# A. Appendix

In this Appendix, we provide **i)** extended quantitative analysis of MSTR capturing HOI detection in a multi-scale environment, **ii)** exploration for various possible decoder architectures, **iii)** implementation details of MSTR, **iv)** details on experimental datasets and metrics, **v)** details of training, **vi)** analysis on convergence speed, **vii)** additional qualitative result on our Dual-Entity attention and Entity-conditioned Context attention, and finally, **viii)** limitations of our work.

## A.1. Additional Quantitative Results for MSTR

First, we perform an extended quantitative analysis on the HICO-DET test set to validate the effectiveness of MSTR in a multi-scale environment. MSTR uses multi-scale feature maps to explore the semantics of HOI existing in different scales. In this section, we provide extensive quantitative results that shows the effectiveness of MSTR in capturing the interactions between humans and objects not only at different scales, but in various distances also (*e.g.*, *adjacent* interaction such as 'holding a book' or *remote* interaction such as 'throwing a frisbee'). To this end, we show quantitative results for multi-scale interactions according to 1) relative area of the human and the object, 2) the size of humans/objects, 3) distance between the human and object. For each criterion, we measure the performance across three bins where each bin has an equal and sufficient amount of HOI ground-truth labels to cover ($\sim$11,000 HOIs). For comparison, we set QPIC [7], the state-of-the-art transformer-based approach that uses a single-scale feature map, as our baseline. Note that in this appendix, the size, area, and distance are all calculated in *normalized* image coordinates.

**Relative area of human vs. object.** To observe how MSTR handles interaction between humans and objects with different scales, we first calculate the average precision (AP) over interaction labels that have different relative areas of humans and objects ($\frac{\text{area(hbox)}}{\text{area(obox)}}$). We cover three main cases according to their relative areas: i) $\text{AP}_{h<o}$ where the object area is significantly larger than the human area (*e.g.*, human *sitting* on a *bench*), ii) $\text{AP}_{h=o}$ where the human and the object exists in comparable sizes, and iii) $\text{AP}_{h>o}$ where the object area is significantly smaller than the human area (*e.g.*, human *throwing* a *ball*). We set the threshold for the relative areas so that each bin has an equal number of ground-truth instances (*i.e.*, $\frac{\text{area(hbox)}}{\text{area(obox)}} < 0.48$ for $\text{AP}_{h<o}$ and $\frac{\text{area(hbox)}}{\text{area(obox)}} > 4.33$ for $\text{AP}_{h>o}$). In Table 1, MSTR outperforms QPIC in all three types of interaction categories. Note that the improvement is more substantial in cases where the human and object have vastly different scales (+3.01p for $\text{AP}_{h<o}$ and +1.85p for $\text{AP}_{h>o}$), verifying that MSTR is effectively utilizing multi-scale feature maps.

| Method | $\text{AP}_{h<o}$ | $\text{AP}_{h=o}$ | $\text{AP}_{h>o}$ |
|---|---|---|---|
| QPIC | 34.10 | 30.57 | 25.22 |
| **MSTR** | **37.11** | **31.68** | **27.07** |
| $\Delta$AP | **+3.01** | +1.11 | **+1.85** |

Table 1. Comparison of MSTR with QPIC under interactions with different human/object scale ratio.

**Human & object size.** Here, we compare the average precision over the sizes of humans and objects. $\text{AP}_\text{L}$, $\text{AP}_\text{M}$, $\text{AP}_\text{S}$ each denotes the average precision for **L**arge, **M**iddle, and **S**mall humans and objects. In Table 2, MSTR outperforms QPIC in all three categories in both human and object scales. For the human scales, the improvement is more recognizable in interactions including small human areas (+3.06p in $\text{AP}_\text{S}$) while for object scales, the improvement is consistent over all three scales.

| | **Human Size** | | | **Object Size** | | |
|---|---|---|---|---|---|---|
| Method | $\text{AP}_\text{L}$ | $\text{AP}_\text{M}$ | $\text{AP}_\text{S}$ | $\text{AP}_\text{L}$ | $\text{AP}_\text{M}$ | $\text{AP}_\text{S}$ |
| QPIC | 28.65 | 35.36 | 24.14 | 33.09 | 28.65 | 24.87 |
| **MSTR** | **30.04** | **37.02** | **27.20** | **34.87** | **30.48** | **26.60** |
| $\Delta$AP | +1.39 | +1.66 | **+3.06** | +1.78 | +1.83 | +1.73 |

Table 2. Comparison of MSTR with QPIC under different sizes of humans and objects.

**Interactions in various distances.** Not only does MSTR capture interactions with various sized participants, but MSTR also captures interactions with various sized contexts, *i.e.*, interaction in various distances. To correctly measure how *remote* an interaction is, we note that the distance between center points [7] should be normalized by both the image size and the size of the human and object box participating in the interaction. Given the interaction between hbox $(hx_1, hy_1, hx_2, hy_2)$ and obox $(ox_1, oy_1, ox_2, oy_2)$, the normalized box area as area (hbox) and area (obox), we define the distance $d_\text{interaction}$ as

$$\begin{aligned} d_\text{center} &= \sqrt{\left(\tfrac{hx_1+hx_2}{2} - \tfrac{ox_1+ox_2}{2}\right)^2 + \left(\tfrac{hy_1+hy_2}{2} - \tfrac{oy_1+oy_2}{2}\right)^2}, \\ d_\text{interaction} &= d_\text{center} / \big(\text{area(hbox)} \cdot \text{area(obox)}\big). \end{aligned} \tag{1}$$

Then, we measure the average precision over three categories: i) $\text{AP}_\text{adjacent}$ where the human is interacting with a nearby object, ii) $\text{AP}_\text{distant}$ where the interacting human/object is within moderate distance, and $\text{AP}_\text{remote}$ where the human is interacting with an object sufficiently far away. As in previous sections, we set the distance threshold so that

| Method | $AP_{adjacent}$ | $AP_{distant}$ | $AP_{remote}$ |
|---|---|---|---|
| QPIC | 31.09 | 31.25 | 21.81 |
| **MSTR** | **32.66** | **33.48** | **23.70** |
| $\Delta$AP | +1.57 | **+2.23** | **+1.89** |

Table 3. Comparison of MSTR with QPIC under interactions with various distances.

each bin has an equal number of ground-truth instances. Table 3 shows the improvement of MSTR over QPIC. Note that while MSTR shows improvement across all three categories, the improvement is more distinguishable in cases where humans are interacting with objects in considerable distance (+2.23p for $AP_{distant}$ and +1.89p for $AP_{remote}$, respectively).

## A.2. Analysis on Decoder Architecture

As MSTR considers multiple semantics with two suggested deformable modules (Dual-Entity attention and Entity-conditioned Context attention), it is important to find a suitable decoder architecture that effectively merges the semantics. Here, we explore the possible combinations and various types of decoder architecture candidates when merging the three kinds of semantics. We empirically verify that MSTR architecture shows the most powerful performance.

**Architecture for Dual-Entity attention.** In Figure 1, we explore different architectures for Dual-Entity attention. We start with the most basic form: (a) is the architecture of QPIC, and (b) shows a straightforward application of the deformable attention [9] to QPIC. However, as we discussed in our main paper, (b) degrades the performance a lot from (a), because unlike its counterpart in object detection, multiple localizations need to be entangled to a single reference point in architecture (b). Therefore, we first use Dual-Entity attention to disentangle sampling points and attention weights for the participating entities (*i.e.*, human and object), respectively, to improve HOI detection performance. In Figure 1, (c) and (d) shows two options of dealing with the dual semantics obtained from dual reference points (each for humans and objects). In (c), each reference point is dealt with a separate stack of decoder layers (*i.e.*, Double-stream), while in (d) they are handled within a single-stream by sharing the self-attention layer where the input is simply the sum of the multiple semantics from the previous decoder layer. In Table 6, we show that our Dual-Entity attention shows a valid improvement (see (d) vs. (b)), while it even shows better performance than (c) requiring twice the number of decoder parameters.

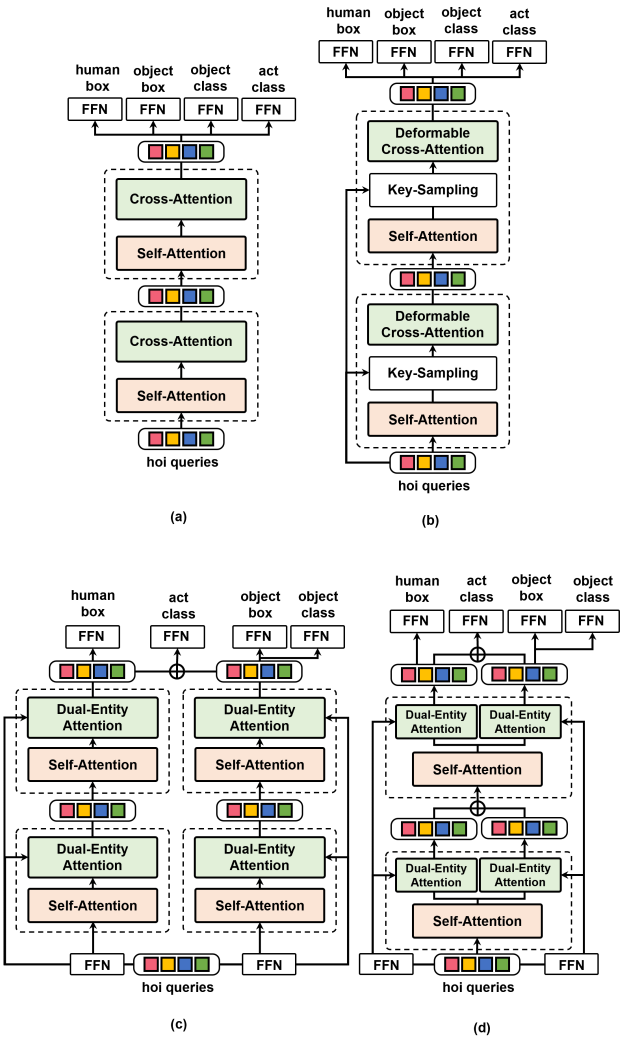


Figure 1. Comparison of a simple 2-layer Decoder architecture for: (a) QPIC, and (b) Direct application of Deformable DETR on QPIC, (c) Dual-Entity attention with two streams of decoder layers and (d) Dual-Entity attention that shares the self-attention layer.

| Method | Default (Full) |
|---|---|
| (a) QPIC | 29.07 |
| (b) QPIC + Deformable attention [9] | 27.52 |
| (c) Double-stream | 28.15 |
| **(d) Dual-Entity attention** | **28.30** |

Table 4. Comparison of Dual-Entity attention performance (d) against architecture in Figure 1 (a-c).

**Modeling Conditional Context attention.** In HOI detection, contextual information often gives an important clue in identifying interactions. In Table 5, we study the two different methods of obtaining context attention using (a) stan-

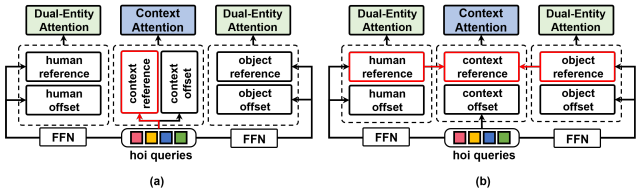


Figure 2. Comparison of: (a) context sampling with deformable attention, and (b) Entity-conditioned Context attention.

dard deformable attention and (b) our Entity-conditioned Context attention; note that in standard deformable attention, context reference points are directly obtained from HOI queries with a linear projection while our method conditionally obtain it from human and object reference points (see Figure 2). It can be observed that despite its simple structure and minimal delay, our Entity-conditioned Context attention achieves an +0.78p improvement compared to its counterpart. This implies that the guidance by human and object points is important to effectively model contextual information.

| Method | Default (Full) |
|---|---|
| (a) Standard Deformable attention | 29.36 |
| **(b) Entity-conditioned Context attention** | **30.14** |

Table 5. Comparison of the performance of Entity-conditioned Context attention against standard deformable attention [9]. Both (a) and (b) leverage Dual-Entity attention and follow the architectural design of Figure 3 (a) for fair comparison.

**Merging the semantics.** Figure 3 shows two different ways of how to merge the three semantics obtained from our Dual-Entity attention and Entity-conditioned Context attention. In MSTR, we merge the multiple semantics after applying self-attention separately to each of the semantic features obtained in the previous layer (Figure 3 (b)) instead of forcedly composing the input features of the self-attention layer (Figure 3 (a)). Table 6 shows that MSTR architecture (b) outperforms (a) by a margin of +1.03p, achieving the final performance. Note that while (b) is better, MSTR outperforms competing algorithms (presented in Table 2 of main paper) even with architecture (a).

| Method | Default (Full) |
|---|---|
| (a) Merge self-attention input | 30.14 |
| **(b) Merge self-attention output** | **31.17** |

Table 6. Comparison of a simple 2-layer Decoder architecture for Transformer-based HOI detectors: (a) Merging the input of the self-attention, and (b) architecture of MSTR (merging the output of self-attention).

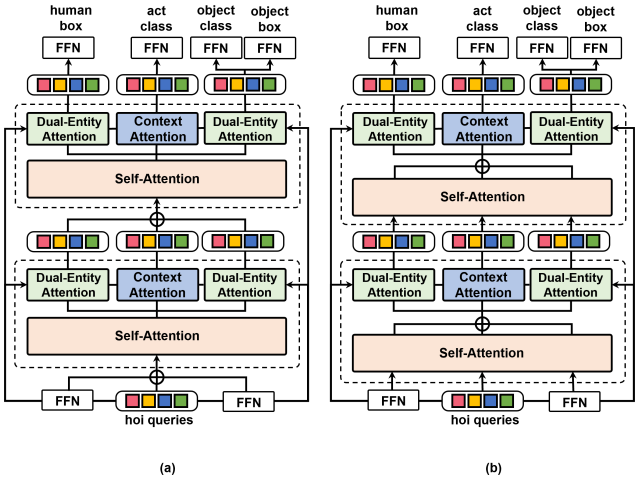


Figure 3. Comparison of a simple 2-layer Decoder architecture for Transformer-based HOI detectors: (a) Merging the input of the self-attention, and (b) architecture of MSTR (merging the output of self-attention).

## A.3. Implementation Details

Following implementation details in Deformable DETR [9], we use ImageNet pre-trained ResNet-50 [5] as our backbone CNN and extract multi-scale feature maps without FPN. The number of attention heads and sampling offsets for deformable attentions are set to $M = 8$ and $K = 4$, respectively. The AdamW optimizer is used with the initial learning rate of 2e-4 and weight decay of 1e-4. All transformer weights are initialized with weights pre-trained in MS-COCO. For a fair comparison with QPIC [7], we use only 100 HOI queries instead of using 300 ones as in Deformable DETR [9].

## A.4. Details on Datasets and Metrics

We evaluate our model on two widely-used public benchmarks: the V-COCO (*Verbs in COCO*) [4] and HICO-DET [2] datasets. V-COCO is a subset of COCO composed of 5,400 trainval images and 4,946 test images. For V-COCO dataset, we report the $\text{AP}_{\text{role}}$ over 25 interactions in two scenarios. In Scenario 1 (denoted as $\text{AP}^{\#1}_{\text{role}}$), detectors should predict an output indicating the non-existence of an object ([0,0,0,0]) when the target object is occluded, while in Scenario 2 (denoted as $\text{AP}^{\#2}_{\text{role}}$), only the localization of human and interaction classification is scored for such cases. HICO-DET contains 37,536 and 9,515 images for each training and test splits with annotations for 600 $\langle verb, object \rangle$ interaction types. In HICO-DET dataset, there are two different evaluation settings: *Default* and *Known object*. The former measures AP on all the test images, while the latter only considers the images with the object class corresponding to each AP. We report our score

with both settings. Note that the *Default* is a more challenging setting as we also need to distinguish background images. We follow the previous settings and report the mAP over three different category sets: (1) all 600 HOI categories in HICO (Full), (2) 138 HOI categories with less than 10 training instances (Rare), and (3) 462 HOI categories with 10 or more training instances (Non-Rare).

## A.5. Training Details of MSTR

In this section, we explain the details of MSTR training. MSTR follows a set prediction approach as in previous transformer-based HOI detectors [3, 6, 7, 10]. We first introduce the cost matrix of Hungarian Matching for unique matching between the ground-truth HOI triplets and HOI set predictions.

**Hungarian Matching for HOI Detection.** MSTR predicts a fixed number $K$ of HOI triplets that consist of a human box, object box, and binary classification for the $a$ types of actions (where $a$=25 in V-COCO and 117 for HICO-DET). Each prediction captures a unique $\langle$human,object$\rangle$ pair with multiple interactions. $K$ is set to be larger than the typical number of interacting pairs in an image (in our experiment, $K = 100$). Let $\mathcal{Y}$ denote the set of ground truth HOI triplets and $\hat{\mathcal{Y}} = \{\hat{y}_i\}_{i=1}^{K}$ as the set of $K$ predictions. As $K$ is larger than the number of unique interacting pairs in the image, we consider $\mathcal{Y}$ also as a set of size $K$ padded with $\varnothing$ (there are no ground-truth that matches the prediction). Let $y = (b^h, b^o, c^o, a)$ where the ground-truth interaction $y_i$ consists of $b_i^h$ and $b_i^o$ which denotes the normalized coordinates for the interacting human/object box, $c_i^o$ denotes the target object class. and $a_i$ denotes the one-hot for multiple actions. To find a bipartite matching between these two sets we search for a permutation of $K$ elements $\sigma \in \mathfrak{S}_K$ with the lowest cost:

$$\hat{\sigma} = \underset{\sigma \in \mathfrak{S}_K}{\operatorname{argmin}} \sum_{i}^{K} \mathcal{C}_{\text{match}}(y_i, \hat{y}_{\sigma(i)}), \tag{2}$$

where $\mathcal{C}_{\text{match}}$ is a pair-wise *matching cost* between ground truth $y_i$ and a prediction with index $\sigma(i)$. Now, the ground-truth is written as $y_i = (b_i^h, b_i^o, c_i^o, a_i)$ and the prediction is written as $\hat{y}_{\sigma(i)} = (\hat{b}^h_{\sigma(i)}, \hat{b}^o_{\sigma(i)}, \hat{c}^o_{\sigma(i)}, \hat{a}_{\sigma(i)})$ where $\hat{y}_{\sigma(i)}$ is the prediction that has the minimal matching cost with $y_i$. $\hat{b}^h_{\sigma(i)}$ and $\hat{b}^o_{\sigma(i)}$ are the normalized box coordinates for humans and objects, respectively, $\hat{c}^o_{\sigma(i)}$ is the classification for the target object of the interaction, and $\hat{a}_{\sigma(i)}$ is the predicted actions.

**Final Cost/Loss function for MSTR.** Based on $\mathcal{C}_{\text{match}}$, we calculate the final loss function for all pairs matched. The cost/loss function for the HOI triplets consists of the localization loss, object classification loss, and the action classification loss as $\mathcal{L}_{\text{H}} = \mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{act}}$ where each function is written as

$$\begin{aligned}
\mathcal{L}_{\text{loc}} &= \sum_{i=1}^{K} \left[\mathcal{L}_{\text{loc}}(b_i^h, \hat{b}^h_{\sigma(i)}) + \mathcal{L}_{\text{loc}}(b_i^o, \hat{b}^o_{\sigma(i)})\right], \\
\mathcal{L}_{\text{cls}} &= \sum_{i=1}^{K} \text{BCELoss}(c_i, \hat{c}_{\sigma(i)}), \\
\mathcal{L}_{\text{act}} &= \sum_{i=1}^{K} \text{BCELoss}(a_i, \hat{a}_{\sigma(i)}).
\end{aligned} \tag{3}$$

Identical to previous works [1, 3, 6, 7, 9, 10], the localization loss is defined by the weighted sum of the L1-loss and the gIoU loss.

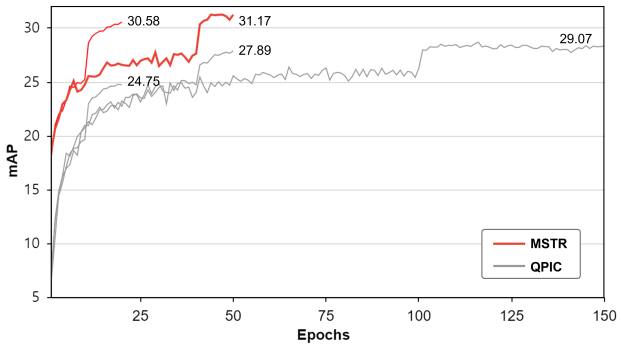


Figure 4. Comparison of convergence curves of QPIC and MSTR in the HICO-DET dataset. MSTR shows faster convergence than QPIC under various training schedules for both methods.

## A.6. Convergence speed

One of the advantages that deformable attention provides is the fast convergence at training. Figure 4 shows the convergence curve of MSTR compared to QPIC. Specifically, MSTR requires a much short number of epochs (50 epochs) compared to QPIC (150 epochs) to reach its best score. Note that MSTR achieves a competitive score to QPIC only with 20 epochs, outperforming QPIC with approximately $\times 4$ shorter training time.

## A.7. Qualitative Analysis for MSTR

In this section, we conduct extensive qualitative analysis of MSTR to observe how Dual-Entity attention and the Entity-conditioned Context attention capture different semantics for interactions in a multi-scale environment.

**MSTR attentions on multi-scale feature maps.** We conduct a qualitative analysis of MSTR on both Dual-Entity attention and the Entity-conditioned Context attention in HOI

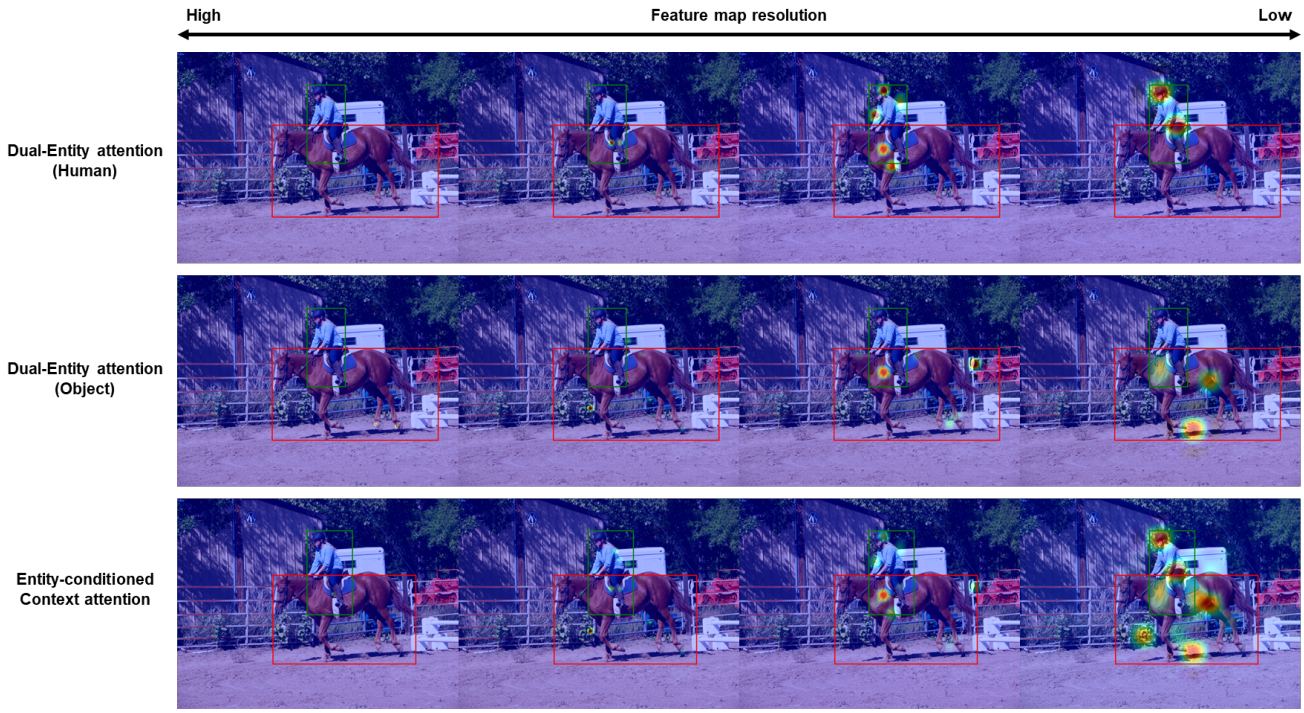


Figure 5. Visualization of the attention for the Dual-Entity attention and Entity-conditioned Context attention of MSTR in multi-scale feature maps for *adjacent* interaction: *ride*.

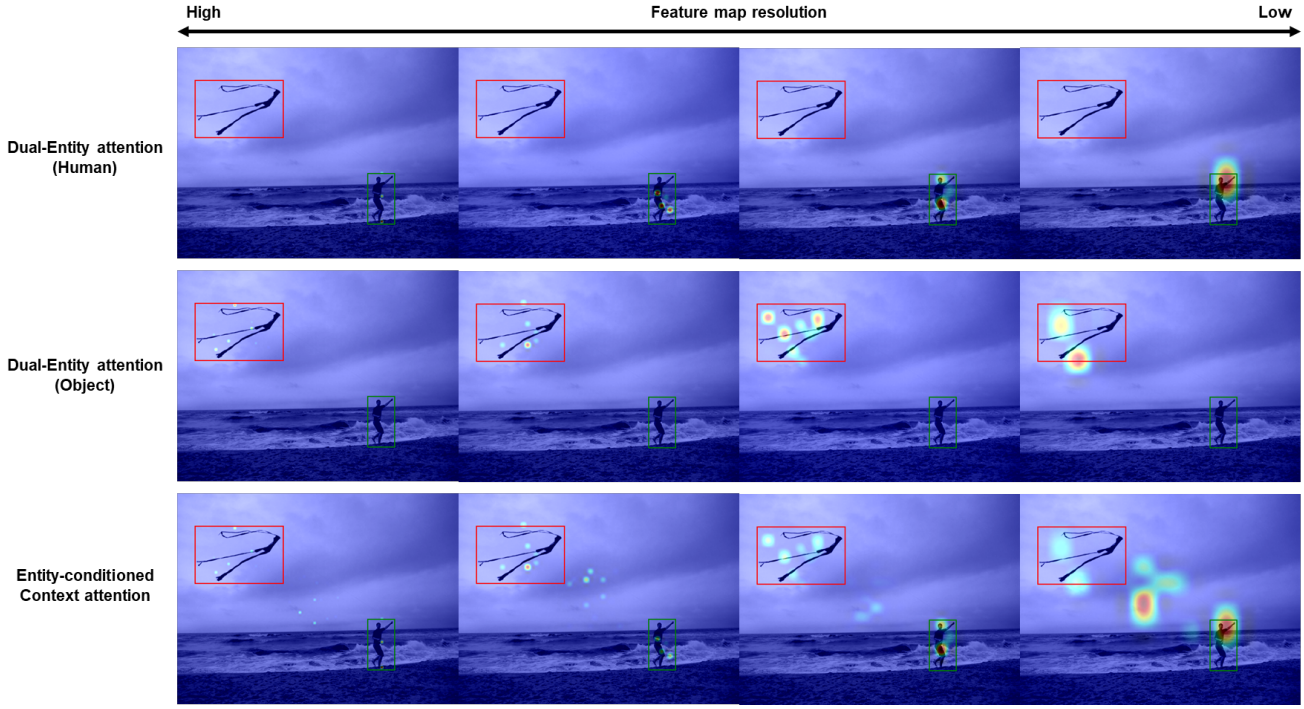


Figure 6. Visualization of the attention for the Dual-Entity attention and Entity-conditioned Context attention of MSTR in multi-scale feature maps for *remote* interaction: *fly*. It can be seen that in both *adjacent* interaction and *remote* interaction, MSTR successfully captures the multiple semantics of the human, object, and contextual information across the multi-resolution feature maps.

Figure 7. MSTR attentions (Dual-Entity attention and Entity-conditioned Context attention) of different scales all visualized at once.

detection to observe how MSTR captures interactions. Figure 5 shows the visualization of each attention in an *adjacent* interaction: *ride*. Figure 6 shows the visualization of each attention in an *remote* interaction: *fly*. For both cases, we can see that the Dual-Entity attention captures the appearance of the human and object across multiple scales of feature maps. In contrast, the Entity-conditioned Context attention tends to capture an inclusive area that covers both two regions and their intermediate background, effectively capturing the context of the interaction.

**MSTR attentions on multi-scale feature maps.** In Figure 7, we provide more qualitative visualizations for the multi-scale attentions of MSTR in various scenes with 1) large human and small object, 2) small human and large object, 3) distant interactions, 4) adjacent interactions.

## A.8. Limitations

The main limitation of our work is the bottleneck caused by the extensive size of the *query* element (multi-scale image features, there are about $\times 20$ more image tokens to process compared to the single-scale feature map). Despite our proposed deformable attentions, MSTR suffers from an estimated $10\%$ increase in parameters and $\sim \times 2$ GFLOPs compared to the single-scale baseline, QPIC [7]. Although recent related works have tackled the efficiency problem in deformable attentions by sampling the query element as well [8], the research scope of this work did not cover this issue.

## References


[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020. 4

[2] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 3

[3] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9004–9013, 2021. 4

[4] Jitendra Gupta, Saurabh Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 3

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[6] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 74–83, 2021. 4

[7] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021. 1, 3, 4, 6

[8] Tao Wang, Li Yuan, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. Pnp-detr: Towards efficient visual analysis with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4661–4670, 2021. 6

[9] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2, 3, 4

[10] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11825–11834, 2021. 4