Supplementary Material for
Visual Relation Diffusion for Human-Object Interaction Detection

1. Zero-shot Settings in HICO-DET Dataset

We adopt four zero-shot settings to evaluate different aspects of generalization in Human-Object Interaction (HOI) detection. In Rare-First Unseen Composition (RF-UC), 120 categories are selected from rare HOI interactions and withheld from training to evaluate compositional generalization on long-tail distributions. Non-Rare First Unseen Composition (NF-UC) follows a similar protocol but selects 120 categories from frequently occurring HOI interactions. For Unseen Object (UO), 100 HOI categories involving 12 object types are excluded from training to assess object-level generalization. In Unseen Verb (UV), 84 HOI categories associated with 20 verb types are withheld to evaluate action-level generalization.

2. Evaluation on V-COCO Dataset

To further validate the effectiveness of our approach, we conduct additional experiments on V-COCO [1], another widely adopted benchmark for HOI detection. V-COCO comprises 10,346 images and 29 verb categories, making it significantly smaller in both dataset scale and interaction variety compared to HICO-DET. This offers a complementary perspective for evaluating the generalization ability of HOI models under limited data and label settings. V-COCO defines two evaluation scenarios: Scenario 1 includes all HOIs, while Scenario 2 filters out "no object" cases. We report results under both settings as AP (S1) and AP (S2). As shown in Tab. 1, our method achieves competitive performance among two-stage approaches using the same ResNet-50 backbone. The consistent results across both large-scale and compact datasets demonstrate the robustness of our proposed framework and its ability to adapt to different data distributions and annotation densities.

3. Parameter Efficiency Analysis.

To more comprehensively evaluate the effectiveness and efficiency of our proposed VRDiff framework, we conduct a detailed comparison of model sizes across representative HOI detection methods. As shown in Fig. 1, our method achieves the best performance under the Unseen Verb (UV) setting while using the fewest parameters among all compared approaches. This demonstrates the model's ability to transfer fine-grained visual knowledge effectively without relying on a large number of parameters. Furthermore, the radar chart provides an intuitive illustration of the trade-off between model complexity and detection accuracy. Compared to other methods that exhibit a performance gap of nearly 10 points between unseen and seen categories, our method achieves a more balanced performance, with a significantly smaller gap, indicating better generalization. These results collectively underscore the effectiveness of our visual relation diffusion strategy, which enables VRDiff to achieve strong accuracy-efficiency trade-offs and robust generalization across varying HOI detection scenarios.

Method	Backbone	AP (S1)	AP (S2)
One-stage Methods			
GEN-VLKT (CVPR22)	ResNet-50	62.4	64.5
HOICLIP (CVPR23)	ResNet-50	63.5	64.8
UniHOI (NeurIPS23)	ResNet-50	65.6	68.3
DIFFUSIONHOI (NeurIPS24)	VQGAN	**66.8**	**70.9**
Two-stage Methods			
UPT (CVPR22)	ResNet-50	59.0	64.5
ADA-CM (ICCV23)	ResNet-50	58.6	64.0
CLIP4HOI (NeurIPS23)	ResNet-50	-	66.3
EZ-HOI (NeurIPS24)	ResNet-50	60.5	66.2
VRDiff (Ours)	ResNet-50	**60.8**	**66.5**

Table 1. Comparison with representative methods on V-COCO.

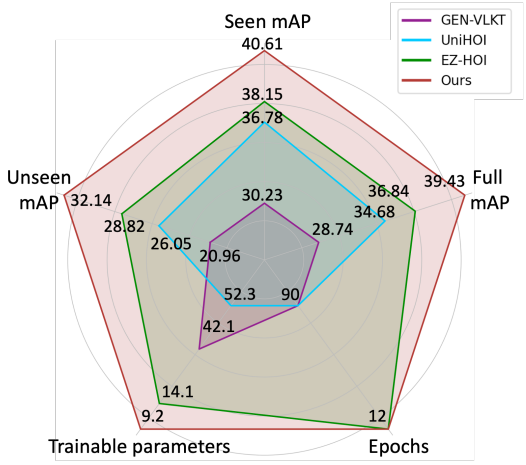

Figure 1. Efficiency-performance comparison of HOI detection models under the UV setting.

References

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