Exploring the Potential of Large Foundation Models for Open-Vocabulary HOI Detection

Ting Lei Shaofeng Yin Yang Liu*
Wangxuan Institute of Computer Technology, Peking University
{ting_lei, yangliu}@pku.edu.cn yin_shaofeng@stu.pku.edu.cn

In this supplementary material, we offer a more comprehensive assessment of our proposed CMD-SE. In Section 1, we further investigate the multi-level decoding mechanism and the effect of different descriptions on our model. In Section 2, we further analyze the generalizability of our CMD-SE by comparing it with the previous state-of-the-art zero-shot HOI detector. In Section 3, we provide additional details regarding the implementation of the proposed approach. In Section 4 we present more qualitative results on images in the wild. In Section 5 we discuss the limitations and potential future directions of our work.

1. Ablation Study

In this section, we empirically investigate the sensitivity of the proposed method to the multi-level decoding mechanism and the effect of different description on the open-vocabulary SWIG-HOI dataset. Specifically, besides the four aspects of CMD-SE we have discussed in section 4.3, we further ablate on (1) utilizing different levels of feature maps to decode HOIs and (2) the importance of different descriptions.

The multi-level decoding mechanism. As shown in Table 1, utilizing the feature maps from level $\{6, 9, 12\}$ brings a 0.73% mAP gain on all categories compared with utilizing the feature maps from level $\{9, 12\}$. It is worth noting that the performance on HOIs with large distances improves by a large margin (4.69%), showing the effectiveness of utilizing multi-level feature maps to model HOIs with different distances. Furthermore, we empirically find that utilizing more levels of feature maps leads to marginally inferior performance. This shows that utilizing more levels of feature maps might lead to a more challenging optimization during the training process. Therefore, in the main paper, we use the feature maps from level $\{6, 9, 12\}$ by default to report all experimental results unless otherwise specified.

The importance of different descriptions. (1) *HOI description:* besides using the annotated verb/object definitions in main paper (line 1 of Table 6) as HOI description, we try 2 different prompts for querying the LLM: "Knowledge retrieve for HOI", and "Describe the visual appearance of HOI" (lines 2&3 of Table 2). We observe that "Knowledge Retrieval" often includes redundant information lacking relevance to visual cues (e.g., "a common recreational activity"), "Visual Appearance" also performs worse compared to ours, suggesting that the underlying states of body parts offer a more broadly applicable comprehension of HOI concepts. (2) *Bodypart description:* To better understand the importance of body parts description, we experiment with a subset of body parts (last row), yielding inferior results compared to using all body parts. This indicates the visual cues from multiple body parts provide complementary and valuable information for open-vocabulary HOID.

2. Comparative Analysis of Open Vocabulary and Zero-shot HOI Detectors

We observe from Table 2 of the main text that all open vocabulary methods perform worse than zero-shot methods on the HICO-DET dataset. However, this comparison is not entirely fair because the zero-shot methods depend on a DETR architecture and often use pretrained weights from COCO [2][1]. To enable a controlled comparison in dealing with unseen objects in a fair manner, we conduct the following experiment to analyze the generalizability of our model and the previous state-of-the-art zero-shot HOI detector [4]. Specifically, we train both methods on the default setting of the HICO-DET dataset and evaluate their performance on a subset of the SWIG-HOI test set. We randomly select a few classes from the SWIG-HOI test set for evaluation. As shown in Table 3, our proposed method outperforms HOICLIP [4] by 5.62% and 4.79% mAP when selecting 20 or 50 classes, respectively. It is worth noting that interactions in SWIG-HOI may involve arbitrary object categories that are not present in the COCO dataset, which is closer to an open-world scenario. Previous methods that rely on a pretrained DETR [1] exhibit inferior performance

*Corresponding author

[1]The COCO pretraining implicitly encompasses all object categories present in HICO-DET.

$\mathbb{L}$	Small	Large	Seen	Unseen	Full
{9, 12}	15.26	13.30	16.79	10.05	15.09
{6, 9, 12}	**15.20**	**17.35**	**16.79**	**10.70**	**15.26**
{3, 6, 9, 12}	15.13	16.93	16.73	10.14	15.07

Table 1. Ablation on the multi-level decoding mechanism. $\mathbb{L}$: the levels of feature maps used for decoding. Small: HOIs with small distances ($\leq$0.33). Large: HOIs with large distances ($\geq$0.67).

Prompts	Unseen
None	9.32
Knowledge Retrieval	9.85[+0.53]
Visual Appearance	10.27[+0.95]
Body Part	**10.70**[+1.38]
Subset of Body Part	10.25[+0.93]

Table 2. Importance of different descriptions.

when directly applied to SWIG-HOI. In contrast, our proposed CMD-SE, which does not rely on a pretrained detector, demonstrates the greater potential for detecting HOIs in open-world scenarios.

3. Implementation Details

In this section, we present a comprehensive description of the implementation details of our model. Our model is built upon the pretrained CLIP and all its parameters are frozen during training. Following the previous work [5], we employ the ViT-B/16 version as our visual encoder to ensure a fair comparison, which processes a 224 $\times$ 224 $\times$ 3 image as input. During training, we first apply a few data augmentation techniques, including RandomHorizontalFlip, RandomCrop, and ColorJitter. Then, the augmented images are resized to the CLIP input resolution, *i.e.*, 224 $\times$ 224. The visual encoder consists of a total of 12 layers, and we extract feature maps from levels $\{6, 9, 12\}$, forwarding them to the HOI decoder. The number of layers of the HOI decoder is set to 4. We employ 10 and 25 HOI queries on the SWIG-HOI and HICO-DET datasets, respectively. We set the cost weights λ_b, λ_{iou}, λ_{cls} and λ_d to 5, 2, 5, and 5 during training. We use focal loss [3] for interaction classification to counter the imbalance between positive and negative examples. We set γ to 2 during inference. We introduce 8 prefix tokens and 2 conjunctive tokens to connect the words of human actions and objects following [5]. We set the learning rate as 10^{-4} and use the Adam optimizer with decoupled weight decay regularization. We train our model for 80 epochs with a batch size of 128 on 2 A100 GPUs.

Method	#(Classes)	mAP
HOICLIP [4]	20	7.41
CMD-SE (Ours)	20	**13.03**
HOICLIP [4]	50	4.39
CMD-SE (Ours)	50	**9.18**

Table 3. Experiments on applying methods trained on HICO-DET to SWIG-HOI dataset. #(Classes): the number of classes we select on the SWIG-HOI test set.

4. Qualitative Examples in The Wild

To provide further evidence of the enhanced performance and generalization capabilities of our model, we present additional qualitative results on data in the wild.

Human Body Parts. As depicted in Figure 1a and 1b, our model effectively captures contextual features from image regions containing human body parts, particularly those involved in interactive actions (*e.g.*, legs in "kicking person," mouth in "drinking drinking-glass"). It accurately predicts interaction categories based on these extracted features.

Different Distances. Our model demonstrates strong performance across varying distances and scales of human-object interactions, as demonstrated in Figure 1c and 1d. Whether dealing with small-scale interactions like "talking telephone" or large-scale interactions like "photographing camera," our model consistently achieves favorable results.

Different art styles. To assess the model's generalizability to different art styles, we conduct tests using images of video game characters. Figure 1e and 1f showcase the model's correct predictions of interactions even when presented with out-of-domain images. This demonstrates its commendable performance for images with diverse art styles.

Multiple HOIs. We also evaluate the model's ability to detect multiple HOIs in a single image. As shown in Figure 1g and 1h, our model correctly identifies two different interactions through separate prediction heads. Although the attention patterns of these heads may appear similar, closer inspection reveals that each head actually pays more attention to the objects related to its predictions.

5. Limitations

Currently, our method is limited in the following two key aspects. Firstly, while we utilize GPT to generate fine-grained descriptions of human body parts for each HOI, our model's ability to distinguish between different HOIs is still constrained by the embedding space of the CLIP text encoder. Secondly, our current approach does not rely on pretrained detectors, which is a strength but may also be a weakness given recent advances in open-vocabulary object detection. In the future, we will explore ways to utilize more advanced

(a) kicking person. (b) drinking drinking-glass. (c) talking telephone. (d) photographing camera.

(e) shooting shotgun. (f) riding horse. (g) reading book. (h) smoking cigarette.

Figure 1. Qualitative results of our method on wild data.

text encoders and integrate high-quality open-vocabulary object detectors into our model. We believe that these improvements will lead to even better performance and make our method more practical for real-world applications.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[3] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[4] Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23507–23517, 2023. 1, 2

[5] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detectors with natural language supervision. In *CVPR*, 2022. 2