–Supplementary Material–
F-HOI: Toward Fine-grained Semantic-Aligned 3D Human-Object Interactions

Jie Yang[1,2,⋆], Xuesong Niu[2,⋆], Nan Jiang[2,3,⋆], Ruimao Zhang[1,†], and Siyuan Huang[2,†]

[1]The Chinese University of Hong Kong, Shenzhen [2]State Key Laboratory of General Artificial Intelligence, BIGAI [3]Institute for AI, Peking University

1 More Qualitative Results

As shown in Fig. 1, we present additional qualitative results for understanding and reasoning tasks. These results demonstrate the potential of fine-grained HOI modeling to effectively understand and reason about the HOI state. However, as discussed in the limitation section of the main article, the restricted data volume used to align HOI-Pose with fine-grained descriptions, coupled with the richness and quality of the textual descriptions, affects performance and generalization ability. From our experiments, we have observed several understanding and reasoning errors, such as inaccurate judgments of interactions and incorrect assessments of the spatial relationships between body parts. In future research, the most impressive improvement can be achieved by merging more HOI datasets, scaling up the number of samples, and enriching the textual descriptions. Additionally, exploring the hierarchy of human body part states, object states, and actions is promising and meaningful. These diverse, hierarchical, and large-scale data can more effectively align the HOI state with fine-grained descriptions, potentially benefiting subsequent downstream applications.

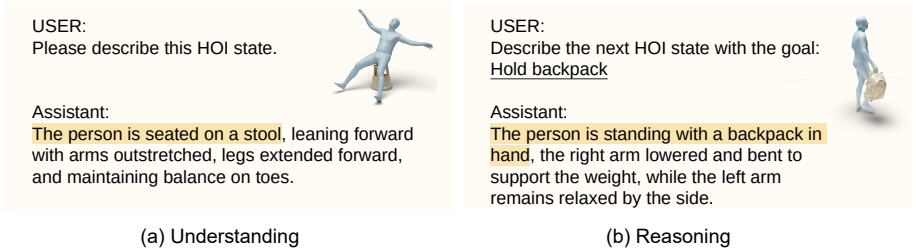

Fig. 1: We show the potential of fine-grained HOI modeling to effectively understand and reason about the HOI state.

⋆Equal contribution †Corresponding author

2 Details in Dataset Construction

As described in Sec. 3.3 of the main article, we prompt GPT-4V [1] to annotate fine-grained textual descriptions for HOI states and the movements that happen between two consecutive states. Specifically, given 2D images for two HOI states, we meticulously design the formats of prompts as follows: (a) decoupled human pose descriptions, including whole-body, head, two arms, two hands, two legs, and two feet; (b) object state descriptions; (c) interaction state descriptions. The complete query message to request GPT-4V is given in Tab. 1.

Table 1: Prompt messages to request GPT-4V for fine-grained description generation.

```
base64_image1 = encode_image(first_image_path)
base64_image2 = encode_image(second_image_path)
object_name = object_name
output_format = "First image human pose: \n First image human object interaction: \n First image human object contact part:\n Second image human pose: \n Second image human object interaction: \n Second image human object contact part:\n Whole body movement: \n Head movement: \n Arm movement: \n Hands movement: \n Legs movement: \n Feet movement: \n Object trajectory: \n Human object interaction: \n Action Description: \n "

prompt = f"You are an expert for human motion and human objection interaction description. You will be given two images. Please describe the human pose, human-object interaction, and contact part of the human and the object in the first and second images. Please also describe the detailed human motion movement from the first image to the second image for the whole body and human body parts including head, arms, hands, legs, and feet. Please also describe the translation, rotation, and moving trajectory of the object. Please also describe the human-object interaction and the action of the human. The object in the image is a {obj_name}. All the descriptions should be in short words, and the output format should be like: {output_formate}"

messages = [{"role": "user", "content": [{"type": "text", "text": prompt},
{"type": "image_url", "image_url": {"url": f"{base64_image1}"}},
{"type": "image_url", "image_url": {"url": f"{base64_image2}"}}]}]
```

References

1. OpenAI: GPT-4 technical report. (2023) 2