A. Our New Tesetsets

A.1. OOD Image Testset

OOD (Our-Of-Distribution) image testset is a testset we built to assess the HOI detection models on images which are different from the training images. We compare the distribution of visual features between 100 images from HICO-DET testset and those from our OOD image test using t-SNE and PCA. The formation of different clusters in Fig. 8 and Fig. 9 implies the two image sets are of different distribution.

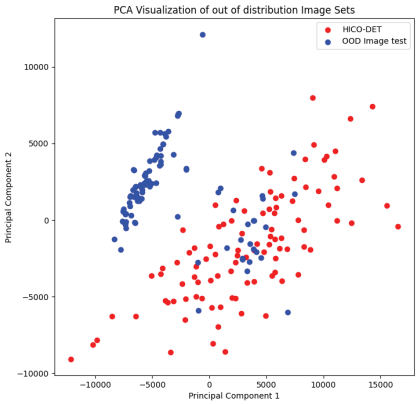

Figure 8. Visualization of the visual feature distributions using PCA. Red dots and blue dots represent the visual features of the images from HICO-DET and our OOD image testset, respectively.

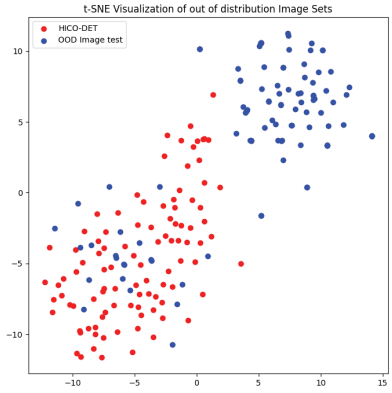

Figure 9. Visualization of the visual feature distributions using t-SNE. Red dots and blue dots represent the visual features of the images from HICO-DET and our OOD image testset, respectively.

1. table	22. basket	43. yogamat
2. hose	23. paint brush	44. jump rope
3. gas pipe	24. headphone	45. golf cart
4. tablet	25. helmet	46. pillow
5. pogo stick	26. mask	47. trampoline
6. scooter	27. ladder	48. stick
7. panda	28. zipline	49. bean bag
8. drone	29. bucket	50. water can
9. canvas	30. trash bag	51. piano
10. lemon	31. sunglasses	52. tree
11. gascan	32. broom	53. helicopter
12. glove	33. sofa	54. stool
13. hat	34. cap	55. wall
14. barbell	35. flowers	56. wagon
15. tiger	36. ice cream	57. headphones
16. television camera	37. playground slide	58. sun lounger
17. groceries bag	38. hammock	59. perfume
18. cart	39. computer monitor	60. towel
19. lunch box	40. punching bag	61. milk glass
20. envelope	41. vending machine	62. lollipop
21. glider	42. money	

Table 5. Novel objects

A.2. Novel Object Testset

Novel object testset is a testset we built to assess the model's ability to generalize to novel objects that do not appear in the HICO-DET. The table 5 shows a list of 62 novel objects we added in this testset.

B. Impact of CLIP Image Encoder

We investigate the impact of CLIP image encoder by employing the different OWLv2 for our instance detection part. We test a larger variant *owlv2-large-patch14-ensemble* which uses a CLIP backbone with a ViT-L/14, in addition to our default variant *owlv2-base-patch16-ensemble* which employs ViT-B/16. Results on both default and zero-shot settings are shown in Table 6 and 7. The results demonstrate that our model with the larger CLIP outperforms our default model on all conditions especially on Unseen condition. At the same time, model size is significantly increased by approximately three times. This larger model does not fit in our GPU used to train the default model (a single NVIDIA A30 with 24GB), thus we use eight V100 with 16GB to train the larger model.

C. Analysis of Instance Detection

To assess the impact of object detection performance on human-object interaction (HOI) detection, we compared our method's object detector against HOICLIP, a model equipped with a DETR-pretrained object detector. We cal-

Methods	CLIP	Model size	Default			Known Object		
			Full	Rare	Non-Rare	Full	Rare	Non-Rare
OWLv2-Base (Default)	ViT-B/16	108M	36.89	32.98	38.06	40.09	36.56	41.15
OWLv2-Large	ViT-L/14	326M	**39.64**	**36.29**	**40.64**	**42.31**	**39.15**	**43.25**

Table 6. HICO-DET Default performances with different CLIP variants.

Methods	RF-UC			NF-UC			UO			UV		
	Unseen	Seen	Full	Unseen	Seen	Full	Unseen	Seen	Full	Unseen	Seen	Full
OWLv2-Base (Default)	25.16	38.16	35.17	28.54	29.30	29.12	26.21	31.50	30.28	30.81	35.41	34.35
OWLv2-Large	**29.85**	**39.70**	**37.43**	**32.15**	**31.07**	**31.32**	**27.96**	**33.28**	**32.06**	**34.84**	**36.87**	**36.41**

Table 7. HICO-DET zero-shot performances with different CLIP variants.

culated the mean average recall (mAR) for all 128 objects (64 subjects and 64 objects) generated by both methods. Our method achieved an mAR of 39.63 under default setting, while HOICLIP attained an mAR of 38.24. Similarly, our mAR under UO setting (33.95) outperforms the one of HOICLIP (31.75). Our object detector performed marginally better compared to HOICLIP. On the other hand, with regard to mAP, our method is inferior to HOICLIP. This can be attributed to the more generalized nature of the OWLv2 object detector used in our method, which detects a wide variety of objects than the HICO-DET-fine-tuned object detector in HOICLIP.

D. Analysis of HO Pair Decoder

We analyze the behavior of the HO pair decoder (Sec. 3.3), which is a key component in our method. We shows a visualization of an attention map of the last cross-attention layer in HO pair decoder, in Fig. 10. The number of queries for human and objects are 64 for each. The number of instances selected in the instance selector (Sec. 3.2) is 100. Fig. 10 shows that each query selectively attends a specific instance. In the case shown in Fig. 10, all the human queries focus on three instances, i.e., they can be a subject. On the other hand, the attention of the object queries is spread out more widely. This is because the same person can interact with multiple objects.

E. Effect of HO Pair Decoder

To verify the importance of HO pair decoder and the effect of changing the number of layers l in it, we test the performance with the different number of its Transformer layers under RF-UC setting on HICO-DET. The results shown in Table 8 demonstrate that the performance gradually decreases when the number of layers is reduced. We select $l = 3$ as our default. This result also shows the importance of the HO pair decoder for HOI detection task.

# of layers	RF-UC		
	Full	Unseen Object	Seen Object
1	31.56	24.21	33.39
2	33.80	24.44	36.14
3	**35.15**	**24.51**	**37.81**

Table 8. Effect of the number of Transformer layers in HO pair decoder under RF-UC setting.

Thresholds	RF-UC		
	Full	Unseen Object	Seen Object
0	**35.44**	**25.36**	**37.96**
0.5	35.15	24.51	37.81
1.0	33.61	23.89	36.04

Table 9. Effect of a threshold in instance selector under RF-UC setting. With a lower threshold, the filtered instance features include more noisy instances.

F. Threshold for Instance Selection

Instance selector is a key component of our model, and the instance selector filters the input visual features with a threshold T to select instances. Table 9 shows the performances with different thresholds. With a lower threshold, the filtered instance features include more noisy instances. $T = 0$ means there is no filter and M features are selected in descending order of similarity. We can see a large performance gap between 0.5 and 1.0. We select $T = 0.5$ because of the small performance gap between 0.5 and 0.

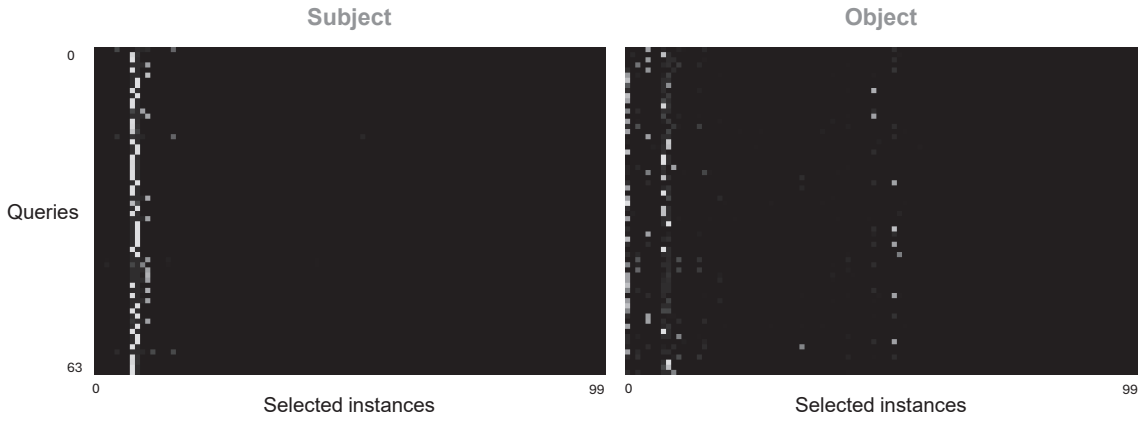

Figure 10. Visualization of an attention map of the last cross-attention layer in HO pair decoder.

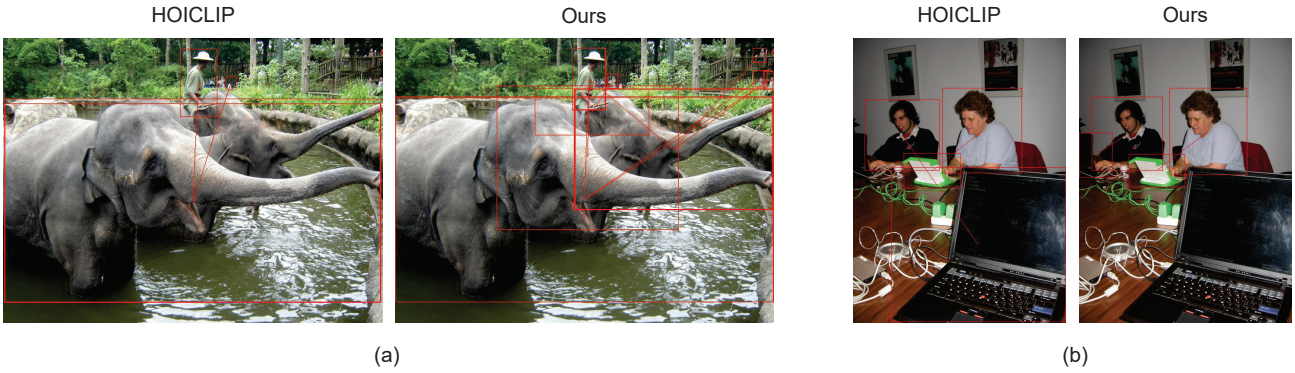

Figure 11. Qualitative comparisons between HOICLIP (left) and ours (right).

G. Qualitative Analysis

We conducted some qualitative analysis. Figure 11 shows comparisons of HOI detection results of HOICLIP (left) and our method (right). In Fig. 11 (a), we visualize the detected HOIs with the same threshold. Our method detects more HOIs with high confidence. In Fig. 11 (b), HOICLIP failed to infer more noteworthy relationships (i.e., a man on the left and a computer in front of him), while our method successfully detects the relationships between the man and the computer that is harder to detect than the computer in the foreground in the image.

We also show some failure cases in Fig. 12. In Fig. 12 (a), two men are each handling different suitcases. However, our method predicted that both men are handling the same suitcase, possibly due to the lack of depth perception. In Fig. 12 (b), although our method detected all the umbrella, it did not recognize the subjects correctly. In this case, common sense may help the model to understand since the image is very ambiguous and the persons are difficult to detect.

(a) (b)

Figure 12. Examples of failure cases.