# Separate Guided Denoising Training for Human-Object Interaction Detection

Yuki Isoda[1] and Daisuke Kobayashi[2]

Corporate Research and Development Center, Toshiba Corporation,
[1]yuki2.isoda@toshiba.co.jp,[2]daisuke32.kobayashi@toshiba.co.jp

**Abstract.** Understanding scenes requires not only the detection objects but also the recognition of the interactions between them. Human-Object Interaction (HOI) detection plays a crucial role in enhancing contextual comprehension by identifying the interactions between humans and objects, which is essential for building more robust and intelligent vision systems. While DETR-based models have shown significant success in HOI detection, they are hindered by slow training convergence. The SOV-STG method has attempted to address this challenge in previous research. To further improve the learning efficiency and accuracy of SOV-STG, we introduce a novel Separate Guided Denoising training strategy specifically designed for HOI detection. Our approach separates the denoising of noised ground truth data for both the human-object decoder and the verb decoder, enabling more efficient and targeted training. Furthermore, we enhance training performance by merging redundant human-object pair annotations, and filtering and regenerating noised bounding boxes. The proposed method was validated on the HICO-DET dataset, achieving state-of-the-art results. Our contributions include a novel training strategy that improves accuracy and ablation studies demonstrating its effectiveness.

**Keywords:** Human-Object Interaction Detection · Denoising training · Visual Relationship Detection

## 1 Introduction

Recent advancements in computer vision and machine learning have significantly enhanced our ability to understand the visual semantics of complex interactions between humans and objects in images. Human-Object Interaction (HOI) detection, which aims to identify triplets of <human, verb, object>, is a crucial task with applications in various domains such as human activity recognition [43], image retrieval [17], and visual question answering [38].

Models based on DETR [3] (Detection Transformer) have recently achieved impressive results in HOI detection. However, DETR-based models are known for their slow training convergence and require many epochs to achieve good performance. Originally designed for object detection, DETR framework has prompted extensive research to improve its efficiency, resulting in approaches

such as DAB-DETR [29] and DN-DETR [23]. DAB-DETR enhances performance by leveraging 4D coordinates (x, y, w, h) as anchor boxes and refining the mode through spatial positional training. Meanwhile, DN-DETR focuses on the instability of bipartite graph matching. Denoising (DN) training of queries is introduced to learn one-to-one matching between the outputs of DN queries and ground truth data, which stabilizes the bipartite graph matching during the training process.

HOI detection faces similar challenges with the slow convergence of DETR training. SOV-STG [5] has been proposed to extend these studies to HOI detection. SOV assigns a single decoder to detect humans, objects, and recognize verbs, while STG utilizes learnable object and verb label embeddings to guide training. SOV stabilizes training by focusing the decoders on specific targets, and STG speeds up convergence by connecting ground truth labeling information with predefined dataset labels. However, the STG DN training strategy presents challenges when applied to HOI detectios Fig. 1 show the differences in noise handling between the previous method and the proposed method. Specifically, it is unclear what is removed as noise and what remains for DN training. To address this, we propose a separate guided DN training strategy, where the detection of human-object pairs and the DN training for verb recognition are performed separately. This approach clarifies what should be denoised for each decoder, making the DN training is more effective. Furthermore, we focus on the existing dataset and add noise to make the DN training for HOI detection more effective. There are cases in existing datasets where verbs are annotated in redundant bounding boxes for the same human-object pair. As shown in **??**, one output may recognize only "straddle" while another recognizes "ride" and "sit on" even though they refer to the same human-object pair. It is redundant to recognize split verb annotations for the same human-object pair, so we merge them into a single pair. Furthermore, we filter and regenerate noised bounding boxes to ensure they are closest to their respective reconstruction targets.

In summary, our contributions are threefold:

· We proposed a separate guided DN training strategy that allows each decoder to concentrate on its specific denoising task.

· We introduced methods for merging redundant human-object pair annotations and filtering and regenerating noised bounding boxes to enhance the effectiveness of DN training.

· We achieved state-of-the-art results in the HICO-DET benchmark.

## 2 Related Work

**Two-stage Methods.** The two-stage approach [4, 10, 12, 14, 15, 21, 22, 34, 41, 42, 44–46, 52, 53, 55, 57], using off-the-shelf detectors, first detects humans and objects and recognize interactions for detected human-object pairs. Since the introduction of the multi-stream architecture in HO-RCNN [4], many methods have been proposed based on this framework. In HO-RCNN, human appearance features, object appearance features, and spatial features are extracted in each

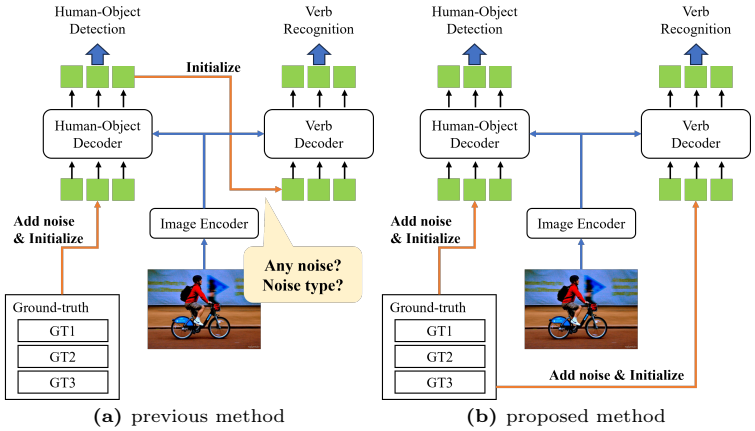
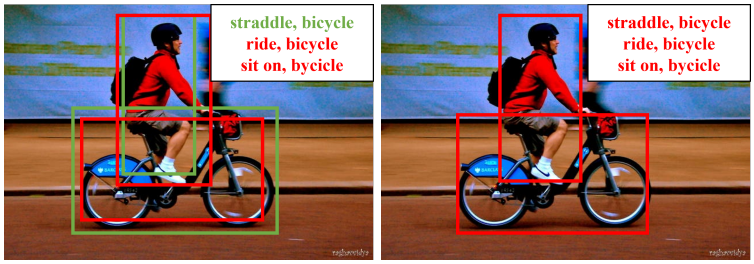



**Fig. 1:** Comparison between the previous method and the proposed method. In the previous method, DN queries initialized using noised ground truth and fed into the human-object decoder. The output of the human-object decoder is then passed to the verb decoder, making unclear which part of the noise is removed during the human-object decoding and what remains for the verb decoder.In contrast, our proposed method separately initializes DN queries for both the human-object decoder and the verb decoder from the noised ground truth, ensuring that each decoder handles its specific denoising task independently.

stream and then recognize interactions. Subsequent methods have incorporated human pose features [12,25,41,55], linguistic features [9,31], and graph structures [9, 35, 40, 52, 57] to improve recognize interactions.

**One-stage Methods.** One-stage methods [1,2,5,6,19,24,26,27,33,36,36,37,39, 47–51,54,59,60] typically perform object detection and recognize interactions simultaneously. Early methods used interaction keypoints [26,47] and join regions [18] as predefined anchors. Recently, DETR-based HOI detectors have gained attention, leading to significant performance improvements. However, these methods often suffer from slow learning convergence. Some approaches [7, 27, 51, 56] have introduced multiple decoders for each subtask to address this issue, but they still face challenges in achieving fast convergence.

**Effective Learning Methods with Ground Truth.** In the DETR family of object detection methods [3, 29, 58], [23] DN-DETR introduces query denoising(DN) to accelerate training by addressing the instability of bipartite graph matching. The DN queries are initialized by adding noise to both the ground truth bounding boxes and their associated labels. These noised queries are then fed into the Transformer decoder. The model is trained to reconstruct the original bounding boxes and labels, stabilizing the training process and improving convergence speed. In the HOI detection task, HQM [54] encodes shifted ground truth boxes as hard positive queries to guide training. However, HQM does not consider ground truth label information. DOQ [36] introduces an oracle query that implicitly encodes the ground truth boxes and object labels of human-




**(a)** Original annotation **(b)** Merged annotation

**Fig. 2:** Illustration of the problem of redundant human-Object pairs annotations in the HICO-DET [4] dataet for HOI detection. The left side displays the original annotations of the dataset, where human-object pairs are represented by red and green bounding boxes. In this example, the same human-object pair is annotated twice with different verbs, causing redundant annotations. The right side displays the result after merging these redundant annotations, combining the verb labels into a single instance for each human-object pair. This merging reduces redundancy and improves the accuracy of the model during denoising training by avoiding the learning of conflicting annotations.

object pairs, guiding the decoder to reconstruct the ground truth HOI instances during training. SOV-STG [5] encodes DN training queries from noised ground truth data to guide the reconstruction of the original ground truth. However, SOV-STG uses two decoders and inputs the DN training query to the human-object decoder. Since the verb decoder receives the output of the human-object decoder, the noise introduced to the verb decoder is not clearly defined. The unstable output of the human-object decoder during the early stages of training destabilizes the verb decoder's training. In addition, the unstable output of the human-object decoder during the early stages of training may destabilize the verb decoder's training. Furthermore, in the later stages of DN training process applied to the human-object decoder for verbs can interfere with the training of the verb recognition decoder. We propose a DN training method for the verb decoder that is not dependent on the results of the human-object decoder. Furthermore, we use annotation cleaning and added noise filtering to make the DN training more effective for HOI detection.

## 3 Method

Fig. 3 shows the training pipeline of our framework. In the normal training and inference phases, learnable anchor boxes and label queries are used as inputs to the human-object decoder, which is responsible for detecting human-object pairs. The embeddings and anchor pairs generated by the human-object decoder are subsequently fed into the verb decoder to predict verb classes. The human-object embeddings and anchor pairs output from the human-object decoder are then input to the verb decoder to predict verb classes. For denoising(DN) training, we utilize DN queries that are initialized from ground truth HOI instances with

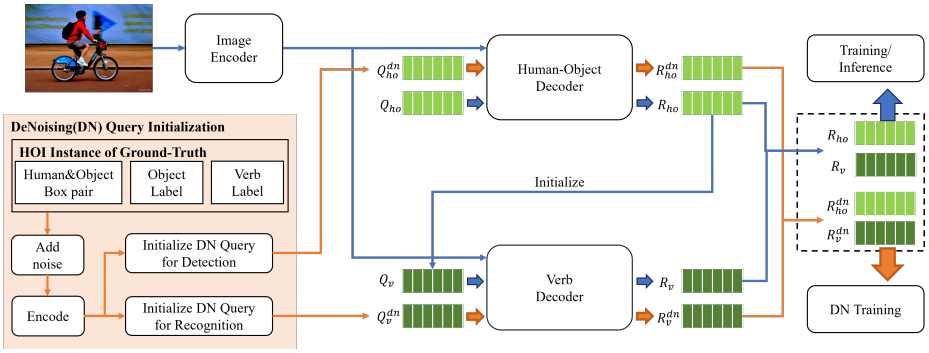



**Fig. 3:** The training pipeline of our proposal framework. During the inference and training phase, The human-object decoder and the verb decoder are connected in series to predict HOI. Queries $Q_{ho}$ are initialized learnable parameters and queries $Q_v$ are initialized Human-Object Decoder output. However, in the denoising training phase, DN queries $Q_{ho}^{dn}$ and $Q_v^{dn}$ are initialized and input in parallel to the human-object decoder and the verb decoder. This separation enables each decoder to focus on its specific task, improving training efficiency by reducing inference between decoders. By isolating the tasks of detection human-object pairs and recognizing verbs, the model effectively reconstruct ground truth HOI instances.

added noise. These DN queries are separately input into both the human-object decoder and the verb decoder, allowing for the reconstruction of the ground truth HOI instances. This separation of DN queries enhances learning efficiency by clearly defining the reconstruction targets for each decoder.

Sec. 3.1 provides a detailed description of our framework, while Sec. 3.2 explains the methods for redundant human-object pair annotations and filter and regenerate noised bounding boxes.

### 3.1 Separate Guided Denoising Training

In our training pipeline, we populate the human-object decoder and the verb decoder with their respective DN queries, thereby enhancing training efficiency by clearly defining the training targets for each decoder.

**Human-Object Decoder and Verb Decoder.** Our framework uses a image encoder extract global features, which are then input to a human-object decoder and a verb decoder. The image encoder leverages a hierarchical backbone and a deformable transformer encoder [58] to capture multi-scale global features $\boldsymbol{f}g \in \mathbb{R}^{Ng \times D}$, where $N_g$ represents the total number of pixels in the multi-scale feature map and $D$ denotes the hidden dimension of the embedding within the entire transducer architecture. For the decoding process, we utilize the deformable transformer decoder as proposed in [29], capable of handling label queries and anchor boxes. The human-object decoder uses a label query $\boldsymbol{Q}o \in \mathbb{R}^{Nq \times D}$ as its input query, which is initialized from learnable parameters. The output from the human-object decoder is then used to predict the object class, object bound-

ing box, and human bounding box. The output from the human-object decoder is then utilized to predict the object class, object bounding box, and human bounding box. Subsequently, the output is fed into the verb decoder to predict verb classes. During DN training, the human-object decoder and verb decoder receives a DN queries, which is generated by adding noise to the ground truth bounding boxes and verb labels.

**Label-specific Priors.** We initialize the object and verb label embeddings based on the SOV-STG [5] framework. Specifically, the object label embeddings $\boldsymbol{t}_o \in \mathbb{R}^{Co \times D}$ serve as the object label priors, and the verb label embeddings $\boldsymbol{t}v \in \mathbb{R}^{Cv \times D}$ serve as the verb label priors. The query embeddings for object labels $\boldsymbol{q}_o \in \mathbb{R}^{Nq \times D}$ are initialized through a linear combination of the object label embeddings $\boldsymbol{t}o$ and the object coefficient matrix $\boldsymbol{A}_o \in \mathbb{R}^{N_q} \times C_o$. Similarly, the query embeddings for verb labels $\boldsymbol{q}_v \in \mathbb{R}^{Nq \times D}$ are initialized through a linear combination of $\boldsymbol{t}_v$ and $\boldsymbol{A}_v \in \mathbb{R}^{N_q} \times C_v$. These label embeddings $\boldsymbol{t}_o$ and $\boldsymbol{t}_v$ are utilized in both the DN training and inference stages, and are trained jointly to enhance training efficiency.

**Separate Guided Denoising Training Strategy.** As shown in Fig. 3, DN training is conducted separately for the human-object decoder and the verb decoder. This separation clarifies the training targets for each decoder and enhances training efficiency. DN queries are generated for each decoder's input. The process of adding noise to the ground truth and creating DN label embeddings follows the SOV-STG framework. First, we explain the method of adding noise to the ground truth. Given a set of ground truth object labels $\boldsymbol{O}_{gt} = \boldsymbol{o}_i{}_{i=1}^{K}$ and a set of verb labels $\boldsymbol{V}_{gt} = \boldsymbol{v}_i{}^{K} i = 1$ for an image, where $\boldsymbol{o}_i$ and $\boldsymbol{v}_i$ are the object class and verb class labels, respectively, and $K$ is the number of ground truth HOI instances. For the $i$th ground truth HOI instance, the noised object label is obtained by randomly changing the ground truth index of the object class $\boldsymbol{o}_i$ to another object class index. Since the verb labels $\boldsymbol{v}_i$ consist of co-occurring ground truth classes, the indices other than the ground truth verb labels are randomly changed to preserve the co-occurring ground truth indices that appear in the noised verb labels. Two flipping rate hyperparameters $\eta_o \in (0, 1)$ and $\eta_v \in (0, 1)$ control the percentage of noised HOI instances for object and verb labels, respectively. In addition, the verb class flipping rate hyperparameter $\lambda_v \in (0, 1)$ controls the class-specific flipping rate of verb labels.

Next, we describe how to initialize DN label embeddings using noised object and verb labels. The DN query embeddings are initialized using the indices of the noised label and label embeddings $\boldsymbol{t}o$ and $\boldsymbol{t}v$. The DN query embedding of the object label $\boldsymbol{q}\tilde{o}^i$ is initialized based on the object label embeddings $\boldsymbol{t}o$, which correspond to the index of the noised object label. Similarly, the DN query embedding of the verb label $\boldsymbol{q}_{\tilde{v}}^i$ is initialized based on the sum of the verb label embeddings $\boldsymbol{t}_v$, which correspond to the indices of the noised verb labels.

Finally, we explain the method of initializing the DN queries for DN training. The detection DN queries $\boldsymbol{Q}_{ho}^{dn}$ used for training the human-object decoder are initialized from the DN query embeddings of the object labels $\boldsymbol{q}_{\tilde{o}}^i$ and the DN query embeddings of the verb labels $\boldsymbol{q}_{\tilde{v}}^i$. Here, the DN process is trained to

reconstruct the ground truth of the object while handling noised verb labels as anything. The detection DN queries $\boldsymbol{Q}_{ho}^{dn}$ used for training the human-object decoder is initialized from the DN query embeddings of object label $\boldsymbol{q}_{\hat{o}}^{i}$ and the DN query embeddings of verb labels $\boldsymbol{q}_{\hat{v}}^{i}$. Here, the DN is trained to reconstruct the ground truth of the object. The recognition DN queries $\boldsymbol{Q}_{v}^{dn}$ used for training the verb decoder are initialized from the DN query embeddings of the verb labels $\boldsymbol{q}_{\hat{v}}^{i}$ and the query embeddings of the ground truth object labels $\boldsymbol{q}_{ogt}^{i}$. The query embeddings of the ground truth object labels enable DN training of verb decoder given the ground truth object information. Thus, by inputting detection DN queries into the human-object decoder and recognition DN queries into the verb decoder, the DN training of each decoder can be clarified to reconstruct targets.

### 3.2 For effective denoising training of HOI detection

**Merging redundant human-object pair annotations.** In object detection, a single object in an image is typically assigned a single instance with a bounding box and class label. However, in HOI detection, a single human-object pair may have multiple bounding box pairs, as shown in **??**. During DN training, the data is trained using one-to-one matching to reconstruct ground truth from noised data. If multiple bounding box pairs exist for the same human-object pair and if the verbs are split, only the verbs associated with each bounding box pair are learned as positive, while the rest are learned as negative. To ensure all verbs associated with a human-object pair are recognized and prevent them from being incorrectly learned as negative, we propose a method to unify ground truth instances for the same human-object pair. If the object labels are the same and the minimum value of Intersection over Union (IoU) for a human-object bounding box pair is above the threshold and does not contain the same verb label, it can be considered a split verb annotation for the same human-object pair. The formula for $IoU$ is as follows.

$$IoU_{min}^{(i,j)} = min(IoU_{hum}^{(i,j)}, IoU_{obj}^{(i,j)}) \tag{1}$$

$$IoU_{hum}^{(i,j)}(\boldsymbol{B}_{hum}^{i}, \boldsymbol{B}_{hum}^{j}) = \frac{|\boldsymbol{B}_{hum}^{i} \cap \boldsymbol{B}_{hum}^{j}|}{|\boldsymbol{B}_{hum}^{i} \cup \boldsymbol{B}_{hum}^{j}|} \tag{2}$$

$$IoU_{obj}^{(i,j)}(\boldsymbol{B}_{obj}^{i}, \boldsymbol{B}_{obj}^{j}) = \frac{|\boldsymbol{B}_{obj}^{i} \cap \boldsymbol{B}_{obj}^{j}|}{|\boldsymbol{B}_{obj}^{i} \cup \boldsymbol{B}_{obj}^{j}|} \tag{3}$$

where $i, j$ denote the $i$th or $j$th HOI instance in the same image, and $\boldsymbol{B}_{hum}$ and $\boldsymbol{B}_{obj}$ represent the human and object bounding box. If there are split verb annotations for the same human-object pair, the human and object bounding boxes are averaged, and the set of verb labels is unified into a single instance. Eventually, the process is repeated for HOI instances in the image, ensuring there are no more split annotations for the same human-object pair.

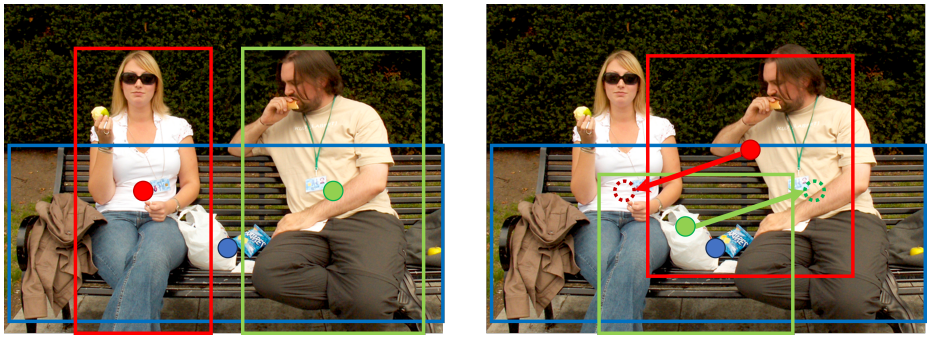


**Fig. 4:** The effect of noised bounding boxes in the HICO-DET [4] dataset. The left side shows the ground truth instances, while the right side shows the effect of adding noise. The red and green boxes represent humans, and the blue box represents the object (a bench). The dashed circle represents the original center, and the arrow represents the displacement required adjust the noised bounding box to match correct one. When a bounding box is used as an anchor, the model tends to detect the object closest to it. During denoising training, the ground truth and noised bounding boxes are learned through one-to-one matching. If the noised box is closer to non-target object than the actual target, the model learns incorrect predictions by focusing on the wrong object. To resolve this, we regenerate the noised bounding boxes to ensure that the target object is closest to the anchor, improving training efficiency.

**Filtering and regenerating noised bounding boxes.** In DN training for HOI detection, ground truth bounding boxes of human-object pair are reconstructed from noised bounding boxes. When object detection is based on anchors, it is common to detect objects that are close to the anchor. However, adding noise to the bounding box may bring it closer to objects other than the target object for reconstruction. For example, if the noised bounding box as shown in Fig. 4, is used as an anchor for DN training, it will learn to detect objects that are not the closest to the anchor. To address this issue, we introduce to filter out the noised bounding box and regenerate it.

We compute the $IoU$ between the ground truth bounding boxes and the noised bounding boxes. If the $IoU$ of a noised bounding box with any bounding box other than the target for reconstruction is the largest, regenerate it. We distinguish between the bounding boxes of humans and objects, and repeat this calculation only for the bounding boxes of humans and only for the objects. Continue regenerating the noised bounding boxes until all are closest to their respective reconstruction targets. In this way, DN training more effective.

### 3.3 Training and Inference

As shown in Fig. 3, our proposed method performs DN training simultaneously with normal training. For the inference queries $\boldsymbol{Q}_{ho}$ and $\boldsymbol{Q}_{v}$, the Hungarian algorithm is used to match ground truth HOI instances with predicted HOI instances, and the matching cost and the learning loss of predicted HOI instances

are similar to previous transformer-based methods [6]. For DN queries $\boldsymbol{Q}_{ho}^{dn}$ and $\boldsymbol{Q}_{v}^{dn}$, the ground truth index used in query initialization is used for matching with predictive HOI instances, and the loss function is the same as for inference queries.

## 4 Experiment

Based on the HICO-DET [4] and V-COCO [11] datasets, the proposed method was evaluated and compared with the previous method, STG. In addition, experiments were carried out on state-of-the-art method RLIPv2. Furthermore, an ablation study was conducted to analyse the contribution of each element and to demonstrate the effectiveness of the proposed method. Through these experiments, we were able to validate the improvements brought by our proposed approach.

### 4.1 Experimental Settings

**Dataset and Metrics.** The HICO-DET dataset contains 38,118 images for training and 9,658 images for testing. The 117 verb classes and 80 object classes in HICO-DET form a total of 600 HOI classes. Based on the number of HOI instances appearing in the dataset, the HOI classes in HICO-DET are classified into two categories, '*rare*' and '*non-rare*'. The V-COCO dataset contains 5,400 training images and 4,946 test images. 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. The mean Average Precision (mAP) scores are reported according to standard evaluations [4].
**Details of implementation.** We have applied and investigated the proposed method in the SOV-STG [5] and RLIPv2 [50] frameworks in order to develop an optimal approach for denoising(DN) training for HOI detection. All experiments were performed on 8 NVIDIA A40 GPUs.
**SOV-STG setups.** The SOV-STG framework comprises a human decoder, an object decoder, and a verb decoder, along with the STG DN training strategy. The weights of the image encoder, human decoder, and object decoder were initialized using the DAB-DeformableDETR model trained on the COCO dataset [28]. The human and object decoders were fed the same detection DN query, and the corresponding indices of the decoder outputs represented human-object pairs. The verb decoder, which combines the outputs of the human and object decoders using the SO-Attention module, was then used to predict verb classes. The feature image encoder consists of a ResNet-50 [13] backbone and a 6-layer deformable transformer encoder. The total number of backbones and decoders is based on the SOV-STG paper set-up, while ResNet-50 and 3-layer decoders were validated in SOV-STG-S. The hidden dimension of the transformer is $D = 256$ and the number of queries is $N_q = 64$. In the DN part, a $2N_p = 6$ group of noised labels is generated for each ground truth HOI instance. The dynamic DN scale is set to $\gamma = \frac{2}{3}$, the box noisification rate is set to

$\delta_b = 0.4$, the object label flipping rate to $\eta_o = 0.3$ and the verb noisification rate to $\eta_v = 0.6$, The maximum noisification level is defined by setting the flipping rate of verb labels to $\lambda_v = 0.6$. The model is trained by the AdamW optimiser with a learning rate of 2e-4 and weight decay of 1e-4. The backbone was fixed in the SOV validation to reduce training time. The batch size is set to 32, the training epochs are 30, and learning rate drops at the 20th epoch.

**RLIPv2 setups.** Since RLIPv2 does not use the label-specific priors of SOV-STG, we replaced the label embeddings used for initializing the DN queries with language features obtained from RLIPv2's Asymmetric Language-Image Fusion (ALIF). Without altering the content of the inference queries, we added new DN queries and fine-tuned the pre-trained model on the dataset. The basic setup was similar to that of SOV-STG. We verified this by finetuning the pre-trained models of RLIPv2-ParseDA ResNet-50 and Swin-Large [32] on each dataset. The batch size is set to 16, the training epochs are 20, and learning rate drops at the 15th epoch. The other setups was similar to that of SOV-STG setups.

### 4.2 Comparison to State-of-the-Arts

Tab. 1 presents a comparison of our proposed method with recent state-of-the-art (SOTA) methods on the HICO-DET dataset. Our method, when integrated with SOV, shows an improvement of 0.4 percentage points in mean Average Precision (mAP) over the experimental results of SOV-STG in the full category under default settings. Furthermore, when our method is applied to the pre-trained model of the SOTA method RLIPv2-ParSeDA and fine-tuned on the HICO-DET dataset, we achieve an improvement of 0.80 percentage points for the ResNet-50 (R50) model and 1.00 percentage points for the Swin-Large (Swin-L) model. Tab. 2 compares the results on the V-COCO dataset, demonstrating that our proposed method improves accuracy in both scenario 1 and scenario 2. Specifically, our method enhances the performance of RLIPv2-ParSeDA, leading to higher accuracy scores in both scenarios.

### 4.3 Ablation Study

**Contributions of proposed component.**

Tab. 3 shows the contributions of each proposed component using the HICO-DET dataset. The columns "Separate Guided" "Merge Annotations" and "Noise Filtering" indicate whether DN training is separated, merging redundant human-object pair annotations, filtering and regenerating noised bounding boxes, respectively. Row (1) represents the baseline result without the proposed method, using the SOV-STG R50 model. Both "Separate Guided" and "Merge Annotations" were effective on their own and improved accuracy. "Noise Filtering" needed to be combined with "Merge Annotations" to be effective. The combination of all elements resulted in the highest accuracy improvement.

**Contributions of merging redundant human-object annotations on HICO-DET.** In Tab. 4, the effect of merging redundant human-object annotations is investigated. During DN training, the data is trained using one-to-one

**Table 1: Comparisons with previous methods on HICO-DET.** R50 denote ResNet-50 [13]. Swin-L denote Swin-Large [32]. * denotes evaluation results using publicly available models or models we have trained, and unmarked denotes results from paper.

| Method | Backbone | Default Setting | | |
|---|---|---|---|---|
| | | *Full* | *Rare* | *Non-Rare* |
| CATN [8] | R50 | 31.86 | 25.15 | 33.84 |
| Liu et al. [30] | R50 | 33.51 | 30.30 | 34.46 |
| QAHOI [6] | R50 | 26.18 | 18.06 | 28.61 |
| QPIC [37] | R50 | 29.07 | 21.85 | 31.23 |
| CDN-S [51] | R50 | 31.44 | 27.39 | 33.53 |
| DOQ(CDN-S) [36] | R50 | 33.28 | 29.19 | 34.50 |
| GEN-VLKT-S [27] | R50 | 33.75 | 29.25 | 35.10 |
| DiffHOI-S [48] | R50 | 34.41 | 31.07 | 35.40 |
| CLIP4HOI [33] | R50 | 35.33 | 33.95 | 35.74 |
| LOGICHOI [24] | R50 | 35.47 | 32.03 | 36.22 |
| PViC w/DETR [53] | R50 | 34.69 | 32.14 | 35.45 |
| DiffHOI-L [48] | Swin-L | 41.50 | 39.96 | 41.96 |
| PViC w/$\mathcal{H}$-DETR [53] | Swin-L | 44.32 | 44.61 | 44.24 |
| SOV-STG* [5] | R50 | 33.19 | 29.39 | 34.32 |
| SOV+Ours* | R50 | 33.59 | 29.20 | 34.90 |
| RLIPv2-ParSeDA* [50] | R50 | 34.60 | 30.07 | 36.82 |
| RLIPv2-ParSeDA+Ours* | R50 | 35.40 | 31.43 | 37.36 |
| RLIPv2-ParSeDA* [50] | Swin-L | 45.12 | 45.33 | 44.70 |
| RLIPv2-ParSeDA+Ours* | Swin-L | **46.12** | **45.58** | **47.22** |

matching to reconstruct ground truth from noised data. If multiple bounding box pairs exist for the same human-object pair and the verbs are split, only the verbs associated with each bounding box pair are learned as positive, while the rest are learned as negative. Despite the same person-object pair annotations, it learns to recognise only different verbs for each bounding box pair. In merging annotations, if the minimum value of $IoU$ for a rectangular human-object pair with the same object labels is above a threshold and does not contain the same verb labels, it is considered a segmented annotation for the same human-object pair and is combined into a single annotation. merging threshold represents the threshold of $IoU$ and (1) represents the result without merging duplcate human-object annotations. The accuracy improved the most when $threshold = 0.8$. Otherwise, for example, when $threshold = 0.4$, the object class and verb class are looked at for pairs with $IoU$ greater than 0.4 and a decision is made whether to combine them into one. If the threshold is low, different human-object pairs are combined, which reduces accuracy. In HICO-DET, there is noise in the form of redundant human-object pairs in the dataset annotations. The proposed method reduces this noise and improves accuracy.

**Contributions of filter and regenerate noised bounding boxes on HICO-DET.** Tab. 5 examines the effects of noise filtering. As noised bounding boxes

**Table 2: Comparisons with previous methods on V-COCO.**

| Method | Backbone | $AP^{S1}_{role}$ | $AP^{S2}_{role}$ |
|---|---|---|---|
| RLIP-ParSe [49] | R50 | 61.9 | 64.2 |
| MSTR [20] | R50 | 62.0 | 65.2 |
| ParSe [49] | R50 | 62.5 | 64.8 |
| GEN-VLKT-M [27] | R101 | 63.3 | 65.6 |
| GEN-VLKT-L [27] | R101 | 63.6 | 65.9 |
| CDN-L [51] | R101 | 63.9 | 65.9 |
| SSRT [16] | R101 | 65.0 | 67.1 |
| SOV-STG* [5] | R50 | 63.1 | 64.6 |
| SOV+Ours* | R50 | 63.4 | 65.2 |
| RLIPv2-ParSeDA* [50] | R50 | 65.9 | 68.1 |
| RLIPv2-ParSeDA+Ours* | R50 | 66.4 | 68.5 |
| RLIPv2-ParSeDA* [50] | Swin-L | 72.0 | 74.1 |
| RLIPv2-ParSeDA+Ours* | Swin-L | **72.4** | **74.8** |

**Table 3:** Ablation studies for proposal component on HICO-DET.

| # | Separate Guided | Merging Annotations | Noise Filtering | Default Setting *Full* | *Rare* | *Non-Rare* |
|---|---|---|---|---|---|---|
| (1) | | | | 33.19 | 29.39 | 34.32 |
| (2) | ✓ | | | 33.48 | 29.23 | 34.75 |
| (3) | | ✓ | | 33.48 | **29.82** | 34.44 |
| (4) | | | ✓ | 33.34 | 29.61 | 34.46 |
| (7) | ✓ | ✓ | ✓ | **33.59** | 29.20 | **34.90** |

are used as anchors, undesired training may occur if they are close to a human or object other than the reconstruction target. Therefore, the noise-added bounding box is monitored, and the noise is regenerated when a human or object other than the reconstruction target is closest. Rows (1) and (3) show the results when noise filtering is not applied. Applying noise filtering improved accuracy. Additionally, in row (4), where the parameter that generates noise in the bounding box is increased, the improvement in accuracy is greater, verifying the effectiveness of noise filtering.

**Table 4:** Ablation studies for Merging Annotations on R50 model.

| Method | Merging threshold | Default Full mAP |
|---|---|---|
| (1) | 1.0 | 33.19 |
| (2) | 0.8 | **33.48** |
| (3) | 0.6 | 33.37 |
| (4) | 0.4 | 33.16 |

**Table 5:** Ablation studies for Noise Filtering on HICO-DET.

| Method | Noise Filtering | Noise Parameter$\delta$ | Default Full mAP |
|---|---|---|---|
| (1) | | 0.4 | 33.19 |
| (2) | ✓ | 0.4 | **33.34** |
| (3) | | 0.8 | 32.66 |
| (4) | ✓ | 0.8 | 33.21 |

## 5 Conclusion

This paper introduces a novel separate guided denoising (DN) training strategy for Human-Object Interaction (HOI) detection, where the human-object decoder and the verb decoder are trained independently. This approach allows for the application of explicit noise to each decoder, enhancing the effectiveness of DN training and demonstrating superior performance compared to previous methods. Additionally, our method includes merging redundant human-object annotations and filtering and regenerating noised bounding boxes, which further improve the efficiency of DN training for HOI detection. This strategy can be seamlessly integrated into DETR-based one-stage methods, incorporating both a human-object decoder and a verb decoder, thereby enhancing the performance of state-of-the-art models on relevant benchmarks.

## References

1. Cao, Y., Tang, Q., Su, X., Chen, S., You, S., Lu, X., Xu, C.: Detecting any human-object interaction relationship: Universal HOI detector with spatial prompt learning on foundation models. In: NeurIPS (2023)
2. Cao, Y., Tang, Q., Yang, F., Su, X., You, S., Lu, X., Xu, C.: Re-mine, learn and reason: Exploring the cross-modal semantic correlations for language-guided HOI detection. In: ICCV (2023)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229 (2020)
4. Chao, Y., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: WACV (2018)
5. Chen, J., Wang, Y., Yanai, K.: Focusing on what to decode and what to train: Efficient training with HOI split decoders and specific target guided denoising. CoRR **abs/2307.02291** (2023)
6. Chen, J., Yanai, K.: QAHOI: query-based anchors for human-object interaction detection. In: MVA (2023)
7. Chen, M., Liao, Y., Liu, S., Chen, Z., Wang, F., Qian, C.: Reformulating HOI detection as adaptive set prediction. In: CVPR (2021)
8. Dong, L., Li, Z., Xu, K., Zhang, Z., Yan, L., Zhong, S., Zou, X.: Category-aware transformer network for better human-object interaction detection. In: CVPR (2022)

9. Gao, C., Xu, J., Zou, Y., Huang, J.: DRG: dual relation graph for human-object interaction detection. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) ECCV (2020)
10. Gao, C., Zou, Y., Huang, J.: ican: Instance-centric attention network for human-object interaction detection. In: BMVC (2018)
11. Gupta, S., Malik, J.: Visual semantic role labeling. CoRR **abs/1505.04474** (2015)
12. Gupta, T., Schwing, A.G., Hoiem, D.: No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In: ICCV (2019)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
14. Hou, Z., Peng, X., Qiao, Y., Tao, D.: Visual compositional learning for human-object interaction detection. In: ECCV (2020)
15. Hou, Z., Yu, B., Qiao, Y., Peng, X., Tao, D.: Affordance transfer learning for human-object interaction detection. In: CVPR (2021)
16. Iftekhar, A.S.M., Chen, H., Kundu, K., Li, X., Tighe, J., Modolo, D.: What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions. In: CVPR (2022)
17. Johnson, J., Krishna, R., Stark, M., Li, L., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Image retrieval using scene graphs. In: CVPR (2015)
18. Kim, B., Choi, T., Kang, J., Kim, H.J.: Uniondet: Union-level detector towards real-time human-object interaction detection. In: ECCV (2020)
19. Kim, B., Lee, J., Kang, J., Kim, E., Kim, H.J.: HOTR: end-to-end human-object interaction detection with transformers. In: CVPR (2021)
20. Kim, B., Mun, J., On, K., Shin, M., Lee, J., Kim, E.: MSTR: multi-scale transformer for end-to-end human-object interaction detection. In: CVPR (2022)
21. Kim, D., Sun, X., Choi, J., Lin, S., Kweon, I.S.: Detecting human-object interactions with action co-occurrence priors. In: ECCV (2020)
22. Lei, T., Caba, F., Chen, Q., Jin, H., Peng, Y., Liu, Y.: Efficient adaptive human-object interaction detection with concept-guided memory. In: ICCV (2023)
23. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: DN-DETR: accelerate DETR training by introducing query denoising. In: CVPR. pp. 13609–13617 (2022)
24. Li, L., Wei, J., Wang, W., Yang, Y.: Neural-logic human-object interaction detection. In: NeurIPS (2023)
25. Li, Y., Liu, X., Wu, X., Huang, X., Xu, L., Lu, C.: Transferable interactiveness knowledge for human-object interaction detection. IEEE Trans. Pattern Anal. Mach. Intell. **44**(7) (2022)
26. Liao, Y., Liu, S., Wang, F., Chen, Y., Qian, C., Feng, J.: PPDM: parallel point detection and matching for real-time human-object interaction detection. In: CVPR (2020)
27. Liao, Y., Zhang, A., Lu, M., Wang, Y., Li, X., Liu, S.: GEN-VLKT: simplify association and enhance interaction understanding for HOI detection. In: CVPR (2022)
28. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: ECCV (2014)
29. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: DAB-DETR: dynamic anchor boxes are better queries for DETR. In: ICLR (2022)
30. Liu, X., Li, Y., Wu, X., Tai, Y., Lu, C., Tang, C.: Interactiveness field in human-object interactions. In: CVPR (2022)
31. Liu, Y., Yuan, J., Chen, C.W.: Consnet: Learning consistency graph for zero-shot human-object interaction detection. In: ACMMM (2020)

32. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
33. Mao, Y., Deng, J., Zhou, W., Li, L., Fang, Y., Li, H.: CLIP4HOI: towards adapting CLIP for practical zero-shot HOI detection. In: NeurIPS (2023)
34. Park, J., Park, J., Lee, J.: Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In: CVPR (2023)
35. Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.: Learning human-object interactions by graph parsing neural networks. In: ECCV (2018)
36. Qu, X., Ding, C., Li, X., Zhong, X., Tao, D.: Distillation using oracle queries for transformer-based human-object interaction detection. In: CVPR (2022)
37. Tamura, M., Ohashi, H., Yoshinaga, T.: QPIC: query-based pairwise human-object interaction detection with image-wide contextual information. In: CVPR. pp. 10410–10419 (2021)
38. Tsai, T.J., Stolcke, A., Slaney, M.: A study of multimodal addressee detection in human-human-computer interaction. IEEE Trans. Multim. **17**(9), 1550–1561 (2015)
39. Tu, D., Sun, W., Zhai, G., Shen, W.: Agglomerative transformer for human-object interaction detection. In: ICCV (2023)
40. Ulutan, O., Iftekhar, A.S.M., Manjunath, B.S.: Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In: CVPR (2020)
41. Wan, B., Zhou, D., Liu, Y., Li, R., He, X.: Pose-aware multi-level feature network for human object interaction detection. In: ICCV (2019)
42. Wang, H., Zheng, W., Ling, Y.: Contextual heterogeneous graph network for human-object interaction detection. In: ECCV (2020)
43. Wang, L., Zhao, X., Si, Y., Cao, L., Liu, Y.: Context-associative hierarchical memory model for human activity recognition and prediction. IEEE Trans. Multim. **19**(3), 646–659 (2017)
44. Wang, S., Yap, K., Ding, H., Wu, J., Yuan, J., Tan, Y.: Discovering human interactions with large-vocabulary objects via query and multi-scale detection. In: ICCV (2021)
45. Wang, S., Yap, K., Yuan, J., Tan, Y.: Discovering human interactions with novel objects via zero-shot learning. In: CVPR (2020)
46. Wang, T., Anwer, R.M., Khan, M.H., Khan, F.S., Pang, Y., Shao, L., Laaksonen, J.: Deep contextual attention for human-object interaction detection. In: ICCV (2019)
47. Wang, T., Yang, T., Danelljan, M., Khan, F.S., Zhang, X., Sun, J.: Learning human-object interaction detection using interaction points. In: CVPR (2020)
48. Yang, J., Li, B., Yang, F., Zeng, A., Zhang, L., Zhang, R.: Boosting human-object interaction detection with text-to-image diffusion model. CoRR **abs/2305.12252** (2023)
49. Yuan, H., Jiang, J., Albanie, S., Feng, T., Huang, Z., Ni, D., Tang, M.: RLIP: relational language-image pre-training for human-object interaction detection. In: NeurIPS (2022)
50. Yuan, H., Zhang, S., Wang, X., Albanie, S., Pan, Y., Feng, T., Jiang, J., Ni, D., Zhang, Y., Zhao, D.: Rlipv2: Fast scaling of relational language-image pre-training. In: ICCV. pp. 21592–21604 (2023)
51. Zhang, A., Liao, Y., Liu, S., Lu, M., Wang, Y., Gao, C., Li, X.: Mining the benefits of two-stage and one-stage HOI detection. In: NeurIPS (2021)
52. Zhang, F.Z., Campbell, D., Gould, S.: Spatially conditioned graphs for detecting human-object interactions. In: ICCV (2021)

53. Zhang, F.Z., Yuan, Y., Campbell, D., Zhong, Z., Gould, S.: Exploring predicate visual context in detecting of human-object interactions. In: ICCV (2023)
54. Zhong, X., Ding, C., Li, Z., Huang, S.: Towards hard-positive query mining for detr-based human-object interaction detection. In: ECCV (2022)
55. Zhong, X., Ding, C., Qu, X., Tao, D.: Polysemy deciphering network for robust human-object interaction detection. Int. J. Comput. Vis. **129**(6) (2021)
56. Zhou, D., Liu, Z., Wang, J., Wang, L., Hu, T., Ding, E., Wang, J.: Human-object interaction detection via disentangled transformer. In: CVPR (2022)
57. Zhou, P., Chi, M.: Relation parsing neural network for human-object interaction detection. In: ICCV (2019)
58. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: ICLR (2021)
59. Zhuang, Z., Qian, R., Xie, C., Liang, S.: Compositional learning in transformer-based human-object interaction detection. In: ICME (2023)
60. Zou, C., Wang, B., Hu, Y., Liu, J., Wu, Q., Zhao, Y., Li, B., Zhang, C., Zhang, C., Wei, Y., Sun, J.: End-to-end human object interaction detection with HOI transformer. In: CVPR. pp. 11825–11834 (2021)