HORP: Human-Object Relation Priors Guided HOI Detection

Supplementary Material

1. Additional Ablations

In this supplement, we provide additional experimental and qualitative results to support our approach.

Impact of Decoder Layers. To evaluate the relationship between the number of Interaction Decoder layers and model performance, we conduct ablation experiments on both datasets and keep other modules and parameters constant. The experimental results are shown in Table 1. It can be observed that, although changing the number of decoder layers does not significantly impact performance, using fewer layers tends to achieve better efficiency. Specifically, the proposed HORP model achieves its best performance with two interaction decoder layers.

Table 1. Ablation study on the number of decoder layers on the HICO-DET and V-COCO datasets.

# layer	HICO-DET(Default)			V-COCO	
	Full	Rare	Non-rare	S_1	S_2
1	38.19	35.85	38.89	68.5	70.7
2	**38.61**	**36.14**	**39.34**	**68.9**	**71.1**
3	38.39	35.76	39.18	68.7	70.9
4	38.27	35.44	39.12	68.6	70.6

Impact of Different Decoder Component. The interaction decoder infers interaction relationships to predict different interaction categories. We conduct an ablation study of components in the interaction decoder. Concretely, we remove the self-attention, cross-attention, and FFN modules from the decoder to analyze the impact of each component on the recognition results, respectively. In this study, the keys/values in the decoder are derived from the C5 features in ResNet. The results are listed in Table 2. It can be seen that removing any of the components leads to a performance degradation, which indicates the modules are essential in capturing human-object interaction cues. The best performance is obtained by using the features in the backbone compared to the feature sources in the encoder.

Table 2. Ablation study of the interaction decoder components under the HICO-DET Default Setting.

#	Decoder				Deafult Setting		
	Self	Cross	FFN	C.A.src.	Full	Rare	Non-rare
B1				None	34.47	32.25	35.13
B2			✓	None	35.92	33.56	36.71
B3	✓		✓	None	36.58	34.71	37.15
B4	✓	✓	✓	Encoder	37.14	35.62	37.59
B5	✓	✓	✓	Backbone	**38.61**	**36.14**	**39.34**

The influence of the priors on other interaction types. We show improvements for several other interaction types in the Table 3, indicating that our model benefits various interactions by better distinguishing them.

Table 3. Results under the HICO-DET default setting.

Method	hold	carry	type_on	watch	inspect	blow
baseline [46]	0.79	0.69	0.60	0.30	0.21	0.56
ours	**0.91**	**0.84**	**0.77**	**0.51**	**0.49**	**0.69**

Efficiency Analysis. Comparisons in terms of model size, inference time. Our method runs at a comparable speed to baseline (11 vs. 14 FPS), as shown in Table 4. The slight slowdown is due to the additional gaze estimator, but we will accelerate it by integrating the gaze model in future work.

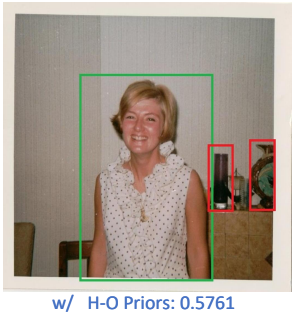
Table 4. Results under the HICO-DET default setting.

Method	Backbone	Params(M)	FPS
baseline [46]	R-50	53.50	14
ours	R-50	137.70	11

2. Additional Qualitative Results

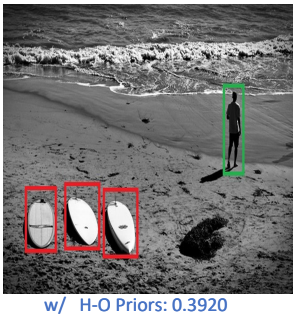
Improvement in Triplet Performance. We provide additional visualization results, as shown in Figure 1. We present the improvement in the AP of the triplet categories. In Figure 1(b), after using human-object priors, the performance of the triplet <person, no-interaction, surfboard> improved by approximately 0.20. The proposed HORP effectively enhances the accuracy of the no-interaction category.

Visualization of Confusion Changes. We provide more qualitative results for the no-interaction category in Figures 2 and 3. As can be seen, our HORP effectively distinguishes the confusion between no-interaction and with-interaction categories.



w/ H-O Priors: 0.5761
w/o H-O Priors: 0.4081

(a) person no-interaction vase



w/ H-O Priors: 0.3920
w/o H-O Priors: 0.1964

(b) person no-interaction surfboard

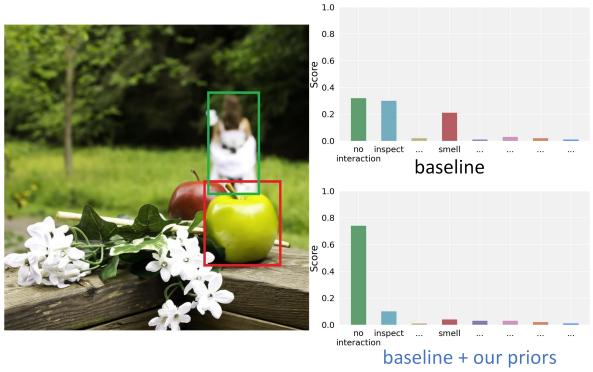
w/ H-O Priors: 0.6244
w/o H-O Priors: 0.3075

(c) person no-interaction cake

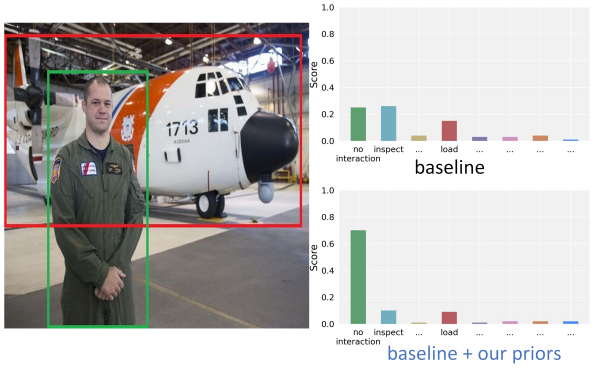
w/ H-O Priors: 0.3514
w/o H-O Priors: 0.1466

(d) person no-interaction stop_sign

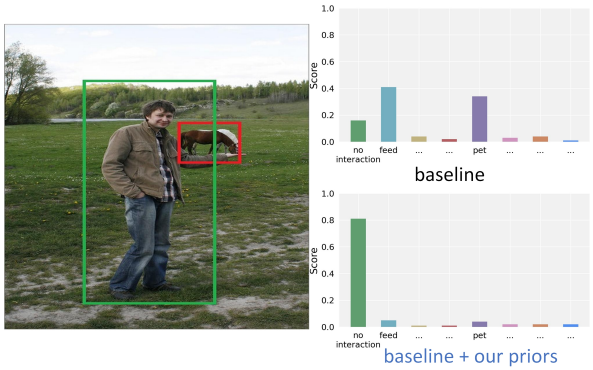
Figure 1. The performance improvement of triplet categories. The black font represents the results without integrating human-object priors, and the blue font indicates the results after incorporating human-object priors.



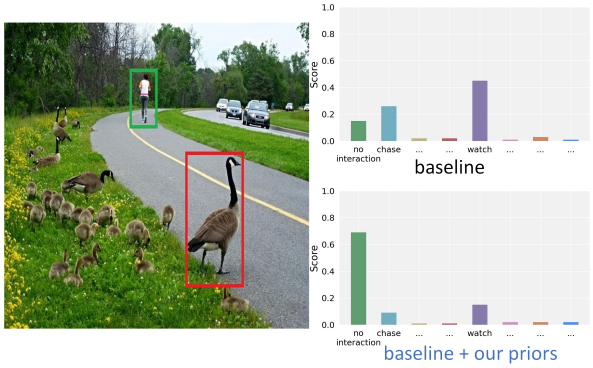
(a) person no-interaction apple



(b) person no-interaction airplane

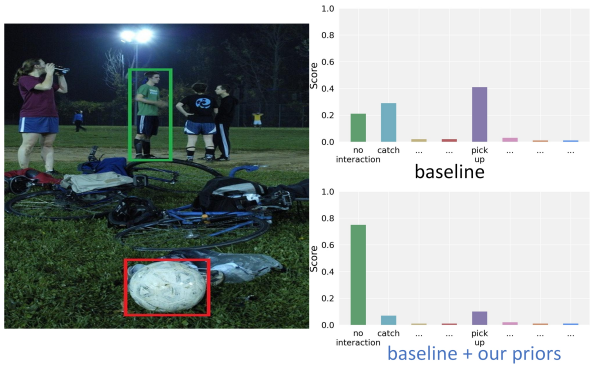


(c) person no-interaction horse

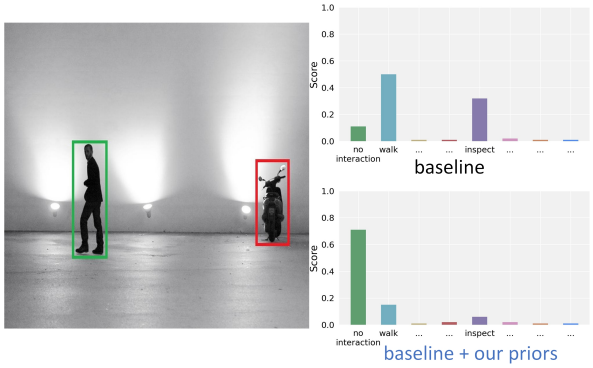


(d) person no-interaction bird

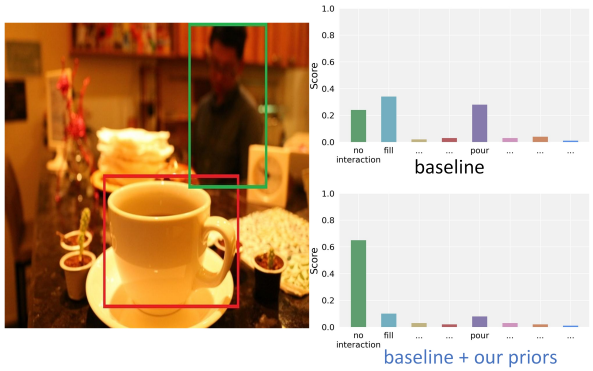
Figure 2. Qualitative results on the HICO-DET test set.



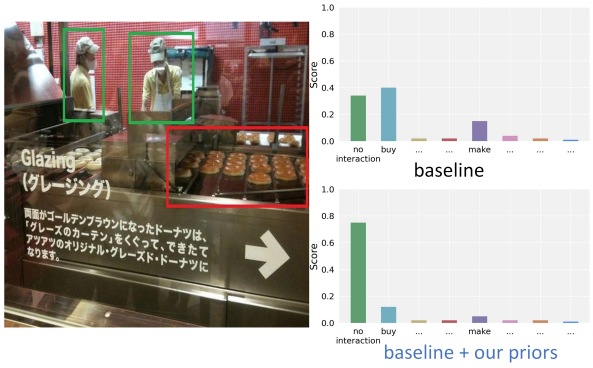
(a) person no-interaction sports_ball



(b) person no-interaction motorcycle



(c) person no-interaction cup



(d) person no-interaction donut

Figure 3. Qualitative results on the HICO-DET test set.