What to Look at and Where: Semantic and Spatial Refined Transformer for Detecting Human-Object Interactions
Supplementary Material

A S M Iftekhar*†, Hao Chen*‡, Kaustav Kundu‡, Xinyu Li‡, Joseph Tighe‡, and Davide Modolo‡
‡AWS AI Labs; †University of California, Santa Barbara
{hxen,kaustavk,-,tighej,dmodolo}@amazon.com; iftekhar@ucsb.edu

In this supplementary material, we first analyze the per interaction category results of our SSRT model, then we provide additional qualitative results.

1. Per Interaction Category Results

In Table 1, we compare the per interaction category results of our SSRT (with ResNet-50 as the backbone) to QPIC [2] on V-COCO [1] dataset under the Scenario 1 setting. Results show that our SSRT improves the performance of all categories over the QPIC without any regression. Closely looking at the table, the top improvements are from: (1) read-obj (+16.04); (2) drink-instr (+11.43); (3) talk-on-phone-instr (+11.39); (4) cut-instr (+9.27); and (5) eat-obj (+7.81). Most of them are interactions with small objects (e.g., books, bottles, phones, scissors, knifes, sandwiches, etc.). While QPIC has low performance in detecting and predicting those categories, SSRT, with enriched semantic and spatial features and refined queries, is able to better focus on these small objects and corresponding interactions, hence improve the performance significantly. On the other hand, we also look into categories that are with smallest improvements. They are: (1) look-obj (+0.57); (2) lay-instr (+0.82); (3) skateboard-instr (+ 0.99); and (4) ride-instr (+1.06). These are interactions either with abstract concepts (e.g., look-obj), or with relatively large objects (e.g., beds, horses, motorcycles, skateboards). Though SSRT still improves performance on these categories, the improvement comparing to QPIC is not as significant as on categories with small and hardly visible objects. As a future work we will explore how to further improve the performance on these categories.

Interaction Category	QPIC	SSRT
hold-obj:	50.61	**55.45**
sit-instr:	51.98	**56.14**
ride-instr:	67.37	**68.43**
look-obj:	47.30	**47.87**
hit-instr:	74.66	**79.02**
hit-obj:	66.52	**72.22**
eat-obj:	58.80	**66.61**
eat-instr:	72.64	**76.06**
jump-instr:	77.81	**80.41**
lay-instr:	54.62	**55.44**
talk-on-phone-instr:	40.26	**51.65**
carry-obj:	41.45	**44.53**
throw-obj:	53.21	**54.78**
catch-obj:	54.47	**57.52**
cut-instr:	38.07	**47.34**
cut-obj:	58.15	**63.78**
work-on-computer-instr:	68.18	**73.05**
ski-instr:	49.31	**52.59**
surf-instr:	70.4	**75.25**
skateboard-instr:	84.43	**85.42**
drink-instr:	44.26	**55.69**
kick-obj:	81.93	**84.14**
read-obj:	35.76	**51.80**
snowboard-instr:	68.68	**74.27**
mAP	58.79	**63.73**

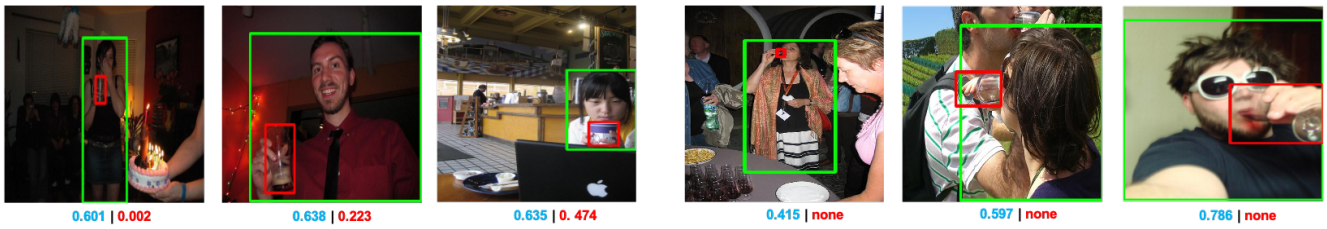
Table 1. Our network's category-wise performance compare to QPIC [2]. The best performance in each row are marked with **bold**. Here, "instr" means instrument and "obj" means object [1].

2. Additional Qualitative Results

Fig. 1 and Fig. 2 include additional qualitative results of SSRT (with ResNet-50 as the backbone) and the comparison of it with QPIC. Specifically, in Fig. 1, we show samples from the top-5 interaction categories with the largest performance improvement from SSRT, as introduced in Sec. 1. We observe that most of samples from these top-5 categories are persons interacting with small or hardly visible objects in some complex scenes. As mentioned in the main paper, SSRT improves over QPIC mainly in two types: (1) increasing the confidence scores of the action predictions - which is

*Equal contribution.
†Work done during an internship at Amazon.

(a) Read Object.



(b) Drink Instrument.

(c) Talk on Phone Instrument.

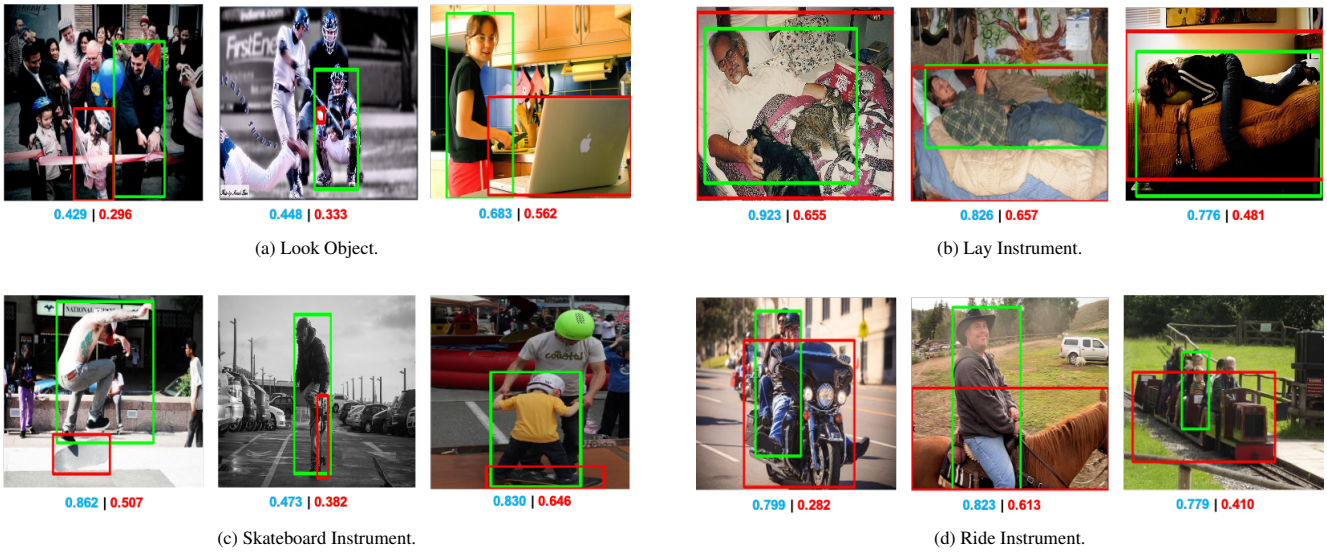
(d) Cut Instrument.

(e) Eat Object.

Figure 1. Qualitative results for top-5 interaction categories with the biggest improvements from SSRT compared to QPIC. For each image, the predicting score of SSRT is marked in blue while the score of QPIC is marked in red. If no matched bounding box pairs are detected then the result is marked as none.

shown in the first three samples of each row in Fig. 1; and (2) successfully detecting the person, object and actions that are completely missed (no bounding box output matches with GT) in QPIC - which is shown in the last three samples of each row in Fig. 1).

Fig. 2 shows samples from the interaction categories that



(a) Look Object.

(b) Lay Instrument.

(c) Skateboard Instrument.

(d) Ride Instrument.

Figure 2. Qualitative results for interaction categories with the least improvements from SSRT compared to QPIC. For each image, the predicting score of SSRT is marked in blue while the score of QPIC is marked in red.

are with least improvement from SSRT, as introduced in Sec. 1. Note that, though the improvements on these categories are not as big as the top-5 categories, SSRT still performs better than QPIC on all of these categories. We notice that samples from these classes are either with abstract interaction concepts, or contain interactions with relatively larger objects comparing to those in Fig. 1.

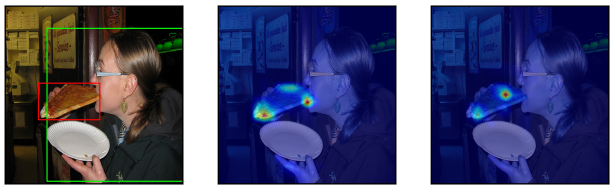
Fig. 3 shows some more visualizations of the attention maps. Recall that the attention map is of the query that predicts the marked person and object bounding boxes, generated from the last layer of the decoder. For samples in Fig. 3a - Fig. 3d, both QPIC and SSRT can localize the persons and the objects, but QPIC fails to predict the actions with high confidences while SSRT does. The attention maps clearly show that attentions from SSRT are more refined and focused on the area of the interaction, while attentions from QPIC are on the roughly correct regions but very coarse and noisy. For samples in Fig. 3e - Fig. 3h, QPIC completely fails in even detecting the object and interaction locations, while SSRT is able to detect to the right area. Most of these samples are persons interacting with small or hardly visible objects in some complex scenes, indicating that SSRT is especially better at handling such scenarios than QPIC.

References

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

[2] 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

(a) Hold a tennis racket. Scores: SSRT: 0.856 | QPIC: 0.001.



(b) Eat a pizza. Scores: SSRT: 0.761 | QPIC: 0.512.

(c) Carry a backpack. Scores: SSRT: 0.588 | QPIC:0.001.

(d) Talk on the phone. Scores: SSRT: 0.880 | QPIC: 0.001.

(e) Talk on the phone. Scores: SSRT: 0.612 | QPIC: none.

(f) Cut with knife. Scores: SSRT: 0.757 | QPIC: none.

(g) Work on the computer. Scores: SSRT: 0.761 | QPIC: none.

(h) Play a surfboard. Scores: SSRT: 0.873 | QPIC: none.

Figure 3. Visualization of the attention. We extract the attention map from the last layer of the decoder. In each sub-figure, from the left to the right are (1) the original image with the ground truth; (2) the attention map of our SSRT, and (3) the attention map of QPIC.