

Supplementary Materials: A Plug-and-Play Method for Rare Human-Object Interactions Detection by Bridging Domain Gap

Anonymous Authors

1 OVERVIEW

In this supplemental file, we provide more details of our work to supply the main paper. Specifically,

- **Generated images** are shown in Sec. 2.
- **Additional ablation studies** are presented in Sec. 3, which includes the ablation studies on the mask ratio, the number of prototype nodes and pseudo label threshold;
- **Additional qualitative results** are presented in Sec. 4.
- **Potential limitations and future developments** are discussed in Sec. 5.

2 GENERATED IMAGES

In our paper, we employ blip-diffusion to generate rare class HOI images. Specifically, for a given HOI triplet <Human, Verb, Object>, we follow the template “a photo of human verbing object” and expand triplet into a sentence as prompt. Then the blip-diffusion model receives prompt and the original image as input to generate the image. As shown in Fig. 1, we visualize some generated images, where the first row represents images from the original dataset, and the subsequent four rows depict the images generated by blip-diffusion.

3 ADDITIONAL ABLATIONS

In this section, We conducted some additional ablation studies on HICO-Det to further investigate the model details.

3.1 The Mask Ratio

As shown in Table 1, we studied the impact of different mask ratios σ in the context enhancement module. We observe a gradual improvement in the mAP for rare classes as the mask ratio increased. The main reason is that the reconstruction of masked patches enhances the model’s ability to learn import context clues. While excessively high reconstruction rate will cause the model to focus too much on the reconstruction task, weakening the effect of feature alignment, resulting in weakened performance. So we can observe that When the mask ratio reaches 0.9, the performance of rare classes decreases by 0.29 mAP compared to the mask ratio of 0.8. To strike a balance between feature alignment and context clue extraction, we select for a mask ratio of 0.8.

3.2 The Pseudo Label Threshold

In this part, we study the impact of the pseudo-label threshold on the model’s performance. As shown in Table 2, we can find that when the pseudo-label threshold is 1.4, the mAP for rare classes increase 1.30 relative to the threshold of 1.0, achieving a +0.39 mAP improved for all classes. The main reason is that a low pseudo-label threshold causes the model to incorporate many incorrect predictions as labels during supervised training, leading to suboptimal results and impacting the model’s performance. Furthermore, when

Table 1: The effects of different mask ratios in the context enhancement module.

	Ratio	Full	Rare	Non-rare
0	0.4	34.77	31.78	35.73
1	0.6	34.86	32.02	35.75
2	0.8	35.00	32.30	35.81
3	0.9	34.85	32.01	35.71

Table 2: The effects of different pseudo label thresholds on HICO-Det datasets.

	Pseudo-label	Full	Rare	Non-rare
0	1.0	34.61	31.00	35.67
1	1.2	34.84	31.60	35.76
2	1.4	35.00	32.30	35.81
3	1.6	34.99	32.17	35.80

Table 3: The effects of different numbers of prototype nodes in the instance feature alignment module.

	The Number of Prototype Nodes	Full	Rare	Non-rare
0	4	34.77	32.01	35.63
1	6	34.96	32.26	35.79
2	8	35.00	32.30	35.81
3	10	34.46	31.61	35.35

the pseudo-label threshold was set to 1.6, the model’s performance slightly deteriorated compared to the threshold of 1.4, indicating that excessively high thresholds filter out correct labels, and insufficient supervision signals also lead to performance degradation.

3.3 The Number of Prototype Nodes

As shown in Table 3, we studied the impact of different numbers of prototype nodes k in the instance feature alignment module. We find that when utilizing 6 prototype nodes, there was a relative improvement of 0.78% in mAP on rare classes compared to using 4 prototype nodes. While increasing the number of prototype nodes to 10, there was a decrease in mAP compared to using 6 prototype nodes. The primary reason is that using a small number of prototype nodes is insufficient to effectively capture and aggregate all the instance information in the images, consequently resulting in the loss of human-object pairs features and a subsequent decline in performance. Conversely, an excessive number of prototype nodes leads to the dispersion of human-object pairs features, hindering the proper aggregation of instance information.

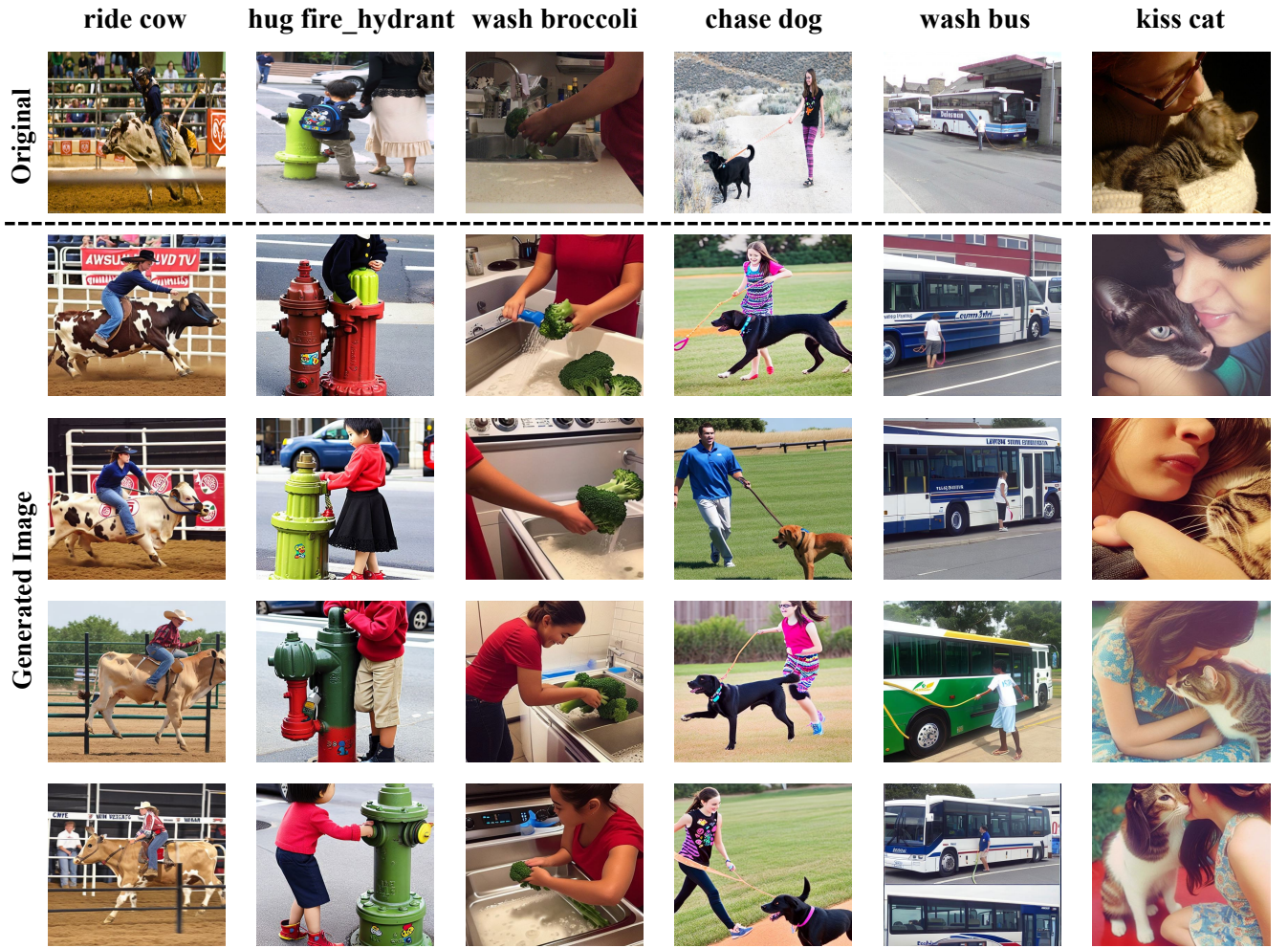


Figure 1: Visualization of images generated by Blip-diffusion. The first row represents images from the original dataset, while the remaining four rows depict the generated images. The HOI categories used for image generation are displayed at the top of each image.

4 ADDITIONAL QUALITATIVE RESULTS

In Fig. 2, we provide more qualitative results in addition to the cases mentioned in the paper. “GT” represents the ground truth, “HOICLIP” shows the predictions from the previous benchmark model HOICLIP, and “Ours” presents the results from our proposed method. In Fig. 2, blue bounding boxes represent person, and green bounding boxes represent objects.

The visualization results above the dashed line demonstrate the correct predictions made by our model. We can observe that our model not only slightly improves the localization of rare human-object pairs but also greatly enhances the classification of rare relationships. The results below the dashed line display the instances where our model made incorrect predictions. We can observe that the main reason for these prediction errors is the incorrect localization and classification of human-object pairs. This is attributed to keeping the instance decoder fixed during fine-tuning, which did

not significantly improve the detection capability for human-object pairs. In future research, we can explore and investigate this aspect further.

5 POTENTIAL LIMITATION AND FUTURE DEVELOPMENTS

In this section, we will discuss the limitations of our model and future directions for development. Firstly, Our method only focuses on rare classes, and we can next explore how to improve the performance of non-rare classes. Secondly, with the advancement of generative models, we can consider adopting more advanced and controllable models to generate higher-quality data. Thirdly, we expect that our method can provide new paradigm for using generated data for practical training and HOI detection.

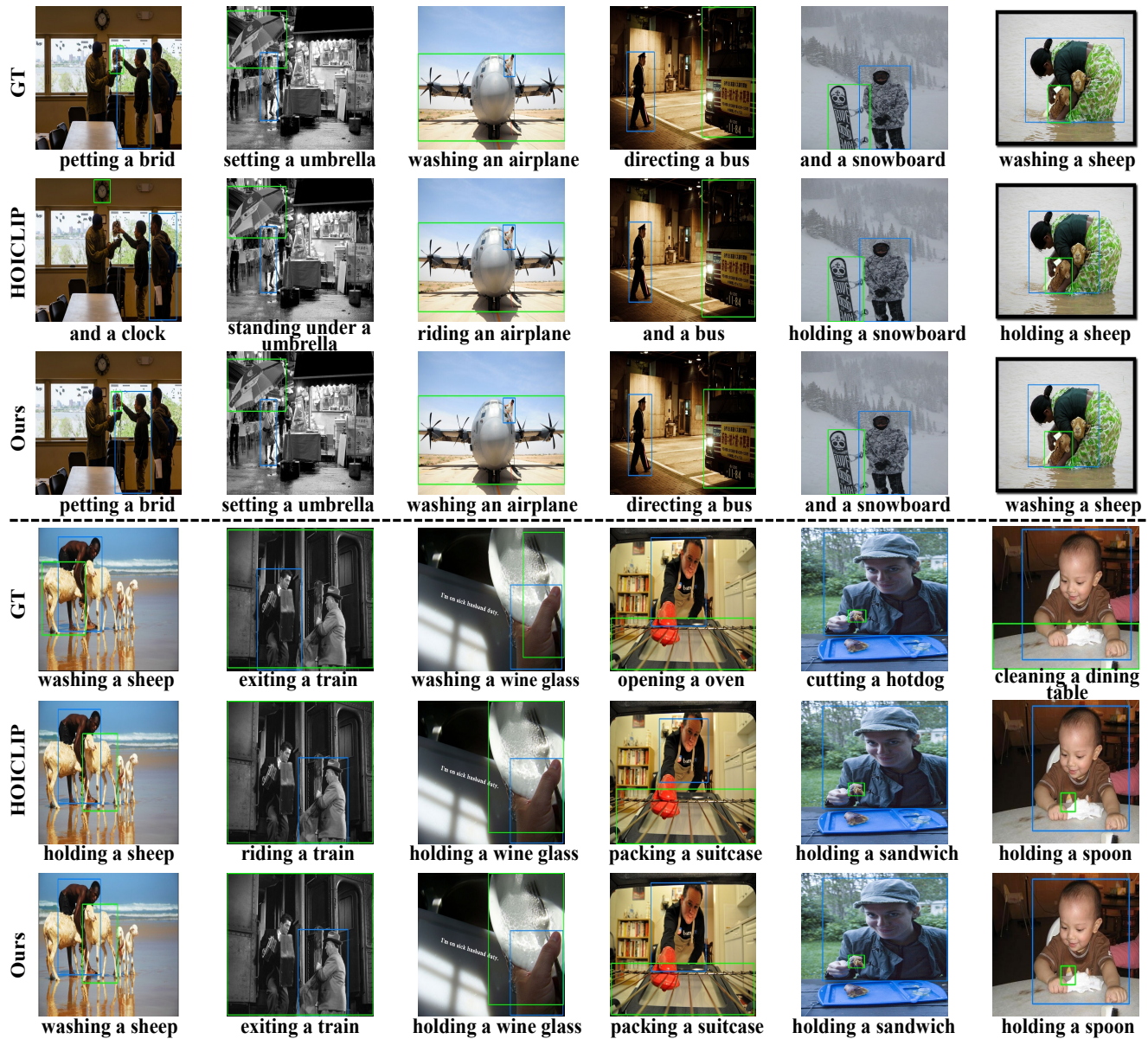


Figure 2: More qualitative evaluation on the HICO-Det dataset. “GT” represents the ground truth, “HOICLIP” shows the predictions from the previous benchmark model HOICLIP, and “Ours” presents the results from our proposed method. In the figure, blue bounding boxes represent person, and green bounding boxes represent objects. Above the dotted line are examples where our model performed well, and below the dotted line are examples where our model makes errors.