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

https://f-hoi.github.io

Abstract. Existing 3D human object interaction (HOI) datasets and models simply align global descriptions with the long HOI sequence, while lacking a detailed understanding of intermediate states and the transitions between states. In this paper, we argue that fine-grained semantic alignment, which utilizes state-level descriptions, offers a promising paradigm for learning semantically rich HOI representations. To achieve this, we introduce Semantic-HOI, a new dataset comprising over 20K paired HOI states with fine-grained descriptions for each HOI state and the body movements that happen between two consecutive states. Leveraging the proposed dataset, we design three state-level HOI tasks to accomplish fine-grained semantic alignment within the HOI sequence. Additionally, we propose a unified model called F-HOI, designed to leverage multimodal instructions and empower the Multi-modal Large Language Model to efficiently handle diverse HOI tasks. F-HOI offers multiple advantages: (1) It employs a unified task formulation that supports the use of versatile multimodal inputs. (2) It maintains consistency in HOI across 2D, 3D, and linguistic spaces. (3) It utilizes fine-grained textual supervision for direct optimization, avoiding intricate modeling of HOI states. Extensive experiments reveal that F-HOI effectively aligns HOI states with fine-grained semantic descriptions, adeptly tackling understanding, reasoning, generation, and reconstruction tasks.

Keywords: 3D Human-Object Interaction · Fine-Grained Semantics

1 Introduction

Modeling human-object interaction (HOI) in 3D space is critical for various downstream applications, such as computer animation, virtual reality, and embodied AI [15, 43, 54, 63, 76]. The HOI process involves a sequence of continuous state transitions, which contains intricate changes in body parts, object trajectories, and interaction contacts. However, existing models [2, 47, 56, 64] only align global descriptions with such processes, making them struggle to comprehend each HOI state and the transitions between states in the fine-grained semantic

⋆Equal contribution †Corresponding author

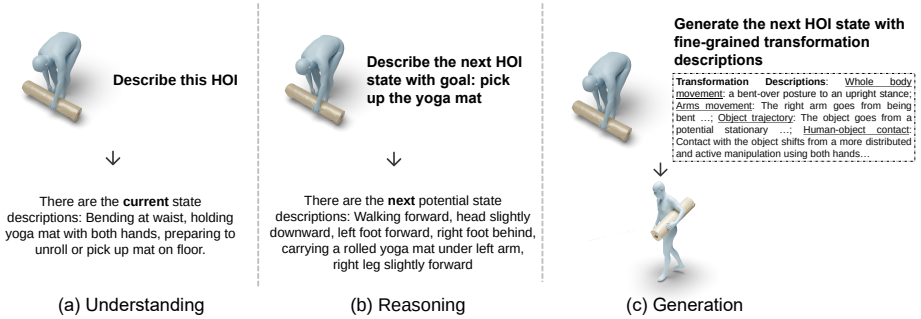


Fig. 1: Illustration of three state-level tasks to achieve fine-grained semantic alignment.

space. In the literature, how to achieve fine-grained semantic-aligned 3D HOI is still a challenging yet under-explored issue due to the following difficulties:

1. **Dataset Gap.** Existing HOI datasets only provide coarse-grained goal descriptions (e.g., "a person picks up a backpack") to depict a long HOI sequence. Thus, the absence of fine-grained semantic descriptions significantly hampers progress in the related field.
2. **Model Capacity.** Aligning the fine-grained semantic descriptions with HOIs is non-trivial. It requires a model that can establish alignments from a limited dataset, possesses powerful semantic comprehension skills to handle extensive textual descriptions, and has prior knowledge of real-world actions.

To address the first issue, we reconstruct a new dataset named Semantic-HOI from three existing datasets, bridging the semantic gap in current datasets by furnishing fine-grained descriptions for HOI states and detailing the movements between two consecutive HOI states. Based on the proposed dataset, we design three state-level tasks to achieve fine-grained HOI modeling from different perspectives: **(1) Understanding:** None of the current tasks explicitly involve understanding an HOI state (*i.e.*, human pose with object pose) via textual descriptions. Thus, we are motivated to propose such a task and aim to achieve fine-grained understanding, as shown in Fig. 1-(a). **(2) Reasoning:** Building upon the understanding task, we further increase the level of difficulty by describing the next HOI state given the current HOI state and the overall HOI goal, as shown in Fig. 1-(b). **(3) Generation:** Beyond the understanding and reasoning tasks, we further explore the fine-grained action control, which aims to leverage the transformation descriptions to generate the next state from the current one, as shown in Fig. 1-(c).

To address the second problem, we introduce F-HOI, a novel unified framework to tackle diverse 3D HOI tasks. Specifically, F-HOI first integrates various input modalities, including 2D images, 3D object meshes, 3D HOI-Pose (comprising human and object poses), and textual descriptions, into a unified architecture. By employing different task instructions in the training phase, it progressively learns consistent HOI representations across 2D, 3D, and linguistic spaces, and realizes the mutual enhancement of different tasks. Once the model

is optimized, it can leverage the powerful language understanding capabilities inherent in the multi-modal large language model to adeptly execute diverse HOI tasks with flexible inputs.

Through extensive experiments, we demonstrate that F-HOI can effectively align HOI states of sequences with fine-grained semantic descriptions, adeptly tackling understanding, reasoning, and generation tasks, along with the traditional reconstruction task. In addition, our ablation studies reveal that our model designs, coupled with the proposed dataset and training strategies, could improve fine-grained 3D HOI modeling. Finally, as a pioneering work, we provide a comprehensive discussion to inspire future research in the related field.

In summary, the contributions of this work are three-fold:

- As far as we know, this is the first work to explore the problem of fine-grained semantic-aligned 3D HOI modeling. To tackle such a problem, we introduce a new dataset named Semantic-HOI to bridge the annotation gap present in current datasets by providing fine-grained descriptions for HOI states and the body movement between two consecutive states.
- To learn and evaluate the fine-grained HOI representation, we define three new state-level HOI tasks from the perspectives of understanding, reasoning, and generation. Furthermore, we present F-HOI, which empowers the MLLM to execute the above HOI tasks with flexible inputs.
- Extensive experiments show F-HOI can effectively align HOI states with fine-grained semantic descriptions. We hope our proposed dataset and tasks can bring in new perspectives to fine-grained semantic-aligned HOI modeling.

2 Related Work

2.1 Human-Object Interaction

Research in Human-Object Interaction (HOI) has traditionally focused on identifying interactions from images [4,16,22,31,74,75,81], 3D interaction reconstruction [6,17,52,62,66,73,83,88] and generation [12,18,19,25,60,64,67,71,89]. Some noteworthy contributions include Phosa [87], which geometrically reconstructs HOIs by utilizing contact priors from different body regions, and GOAL [55], which employs a conditional variational autoencoder (cVAE) to generate full-body motions for object grasping by estimating the grasping pose for the entire body. CHOIS [28] further extends the field by synthesizing HOI motions using a conditional diffusion model based on language descriptions and the initial state of the object and the human involved. On the other hand, the advent of 3D HOI assets [87] and datasets, which include visual recordings [17, 24, 25, 73], text annotations [29, 56] or both [2, 64], have promoted effective HOI modeling across various applications. However, existing models and datasets typically rely on coarse-grained descriptions for interactions, making it challenging to learn fine-grained semantic alignment. Our work represents the **first** attempt to address this limitation by constructing a new dataset with rich descriptions for body parts, objects, and their interactions, and proposing three new fine-grained state-level HOI tasks from different perspectives.

Table 1: Statistics of Semantic-HOI collected from three existing datasets.

Datasets	2D Image	3D HOI		Text Descriptions		# Unified HOI Pairs
		# 3D Object	3D Human Pose	Goal Descriptions	Fine-grained Descriptions	
GRAB [56]	✗	✓	✓	✓	✗	1187
CHAIRS [24]	✓	✓	✓	✓	✗	3368
BEHAVE [2]	✓	✓	✓	✓	✗	15886
Semantic-HOI	✓	✓	✓	✓	✓	20441

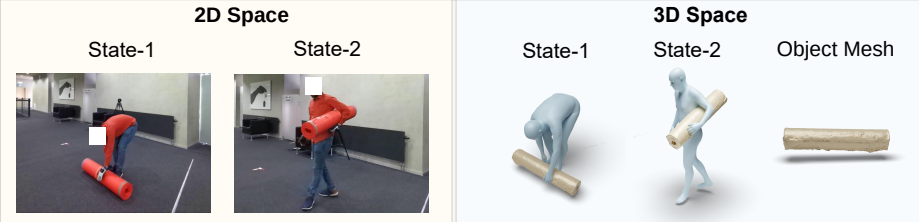
2.2 Multimodal Large Language Models

Large Language Models (LLMs) are rapidly emerging as powerful tools across various domains. While leading models like OpenAI's ChatGPT [44] and GPT-4 [45] remain proprietary, the availability of open-source LLMs such as Vicuna [8], LLaMA [59], and Alpaca [57] is enabling researchers to engage in multimodal research. In general, there are two technical paradigms to leverage the LLMs to solve multi-modality tasks: Firstly, LLMs can serve as effective decision-making agents by interfacing with task-specific models through API calls [21, 40, 48, 53, 61, 79, 80]. Through carefully designed prompt engineering or instruction tuning, LLMs can generate API calls to address multi-modal tasks. However, this approach may not fully comprehend the intricacies of task-specific modalities, leading to potential failures when dealing with complex scenes. Secondly, an advanced approach is to map modality-specific representations into the language embedding space of the LLM. Recent works like LLaVA [37, 38] and MiniGPT-4 [94] incorporate pre-trained visual encoders to obtain image features and train projection layers to align visual representations with the language space of LLM. This approach can also be extended to speech generation [84], image generation [23, 91], video understanding [30, 85], and other perception tasks [13, 27, 49, 70, 86], providing a more comprehensive understanding of multi-modal data. Among these, ChatPose [14] is most relevant to our work, as it explores the use of 3D body pose as a new modality for LLMs to process. However, our work goes further by comprehensively considering humans, objects, and their interactions. More importantly, we emphasize exploring the problem of fine-grained semantic-aligned HOI modeling. To achieve this, we leverage multi-modal instructions to empower MLLMs to complete fine-grained HOI tasks, thereby demonstrating significant potential.

3 Dataset

3.1 Motivation

As illustrated in Tab. 1, existing datasets focus on different subsets of objects and interactions, while providing only goal descriptions for long-term interaction processes. These limitations restrict the effectiveness of models in handling diverse scenes and achieving fine-grained semantic alignments for HOIs. To address these dataset gaps, we introduce a novel dataset called Semantic-HOI, characterized by two primary features: (1) **Diverse objects**. We aggregate and unify data from three established datasets to incorporate a wider range of objects;



Language Space

1. **Goal Description:** pick up yoga mat
2. **Fine-grained Description for State-1:** Bending at waist, holding yoga mat with both hands, preparing to unroll or pick up mat on floor
3. **Fine-grained Description for State-2:** Walking forward, head slightly downward, carrying a rolled yoga mat under left arm, right leg slightly forward.
4. **Transformation Description between State-1 and State-2:** ***Whole body movement:*** a bent-over posture to an upright stance; ***Head movement:*** a shift from looking down at the object to facing forward; ***Arms movement:*** The right arm goes from being bent and engaged in activity to hanging straight down, and the left arm now supports the object's weight; ***Hands movement:*** The hands change from possibly gripping or unrolling the mat to holding the object while walking; ***Legs movement:*** The legs' movement transforms from a static position, providing stability for bending over, to dynamic walking, with one leg forward, indicating motion; ***Feet movement:*** There's a change from the feet being planted on the ground for balance to one foot being in mid-step; ***Object trajectory:*** The object goes from a potential stationary or unrolling phase on the ground to being lifted and carried alongside the body; ***Human-object interaction difference:*** Contact with the object shifts from a more distributed and active manipulation using both hands and potentially the body to a more passive but secure holding pattern as the person walks.

Fig. 2: Illustration of the annotations in Semantic-HOI for a paired HOI sample.

(2) **Detailed descriptions**. Instead of directly describing long-term HOI sequences, which demand extensive language to capture various redundant action changes, we provide detailed descriptions for each HOI state and highlight the body movements between consecutive HOI states.

3.2 Data Collection

We collect data from 3 existing datasets (GRAB [56], CHAIRS [24] and BEHAVE [2]), while carefully considering the following data balancing principles:
Balanced Data for Interaction Diversity: Each dataset in Tab. 1 is video-based and each video contains a single interaction process. To ensure diversity in interaction, we randomly sample a subset of state pairs from these videos.
Balance Data for Images Discrepancies and Object Categories: Since the GRAB dataset lacks natural images, we need to render 3D HOIs into 2D images. Additionally, the CHAIRS dataset predominantly features interactions with a diverse range of chairs. To address the variations in data distribution between rendered and natural images, and to ensure the super-category diversity of the interactive objects, we have restricted the number of samples drawn from

Table 2: Comparison with related works about linking textual descriptions and human pose. Our dataset and annotation strategy are the first to explore the problem of fine-grained semantic-aligned 3D human-object interaction modeling.

Method	Human	Object	Contact	State	Transition	Fine-grained
PoseScript [10]	✓	×	×	✓	×	×
PoseFix [11]	✓	×	×	×	✓	✓
Ours	✓	✓	✓	✓	✓	✓

both the GRAB and CHAIRS datasets to maintain balance. In contrast, we have included an increased number of samples from the BEHAVE dataset.

3.3 Dataset Construction

Annotations. As depicted in Fig. 2, the annotations of an HOI pair comprises the following components:

1. 2D images for the current state and the next state.
2. HOI poses for the current state and the next state, along with object mesh.
3. Goal description for the action.
4. Fine-grained descriptions for both the current state and the next state.
5. Transformation descriptions detailing the changes between the current state and the next state.

where components 1-3 can be derived from the original datasets. For components 4-5, we prompt GPT-4V [45] using the given 2D images for annotations. During this process, we meticulously design the formats for fine-grained descriptions 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. Based on the above prompts, we can offer both part-level state descriptions and part-level movement descriptions. We also provide action descriptions to supplement ambiguous and incomplete goal descriptions.

Statistics Analysis. Considering the potential for response errors in GPT-4V, we conduct manual verification to filter out improperly formatted fine-grained descriptions. In total, our Semantic-HOI comprises $20,441$ pairs, with $1,187$ from GRAB, $3,368$ from CHAIRS, and $15,886$ from BEHAVE. Furthermore, following BEHAVE, we split Semantic-HOI into about 70% for training and 30% for testing to show the potential of fine-grained semantic-aligned HOI.

3.4 Discussion

There are several related works worth discussing, which focus on linking textual descriptions to human poses. We have summarized key differences in Tab. 2: (1) PoseScript [10] utilizes generic rules to annotate textual descriptions from state-level 3D keypoints. In contrast, utilizing GPT-4V with well-designed prompts for annotation is more simple and fine-grained. Additionally, our annotation introduces state-to-state human and object transitions, as well as interactions, which PoseScript overlooks. (2) Although PoseFix [11] annotates text descriptions of human state transitions, it similarly overlooks object state transitions and interaction transitions, which are critical and unique for the HOI problem.

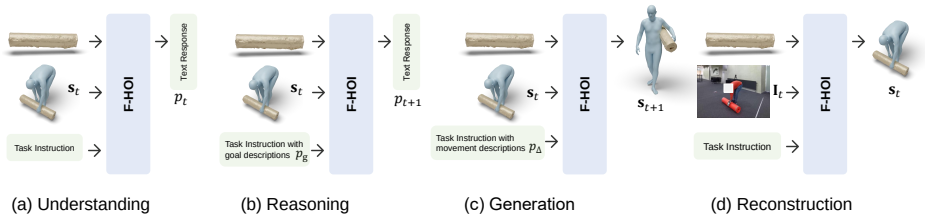


Fig. 3: Input and output definitions for each task.

4 Fine-grained 3D HOI Tasks

Motivation. Building on Semantic-HOI introduced in Sec. 3, we propose three novel tasks to showcase fine-grained semantic-aligned HOI, motivated by varying objectives: **(1) Understanding:** The task of captioning an HOI state (*i.e.*, the poses of both the human and the object) with textual descriptions serve as the most straightforward method for demonstrating alignment. This approach addresses a gap often overlooked in existing tasks. **(2) Reasoning:** Building upon the understanding task, the depth of the problem can be extended to reason about the next HOI state via textual descriptions, when knowing the action goal. **(3) Generation:** Moving forward, beyond simply considering the current or next state, comprehending the fine-grained movement descriptions to generate the next state based on the current one is promising.

Problem Definitions. As shown in Fig. 3, conditioned by the object mesh and each task instruction, the proposed three tasks, along with the traditional reconstruction task, can be formulated as follows: **(1) Understanding.** Given the t-th HOI state $\mathbf{s}_t$, the model aims to produce the corresponding fine-grained textual descriptions p_t. Specifically, the HOI state $\mathbf{s}_t$ can be formulated as $(M(\theta, \beta), O)$. $M(\theta, \beta)$ is the human state obtained by a parametric human body model SMPL M [41]. θ and β are pose and shape parameters, respectively. Following [14], β is by default set to zeros, corresponding to the average body shape. O is the object state represented by 6 degrees of freedom (6DoF) object pose (*i.e.*, translations and rotations). **(2) Reasoning.** Given the goal descriptions p_g (e.g., a person picks up a backup) and the t-th HOI state $\mathbf{s}_t$, we expect the model to reason the next state's text descriptions p_{t+1}. **(3) Generation.** Given the t-th HOI state $\mathbf{s}_t$ and the movement descriptions p_Δ, the model needs to generate the next HOI state $\mathbf{s}_{t+1}$. **(4) Reconstruction.** Given the 2D image $\mathbf{I}_t$ at t-th state, the model output the HOI state $\mathbf{s}_t$. However, directly outputting the object pose is challenging. We perform object-conditioned human reconstruction, which inputs the object pose at t-th state and outputs the human pose at t-th state.

5 Model

5.1 Motivation

As defined in Sec. 4, the consistency across the four tasks essentially involves learning fine-grained alignments across 2D, 3D, and language spaces. Therefore,

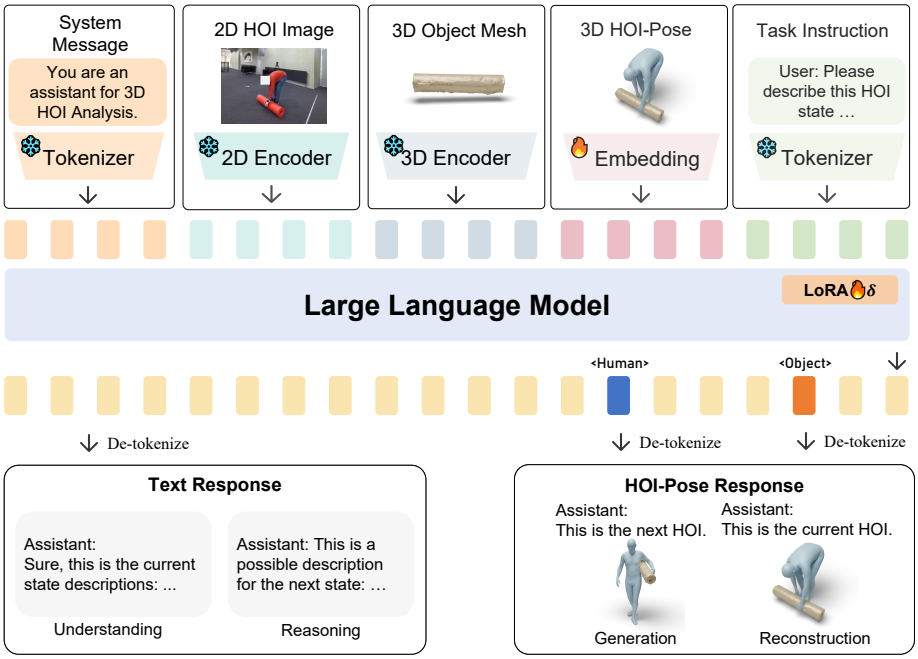


Fig. 4: Overview of our F-HOI framework, which contains the three components: multi-modal encoders, a large language model, and task-specific projectors. Based on different task instructions, F-HOI could support multi-modal inputs and complete diverse HOI tasks, covering understanding, reasoning, generation, and reconstruction tasks.

it is crucial to use a single model to unify the formulations of these four tasks and stimulate potential mutual benefits among them, which is demonstrated in our ablation study. On the other hand, due to the complexity of new tasks, previous technical paradigms [6,17,64,66,73,88,89] cannot meet the requirements of tasks, which necessitate semantic comprehension and cognition capabilities for handling lengthy sentences in fine-grained descriptions. Based on the above motivations, we propose to empower the Multi-modal Large Language Model to complete the proposed HOI tasks with flexible inputs.

5.2 Architecture

As illustrated in Fig. 4, based on different task instructions, F-HOI could support multi-modal inputs and complete diverse HOI tasks containing the three components: multi-modal encoders, LLM Backbone, and task-specific projectors.
Multi-modal Encoders. We leverage different modality encoders to project each input modality into tokens that align together within the language space of the LLM backbone. Specifically, (1) We utilize the SentencePiece tokenizer [26] to encode textual descriptions, including task instructions, goal descriptions p_g in the reasoning task, and fine-grained movement descriptions p_Δ in the generation task, as in Fig. 3; (2) We employ the frozen CLIP image encoder and one

additional projection layer to encode 2D HOI images; (3) We utilize the frozen 3D point encoder in Uni3D [93] with a trainable projection layer to encode 3D object meshes; (4) For HOI-Pose, we use separate projection layers to project the human pose and object pose into the hidden space of the LLM backbone.
LLM Backbone. We utilize Vicuna-7B [92] as the LLM backbone and employ LoRA [20], which initializes a trainable matrix for fine-tuning the LLM.
De-tokenlization. (1) For text response, we also utilize the SentencePiece detokenizer to decode all the text tokens into textual descriptions; (2) For HOI-Pose response, we utilize a trainable human-pose projection layer to decode the special token `<Human>` into human pose, while employing another trainable object-pose projection layer to decode the special token `<Object>` into object pose.

5.3 Training Pipeline

Pretraining for Alignments. Thanks to the versatility of our model in handling various input and output formats, we are able to utilize multiple related datasets for pretraining, thereby improving alignments across different modalities. Specifically, acknowledging the critical role of human pose diversity in HOI tasks, we engage with extensive pose estimation and description datasets to facilitate alignment between images and human poses, as well as text and human poses. To this end, we employ the COCO [36] and Posescript [10] datasets, respectively. We convert these two datasets into different question-answer formats and optimize the model using the following loss functions:

$$\mathcal{L} = \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{pose}}$$

where $\mathcal{L}_{\text{text}}$ is the cross-entropy loss typically applied for prefix language modeling such as GPT. $\mathcal{L}_{\text{pose}} = \|\theta_{\text{gt}} - \theta_{\text{pred}}\|$ is the L1 loss computed between predicted human pose parameters and ground-truth ones.
Multi-Task Instruction Tuning. In this stage, we convert the proposed dataset into task-specific instruction-following format, including understanding, reasoning, generation, and reconstruction tasks. By joint training these tasks, we optimize the model using the following loss functions:

$$\mathcal{L} = \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{hoi}}$$

where $\mathcal{L}_{\text{hoi}} = \|\Delta\theta_{\text{gt}} - \Delta\theta_{\text{pred}}\| + \|\Delta O_{\text{gt}} - \Delta O_{\text{pred}}\|$ is the L1 loss to minimize the translation difference between predicted and ground-truth human and object pose parameters from the start state to the next state, which only applies to the generation (current state as start state) and reconstruction (default state as start state). Therefore, $\mathcal{L}_{\text{hoi}}$ is computed as the offset. This approach could benefit the generation task, which we refer to as **offset regression**.

6 Experiment

6.1 Experimental Setup

Evaluation Metric. Since the output of understanding and reasoning tasks are textual descriptions, we employ two commonly used metrics from the NLP

field for evaluations: (1) BLUE-4 [46] analyzes the co-occurrences of 4-grams between the predicted and ground-truth sentences. (2) ROUGH [33] examines the adequacy and fidelity of the predicted sentences based on recall rate. In contrast, for evaluating generation and reconstruction tasks, we follow previous works [2, 69, 87] to utilize the Chamfer distance for both humans and objects. Specifically, we decouple the human pose into different body parts for evaluation, including the head, two arms, two hands, and two legs, to better demonstrate the performance guided by fine-grained semantic descriptions.

Baseline. Since this work focuses on entirely new HOI tasks that existing models have not been able to tackle, we choose the base multi-modal large language model in F-HOI for comparison. Specifically, we use our Semantic-HOI to finetune LLaVA-1.5V-7B [38] as our baseline model, which incorporates Vicuna-7B [92] as the LLM backbone with a CLIP image encoder [51] for visual encoding. Considering the original LLaVA only takes images and textual descriptions as inputs, we further incorporate 3D HOI-Pose embedding to process HOI poses, enabling the completion of our tasks. In contrast, we do not input the object mesh and instead output the HOI-Pose as a textual response. To obtain a complete and precise HOI-Pose, we decouple the human pose parameters and the object pose, and then perform multi-turn conversation for training and the batch inference to query them separately.

Implementation Details. We employ LLaVA-1.5V-7B to initialize the model weight. During training, we freeze the CLIP image encoder while fine-tuning the LLM using LoRA [20]. Additionally, we train projection layers to adapt the LLM to our proposed HOI tasks. For fine-tuning with LoRA, we set the rank to 128 and the alpha to 256. We employ AdamW [42] for network optimization, with a learning rate of $2e-4$ and weight decay of 0. During training, we utilize 8 Nvidia A100 GPUs, each with 80G of memory. We set the batch size to 16 for each GPU and configure a gradient accumulation step of 1.

6.2 Main Results

For quantitative results, as illustrated in Tab. 3, Tab. 4, Tab. 5, and Tab. 6, we conduct a comprehensive comparative analysis of our proposed model, F-HOI, across all four tasks against the established baseline, utilizing our Semantic-HOI dataset. F-HOI significantly and consistently outperforms its baseline, demonstrating its effectiveness in achieving fine-grained semantic alignment. Moreover, we provide the qualitative results as shown in Fig. 5 for the generation task. Overall, F-HOI can learn rich semantic representations at the state level, adeptly tackling our proposed understanding, reasoning, and generation tasks.

Table 3: Understanding Task.

Method	BLEU-4↑ [46]	ROUGE↑ [33]
Baseline	20.09	35.20
F-HOI	**26.78**	**45.29**

Table 4: Reasoning Task.

Method	BLEU-4↑ [46]	ROUGE↑ [33]
Baseline	19.51	35.69
F-HOI	**25.56**	**41.84**

Table 5: Generation Task.

Method	Head	Left Arm	Right Arm	Left Hand	Right Hand	Left leg	Right Leg	Object	Averaged↓
Baseline	36.8	28.1	34.0	55.1	73.7	23.4	30.9	72.9	44.4
F-HOI	**16.8**	**13.5**	**16.2**	**35.2**	**42.3**	**10.6**	**14.1**	**34.8**	**22.9**

Table 6: Object-conditioned Reconstruction Task.

Method	Head	Left Arm	Right Arm	Left Hand	Right Hand	Left leg	Right Leg	Averaged↓
Baseline	45.8	37.3	42.8	62.5	79.4	32.9	39.4	48.6
F-HOI	**20.8**	**15.1**	**19.5**	**38.7**	**50.4**	**11.6**	**16.9**	**24.7**

6.3 Ablation Study

Offset Regression. For the generation task, F-HOI designs an offset-based regression method, which expects the model to predict offsets from the input HOI-Pose, thereby supervising the HOI-Pose offsets to achieve better alignment with movement descriptions. The results in Tab. 7 indicate that this regression approach can notably enhance the generation task by achieving improved alignment with movement descriptions.

Table 7: Effect of offset regression on different HOI tasks. We adopt the BLUE-4 to evaluate understanding and reasoning tasks and use the averaged Chamfer distance for generation and reconstruction tasks.

Offset	Understanding↑	Reasoning↑	Generation↓	Reconstruction↓
✗	**26.91**	24.73	27.9	25.5
✓	26.78	**25.56**	**22.9**	**24.7**

Image-to-Pose Alignment. Thanks to the input and output flexibility of our model, we can use large-scale image-pose paired dataset COCO [36] to perform image-to-pose alignment. The results in Tab. 8, #1 vs. #2, demonstrate significant improvements in both the reconstruction and generation tasks. The enhanced human pose diversity is crucial for effective HOI, thereby contributing to the observed improvements.

Text-to-Pose Alignment. We utilize the text-pose paired dataset PoseScript [10] to achieve text-to-pose alignment. The results in Tab. 8, comparing methods #2 and #3, demonstrate notable benefits for understanding and reasoning tasks. This approach effectively aligns poses with diverse descriptions, contributing to improved performance.

Multi-task Joint Training. In the previous ablation studies, a consistent phenomenon is observed: an improvement in one task's performance often leads to improvements in other tasks as well. Furthermore, we conduct ablations focusing on single tasks. The results in Tab. 9 demonstrate the presence of mutual benefits across different tasks in multi-task training, significantly outperforming single-task training. In addition, training with multiple modalities input of the same sample (e.g., images, HOI-Pose, and textual descriptions) can also provide more information to improve performance.

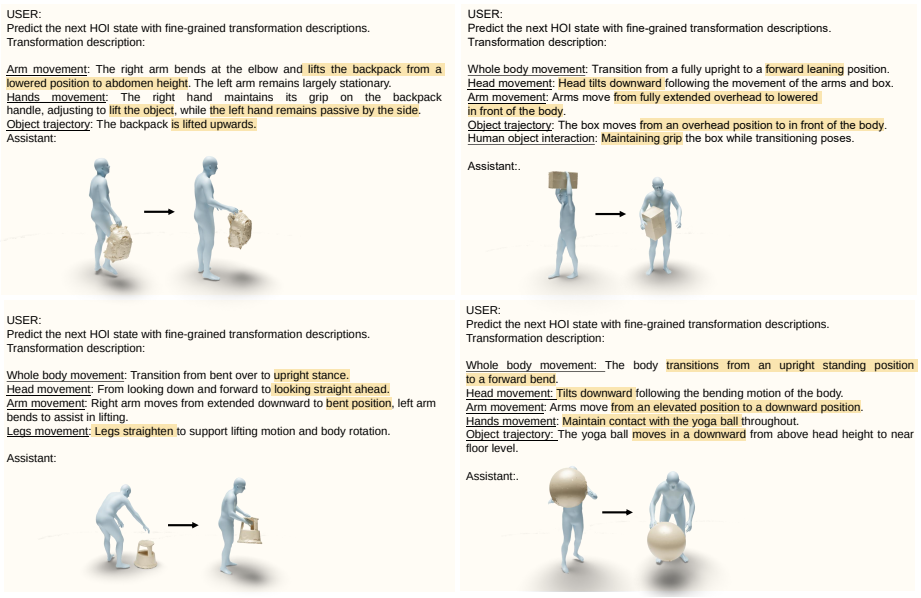


Fig. 5: Qualitative results of F-HOI on generation task.

Table 8: Effect of image-to-pose and text-to-pose alignment on different HOI tasks.

No.	Image-to-Pose	Text-to-Pose	Understanding↑	Reasoning↑	Generation↓	Reconstruction↓
#1	✗	✗	21.21	20.98	32.7	33.9
#2	✓	✗	23.76	23.43	24.8	25.8
#3	✓	✓	**26.78**	**25.56**	**22.9**	**24.7**

7 Discussion

State-by-state Generation for an HOI process. Although the primary focus of this work is to align fine-grained semantic details with HOI at the state level, we demonstrate the potential of this paradigm to generate long sequences for the interaction process while maintaining a detailed understanding of each intermediate state and the transitions between states, as shown in Fig 6.

Failure Case Analysis. We show three types of failure cases in our method, as shown in Fig. 7. **(1) Interpenetration:** F-HOI employs fine-grained textual supervision (e.g., body part with contact) for implicit human-object optimization. Compared with previous methods [6, 17, 62, 64, 66, 73, 73, 88, 89], it is conceptually simple by eliminating complex structured HOI modeling and explicit contact supervision [12, 67]. However, in cases where there are conflicts between human actions and objects, the interpenetration still persists, as illustrated in Fig. 7-(a). **(2) Physics Gap:** Since F-HOI does not explicitly incorporate the physical laws [63, 67, 68, 72, 82], the generation of the next HOI state merely aligns with linguistic descriptions without constraints imposed by physical reality. For instance, as depicted in Fig. 7-(b), without contact, the "keyboard" would actually fall to the ground due to gravity. **(3) Difficulty with Complex Movements:**

Table 9: Effect of joint training across multiple tasks on each HOI task.

Task	Understanding↑	Reasoning↑	Generation↓	Reconstruction↓
Understanding	24.67	-	-	-
Reasoning	-	23.89	-	-
Generation	-	-	24.0	-
Reconstruction	-	-	-	26.8
All	**26.78**	**25.56**	**22.9**	**24.7**

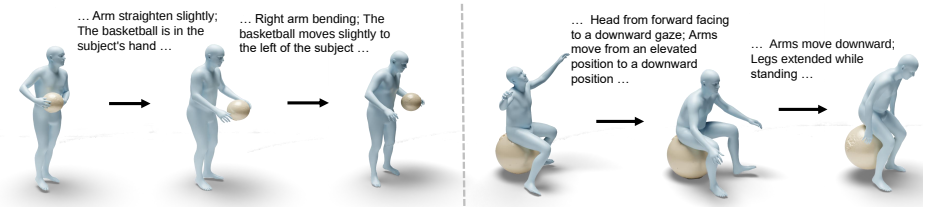


Fig. 6: We show that F-HOI has the potential to utilize fine-grained descriptions at the state level for performing sequence generation state-by-state.

F-HOI fails to generate the next HOI state when the trajectory of object states between the current and next states is complex or highly variable, as exemplified in Fig. 7-(c). This is primarily due to our dataset lacking sufficient examples of long-term and complex movements. One direct solution to address this issue is to decompose the description of movements into multiple fine-grained stages.

Limitations. Overall, our work serves as a pioneering work that provides the community with a new perspective for fine-grained semantic-aligned 3D human-object interaction modeling. However, as an early-stage effort, our work leaves ample room for further exploration in this field. Here, we discuss several limitations to inspire future research: **(1)** As shown in Fig. 3, our proposed tasks require inputs including a 3D object mesh, 3D HOI-Pose, and textual descriptions. These input requirements significantly hurt the convenience of inference. Reducing the strict requirements of input modalities is an area worth exploring. **(2)** Our work only evaluates the effectiveness of fine-grained textual descriptions and HOI-Pose alignment in closed-set scenarios within existing datasets. However, such a model lacks generalization to open-set scenarios, which is limited by interaction diversity and unseen object meshes. **(3)** Based on our task and problem definitions, F-HOI can only perform long sequence generation state-by-state through fine-grained textual descriptions for addressing the HOI motion generation task. This approach introduces complexity and error accumulation, and it does not address the issue of smooth transitions between states. **(4)** The design of F-HOI follows a simple and intuitive principle and could be considered as a baseline model. It may perform poorly compared to previous methods [12, 69, 87, 89] in generation and reconstruction tasks. **(5)** We expect our model to predict hand parameters to better demonstrate the alignment between fine-grained textual descriptions and HOIs. However, our preliminary results indicate that our model underperforms in capturing the details of the hands. **(6)** For understanding and reasoning tasks, F-HOI heavily relies on the priors of large language models. Despite showing significant potential, we still identify

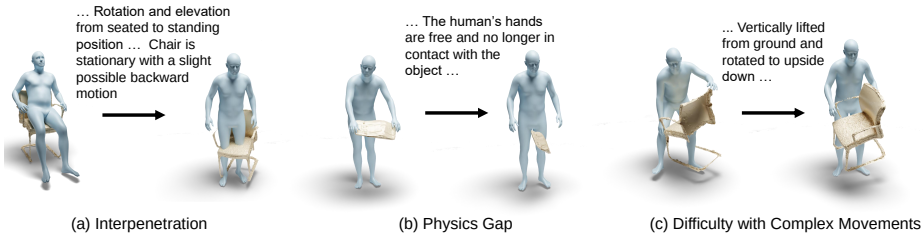


Fig. 7: We show three types of failure cases in our method.

several understanding and reasoning errors, such as inaccurate judgments of interactions and incorrect assessments of the spatial relationships between body parts. These limitations arise from the restricted data volume used to align HOI-Pose with fine-grained descriptions, while the richness and quality of the textual descriptions also affect performance [5].

Future Directions. (1) Flexiable Input Modality. As discussed in the limitations above, reducing the required input modalities is worth considering. For instance, the object mesh could potentially be obtained through a text-to-3D approach [7, 32, 50]. Furthermore, the initial HOI-Pose could be directly derived from image input, as the human SMPL parameters and object 6DoF pose can be obtained by other powerful models [3, 9, 34, 35, 65, 77, 78]. **(2) Increase Data Scale.** Our Semantic-HOI currently covers only three datasets with a limited number of samples. Merging more HOI datasets [39, 90], scaling up the number of samples, and enriching the textual descriptions are also worth exploring. Moreover, exploring the hierarchy of human body part states, object states, and actions are promising and meaningful [1]. **(3) Diverse Model Architectures.** Due to the complexity of new tasks and the limited data samples, our model is built on the Multi-modal Large Language Model, which brings semantic comprehension and cognitive capabilities for handling lengthy sentences in fine-grained descriptions. However, compared to previous HOI models, our model has a significantly larger amount of parameters, making it less lightweight for addressing HOI tasks. Thus, as data scales up, exploring other model architectures [58] also becomes an important consideration.

8 Conclusion

This paper proposes the overlooked challenge of fine-grained semantic-aligned 3D human-object interaction (HOI), which is inadequately addressed by current HOI datasets and models. To bridge this gap, we introduce Semantic-HOI, a new dataset featuring over 20K meticulously annotated HOI state pairs, each equipped with detailed descriptions and corresponding body movements between consecutive states. Leveraging this dataset, we formulate three state-level HOI tasks aimed at achieving fine-grained semantic alignment within the HOI sequence. Moreover, we present F-HOI, which empowers the MLLM to proficiently tackle the proposed HOI tasks. Extensive experiments showcase F-HOI's prowess in aligning HOI states with fine-grained semantic descriptions.

Acknowledgement

The work is partially supported by the Young Scientists Fund of the National Natural Science Foundation of China under grant No.62106154, by the Natural Science Foundation of Guangdong Province, China (General Program) under grant No.2022A1515011524, and by Shenzhen Science and Technology Program JCYJ20220818103001002, and by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong (Shenzhen).

References

1. Belkhale, S., Ding, T., Xiao, T., Sermanet, P., Vuong, Q., Tompson, J., Chebotar, Y., Dwibedi, D., Sadigh, D.: Rt-h: Action hierarchies using language. arXiv preprint arXiv:2403.01823 (2024) 14
2. Bhatnagar, B.L., Xie, X., Petrov, I.A., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Behave: Dataset and method for tracking human object interactions. In: CVPR (2022) 1, 3, 4, 5, 10
3. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: ECCV (2016) 14
4. Chao, Y.W., Wang, Z., He, Y., Wang, J., Deng, J.: Hico: A benchmark for recognizing human-object interactions in images. In: ICCV (2015) 3
5. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793 (2023) 14
6. Chen, Y., Huang, S., Yuan, T., Qi, S., Zhu, Y., Zhu, S.C.: Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In: CVPR (2019) 3, 8, 12
7. Chen, Z., Wang, F., Wang, Y., Liu, H.: Text-to-3d using gaussian splatting. In: CVPR (2024) 14
8. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), https://lmsys.org/blog/2023-03-30-vicuna/ 4
9. Corsetti, J., Boscaini, D., Oh, C., Cavallaro, A., Poiesi, F.: Open-vocabulary object 6d pose estimation. In: CVPR (2024) 14
10. Delmas, G., Weinzaepfel, P., Lucas, T., Moreno-Noguer, F., Rogez, G.: Posescript: 3D human poses from natural language. In: ECCV (2022) 6, 9, 11
11. Delmas, G., Weinzaepfel, P., Moreno-Noguer, F., Rogez, G.: Posefix: correcting 3d human poses with natural language. In: CVPR (2023) 6
12. Diller, C., Dai, A.: Cg-hoi: Contact-guided 3d human-object interaction generation. arXiv preprint arXiv:2311.16097 (2023) 3, 12, 13
13. Feng, Y., Lin, J., Dwivedi, S.K., Sun, Y., Patel, P., Black, M.J.: Posegpt: Chatting about 3d human pose. arXiv preprint arXiv:2311.18836 (2023) 4
14. Feng, Y., Lin, J., Dwivedi, S.K., Sun, Y., Patel, P., Black, M.J.: Chatpose: Chatting about 3d human pose. In: CVPR (2024) 4, 7
15. Geijtenbeek, T., Pronost, N.: Interactive character animation using simulated physics: A state-of-the-art review. In: Computer graphics forum. vol. 31, pp. 2492–2515. Wiley Online Library (2012) 1

16. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: CVPR (2018) 3
17. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3d human pose ambiguities with 3d scene constraints. In: ICCV (2019) 3, 8, 12
18. Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M.J.: Populating 3d scenes by learning human-scene interaction. In: CVPR (2021) 3
19. Holden, D., Komura, T., Saito, J.: Phase-functioned neural networks for character control. ACM TOG **36**(4), 1–13 (2017) 3
20. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022) 9, 10
21. Huang, R., Li, M., Yang, D., Shi, J., Chang, X., Ye, Z., Wu, Y., Hong, Z., Huang, J., Liu, J., Ren, Y., Zhao, Z., Watanabe, S.: AudioGPT: Understanding and generating speech, music, sound, and talking head. arXiv preprint arXiv:2304.12995 (2023) 4
22. Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., Zhu, S.C.: Diffusion-based generation, optimization, and planning in 3d scenes. In: CVPR (2023) 3
23. Huang, Y., Xie, L., Wang, X., Yuan, Z., Cun, X., Ge, Y., Zhou, J., Dong, C., Huang, R., Zhang, R., et al.: Smartedit: Exploring complex instruction-based image editing with multimodal large language models. arXiv preprint arXiv:2312.06739 (2023) 4
24. Jiang, N., Liu, T., Cao, Z., Cui, J., Zhang, Z., Chen, Y., Wang, H., Zhu, Y., Huang, S.: Full-body articulated human-object interaction. In: ICCV (2023) 3, 4, 5
25. Jiang, N., Zhang, Z., Li, H., Ma, X., Wang, Z., Chen, Y., Liu, T., Zhu, Y., Huang, S.: Scaling up dynamic human-scene interaction modeling. In: CVPR (2024) 3
26. Kudo, T., Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226 (2018) 8
27. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: LISA: Reasoning segmentation via large language model. arXiv preprint arXiv:2308.00692 (2023) 4
28. Li, J., Clegg, A., Mottaghi, R., Wu, J., Puig, X., Liu, C.K.: Controllable human-object interaction synthesis. arXiv preprint arXiv:2312.03913 (2023) 3
29. Li, J., Wu, J., Liu, C.K.: Object motion guided human motion synthesis. ACM Transactions on Graphics (TOG) (2023) 3
30. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: VideoChat: Chat-centric video understanding. arXiv preprint arXiv:2205.06355 (2023) 4
31. Liao, Y., Zhang, A., Lu, M., Wang, Y., Li, X., Liu, S.: Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In: CVPR (2022) 3
32. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: CVPR (2023) 14
33. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004) 10
34. Lin, J., Liu, L., Lu, D., Jia, K.: Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In: CVPR (2024) 14
35. Lin, J., Zeng, A., Wang, H., Zhang, L., Li, Y.: One-stage 3d whole-body mesh recovery with component aware transformer. In: CVPR (2023) 14
36. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014) 9, 11

37. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023) 4
38. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023) 4, 10
39. Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., Yi, L.: Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In: CVPR (2022) 14
40. Liu, Z., He, Y., Wang, W., Wang, W., Wang, Y., Chen, S., Zhang, Q., Lai, Z., Yang, Y., Li, Q., Yu, J., et al.: InternGPT: Solving vision-centric tasks by interacting with chatbots beyond language. arXiv preprint arXiv:2305.05662 (2023) 4
41. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. In: ACM TOG (2015) 7
42. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019) 10
43. Mourot, L., Hoyet, L., Le Clerc, F., Schnitzler, F., Hellier, P.: A survey on deep learning for skeleton-based human animation. In: Computer Graphics Forum. vol. 41, pp. 122–157. Wiley Online Library (2022) 1
44. OpenAI: Introducing chatgpt (2022) 4
45. OpenAI: GPT-4 technical report. (2023) 4, 6
46. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics (2002) 10
47. Peng, X., Xie, Y., Wu, Z., Jampani, V., Sun, D., Jiang, H.: Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. arXiv preprint arXiv:2312.06553 (2023) 1
48. Pi, R., Gao, J., Diao, S., Pan, R., Dong, H., Zhang, J., Yao, L., Han, J., Xu, H., Kong, L., Zhang, T.: DetGPT: Detect what you need via reasoning. arXiv:2305.14167 (2023) 4
49. Pi, R., Yao, L., Gao, J., Zhang, J., Zhang, T.: Perceptiongpt: Effectively fusing visual perception into llm. arXiv preprint arXiv:2311.06612 (2023) 4
50. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3D using 2D diffusion (2022) 14
51. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021) 10
52. Savva, M., Chang, A.X., Hanrahan, P., Fisher, M., Nießner, M.: Pigraphs: learning interaction snapshots from observations. ACM TOG **35**(4), 1–12 (2016) 3
53. Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y.: HuggingGPT: Solving ai tasks with chatgpt and its friends in huggingface. arXiv preprint arXiv:2303.17580 (2023) 4
54. Stacey, J., Suchman, L.: Animation and automation–the liveliness and labours of bodies and machines. Body & Society **18**(1), 1–46 (2012) 1
55. Taheri, O., Choutas, V., Black, M.J., Tzionas, D.: Goal: Generating 4d whole-body motion for hand-object grasping. In: CVPR (2022) 3
56. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: Grab: A dataset of whole-body human grasping of objects. In: ECCV (2020) 1, 3, 4, 5
57. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model (2023), https://github.com/tatsu-lab/stanford_alpaca 4
58. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: ECCV. Springer (2022) 14

59. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 4
60. Wang, J., Xu, H., Xu, J., Liu, S., Wang, X.: Synthesizing long-term 3d human motion and interaction in 3d scenes. In: CVPR (2021) 3
61. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: VisionLLM: Large language model is also an open-ended decoder for vision-centric tasks. arXiv:2305.11175 (2023) 4
62. Wang, X., Li, G., Kuo, Y.L., Kocabas, M., Aksan, E., Hilliges, O.: Reconstructing action-conditioned human-object interactions using commonsense knowledge priors. In: 2022 International Conference on 3D Vision (3DV). pp. 353–362. IEEE (2022) 3, 12
63. Wang, Y., Lin, J., Zeng, A., Luo, Z., Zhang, J., Zhang, L.: Physhoi: Physics-based imitation of dynamic human-object interaction. arXiv preprint arXiv:2312.04393 (2023) 1, 12
64. Wang, Z., Chen, Y., Liu, T., Zhu, Y., Liang, W., Huang, S.: Humanise: Language-conditioned human motion generation in 3d scenes. NeurIPS (2022) 1, 3, 8, 12
65. Wen, B., Yang, W., Kautz, J., Birchfield, S.: Foundationpose: Unified 6d pose estimation and tracking of novel objects. In: CVPR (2024) 14
66. Weng, Z., Yeung, S.: Holistic 3d human and scene mesh estimation from single view images. In: CVPR (2021) 3, 8, 12
67. Xiao, Z., Wang, T., Wang, J., Cao, J., Zhang, W., Dai, B., Lin, D., Pang, J.: Unified human-scene interaction via prompted chain-of-contacts. arXiv preprint arXiv:2309.07918 (2023) 3, 12
68. Xie, K., Wang, T., Iqbal, U., Guo, Y., Fidler, S., Shkurti, F.: Physics-based human motion estimation and synthesis from videos. In: ICCV (2021) 12
69. Xie, X., Bhatnagar, B.L., Pons-Moll, G.: Chore: Contact, human and object reconstruction from a single rgb image. In: ECCV (2022) 10, 13
70. Xu, J., Zhou, X., Yan, S., Gu, X., Arnab, A., Sun, C., Wang, X., Schmid, C.: Pixel aligned language models. arXiv preprint arXiv:2312.09237 (2023) 4
71. Xu, J., Xu, H., Ni, B., Yang, X., Wang, X., Darrell, T.: Hierarchical style-based networks for motion synthesis. In: ECCV (2020) 3
72. Xu, S., Li, Z., Wang, Y.X., Gui, L.Y.: Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In: ICCV (2023) 12
73. Xu, X., Joo, H., Mori, G., Savva, M.: D3d-hoi: Dynamic 3d human-object interactions from videos. arXiv preprint arXiv:2108.08420 (2021) 3, 8, 12
74. Yang, J., Li, B., Yang, F., Zeng, A., Zhang, L., Zhang, R.: Boosting human-object interaction detection with text-to-image diffusion model. arXiv preprint arXiv:2305.12252 (2023) 3
75. Yang, J., Li, B., Zeng, A., Zhang, L., Zhang, R.: Open-world human-object interaction detection via multi-modal prompts. In: CVPR (2024) 3
76. Yang, J., Wang, C., Li, Z., Wang, J., Zhang, R.: Semantic human parsing via scalable semantic transfer over multiple label domains. In: CVPR (2023) 1
77. Yang, J., Zeng, A., Li, F., Liu, S., Zhang, R., Zhang, L.: Neural interactive keypoint detection. In: ICCV (2023) 14
78. Yang, J., Zeng, A., Zhang, R., Zhang, L.: Unipose: Detecting any keypoints. arXiv preprint arXiv:2310.08530 (2023) 14
79. Yang, R., Song, L., Li, Y., Zhao, S., Ge, Y., Li, X., Shan, Y.: GPT4Tools: teaching large language model to use tools via self-instruction. arXiv preprint arXiv:2305.18752 (2023) 4

80. Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., Wang, L.: MM-ReAct: Prompting chatgpt for multimodal reasoning and action. arXiv preprint arXiv:2303.11381 (2023) 4
81. Yuan, H., Wang, M., Ni, D., Xu, L.: Detecting human-object interactions with object-guided cross-modal calibrated semantics. In: AAAI (2022) 3
82. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: Physdiff: Physics-guided human motion diffusion model. In: ICCV (2023) 12
83. Zanfir, A., Marinoiu, E., Sminchisescu, C.: Monocular 3d pose and shape estimation of multiple people in natural scenes: The importance of multiple scene constraints. In: CVPR (2018) 3
84. Zhang, D., Li, S., Zhang, X., Zhan, J., Wang, P., Zhou, Y., Qiu, X.: SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. arXiv preprint arXiv:2305.11000 (2023) 4
85. Zhang, H., Li, X., Bing, L.: Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023) 4
86. Zhang, H., Li, H., Li, F., Ren, T., Zou, X., Liu, S., Huang, S., Gao, J., Zhang, L., Li, C., et al.: Llava-grounding: Grounded visual chat with large multimodal models. arXiv preprint arXiv:2312.02949 (2023) 4
87. Zhang, J.Y., Pepose, S., Joo, H., Ramanan, D., Malik, J., Kanazawa, A.: Perceiving 3d human-object spatial arrangements from a single image in the wild. In: ECCV (2020) 3, 10, 13
88. Zhang, S., Zhang, Y., Bogo, F., Pollefeys, M., Tang, S.: Learning motion priors for 4d human body capture in 3d scenes. In: ICCV (2021) 3, 8, 12
89. Zhang, X., Bhatnagar, B.L., Starke, S., Guzov, V., Pons-Moll, G.: Couch: Towards controllable human-chair interactions. In: ECCV (2022) 3, 8, 12, 13
90. Zhang, X., Bhatnagar, B.L., Starke, S., Petrov, I., Guzov, V., Dhamo, H., Pérez-Pellitero, E., Pons-Moll, G.: Force: Dataset and method for intuitive physics guided human-object interaction. arXiv preprint arXiv:2403.11237 (2024) 14
91. Zheng, K., He, X., Wang, X.E.: Minigpt-5: Interleaved vision-and-language generation via generative vokens. arXiv preprint arXiv:2310.02239 (2023) 4
92. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena. arXiv preprint arXiv:2306.05685 (2023) 9, 10
93. Zhou, J., Wang, J., Ma, B., Liu, Y.S., Huang, T., Wang, X.: Uni3d: Exploring unified 3d representation at scale. arXiv preprint arXiv:2310.06773 (2023) 9
94. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: enhancing vision-language understanding with advanced large language models. arXiv:2304.10592 (2023) 4