Open-World Human-Object Interaction Detection via Multi-modal Prompts –Supplementary Material–

Jie Yang[1,2*], Bingliang Li[1*], Ailing Zeng[2†], Lei Zhang[2], Ruimao Zhang[1†]
[1]The Chinese University of Hong Kong, Shenzhen [2]International Digital Economy Academy

Contents

A. Detailed Experimental Settings

Datasets. We evaluate our models on four datasets: HICO-DET [1], V-COCO [2], SWiG-HOI [8], HCVRD [11]. HICO-DET consists of $47,776$ images, including $38,118$ training images and $9,658$ test images. It comprises 80 object categories aligned with the categories in MS-COCO [5], along with 117 verb classes. This results in 600 types of HOI triplets. HICO-DET provides annotations for over $150,000$ human-object pairs. V-COCO is a subset of MS-COCO and is significantly smaller in size compared to HICO-DET. It consists of $10,346$ images, with $2,533$ training images, $2,867$ validation images, and $4,946$ test images. V-COCO includes the same 80 object categories and 29 verb classes. SWiG-HOI is based on the SWiG[7] dataset, which consists of $68,189$ images, with $41,320$ training images, $13,588$ development images, $13,281$ test images. SWiG-HOI includes $1,000$ object categories and 407 verb classes, results in more than $13K$ HOI triplets. HCVRD consists of $41,586$ images, with $31,586$ training images, $10,000$ test images. HCVRD includes $1,821$ object categories and 884 verb classes, results in more than $33K$ HOI triplets.

Evaluation Metrics. We adopt the mean Average Precision (mAP) following the standard evaluation [4, 9]. A prediction for an HOI triplet is considered a true positive if it satisfies two conditions: 1) accurate human and object locations (box IoU with reference to GT box is greater than 0.5) and 2) correct object and verb categories. For HICO-DET, we evaluate three different category sets: all 600 HOI categories (Full), 138 HOI categories with fewer than 10 training instances (Rare), and the remaining 462 HOI categories (Non-Rare). For V-COCO, we report the role mAPs for two scenarios: Scenario 1 (S1) for the 29 verb categories, including the 4 body motions, and Scenario 2 (S2) for the 25 verb categories excluding the no-object HOI categories. For SWiG-HOI and HCVRD, we report the mAP for three category sets: Full, Rare and Non-Rare, similar to the settings in HICO-DET.

Implementation Details. The number of channels C is set to 256. We optimize our network with AdamW with a weight decay of 10^{-4}. We train the model for 90 epochs with an initial learning rate of 10^{-4}, decreased by 10 times at the 60th epoch. The Stable Diffusion model and CLIP model are frozen during training for all the settings. The scene-aware adaptor α and β are the trainable linear layer. We set the cost weights λ_b, λ_g, λ_c^o and λ_c^i to 2.5, 1, 1 and 1, respectively, following [4, 9]. We conduct all the experiments with a batch size of 16 on A100 GPUs.

B. The SynHOI Dataset

This section introduces a high-quality synthetic HOI dataset called *SynHOI* to complement *Magic-HOI*. To make the flow of dataset production scalable, we present an automatic pipeline, including the HOIPrompt design, automatic labeling and filtering, and quality verification, designed to continually scale up the generation of diverse and high-precision HOI-annotated data, as shown in Fig. 1.

B.1. Automated Generation and Annotation System

The HOIPrompt Design. As illustrated in Fig. 1 (a), we propose pre-defined HOIPrompts. Complete HOI triplets are formed by sampling verb and noun combinations from the *Magic-HOI* dataset, creating "a {race} {age & gender} verb-ing an object." To describe a person's appearance, we use the format "a {race} {age & gender}," randomly selecting elements from the HOIPrompts. We generate images with combinations likely to occur together by analyzing co-

*Equal contribution.
†Corresponding author.

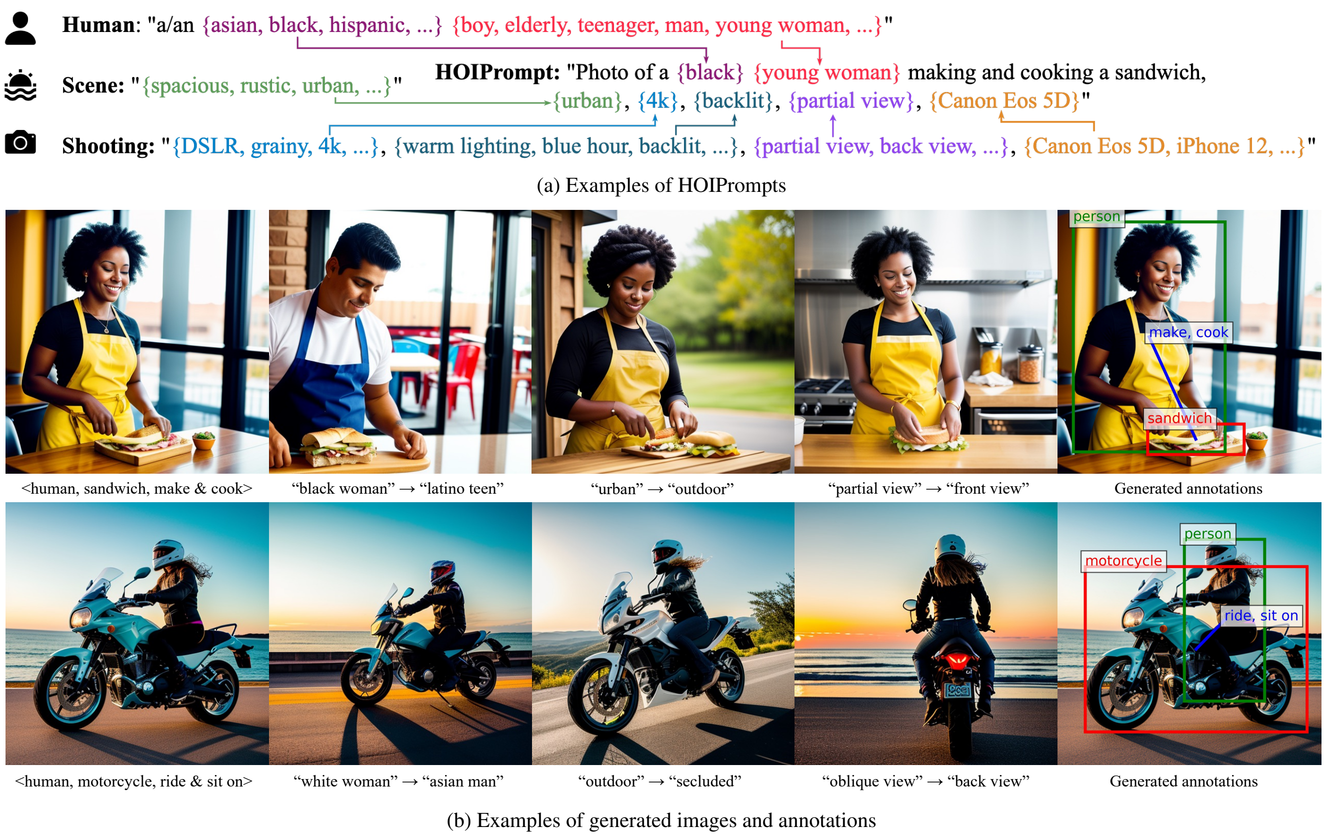

Figure 1. Illustration of a) HOIPrompts and b) How HOIPrompts guide the text-to-image generation process to enhance diversity.

This GPT model is designed to generate detailed captions of human-object interactions in images. Upon receiving an image, it will create a caption describing all the interactions between humans and objects or human and human. Then, after receiving results from an expert model on the same image, the GPT will refine and correct its initial caption based on this new information. The GPT should incorporate additional information from the expert model based on the initial caption. The responses are structured in two sections:

1. Output #1: "Caption without additional information: [Initial detailed caption]"

(After the user inputs of the expert model)

2. Output #2: "Caption with additional information: [Refined and corrected caption]"

The model should focus on accuracy and detail in both the initial caption and the refinement process. It should use the expert model's results to enhance its understanding of the scene and improve the caption's accuracy. The second caption should be similar to the initial caption with proper adjustment based on the expert model. Only change the necessary descriptions in the initial caption to make minimal changes so the refined caption is as close as the initial caption. Keep as much information as possible from the initial caption.

Table 1. Task instructions of caption generation and refinement using GPT-4V via GPTs. The "expert model" refers to *MP-HOI*, which is trained specifically for Human-Object Interaction Detection.

occurring HOI triplets. For interaction environments, we select an adjective from a range of options to describe the atmosphere (denoted as "{environment}"). Photographic information is represented by four components ("{quality}," "{lighting}," "{view}," and "{camera}"), aligning the synthetic images with real HOI data and providing camera angle diversity. We further enhance diversity and quality through negative prompts and random model configurations.

Automatic Labeling and Filtering. We design a three-step process to automatically annotate and filter the synthetic images. Firstly, we utilize a state-of-the-art detection model [10] to detect objects within each image. Secondly, we discard any images in which the confidence score of the detected object specified in the corresponding HOI triplet(s) prompt is below a threshold of 0.5. Thirdly, we associate humans with objects in the images and assign the appropriate HOI category from the prompts to the human-object combination(s). In practice, we assign the HOI category to the person closest to the center of the object's bounding box. If not all humans in the image have an HOI label, we select the object closest to the corresponding human's bounding box. Upon completing the automated labeling and filtering process, we obtain a synthetic dataset, namely *SynHOI*, including over $100K$ annotated images with complete labels associated with the detected human-object interactions.

Visualization and Manual Verification. We develop a visualization tool to facilitate the manual inspection and filtering of any incorrect HOI annotations. We incorporate manual efforts to sample and inspect the annotated results to ensure their quality. During this inspection, we observe that the SOTA detector trained on COCO [5] performs well in detecting humans and objects, indicating that the data distribution in *SynHOI* closely resembles that of natural images in COCO. However, due to the inherent ambiguity of verbs, a small number of synthetic interactions in *SynHOI* may be incompletely accurate, which also exists in *Magic-HOI*. To address this issue, we construct a subset comprising 10% data of *SynHOI* (over $10K$), namely *SynHOI-Sub*, during the sampling and inspection process. This subset has undergone meticulous manual examination, resulting in annotations that are verified to be completely accurate.

B.2. Data Characteristics

SynHOI has three key data characteristics: (1) **High-quality data.** *SynHOI* showcases high-quality HOI annotations. First, we employ CLIPScore [3] to measure the similarity between the synthetic images and the corresponding HOI triplet prompts. The *SynHOI* dataset achieves a high CLIPScore of 0.849, indicating a faithful reflection of the HOI triplet information in the synthetic images. Second, Fig. 1-(b) provides evidence of the high quality of detection annotations in *SynHOI*, attributed to the effectiveness of the SOTA detector [10] and the alignment of *SynHOI* with real-world data distributions. (2) **High-diversity data.** *SynHOI* exhibits high diversity, offering a wide range of visually distinct images. Fig. 1-(b) demonstrates the impact of random variations in person's descriptions, environments, and photographic information within the HOIPrompts on the diversity of synthetic images. (3) **Large-scale data with rich categories.** *SynHOI* aligns *Magic-HOI*'s category definitions to effectively address the long-tail issue in *Magic-HOI*. It consists of over $100K$ images, $130K$ person bounding boxes, $140K$ object bounding boxes, and $240K$ HOI triplet instances.

B.3. Visualization of SynHOI

In *Magic-HOI*, a total of $1.2K$ verbs are defined to depict the relationship between humans and objects. We demonstrate the representation of these verbs in our *SynHOI* dataset. Specifically, our focus lies in presenting the verb-object combinations that present challenges in collecting images from the internet. For instance, these combinations include "A man, smells a carrot", "A person holds a broccoli", or "a kid paints the fire hydrant". As shown in Fig. 2, our proposed pipeline for dataset construction enables us to generate a wide range of diverse and realistic images with high-quality annotations. More importantly, a significant advantage of our approach is its capability to generate interactions in highly rare scenarios.

C. Benefiting GPT-4V with MP-HOI

Motivation. In the rapidly advancing field of Multimodal Large Language Models like GPT-4V, many studies have shown their strong image understanding capabilities. However, for human-object interaction recognition, they still face challenges in open-world scenarios, like imprecise object/interaction recognition and inaccurate object localization. Inspired by these, we explore *MP-HOI*, the powerful HOI generalist model, to improve GPT-4V's accuracy for comprehensive HOI understanding.

Combination MP-HOI with GPT-4V. We set up GPT-4V [6] with predefined instructions as shown in Tab. 1. Through it, we could upload an in-the-wild test image and generate the initial caption. Next, we input the results from *MP-HOI*, e.g. "Person holding an iPhone 15 Pro Max", to refine and correct the initial caption. The results in Fig. 2 highlight the capacity of *MP-HOI* to enhance the accuracy of GPT-4V in comprehensive HOI understanding. It's worth noting that to maintain data integrity and prevent potential information leakage, we initiate a fresh conversation with GPT-4V for each image.

References

[1] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 1

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

[3] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 3

[4] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20123–20132, 2022. 1

[5] 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

Image	Initial Caption	Refined Caption
	A woman in a white bridal gown and veil, holding a bouquet of flowers, is smiling and standing next to a man in a military uniform adorned with medals. The man appears to be kissing his hand, and both individuals are standing behind a balcony ledge draped with a red fabric featuring a decorative gold tassel.	Prince Charles, dressed in a ceremonial military uniform with medals, is kissing the hand of Princess Diana, who is wearing a white bridal gown with a veil and holding a bouquet of flowers. Princess Diana is looking at Prince Charles with a smile, as they stand behind a balcony ledge draped in red fabric with a gold tassel.
	Two men are engaged in an interaction over a handheld electronic device, which one of them, wearing glasses and a grey polo shirt, is holding and possibly demonstrating to the other, who is smiling, wearing a blue checked shirt and a red lanyard. In the background, a third person is visible, slightly out of focus, holding up a yellow object that appears to be a smartphone, capturing the interaction or an unrelated subject.	Two men are engaged in an interaction involving an iPhone 15 Pro Max. The individual in the grey polo shirt and glasses is holding the device, possibly demonstrating it to the other man who is observing the phone with a smile, wearing a blue checked shirt and a red lanyard. In the blurred background, another person is visible, holding up a smartphone, possibly capturing the interaction or another subject.
	In the image, two individuals appear to be engaged in a standoff inside what looks like a spacecraft or futuristic setting. The person on the left is dressed in a light brown robe with a white undergarment and is holding a blue lightsaber, with the blade pointed towards the person on the right. The person on the right, facing the first individual, is clad in a black suit and helmet that suggests a form of armor, possibly for a character with a villainous persona, and is wielding a red lightsaber with the blade directed towards the first person. The two lightsabers are crossed in the middle, indicating a duel. The environment suggests a corridor with metallic walls and several lights, giving the impression of a scene from a science fiction narrative.	The image shows two individuals engaged in a duel, with one person wielding a blue lightsaber and the other countering with a red lightsaber. They are standing in a metallic corridor that resembles the interior of a spacecraft, indicative of a science fiction setting. The first person is dressed in a light brown robe and a white undergarment, suggestive of a wise and skilled character, while the second person is in a black armored suit with a helmet, implying a menacing and formidable opponent. The crossed lightsabers and the focused stances of both individuals suggest that they are in the midst of a combat interaction.
	A man in a striped sweater and apron is standing in a food stall, carving meat from a large rotating spit with a long knife. He appears focused on the task at hand, with a slight grimace possibly due to the heat or intensity of the work. Behind him, another individual is working at a counter with various dishes displayed, seemingly preparing or organizing plates. In the background, the interior of a casual dining establishment with tables and chairs can be partially seen, and various signs and commercial branding are visible, indicating an urban setting likely within a market or busy street.	The first person, wearing a striped sweater and apron, is cooking Döner kebab, attentively carving the meat from the large rotating spit with a long knife. The second person, in the background, looks at the Döner kebab, possibly waiting for his portion or simply observing the cooking process. The third person, dressed in black, is positioned by the serving counter, engaged in the process of buying Döner kebab, as he appears to be examining or choosing items from the displayed dishes.
	An astronaut clad in a white space suit with American flag and mission patches is conducting a spacewalk. The astronaut is using their gloved hands to manipulate a tool attached to their suit, with the Earth's horizon and solar panels of a spacecraft visible in the background.	An astronaut is holding a space drill during a repair operation on the International Space Station. The astronaut's hands are engaged with the tool, performing maintenance work with the Earth and part of the station's solar arrays in the background.

Table 2. Image Captions Comparison. Text colored in red denotes information misinterpreted or oversimplified by GPT-4V; text in blue represents additional insights derived from the results of *MP-HOI*; text in green represents corrections made from the results of *MP-HOI*.

Figure 2. Visualization of the proposed *SynHOI* dataset.

Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014. 1, 3

[6] OpenAI. Introducing gpts. 3

[7] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 314–332. Springer, 2020. 1

[8] Suchen Wang, Kim-Hui Yap, Henghui Ding, Jiyan Wu, Junsong Yuan, and Yap-Peng Tan. Discovering human interactions with large-vocabulary objects via query and multi-scale detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13475–13484, 2021. 1

[9] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. *Advances in Neural Information Processing Systems*, 34:17209–17220, 2021. 1

[10] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 2, 3

[11] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton van den Hengel. Hcvrd: A benchmark for large-scale human-centered visual relationship detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 1