

Unseen No More: Unlocking the Potential of CLIP for Generative Zero-shot HOI Detection

Supplementary Materials

1 MORE ON ALGORITHM PROCEDURE

In addition to the primary details elucidated in the paper, we also present the the algorithm procedure for training HOIGen we propose. It primarily encompasses the following crucial steps: training the feature generation network to produce unseen features, incorporating the generated seen and unseen features to enable supervised training on all categories, and exploiting global features through CLIP and DINO to enhance the contextual understanding.

Algorithm 1: Training Procedure of HOIGen

Input: Images I , Detector $\Phi_{det}(\cdot)$, VAE encoder $E(\cdot)$, generator $G(\cdot)$, multilayer perceptron $MLP(\cdot)$, all category text labels $Text$.

- 1 $I_{bbox} \leftarrow \Phi_{det}(I)$; // Get image bounding box
- 2 Crop the image through I_{bbox} to get union, human and object data set $Data_{u,h,o}$;
- 3 Realistic image feature vector extraction $F_{u,h,o}$;
- 4 **for** $i = 1$ **to** $Epoch$ **do**
- 5 Load data $Data_{u,h,o}$ and train $E(\cdot)$ and $G(\cdot)$ by Eq(1);
- 6 Load features $F_{u,h,o}$ and finetune $MLP(\cdot)$ by Eq(2);
- 7 **end**
- 8 **for** $j = 1$ **to** $Epoch'$ **do**
- 9 // Generate features by $G(\cdot)$ and $MLP(\cdot)$
- 10 $F_{gen} \leftarrow MLP(G(Text))$;
- 11 // Extract image feature map F_{map}
- 12 $F_{vis} \leftarrow ROI - Align(F_{map}, I_{bbox})$;
- 13 Extract image feature vector F_{DINO}, F_{CLIP} ;
- 14 // Fusion of network realistic features and generated features, then calculate the HOI score by Eq(3,6)
- 15 $F_{all} \leftarrow \{F_{vis}, F_{DINO}, F_{CLIP}\} \oplus F_{gen}$;
- 16 Train the whole model by Eq(7);
- 17 **end**

2 DEATILS ON ZERO-SHOT HOI SETTINGS

In this section, we delve into the detailed settings of the zero-shot HOI task. Let $C = \{c_1, c_2, \dots, c_{N_c}\}$ represent the combination of all HOI interactions, $V = \{v_1, v_2, \dots, v_{N_v}\}$ denote the set of action verbs, and $O = \{o_1, o_2, \dots, o_{N_o}\}$ signify the set of interactable objects. C_{seen} and C_{unseen} delineate the seen and unseen sample categories within all HOI interaction combinations. V_{seen} and V_{unseen} represent the seen and unseen sample categories within all actions, and O_{seen} and O_{unseen} signify the seen and unseen sample categories within all objects, respectively.

Following previous research [1–5], we meticulously categorize the 600 HOI categories into five distinct zero-shot settings:

(1) Unseen Composition (UC): Here, we partition the data into 5 different sets. Each set comprises 480 sample categories for C_{seen}

and 120 for C_{unseen} . Both C_{seen} and C_{unseen} encompass both rare and non-rare classes, with all $v_j \in V_{seen}$ and $o_k \in O_{seen}$. The final detection result is computed as the average across these five partitioned datasets.

(2) Rare First Unseen Composition (RF-UC): In this scenario, the sample category of C_{seen} includes 480 samples, encompassing both rare and non-rare samples. Conversely, the sample category of C_{unseen} contains 120 samples, exclusively comprising rare class samples, while all $v_j \in V_{seen}$ and $o_k \in O_{seen}$.

(3) Non-Rare First Unseen Composition (NF-UC): Similar to RF-UC, the sample category of C_{seen} includes 480 samples, consisting of both rare and non-rare samples. However, the sample category of C_{unseen} exclusively consists of non-rare class samples, while all $v_j \in V_{seen}$ and $o_k \in O_{seen}$.

(4) Unseen Object (UO): In this setup, 12 object types are designated as O_{unseen} , while the remaining 68 categories are classified as O_{seen} . Simultaneously, 100 HOI interaction combinations involving these 12 objects are considered as C_{unseen} , with the remaining HOI interaction combinations treated as C_{seen} , with all $v_j \in V_{seen}$.

(5) Unseen Verb (UV): 20 verbs are assigned to V_{unseen} , while the remaining 97 categories constitute V_{seen} . Additionally, 84 HOI interactive combinations involving these 20 verbs are regarded as C_{unseen} , with the remaining HOI interactive combinations considered as C_{seen} , with all $o_k \in O_{seen}$.

Table 1: Different prompting methods to generate features.

Prompting Method	Full	Seen	Unseen
(1)	32.58	32.24	33.97
(2)	32.59	32.28	33.83
(3)	32.44	32.09	33.81
(4)	33.08	32.86	33.98

3 MORE RESULTS AND ANALYSIS

3.1 Ablation Study

In this experiment, we explored the influence of different prompting methods on the feature generation performance. (1) In the first method, we employed a random initialization prompt “XXXX” and generated features based on the human, object, and union branches. Notably, we did not delve into a detailed categorization of the human category according to the 80 object categories. (2) In the second method, we utilized a fixed initialization prompt template “a photo of a”, generating features across the human, object, and union branches. Here, we meticulously categorized the human category based on the 80 object categories. (3) Third, we again employed a random initialization prompt “XXXX”, yet this time we solely trained the union branch, generating human and object features through this branch alone. (4) Finally, in the fourth method,

we utilized a random initialization prompt “XXXX” and generated features across the human, object, and union branches, categorizing the human category in detail based on the 80 object categories.

As shown in Table 1, the optimal performance is achieved when employing the fourth prompting method. This outcome underscores the effectiveness of randomly initialized prompts for feature generation, along with the importance of carefully tailored prompts for human categories, so as to meet the intricate requirements of zero-shot HOI detection.

3.2 More Qualitative Results

In addition to the qualitative findings presented in the main paper, we provide further insights into some additional qualitative results, as depicted in Fig. 1. Notably, we observe a substantial enhancement in the detection performance for unseen samples upon the incorporation of the generated features. Remarkably, this improvement is achieved without compromising the detection accuracy for seen samples, showcasing the robustness and versatility of our approach.

In the main paper, we have presented the the t -SNE distributions of realistic and generated features under the UO setting. In addition to that, we here show more visualization results under various zero-shot HOI settings, as depicted in Figure 2. Comparing these results, we can observe that the features generated by our feature generation method consistently align with the realistic features across various settings. The high-quality feature generation offers a significant support for competing against the seen-unseen bias.

REFERENCES

- [1] Ting Lei, Fabian Caba, Qingchao Chen, Hailin Jin, Yuxin Peng, and Yang Liu. 2023. Efficient Adaptive Human-Object Interaction Detection with Concept-guided Memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6457–6467.
- [2] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. 2022. GEN-VLKT: Simplify Association and Enhance Interaction Understanding for HOI Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20091–20100.
- [3] Yunyao Mao, Jiajun Deng, Wengang Zhou, Li Li, Yao Fang, and Houqiang Li. 2023. CLIP4HOI: Towards Adapting CLIP for Practical Zero-Shot HOI Detection. In *Advances in Neural Information Processing Systems 36*.
- [4] Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. 2023. HOICLIP: Efficient Knowledge Transfer for HOI Detection with Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23507–23517.
- [5] Mingrui Wu, Jiaxin Gu, Yunhang Shen, Mingbao Lin, Chao Chen, and Xiaoshuai Sun. 2023. End-to-End Zero-Shot HOI Detection via Vision and Language Knowledge Distillation. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*. 2839–2846.

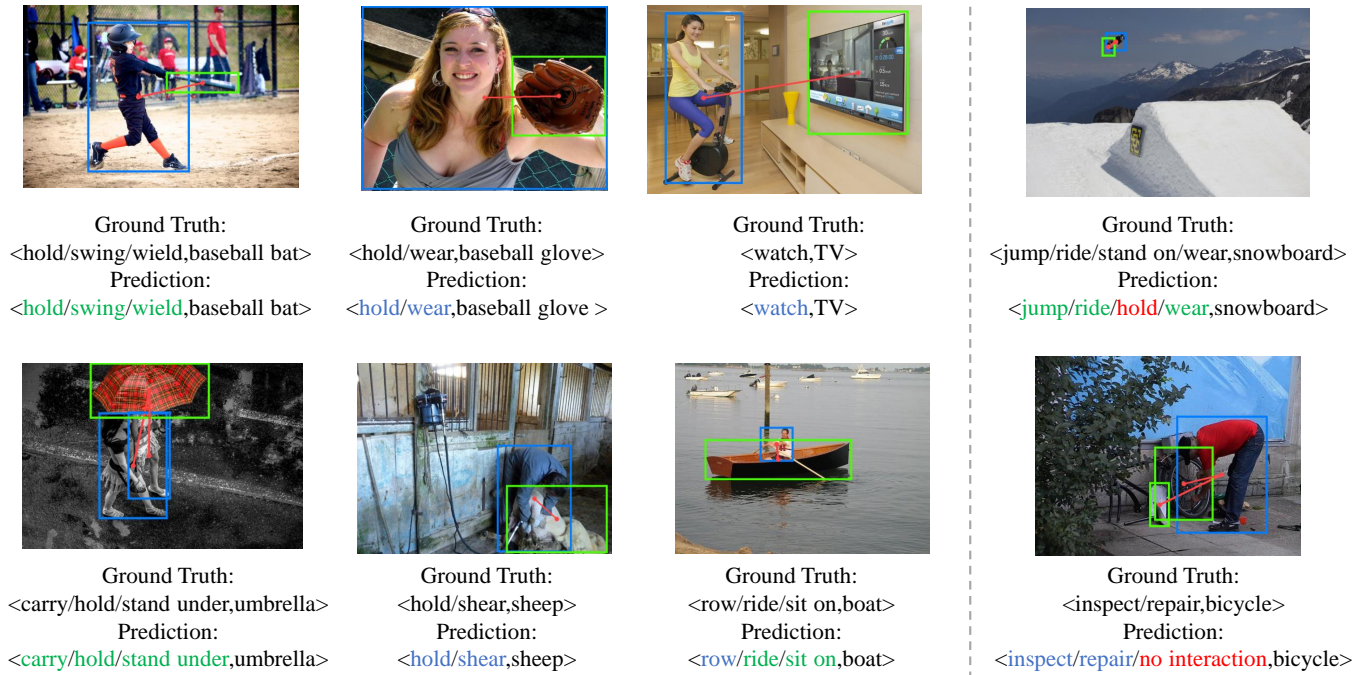


Figure 1: Visualization of zero-shot HOI detection results under NF-UC setting, with the image samples from HICO-DET dataset. Correctly classified seen and unseen category are marked in blue and green respectively, and incorrect recognition results are marked in red.

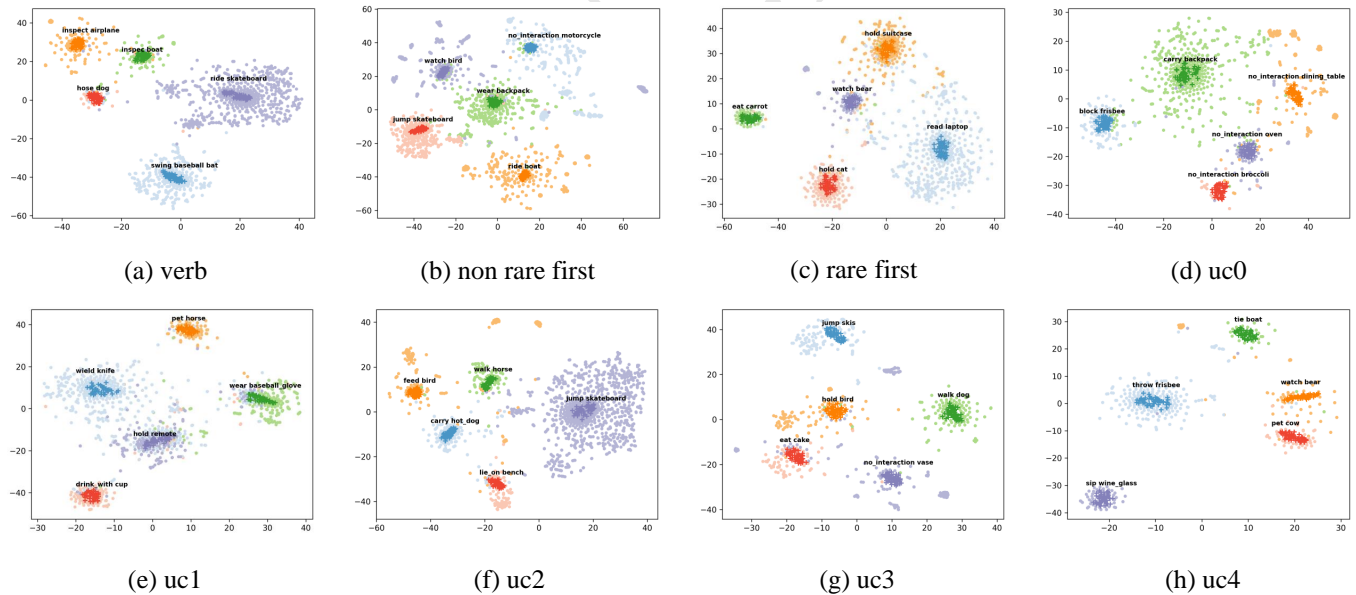


Figure 2: Visualization of realistic features (depicted as light regions) and synthesized features (depicted as dark regions) using t -SNE under various settings, particularly focusing on unseen HOI categories from the HICO-DET dataset. In the RF-UC setting, it pertains to the seen HOI categories.