2021; Zou et al., 2021; Chen & Yanai, 2021; Kim et al., 2022) also adopt different ways to represent the HOI instance according to the attention mechanism of the transformer decoders. A simple way is to use query embedding to represent all the elements of the HOI instance (Tamura et al., 2021; Zou et al., 2021). However, the query embedding is learned to represent the localization and recognition information simultaneously, leading to slow convergence and low accuracy. Subsequent studies (Chen & Yanai, 2021; Kim et al., 2022) attempt to leverage the deformable attention mechanism (Zhu et al.,



(a) QAHOI (b) MSTR (c) SOV

Figure 1: The comparison of recent one-stage transformer-based methods.

2020) to guide the decoding by reference points. In Figure 1a, QAHOI (Chen & Yanai, 2021) views the deformable transformer decoder's reference point as the HOI instance's anchor and uses the anchor to guide the subject and object detection. Although QAHOI splits the embedding for reference points from the HOI query embeddings, the HOI query embeddings are still used to predict all the elements of the HOI instance. In Figure 1b, MSTR (Kim et al., 2022) proposes to use the subject, object, and context reference points to represent the HOI instance and predicts the subject, object, and verb based on the reference points. The context reference point is defined as the center of the subject and object reference point, which follows the idea of the interaction point (Liao et al., 2020; Wang et al., 2020; Zhong et al., 2021b). Nevertheless, the query embedding in MSTR is used to predict the final boxes and labels of the HOI instance and still suffers from ambiguous representations. Besides, QAHOI and MSTR use x-y coordinates as the positional priors to guide the decoding, while the box size priors are not considered.

2.1 Adaptive Shifted MBR for Interaction Detection

To clarify the query embeddings for specific usage, we leverage the attention mechanism of DAB-Deformable-DETR (Liu et al., 2022a) to construct our framework and directly use learnable subject and object anchor boxes to predict the subject and object boxes. The anchor boxes are updated layer by layer during the decoding process, and the subject and object boxes from the last layer are used to form the verb box. As shown in Figure 1c, we introduce the adaptive shifted minimum bounding rectangle (MBR) to generate the verb box while considering the spatial relationship between the subject and object boxes. Unlike the UnionDet, which uses the union box to guide the verb recognition, the verb box of SOV is not learned from any additional module but directly from the subject and object boxes. As shown in Figure 2, given the final 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 adaptive shifted MBR (verb box) of the two boxes is defined as:

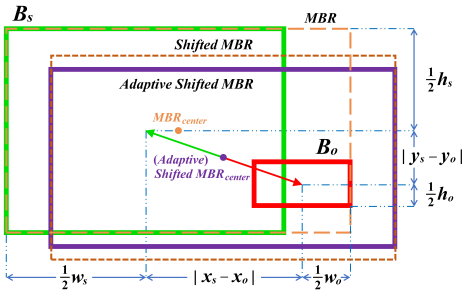


Figure 2: Illustration of adaptive shifted MBR.

$$\boldsymbol{B}_v = \left(\frac{x_s + x_o}{2}, \frac{y_s + y_o}{2}, \frac{w_s + w_o}{2} + |x_s - x_o|, \frac{h_s + h_o}{2} + |y_s - y_o| \right) \tag{1}$$

With the intention of balancing 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. Finally, 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.

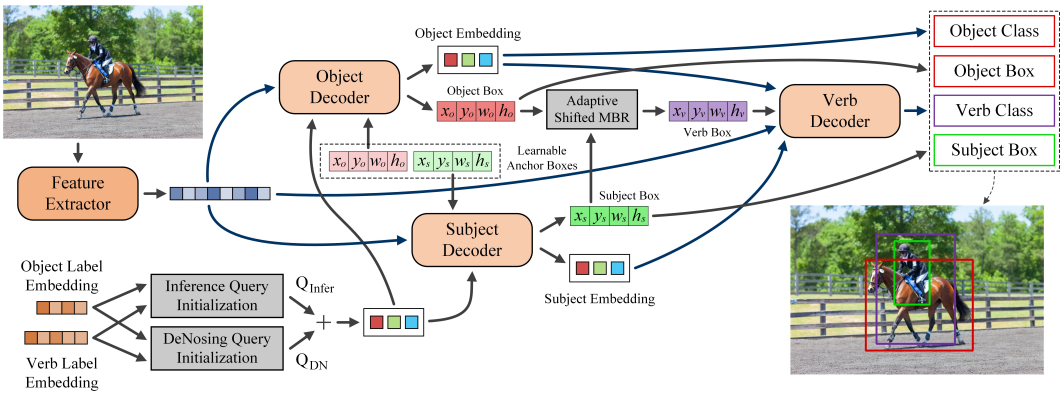


Figure 3: **The overall architecture of SOV-STG.** SOV is composed of the feature extractor and SOV decoders. The learnable anchor boxes and the label embeddings provide HOI-specific priors for inference and denoising training. The entire network follows an encoder-decoder design and can be trained end-to-end.

3 HOI Efficient Decoding and Training

Figure 3 shows the overall architecture of our framework. In this section, we first introduce the HOI efficient decoding architecture, which includes the design of the split decoder in Section 3.1 and the initialization of the label queries in Section 3.2. Then, the STG denoising training strategy built on the efficient decoding architecture is introduced in Section 3.3. Finally, the training and inference details are presented in Section 3.4.

3.1 HOI split decoders

Subject Decoder and Object Decoder. The same as QAHOI (Chen & Yanai, 2021) and MSTR (Kim et al., 2022), we leverage a CNN backbone and deformable transformer encoder (Zhu et al., 2020) to extract the multi-scale global features $\boldsymbol{f}_g \in \mathbb{R}^{N_g \times D}$, where N_g is the number of the total pixels of the multi-scale feature maps and D is the hidden dimension of the embeddings in the whole transformer architecture. As shown in Figure 3, the global features are fed into the subject and object decoder with the learnable anchor boxes. To maintain the detection capability of the object detector, the object decoder with the feed-forward heads is the same as the one trained in the detection task. Furthermore, we clone the object decoder to initialize the subject decoder and alleviate the learning burden of the subject decoder. The subject and object decoder updates the subject anchor box $\boldsymbol{B}_s$ and object anchor box $\boldsymbol{B}_o$ and query embeddings $\boldsymbol{e}$ layer by layer in a parallel manner. Then, the object embedding $\boldsymbol{e}_o$ from the object decoder is used to predict the object class, and the subject box and object box are used to generate the verb box $\boldsymbol{B}_v$. Finally, the object and subject embeddings with the verb box are fed into the verb decoder to predict the verb class.

Verb Decoder with S-O attention. Since our architecture disentangles the detection of subject and object and extracts the object embedding and subject embedding separately, we design a verb decoder to fuse the subject and object embeddings. In the verb decoder, we replace the original self-attention with our **S**ubject-**O**bject (S-O) attention, which is illustrated in Figure 4. Following the idea of two-stage methods (Zhang et al., 2021b; Ulutan et al., 2020) use element-wise multiplication to fuse different features from different streams, we adopt the multi-branch element-wise multiplication (Zhang et al., 2021b) as the multiplication attention in our S-O attention. However, unlike the two-stage method, we fuse two embeddings in a multi-layer manner and share the weight in the multiplication attention module across different layers. Moreover, we introduce a bottom-up path in S-O attention to integrate the

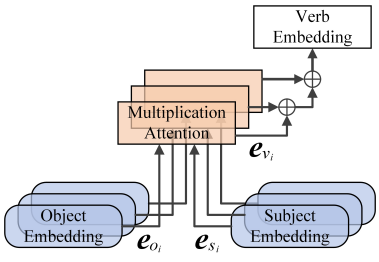


Figure 4: Illustration of S-O attention.

information from the bottom to the top layer. Given the subject embedding $\boldsymbol{e}_{s_i} \in \mathbb{R}^{N_q \times D}$ and object embedding $\boldsymbol{e}_{o_i} \in \mathbb{R}^{N_q \times D}$ from the i-th layer ($i > 1$), where N_q is the number of queries, the verb embedding $\boldsymbol{e}_{v_i}$ after the bottom-up path can be defined as:

$$\boldsymbol{e}_{v_i} = \frac{\text{MulAttn}(\boldsymbol{e}_{s_{i-1}}, \boldsymbol{e}_{o_{i-1}}) + \text{MulAttn}(\boldsymbol{e}_{s_i}, \boldsymbol{e}_{o_i})}{2} \tag{2}$$

Then, the verb embedding output from the top layer is fed into the cross-attention module to further extract the global semantic information based on the global feature $\boldsymbol{f}_g$ and the verb box.

3.2 Split Label Embeddings

As shown in Figure 3, two kinds of learnable label embeddings are used to initialize the query embeddings for SOV decoders. Different from the original denoising method (Li et al., 2022), we use the label embeddings both in the denoising and inference parts and enable the inference part obtain the input query with label-specific information from the beginning. We define the object label embeddings $\boldsymbol{t}_o \in \mathbb{R}^{C_o \times D}$ as the object label priors, which consist of C_o vectors with D dimensions, where C_o is the number of object classes and D is the hidden dimension of the transformer. Similarly, the verb label embeddings $\boldsymbol{t}_v \in \mathbb{R}^{C_a \times D}$ 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:

$$\begin{aligned} \boldsymbol{q}_o = \boldsymbol{A}_o \boldsymbol{t}_o, \quad \boldsymbol{q}_v = \boldsymbol{A}_v \boldsymbol{t}_v \\ \boldsymbol{q}_{ov} = \boldsymbol{q}_o + \boldsymbol{q}_v \end{aligned} \tag{3}$$

3.3 Split Target Guided Denoising

As the object and verb labels are the targets of HOI detection, the two label embeddings can be viewed as the split target priors. Since the denoising query embeddings are generated from the split target priors and used to guide the denoising training, thus, we call our denoising strategy as **S**plit **T**arget **G**uided (STG) denoising. In Figure 5, we show the initialization of the DN query embeddings and visualize the process of adding noise to one of the ground truth HOI instances. 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 one-hot label of the object and verb classes, k is the number of ground-truth HOI instances, two kinds of label DN query embeddings are initialized. Following the DN-DETR (Li et al., 2022), for the k-th ground-truth HOI instance, the noised object label $\boldsymbol{o}'_k$ is obtained by randomly flipping the ground-truth index of the object label $\boldsymbol{o}_k$ to another object class index, and N_p groups of noised labels are generated. Next, the object DN query embeddings $\boldsymbol{q}_{dn}^{(o)} \in \mathbb{R}^{N_p \cdot k \times D}$ are gathered from the object label embeddings $\boldsymbol{t}_o$ according to the indexes of the noised object labels $\boldsymbol{O}'_{gt}$. Because the verb label consists of co-occurrence ground-truth classes, to keep the co-occurrence ground-truth indexes appearing in the noised verb label, we randomly flip the other indexes of the ground-truth verb label to generate the noised verb label $\boldsymbol{v}'_k$. Then, the verb label DN query embeddings $\boldsymbol{q}_{dn}^{(v)} \in \mathbb{R}^{N_p \cdot k \times D}$ are the sum of the verb label DN embeddings selected from the verb label embeddings $\boldsymbol{t}_v$ according to the indexes of the noised verb labels $\boldsymbol{V}'_{gt}$. Finally, we concatenate the object DN query embeddings and verb DN query embeddings to form the DN query embeddings $\boldsymbol{q}_{dn} \in \mathbb{R}^{2N_p \cdot k \times D}$ for the denoising training. In this way, the split target priors can be learned by the denoising training separately and can be used to guide the inference part of SOV.

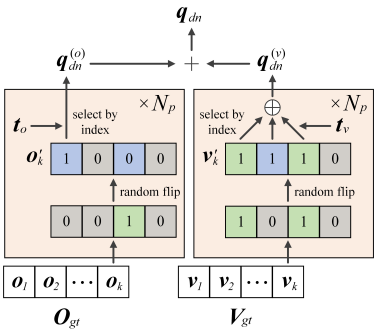


Figure 5: Illustration of DN query initialization.

3.4 TRAINING AND INFERENCE

Our proposed framework SOV-STG is trained in an end-to-end manner. For inference query embeddings, the Hungarian algorithm (Kuhn, 1955) is used to matching the ground-truth HOI instances with the predicted HOI instances, and the matching cost and the training loss are the same as QAHOI (Chen & Yanai, 2021). Moreover, the denoising and inference parts are trained with the same loss function. With the basic concept that the same ground-truth label flip rate is difficult for the model to denoise at the beginning of the training but becomes acceptable during the training, we further improve the denoising strategy by introducing a dynamic DN scale factor $\gamma \in (0, 1)$ to control the object label flip rate $\eta_o \in (0, 1)$ and the verb label flip rate $\eta_v \in (0, 1)$ according to the training epochs. With the dynamic DN scale strategy, the label flip rate η will be set to $\gamma \cdot \eta$ at the beginning of the training and linearly increase to η during the training. The box denoising is the same as the DN-DETR (Li et al., 2022). However, the dynamic DN scale strategy is also implemented to the box noising rate δ_b to improve the denoising performance. As our STG moves the label encoding embeddings out of the denoising part as the split target priors, SOV-STG uses all of the parameters in training and inference.

4 EXPERIMENTS

We evaluate our proposed SOV-STG on the HICO-DET (Chao et al., 2018) and V-COCO (Gupta & Malik, 2015) datasets to compare with current SOTA methods and conduct extensive ablation studies to analyze the contributions of each component and show the effectiveness of our proposed method.

4.1 EXPERIMENTAL SETTINGS

Dataset and Metric. The HICO-DET (Chao et al., 2018) dataset contains 38,118 images for training and 9,658 images for the test. The 117 object and 80 verb 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 (Gupta & Malik, 2015) 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 (Chao et al., 2018) and report the mAP scores.

Implementation Details. We adopt the DAB-Deformable-DETR trained on the COCO (Lin et al., 2014) dataset to initialize the weight of the feature extractor, the subject decoder, and the object decoder. The feature extractor consists of a ResNet-50 (He et al., 2016) backbone and a 6-layer deformable transformer encoder. The subject decoder and the object decoder are both 6-layer deformable transformer decoders, and the cross-attention in the verb decoder also has six layers. The hidden dimension of the transformer is $D = 256$, and the number of the query is set to $N_q = 64$. For the DN part, $2N_p = 6$ groups of noised labels are generated for each ground-truth HOI instance and follow the DN-DETR. The dynamic DN scale is set to $\gamma = \frac{2}{3}$, and we maintain the same denoising level as the DN-DETR at the start of the training by setting the noising rate of the box to $\delta_b = 0.6$, the object flip rate to $\eta_o = 0.3$, and the verb flip rate to $\eta_v = 0.6$. We train the model with the AdamW optimizer (Loshchilov & Hutter, 2018) with a learning rate of 2e-4 (except for the backbone, which is 1e-5) and a weight decay of 1e-4 for the HICO-DET dataset. The batch size is set to 32 (4 images per GPU), and the training epochs are 30 (learning rate drops at the 20th epoch), which is one-third of the CDN (Zhang et al., 2021a), and one-fifth of the QPIC (Tamura et al., 2021) and QAHOI (Chen & Yanai, 2021). For the V-COCO dataset, we freeze the backbone to prevent overfitting and set the learning rate to 1e-4 and the batch size to 16 (2 images per GPU). All of the experiments are conducted on 8 NVIDIA A6000 GPUs.

4.2 COMPARISON TO STATE-OF-THE-ARTS

In Table 1, we compare our proposed SOV-STG with the recent SOTA methods on the HICO-DET dataset. Our SOV-STG with ResNet-50 backbone achieves 33.57 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

Method	Backbone	Default			Known Object		
		Full	*Rare*	*Non-Rare*	*Full*	*Rare*	*Non-Rare*
Two-stage							
IP-Net (Wang et al., 2020)	Hourglass-104	19.56	12.79	21.58	22.05	15.77	23.92
VSGNet (Ulutan et al., 2020)	ResNet-152	19.80	16.05	20.91	-	-	-
ACP (Kim et al., 2020b)	ResNet-152	20.59	15.92	21.98	-	-	-
DJ-RN (Li et al., 2020)	ResNet-50	21.34	18.53	22.18	23.69	20.64	24.60
PD-Net (Zhong et al., 2021a)	ResNet-152	22.57	17.61	23.79	26.86	21.70	28.44
DRG (Gao et al., 2020)	ResNet-50-FPN	24.53	19.47	26.04	27.98	23.11	29.43
SCG (Zhang et al., 2021b)	ResNet-50-FPN	31.33	24.72	33.31	34.37	27.18	36.52
CATN (Dong et al., 2022)	ResNet-50	31.86	25.15	33.84	34.44	27.69	36.45
UPT (Zhang et al., 2022)	ResNet-101-DC5	32.62	28.62	33.81	36.08	31.41	37.47
One-stage							
UnionDet (Kim et al., 2020a)	ResNet-50-FPN	17.58	11.72	19.33	19.76	14.68	21.27
PPDM (Liao et al., 2020)	Hourglass-104	21.73	13.78	24.10	24.58	16.65	26.84
GGNet (Zhong et al., 2021b)	Hourglass-104	23.47	16.48	25.60	27.36	20.23	29.48
HOITrans (Zou et al., 2021)	ResNet-101	26.61	19.15	28.84	29.13	20.98	31.57
HOTR (Kim et al., 2021)	ResNet-50	25.10	17.34	27.42	-	-	-
QAHOI (Chen & Yanai, 2021)	ResNet-50	26.18	18.06	28.61	-	-	-
AS-Net (Chen et al., 2021)	ResNet-50	28.87	24.25	30.25	31.74	27.07	33.14
QPIC (Tamura et al., 2021)	ResNet-101	29.90	23.92	31.69	32.58	26.06	34.27
MSTR (Kim et al., 2022)	ResNet-50	31.17	25.31	32.92	34.02	28.83	35.57
(Zhou et al., 2022)	ResNet-50	31.75	27.45	33.03	34.50	30.13	35.81
CDN-L (Zhang et al., 2021a)	ResNet-101	32.07	27.19	33.53	34.79	29.48	36.38
SOV-STG	ResNet-50	**33.57**	**29.82**	**34.69**	**36.04**	**31.79**	**37.30**

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

Method	AP^{S1}_{role}	AP^{S2}_{role}
UnionDet (Kim et al., 2020a)	47.5	56.2
SCG (Zhang et al., 2021b)	54.2	60.9
GGNet (Zhong et al., 2021b)	54.7	-
HOTR (Kim et al., 2021)	55.2	64.4
QPIC (Tamura et al., 2021)	58.8	61.0
UPT (Zhang et al., 2022)	61.3	67.1
SOV-STG	61.5	63.0

Table 2: Comparison on V-COCO.

Method	Default		
	Full	*Rare*	*Non-Rare*
SOV-STG	33.57	29.82	34.69
-STG	33.18	28.71	34.52
-Subject Decoder	32.53	28.31	33.53
-Verb Decoder	32.41	28.04	33.72
-DN	26.39	20.87	28.04

Table 3: Contributions of each module in SOV-STG.

achieves 28.23% and 7.70% mAP improvements, respectively. Compared with the transformer-based two-stage method, UPT, which uses ResNet-101-DC5 as the backbone, SOV-STG outperforms UPT by 2.91%. Although we do not optimize the hyper-parameters of the architecture and the denoising strategy, we only adjust the learning rate and the batch size and freeze the backbone for the training on the V-COCO dataset. Similarly, in Table 2, SOV-STG achieves 61.5 mAP on AP^{S1}_{role} and surpasses QPIC and UPT by 4.59% and 0.33%, respectively.

4.3 Ablation Study

We conduct all the ablation experiments on the HICO-DET dataset, and if not explicitly noticed, the same training setup is used as the training of our SOTA 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 Table 3, we remove the proposed modules one by one and conduct ablation studies on the HICO-DET dataset. Line 2 indicates the experiment replacing the STG strategy with the standard DN strategy, which uses the label embeddings only in the DN part and initializes the inference query embeddings from independent learnable embeddings. From the result, the STG strategy improves the performance by 1.18% in the *Full* category. Next, in Line3, we remove the subject decoder of Line 2 and replace the verb decoder with a deformable transformer decoder. In this way, the experiment of Line 3 forms a similar architecture as CDN-B (Zhang et al., 2021a). However, from the result of Line 4, which removes the verb decoder of Line 3, with the DN strategy, the pure DAB-Deformable-DETR model performs better than adding an additional verb decoder. With the standard DN strategy, Line 2, SOV architecture outperforms Line 4, pure DAB-Deformable-DETR architecture, by 2.58%. In Line 5, we remove the DN strategy of Line 4, which is similar to the QAHOI (ResNet-50) (Chen & Yanai, 2021), and this experiment is viewed as the base model of our framework. The result of Line 5 shows that SOV-STG achieves a 27.21% gain by improving the architecture designs and training strategies.

Verb Box	Default		
	Full	*Rare*	*Non-Rare*
Subject Box	32.90	28.04	34.35
Object Box	32.62	27.02	34.29
MBR	32.29	26.66	33.97
Shifted MBR	32.63	26.64	34.41
Adaptive Shifted MBR	**33.57**	**29.82**	**34.69**

Table 4: Different designs for the verb box.

#	Verb Decoder Designs					Default		
	S-O Attention		Feature Fuse	Cross-attention				
	last layer	multi-layer		w/o self-attn	with self-attn	*Full*	*Rare*	*Non-Rare*
(1)	✓		S-O Fuse	✓		**33.57**	29.82	34.69
(2)	✓		S-O Fuse		✓	33.24	28.17	**34.75**
(3)	✓		Mul Fuse	✓		32.68	28.29	33.99
(4)	✓		Sum Fuse	✓		33.30	28.70	34.67
(5)		✓	Sum Fuse	✓		32.53	**29.97**	33.30
(6)		✓	Sum Fuse		✓	32.51	28.45	33.46
(7)	✓		Sum Fuse			31.98	27.39	33.36

Table 5: Ablation studies for verb decoder designs.

Formulations of the verb box. The proposed adaptive shifted MBR is a flexible verb box that dynamically considers the spatial relationship between the subject and object box and guides the verb decoder to extract semantic features from the corresponding region. To verify the effectiveness of the

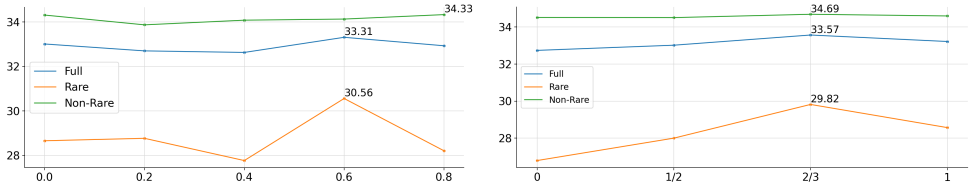
proposed adaptive shifted MBR, we use the verb box degraded from the adaptive shifted MBR to conduct ablation studies, and the results are shown in Table 4. From Line 3 to Line 5 of the results, the adaptive and shift operations for the MBR promote the performance of the verb box, by 3.96% in the *Full* category and 11.85% in the *Rare* category. Furthermore, in Line 1 and 2, we directly use the object or subject box as the verb box, and the results show that the region of the subject plays a more critical role in the verb prediction.

Verb Decoder. The verb decoder is the core module of the SOV model, which is responsible for verb prediction. To illustrate the strength of the verb decoder in SOV, different variants of the verb decoder we have attempted are shown in Table 5. Line 1 indicates the verb decoder used in SOV, where the last layer means using the fused embeddings from the last last layer of the S-O Attention. In Line 2, we restore the self-attention in cross-attention module, however, the accuracy on *Full* and *Non-Rare* categories dropped. In line 3, we also attempt to use multiplication attention without bottom-up. From the results of Line 1 and 3, the bottom-up path increases the performance of the multiplication attention by 2.72%. Moreover, in Line 4, we find that simple sum attention can also achieve a good performance. Furthermore, in Line 5 and 6, similar to Yue *et al.* Liao et al. (2022), we initialize learning verb query embeddings for the cross-attention module and add the multi-layer fused embeddings after the sum attention to the verb query embeddings on corresponding layers. As the results of Line 4, 5, and 6, for our framework, using the last-layer fused embeddings is better. In Line 7, we also explore the contribution of the cross-attention module. Compared with Line 7 and 4, with the cross-attention module, the performance increased by 4.13%.

Denoising Strategies. In Table 6, we investigate the denoising strategies of three parts of the targets, i.e., the box coordinates, the object labels, and the verb labels. In Line 1, we set the noise rate of box coordinates to $\delta_b = 0$, the object label filp rate to $\eta_o = 0$, and the verb label filp 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 3.13% compared with the full denoising training in Line 6. From the results between Line 1 and 2 and Line 5 and 6, by using the box denoising, the accuracy increased by 1.54% and 4.03%, respectively. For the results of Line 3 and 4, only denoising the object labels is better than only denoising the verb labels, and by adding the denoising of the object labels to Line 4, in Line 6, the accuracy increase by 2.50%.

#	Denoising Strategies			Default		
	Box	Obj	Verb	*Full*	*Rare*	*Non-Rare*
(1)				32.55	27.99	33.66
(2)	✓			33.05	27.97	34.57
(3)	✓	✓		33.13	29.08	34.34
(4)	✓		✓	32.75	27.59	34.29
(5)		✓	✓	32.27	27.56	33.68
(6)	✓	✓	✓	**33.57**	**29.82**	**34.69**

Table 6: Ablation studies for denoising strategies.



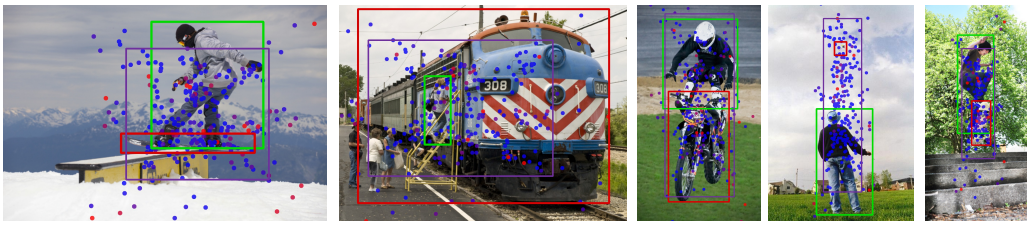
(a) Different verb flip rates (η_v). (b) Different dynamic DN scales (γ).

Figure 6: The effects of the verb flip rate and the dynamic noise scale.

Different verb flip rates and dynamic DN scales. In Figure 6a, we fix the noising rate of the box and the object label flip and adjust the verb flip rate to investigate the effects of the verb flip rates (η_v) on the accuracy without the dynamic DN scale. From the results, $\eta_v = 0.6$ is an appropriate value for the verb flip rate, and the verb flip rate boosts the accuracy of the *Rare* category. Next, in Figure 6, we fix all the noise rate parameters and adjust the dynamic DN scale (γ) to reveal the effects of different dynamic DN scales. As the dynamic DN scale increases, the denoising training becomes more difficult at the beginning, and while using the $\gamma = \frac{2}{3}$, the accuracy on all the categories is highest.

4.4 Qualitative Results

Because our method offers each element of the HOI instance a particular representation, we can easily analyze the inference results from an HOI-specific perspective. Thus, to better understand the



(a) (b) (c) (d) (e)

Figure 7: The visualization of sampling points in the last layer of the verb decoder's cross-attention. We visualize the sampling points of the top-1 score query and draw all the sampling points from different scales and attention heads in one image. The sampling points with high attention weights are colored in red.

strength of our proposed method, we visualize the inference results on the samples from the test set of the HICO-DET dataset in this section.

HOI-specific priors. In our framework, following the definition of the HOI instance, three kinds of priors are learned after the training: the box prior, the object label prior and the verb label prior. During the inference, these priors are updated gradually to represent specific HOI instances according to the global features from the feature extractor. As the box priors are used to obtain the spatial representation of pairs of objects and subjects, we visualize the box priors in Figure 8 to show the relationships between the objects and subjects in the same box priors. We visualize the first eight pairs of object and subject anchor boxes, every two boxes in the same color represent one pair of object and subject. From the results, the subject and object anchor boxes in the same pair are almost the same after the training. We doubt that the layer-by-layer decoding process requires the object and subject boxes to start from the same position to ensure they are in the same pair and share the same semantic feature to recognize the verb label. Moreover, we also visualize the object and verb label priors in Appendix A.1.

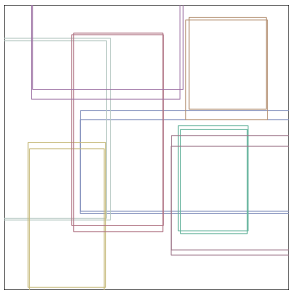


Figure 8: Part of Anchor box priors.

Sampling points in the verb decoder. In the verb decoder, the multi-scale deformable cross-attention is used to aggregate the features from the global semantic feature according to the verb box. In Figure 7, we visualize the attention based on the verb box in several typical cases. From all of the images, the sampling points mainly locate inside the verb box and concentrate on the interaction region of the subject and object. Besides, in Figure 7a and 7d, when the object is smaller than the subject, the sampling points can aggregate the features outside the verb box or evenly cover the verb box to get more information. Moreover, in Figure 7c and 7e, the sampling points on the interaction part of the subject and object are more densely distributed. In Figure 7b, when the object is far larger than the subject, the verb box obtained by the adaptive shifted MBR shrinks the attention region, then the sampling points are more concentrated on the subject, which is the smaller one.

5 Conclusion and Future Work

In this paper, we proposes a novel one-stage framework, SOV with HOI split decoders for target-specific decoding and a split target-guided denoising strategy, STG, for target-specific training. Our framework SOV-STG adopts a new format to represent HOI instances in boxes and learns HOI-specific priors for decoding. With the well-designed architecture and efficient training strategy, our framework achieves state-of-the-art performance with less training cost. Since our architecture disentangles the HOI detection by specific priors and decoders, it is easy to improve any one of them. In the future, we are going to explore the object and verb label priors initialized from language models to improve performance.

REFERENCES

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

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

Junwen Chen and Keiji Yanai. Qahoi: Query-based anchors for human-object interaction detection. *arXiv preprint arXiv:2112.08647*, 2021.

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

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.

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.

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

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

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

Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

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

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.

Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. UnionDet: Union-level detector towards real-time human-object interaction detection. In *ECCV*, 2020a.

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

Bumsoo Kim, Jonghwan Mun, Kyoung-Woon On, Minchul Shin, Junhyun Lee, and Eun-Sol Kim. Mstr: Multi-scale transformer for end-to-end human-object interaction detection. In *CVPR*, 2022.

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

Harold W Kuhn. The hungarian method for the assignment problem. *Naval Res. Logist. Quart*, pp. 83–97, 1955.

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.

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.

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.

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.

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.

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*, 2022a.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018.

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

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

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

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.

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

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*, 2021a.

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

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.

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

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*, 2021b.

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.

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.

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.

A APPENDIX

A.1 LABEL PRIORS

(a) The object label embeddings $\boldsymbol{t}_o$. (b) The verb label embeddings $\boldsymbol{t}_v$.

(c) The object label coefficient matrix $\boldsymbol{A}_o$. (d) The object label coefficient matrix $\boldsymbol{A}_v$.

Figure 9: The label priors of object and verb labels visualized according to the Principal Component Analysis (PCA) algorithm.

As shown in Figure 9a and 9b, we visualize the object and verb label embeddings. From the results, the object labels are evenly divided in the embedding space, and the verb labels are clustered in some small groups in the embedding space. The coefficient matrices $\boldsymbol{A}_o$ and $\boldsymbol{A}_v$ control the combination of the object label embeddings $\boldsymbol{t}_o$ and the verb label embeddings $\boldsymbol{t}_v$ for N_q queries, respectively. Thus, we split the coefficient matrices $\boldsymbol{A}_o$ and $\boldsymbol{A}_v$ by the first dimension and visualize them in Figure 9c and 9d. As shown in Figure 9c, the vectors in the first dimension of $\boldsymbol{A}_o$ locate sparse in the embedding space, which means the combination of each query for the object label is not similar. However, in Figure 9d, the vectors in the first dimension of $\boldsymbol{A}_v$ are gathered in some small groups, which means the combination of each query for the verb label is similar due to the co-occurrence of the verb labels.

A.2 QUALITATIVE RESULTS

In Figure 10, we show the inference process starting from the box priors to the final HOI detection results. From the results, the subject and object decoder can localize the subject and object in the first layer and refine the box within the after two layers. Even in Column 6 and 8, the anchor boxes are not accurate, the subject and object decoder can still localize the subject and object in the first two

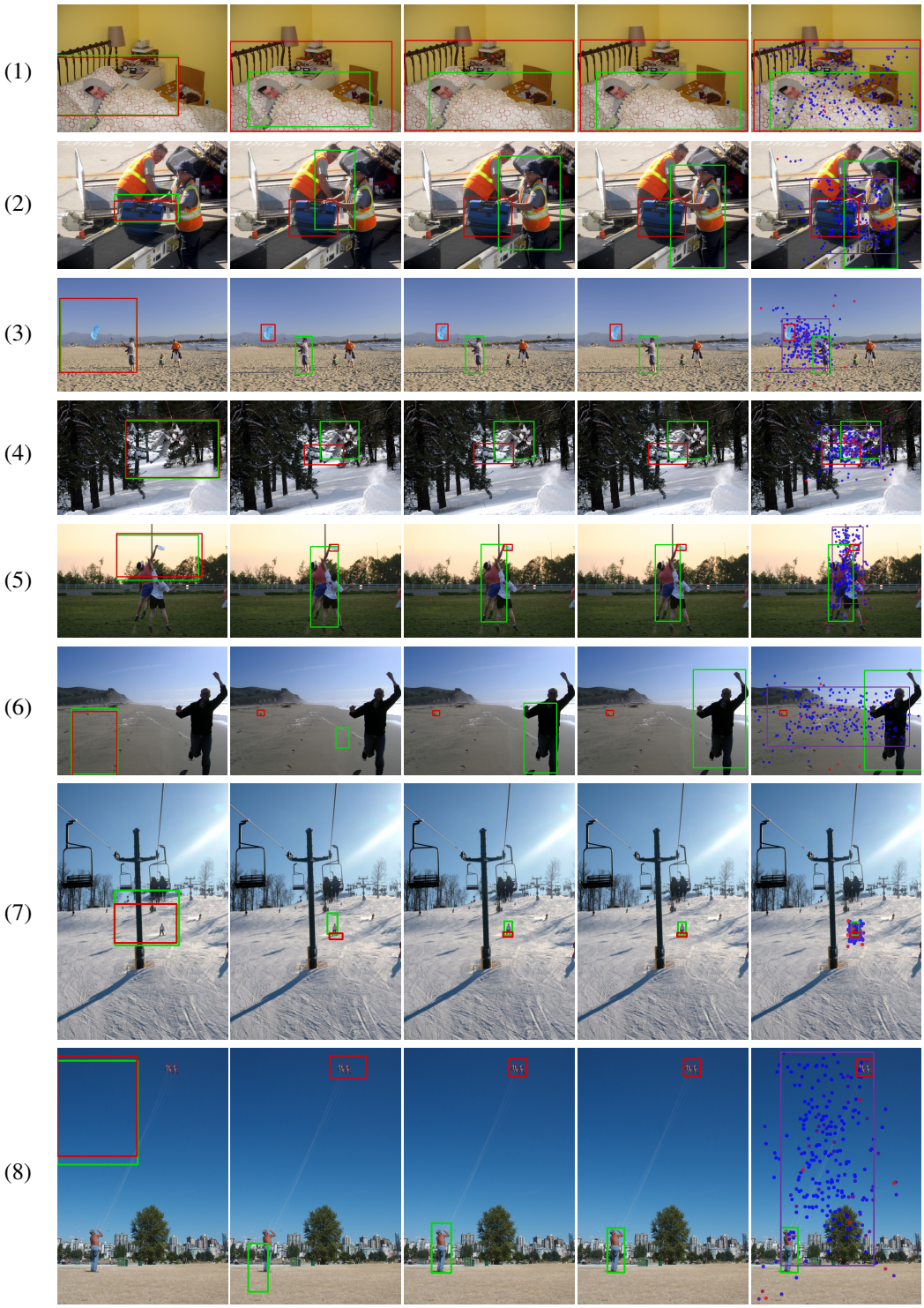


Figure 10: Visualization of the inference process. We show the box updating and the attentions of the query with the highest score.

layers. Furthermore, in Column 7 when the object and subject are small and close, the subject and object decoder can distinguish the subject and object in the first layer.