Human-Object Interaction from Human-Level Instructions

Zhen Wu Jiaman Li Pei Xu C. Karen Liu
Stanford University
{zhenwu,jiamanli,peixu,karenliu}@cs.stanford.edu

Figure 1. Given human-level instructions (top), we first use a diffusion-based motion generator to synthesize natural full-body human motion, finger motion, and object motion to accomplish the task (middle). Next, we employ a physics-based tracking method to imitate the generated interactions which produces physically plausible interactions (bottom). Please refer to our project page for more results.

Abstract

Intelligent agents must autonomously interact with the environments to perform daily tasks based on human-level instructions. They need a foundational understanding of the world to accurately interpret these instructions, along with precise low-level movement and interaction skills to execute the derived actions. In this work, we propose the first complete system for synthesizing physically plausible, long-horizon human-object interactions for object manipulation in contextual environments, driven by human-level instructions. We leverage large language models (LLMs) to interpret the input instructions into detailed execution plans. Unlike prior work, our system is capable of generating detailed finger-object interactions, in seamless coordination with full-body movements. We also train a policy to track generated motions in physics simulation via reinforcement learning (RL) to ensure physical plausibility of the motion. Our experiments demonstrate the effectiveness of our system in synthesizing realistic interactions with diverse objects in complex environments, highlighting its potential for real-world applications.

1. Introduction

Synthesizing physically plausible, long-horizon, human-object interactions in contextual environments is crucial for applications in computer graphics, embodied AI, and robotics. When given human-level language instructions, intelligent agents should be able to comprehend these high-level commands and map them to executable action sequences. For instance, when instructed to set up a workspace, an agent must understand the abstract concept of "workspace", identify the relevant furniture and objects in the room, and translate this concept into a series of concrete actions, such as moving and orienting a desk, placing a chair nearby, and positioning a monitor on the desk. In addition to understanding the instructions, the agent must execute actions that lead to realistic and functional human movements, including navigating cluttered spaces, manipulating various objects for the task, and appropriately releasing them after use.

Prior research has made significant progress in synthesizing human-scene interactions using large language model (LLM) planners [57]. However, most of these interactions have been limited to static objects without hand manipulation. Some studies [24] have demonstrated the manipulation of dynamic objects, but the resulting interactions still lack realistic finger movements. Recently, physics-based RL policies has demonstrated promising results in human-object interactions [31, 52]. However, generating long-horizon motions that interact with various objects in a scene-aware manner remains an ongoing challenge. Build upon these prior work, we introduce the first complete system capable of synthesizing physically accurate, long-horizon human-object interactions with synchronous full-body and intricate finger movements, from human-level language instructions.

Synthesizing coordinated full-body and finger movements from human-level instructions presents two main challenges. First, human-level instructions typically provide only a high-level task outline, requiring common sense and world knowledge to translate them into precise object arrangements and detailed execution plans. Second, existing datasets with paired full-body and finger motion data are limited to interactions with small objects and lack locomotion during object manipulation [8, 43]. Although a recent dataset [25] includes locomotion and manipulation for large objects, it lacks detailed finger movements due to the difficulty of capturing both modalities simultaneously.

To tackle the challenge of comprehending human-level instructions, we utilize the latest advancements in LLMs, which are highly effective in interpreting high-level commands[28, 34]. However, LLMs are not well-suited for directly predicting precise scene layouts, such as the exact positions and orientations of objects [1, 16]. Instead of prompting LLMs to directly specify object placements, we use object spatial relationships as intermediate representations and ask LLMs to derive these relationships. We then propose an algorithm to calculate the precise object poses based on these relationships.

For the low-level motion synthesis, we take the approach of modeling human motion from real-world human data. To address the lack of large-scale object manipulation datasets with synchronized full-body and finger motions, one could use two separate datasets to train a full-body manipulation model and another detailed finger manipulation model. However, simply fusing the two motions together will cause glaring artifacts due to the disagreement in two separately trained models. Our solution is to break this process into three steps. We first generate full-body and object motions without detailed fingers. The object motion and rough wrist poses provide guidance for generating contact-consistent finger motion in the second step. Finally, we generate full-body and object motions again conditioned on precise finger motions. This approach enables our framework to successfully produce synchronized object motion, full-body motion, and finger motions without requiring a dataset that includes all data modalities.

The generated motion, however, is not guaranteed to be physically accurate. To ensure physical plausibility, we further use reinforcement learning (RL) to track the generated motion in a physics simulator. Our method adopts an importance sampling technique to improve the training efficiency of long sequence tracking and supports tracking diverse interactions with multiple distinct objects within a single sequence.

We evaluate our method through extensive quantitative and qualitative comparisons, as well as human studies, demonstrating that our approach outperforms all others. This underscores the effectiveness of our method and its potential for real-world applications.

2. Related Work

Contextual Interaction Synthesis. Recent work on modeling human-object interactions falls into two categories: interactions with static and dynamic objects. Synthesizing human motions in static 3D scenes has been extensively studied [2, 13, 14, 21, 53, 68]. Prior research has explored regression models [2, 14, 32, 48, 49] and diffusion models [18, 21, 54, 65] to generate human motions interacting with 3D scenes, such as sitting on a chair or lying down on a sofa. Some studies have also explored reinforcement learning methods to generate physically plausible human motions [15, 57, 70]. These works focus on static objects, whereas our work focuses on dynamic interactions.

There has been increasing attention on synthesizing human-object interactions with dynamic objects [6, 35, 55, 61]. OMOMO [25] collected a dataset for dynamic human-object interactions and proposed a framework to synthesize human motions from object motions. CHOIS [24] synthesizes interactions in 3D scenes from text and sparse waypoints. Despite these advancements, all of these works focus on full-body human motion generation without detailed finger motions. In this work, we aim to generate synchronized object motions, human motions, and finger motions.

Hand-Object Interaction Synthesis. Hand-object interaction synthesis has been extensively explored. In the realm of static pose generation, traditional methods often rely on physics-based control or optimization techniques to generate grasp candidates [27, 38, 41]. Some recent work leverages deep learning to directly estimate hand grasps from extensive interaction datasets [3, 5, 8, 19, 23, 43]. Beyond static grasping, another line of work also explores dynamic object manipulation [30, 64, 67].

Recent works further explore integrating whole-body motion and hand motion [44–46, 56]. IMoS [10] generates hand-object interaction using paired human motion and hand motion, while [4] leverage reinforcement learning to achieve physically plausible motion. However, these methods are constrained to manipulate small-sized objects in the GRAB dataset [43], unable to generalize to large objects. In this work, we aim to generate realistic full-body and finger motions for manipulating large objects. Since no such paired dataset exists, we leverage DexGraspNet [50] to generate realistic grasp poses, train a conditional diffusion model on GRAB [43] to generate approaching and releasing finger motions, and then integrate these with our full-body motion generation.

Physics-Based Motion Synthesis. Physics-based methods control a simulated character to perform motion in a simulator, ensuring the physical realism of the generated motion. To make the motion more natural, many approaches leverage motion capture data to produce movements that match the style of the data [29, 35–37]. DeepMimic [36] proposes to track motion capture data using RL. Subsequent works [37, 59] leverage generative adversarial networks (GANs) to learn motion styles. Recent research also explores this approach for human-object interactions, such as playing musical instruments [51, 60] or playing soccer [58]. PhysHOI [52] learns basketball-playing skills from human demonstrations. OmniGrasp [31] enables a character to grasp and move an object along a specified object trajectory. In this work, we leverage RL to track generated kinematic motion, successfully producing long-horizon motions that interact with various objects.

LLM-based Planning. LLMs have been widely used in reasoning, planning tasks [7, 16, 17, 47, 66] for their high-level planning capabilities. They have also been adopted in motion synthesis tasks [57, 61, 63, 69]. UniHSI [57] uses LLMs to extract action plans in a chain-of-contact format, guiding a low-level controller to synthesize corresponding motions. InterDreamer [61] employs LLMs to identify object parts that humans interact with and uses this information to retrieve initial human poses. In this work, we leverage LLMs to translate abstract human-level instructions into target scene maps and detailed execution plans, which guide the low-level interaction synthesis.

3. System Overview

Given a human-level instruction and a scene description, we develop a system that generates synchronized human motion and object motion to accomplish the task. An overview of our system is depicted in Fig. 2. The system consists of a high-level planner, a low-level motion generator, and a physics tracker. The high-level planner leverages LLMs to reason about instructions and the scene description, generating a *scene map* along with a detailed *execution plan*. The scene map specifies the desired positions and orientations of the objects, while the execution plan outlines the steps to interact with these objects. The low-level motion generator unites conditional diffusion models trained on separate datasets for full body motion and dexterous hand motion, producing synchronous object and human movements. The physics tracker then uses RL to track the generated motion in a physics simulator, ensuring physical plausibility.

4. High-Level LLM Planner

Inputs to the high-level planner consist of an initial scene description and human-level instructions. The scene description includes a list of objects in a 3D scene and their layout, while the human-level instruction describes a task that requires grounded reasoning based on world knowledge and common sense. For example, the instruction "I want to do yoga in front of the TV" implies that any obstacles in front of the TV should be moved elsewhere.

Outputs of the high-level planner comprise the scene map and the execution plan. The scene map specifies the target position p and orientation q for each object o that requires movement, represented as $\{\{o_1, p_1, q_1\}, \cdots, \{o_n, p_n, q_n\}\}$. The execution plan is a sequence of text actions associated with objects in the scene map. For example, "lift the monitor, move the monitor, put down the monitor".

4.1. Scene Map Generation

To obtain a reliable scene map, inspired by recent works in room layout generation [1, 62], we first instruct LLMs to reason about spatial relationships of the objects as an intermediate representation from which the 3D positions and orientations of the objects are calculated subsequently. We define the following relation functions to capture spatial relationships, with flexibility to include additional relationships if needed.

1. `on(object1, object2)`.
2. `adjacent(object1, object2, direction, distance)`.
3. `facing(object1, object2)`.

For example, if a monitor needs to be put on the table facing the chair, we can use `on(monitor, table)` and `facing(monitor, chair)` to describe its pose. If a table is positioned 1 meter to the north of the door, we can use `adjacent(table, door, north, 1)` to describe the relationship. We prompt LLMs to generate the spatial relationships of the objects according to the given instruction.

Given the spatial relationships, we propose an algorithm to calculate the 3D positions of each object in the scene. The orientations of objects, enforced by "facing" relations, will be addressed in the subsequent step.

The algorithm first constructs a scene graph [22] to organize the objects (nodes) and their spatial relationships (edges), which are directional, pointing from `object2` to `object1`. The algorithm employs a method similar to topological sorting, assuming no cycles are present in the graph. The scene graph is initialized by visiting the nodes associated with static objects and setting their 3D positions as the given values. In each iteration, the algorithm looks for a new node whose 3D position has not been determined but all its predecessors have. By calculating an offset from the predecessors' known positions, we can determine the position of the new node. For example, if object o_1 is "on" object o_2, then o_1's height is set to match the top surface of o_2, while the horizontal position is sampled within the

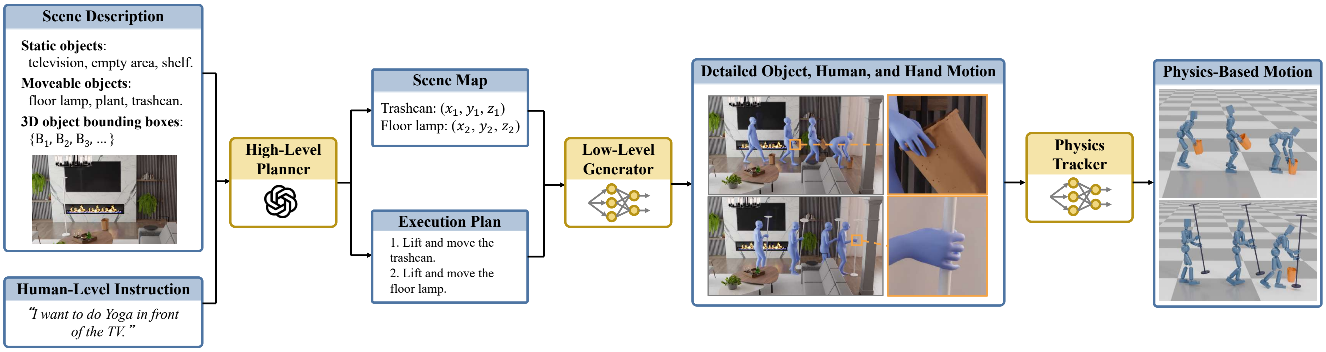


Figure 2. Our system takes the scene description and human-level instruction as input and uses a high-level planner to obtain the scene map and a detailed execution plan. The low-level motion generator then generates synchronized object motion, full-body human motion, and finger motion. Finally, the physics tracker uses RL to track the generated motion, producing physically plausible motion.

horizontal range of o_2. If o_1 is "adjacent" to o_2, we use the direction and distance given by the LLMs as the offset. We have included pseudo-code in the supplementary material.

To compute the orientation enforced by `facing(`o_1`,` o_2`)`, we rotate the canonical direction of o_1 in its own local frame to align with the vector from the position of o_1 to the position of o_2. The canonical direction of each object is part of the input 3D geometry information.

4.2. Execution Plan Generation

After determining the scene map, the planner must specify the interaction order with each object to ensure natural motion. For instance, if a vase is on a table, the vase should be moved before the table. LLMs are asked to generate this action sequence in a natural order and provide reasoning for it. The resulting execution plan consists of textual actions, $\{l_1, l_2, \ldots, l_T\}$, where each action l_t defines interaction involving one object. Each l_t serves as the textual input for the low-level generator, following the format: "lift the *object*, move the *object*, put down the *object*". This format is designed to closely match the textual style of the training dataset of the low-level generator.

With the scene map and execution plan in place, an A* planner generates collision-free paths to execute the plan. The path is a series of 2D waypoints in the scene, which then serve as inputs for the low-level motion generator.

5. Low-Level Motion Generator

Given the waypoints and textual instructions generated by the high-level planner, the low-level motion generator executes an interaction module and a navigation module sequentially to generate long-horizon human-object interaction motions.

5.1. Background: Conditional Diffusion Model

The diffusion model involves a forward diffusion process and a reverse diffusion process. Starting from clean data $\boldsymbol{x}_0$, noise is progressively introduced until, after N steps, the data becomes approximately $\boldsymbol{x}_N \sim \mathcal{N}(0, \boldsymbol{I})$. The denoising network is trained to reverse this process so that sampling a noise $\boldsymbol{x}_N$ and applying N denoising steps yields the clean data representation $\hat{\boldsymbol{x}}_0$. Each denoising step is defined as

$$p_\theta(\boldsymbol{x}_{n-1}|\boldsymbol{x}_n, \boldsymbol{c}) := \mathcal{N}(\boldsymbol{x}_{n-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_n, n, \boldsymbol{c}), \sigma_n^2 \boldsymbol{I}), \quad (1)$$

where p_θ represents a neural network, $\boldsymbol{c}$ represents the conditions, $\boldsymbol{\mu}_\theta(\boldsymbol{x}_n, n, \boldsymbol{c})$ represents the learned mean, σ_n^2 denotes a fixed variance. The training loss is defined as $\mathcal{L} = \mathbb{E}_{\boldsymbol{x}_0, n} ||\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_n, n, \boldsymbol{c}) - \boldsymbol{x}_0||_1$.

5.2. Interaction Module

We introduce a multi-stage interaction module that synthesizes synchronized full-body, finger, and object motion, as shown in Fig. 3. The first stage, *CoarseNet*, generates initial human and object motions without detailed finger movement. Recognizing that finger poses typically remain static during interaction, the second stage generates a *grasp* pose based on the initial results. Then *RefineNet* re-generates the human and object motions to seamlessly align with the optimized grasp pose. Finally, *FingerNet* generates smooth finger motions for approaching and releasing the object, completing the interaction sequence.

5.2.1. CoarseNet

Given the initial human and object poses, waypoints, and final object pose, along with a text instruction, CoarseNet generates synchronized human motion and object motion. This is achieved using a pre-trained diffusion model CHOIS [24].

Data Representation. We denote human motion as $\boldsymbol{H} \in \mathbb{R}^{T \times D}$, where T represents the sequence length and D represents the dimension of the human pose. Each $\boldsymbol{H}_t$ includes global joint positions and 6D rotations [71], excluding finger joints. Object motion $\boldsymbol{O} \in \mathbb{R}^{T \times 12}$ includes the object's global position and rotation matrix. CoarseNet predicts human motion, object motion, and contact labels $\boldsymbol{L} \in \mathbb{R}^2$ for both hands, resulting in $\boldsymbol{x} = \{\boldsymbol{H}, \boldsymbol{O}, \boldsymbol{L}\}$.

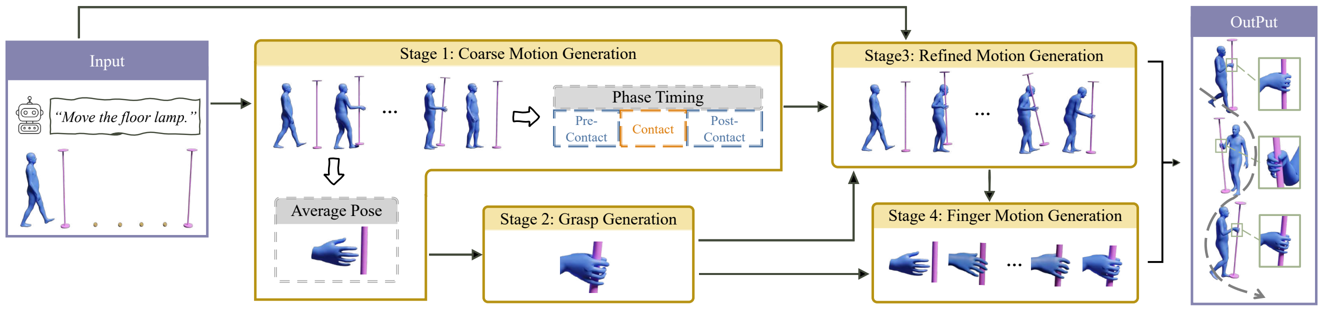


Figure 3. The interaction module consists of four stages. Starting with the initial human and object poses, waypoints, final object pose, and a text instruction, the first stage generates coarse object and human motion. The relative wrist pose from this motion is then used to generate a valid grasp pose. Using this as a condition, the next stage re-generates the motion to align with the grasp constraint. Finally, the finger motion generation module produces finger motions during the approach and release of objects.

The input condition is represented as $\boldsymbol{c} = \{\boldsymbol{S}, \boldsymbol{G}, \boldsymbol{T}\}$. $\boldsymbol{S} \in \mathbb{R}^{T\times(D+12)}$ is a masked motion data representation. It includes the human and object pose in the first frame, the object pose in the last frame, and 2D waypoints sampled every 30 frames. Any remaining entries in $\boldsymbol{S}$ are padded with zeros. The object geometry $\boldsymbol{G}$ represents the BPS representation [39] to encode object geometry, and $\boldsymbol{T}$ refers to the text embeddings extracted using CLIP [40].

Segmenting Contact Phases. We segment the generated interaction sequence into *pre-contact*, *contact*, and *post-contact* phases based on the predicted contact labels $\boldsymbol{L}$. The contact phase is determined separately for each hand, dividing the model output $\boldsymbol{x}_{\text{coarse}}$ into $\boldsymbol{x}_{\text{pre-contact}} \oplus \boldsymbol{x}_{\text{contact}} \oplus \boldsymbol{x}_{\text{post-contact}}$. We then compute the average wrist pose $\boldsymbol{w}$ from $\boldsymbol{x}_{\text{contact}}$, serving as input for the next stage.

5.2.2. Grasp Pose Generation

The second stage generates the finger pose for the contact phase using a state-of-the-art grasp pose generation method [50]. Given an object, wrist pose $\boldsymbol{w}$ and rest finger pose, an optimization is conducted to generate a physically plausible grasp pose. We modified [50] to fit our needs by: (1) increasing the object surface sample points from 2,000 to 20,000 to maintain sample density for larger objects and (2) omitting the force closure term for tasks involving both hands, such as carrying a box, where stability doesn't solely rely on a single hand.

The optimized grasp pose is denoted by $\hat{\boldsymbol{g}} = (\hat{\boldsymbol{w}}, \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{w}}$ is the optimized wrist pose, and $\hat{\boldsymbol{\theta}}$ is the finger pose. This grasp pose $\hat{\boldsymbol{g}}$ is maintained throughout the contact phase. However, as CoarseNet's generated motion may not fully align with $\hat{\boldsymbol{w}}$, we introduce RefineNet to address this misalignment.

5.2.3. RefineNet

The goal of RefineNet is to re-generate human-object interaction motions that align with the optimized grasp pose $\hat{\boldsymbol{g}}$. We adopt a conditional diffusion model based on CoarseNet, adding two conditions: *wrist-object relative pose* to align the wrist with $\hat{\boldsymbol{w}}$ and *object static* to keep the object stationary during non-contact phases. RefineNet's input condition, $\boldsymbol{c}_r = \{\boldsymbol{W}, \boldsymbol{S}_r, \boldsymbol{G}, \boldsymbol{T}\}$, includes $\boldsymbol{G}$ and $\boldsymbol{T}$ from CoarseNet, with $\boldsymbol{W}$ and $\boldsymbol{S}_r$ detailed as follows.

Wrist-Object Relative Pose Condition. The condition $\boldsymbol{W} \in \mathbb{R}^{T\times 18}$ represents the position and 6D rotation of both wrists in the object's frame, including only contact-phase wrist poses, with other entries set to zero.

Object Static Condition. In CoarseNet, the masked motion condition $\boldsymbol{S}$ was zero-padded except at the start, end, and waypoints. RefineNet uses contact phase information from CoarseNet to adjust $\boldsymbol{S}$ to $\boldsymbol{S_r}$, assigning a static object pose in pre-contact and post-contact frames.

Wrist-Object Relative Pose Loss. We incorporate an additional loss to enforce the wrist-object relative pose condition. We sample 100 points $\boldsymbol{K}_{\text{rest}} \in \mathbb{R}^{100\times 3}$ uniformly on the rest object's surface and pre-calculate their positions in both wrists' coordinate frames, denoted as $\boldsymbol{K}_{\text{w}} \in \mathbb{R}^{2\times T\times 100\times 3}$. At each time step, we compute the points' positions in the wrist frame, $\hat{\boldsymbol{K}}_{\text{w}}$, using the predicted orientations and positions of the object and wrist, $\boldsymbol{R}_{\text{o}}, \boldsymbol{T}_{\text{o}}, \boldsymbol{R}_{\text{w}}, \boldsymbol{T}_{\text{w}}$:

$$\boldsymbol{K}_{\text{global}} = \boldsymbol{R}_{\text{o}}\boldsymbol{K}_{\text{rest}} + \boldsymbol{T}_{\text{o}}, \tag{2}$$

$$\hat{\boldsymbol{K}}_{\text{w}} = \boldsymbol{R}_{\text{w}}^{-1}\left(\boldsymbol{K}_{\text{global}} - \boldsymbol{T}_{\text{w}}\right). \tag{3}$$

The loss is computed as:

$$\mathcal{L}_{\text{relative}} = \sum_{t=1}^{T} \boldsymbol{L}_t \left\| \hat{\boldsymbol{K}}_{\text{w},t} - \boldsymbol{K}_{\text{w},t} \right\|_1, \tag{4}$$

where $\boldsymbol{L}_t$ represents the contact labels, masking out the loss when the hand and object are not in contact.

Post-Processing. RefineNet generates human-object motions that align closely with input conditions, though minor adjustments may be required. To refine alignment, we apply post-processing: replacing object poses in pre- and post-contact phases with static poses and recalculating wrist trajectories during the contact phase based on object motion and the optimized wrist pose $\hat{\boldsymbol{w}}$. Additional implementation details are in the supplemental materials.

5.2.4. FingerNet

In this stage, we generate finger motions for the pre-contact and post-contact phases to create natural transitions between the rest and grasp poses, forming a complete interaction sequence along with predictions in prior stages. Given the start and end finger poses and the wrist trajectory in between, we employ a conditional diffusion model to predict the finger motions denoted as $\boldsymbol{F} \in \mathbb{R}^{T \times D'}$, representing local 6D rotations [71].

The model is conditioned on $\boldsymbol{c}_f = \{\boldsymbol{P}, \boldsymbol{F}_s, \boldsymbol{F}_e\}$, where $\boldsymbol{P}$ captures spatial hand-object relationships and $\boldsymbol{F}_s, \boldsymbol{F}_e$ are the start and end finger poses. Following prior work [45, 67], $\boldsymbol{P} \in \mathbb{R}^{T \times 100}$ is computed by sampling 100 palm-side mesh vertices and measuring their closest distances to the object, with the finger pose set to mean pose when calculating $\boldsymbol{P}$.

5.3. Navigation Module

To generate long sequences of interactions with multiple objects, the human needs to navigate through the scene based on waypoints. We adopt a conditional diffusion model to generate human navigation motion $\boldsymbol{H} \in \mathbb{R}^{T \times D}$. The input conditions consist of the initial human pose, 2D waypoints, and waypoint orientations (expressed as normalized direction vectors) to maintain consistent facing directions. Smooth transitions between the interaction and navigation modules are achieved by using the final pose of one module as the starting pose for the next.

6. Physics Tracker

After obtaining a sequence of kinematic motions, we train a tracking policy using reinforcement learning to control a physically simulated character that imitates these motions. This approach helps eliminate artifacts (e.g., hand-object penetration and foot sliding) in the kinematic motions and ensures the physical plausibility of the generated results.

We use PPO [42] as the backbone reinforcement learning algorithm, and adopt an importance sampling strategy to facilitate the tracking of long motion sequences involving interactions with multiple objects. Policies trained by our approach can accurately track the kinematics motions, enabling the character to grasp and move distinct objects along the generated trajectories. Implementation and training details for the physics tracker are provided in the supplementary materials.

7. Experiment

We conduct extensive evaluation to assess the overall performance of our system, as well as each individual component.

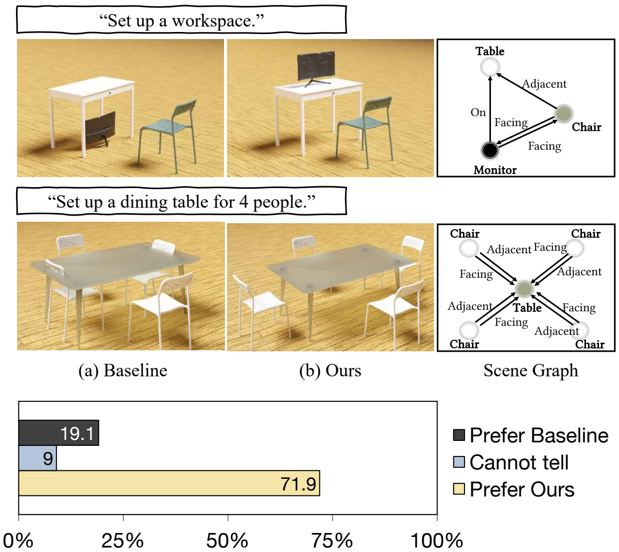


Figure 4. Comparison between (a) the baseline and (b) our high-level planner. The baseline generates an incorrect monitor position and orientation, and places the chair in a position that intersects with the table. The human perception study results (bottom) show that the majority of participants preferred our results.

We use quantitative metrics and user perception studies to evaluate our results compared to the baselines. We also conduct ablation studies to analyze the effects of each system module. We present the key results in this section and highly recommend readers refer to the supplementary video and documents for more detailed results and descriptions of the studies.

7.1. Evaluation of High-Level Planner

The core component of the high-level planner is the scene map calculation. To assess its effectiveness, we quantitatively measure the geometric accuracy of the resulting plans and conduct a human perception study. For evaluation, we compare our method to a baseline that uses LLMs to directly predict object positions and orientations. We design 25 human-level instructions covering common everyday tasks for assessment, using OpenAI's GPT-4o [33] for both our method and the baseline.

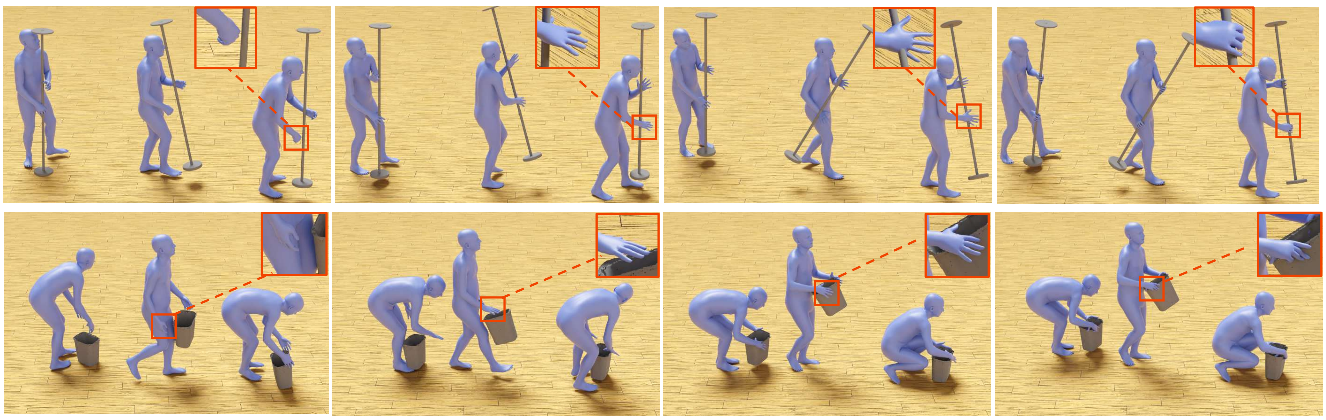
For geometric accuracy, we measure the proportion of objects with positional ($\mathrm{PE_p}$) or orientation errors ($\mathrm{PE_o}$) in the scene. Positional errors include objects placed at incorrect heights or with penetration into other objects, while orientation errors refer to objects being directed in the wrong direction. The results, shown in Tab. 2, demonstrate that our method outperforms the baseline by a large margin. A qualitative comparison is presented in the upper part of Fig. 4,

Table 2. Comparison of geometric accuracy between our method and the baseline.

	$\mathrm{PE_p} \downarrow$	$\mathrm{PE_o} \downarrow$
Baseline	21.9%	12.5%
Ours	**3.1%**	**1.6%**

Method	Condition	Human Motion		Interaction					GT Difference		
	$T_{xy}\downarrow$	$H_{\text{feet}}\downarrow$	FS $\downarrow$	$\text{C}_{\text{prec}}\uparrow$	$\text{C}_{\text{rec}}\uparrow$	$\text{C}_{\text{F1}}\uparrow$	CC $\uparrow$	IV $\downarrow$	MPJPE $\downarrow$	$T_{\text{obj}}\downarrow$	$O_{\text{obj}}\downarrow$
CNet+GRIP [45]	3.58	**2.50**	0.45	0.90	0.64	0.70	3.99%	49.00	14.33	11.55	0.93
CNet	3.26	2.68	0.43	0.91	0.81	0.84	3.54%	48.56	**14.32**	**10.56**	1.06
C+RNet	**2.81**	2.89	**0.33**	**0.91**	**0.95**	**0.92**	3.84%	24.78	15.96	12.57	**0.73**
C+R+FNet (ours)	**2.81**	2.89	**0.33**	**0.91**	**0.95**	**0.92**	**5.53%**	**19.06**	15.96	12.57	**0.73**

Table 1. Interation synthesis on the FullBodyManipulation dataset [25]. Our full model achieves superior interaction quality compared to all others.



(a) CNet+GRIP (b) CNet (c) C+RNet (d) C+R+FNet (ours)

Figure 5. Qualitative comparisons of each method. All baselines and ablations fail to produce precise hand and finger motions, while our method generates natural hand-object interactions.

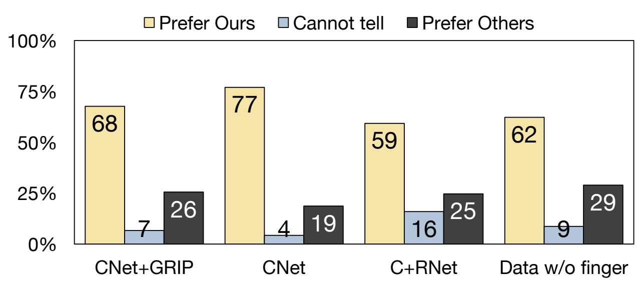


Figure 6. Comparison of our full model and baseline models through human perception studies.

where the baseline generates incorrect object positions and orientations, while our method produces more reasonable layouts.

We also conducted a human perception study, with results presented in the lower part of Fig. 4. Each pair was reviewed by 20 Amazon Mechanical Turk (AMT) workers. They were asked to choose the scene that was more reasonable and consistent with the given instruction. The majority of participants preferred our method.

7.2. Evaluation of Low-Level Motion Generator

Datasets. CoarseNet and RefineNet are trained on the FullBodyManipulation dataset [25], which includes 10 hours of human-object interaction with 15 objects but lacks finger motion. FingerNet uses the GRAB dataset [43], containing full-body and finger motion for 10 subjects with 51 objects. The navigation module is trained on HumanML3D [12], featuring 28 hours of diverse motions with language descriptions. For all datasets, we follow standard data partitioning from prior work [12, 25, 43].

Evaluation Metrics. Following prior work [24], we evaluate our approach from various aspects. All distance errors are measured in centimeters. For condition matching, we report the waypoint following errors (T_{xy}). For human motion quality, we report the foot sliding score (FS) and foot height (H_{feet}). For interaction quality, we use contact precision (C_{prec}), recall (C_{rec}), and F1 score (C_{F1}). To better assess finger-object interaction, we also report contact coverage (CC) – the percentage of hand points within -2mm to +2mm of the object surface – and intersection volume (IV), the overlap volume of hand and object meshes in cm^3, following [11]. For ground truth difference, we report the mean per-joint position error (MPJPE), object position error (T_{obj}), and the object orientation error (O_{obj}). We also conduct a human perception study to measure the motion quality.

Baselines. For generating full-body human motion, finger motion, and object motion from text, prior work [9, 26] trained models on GRAB [43] to predict motions for small-object manipulation. However, GRAB contains only small-object manipulation data without locomotion, so models trained on it cannot generalize to scenarios involving large objects that require full-body coordination and locomotion. Moreover, the approach from [9, 26] is not applicable to the FullBodyManipulation dataset [25] used in our work, as this dataset lacks finger motions, making it infeasible to learn a model via direct supervision. Due to these differences, a directly comparable baseline does not exist. Thus, we establish our baseline by combining two complementary methods: CHOIS [24] (referred to as CoarseNet in our framework), which generates object motion and full-body motion from text, and GRIP [45], which refines arm movements and generates finger motions based on CHOIS's output.

In addition, to evaluate the impact of each component, we compare our full system against two ablations: (1) CoarseNet alone (CNet) and (2) CoarseNet + RefineNet (C+RNet), which represents the full system without finger motion. The full system is denoted as C+R+FNet. This evaluation focuses on the motion synthesis quality before physics tracking. We present the evaluation for the physics tracker in Sec. 7.3.

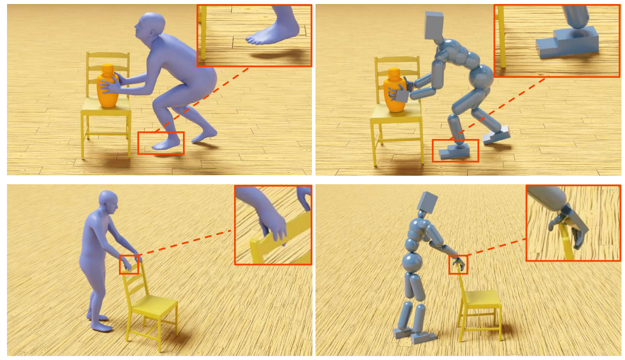
Results. The qualitative comparison in Fig. 5 clearly demonstrates the superiority of our method. Both ablated variants show a lack of finger detail, and GRIP [45] fails to generate accurate hand motion, resulting in significant artifacts. In contrast, our method produces precise finger-object interactions. We also observe the impact of RefineNet: compared to CNet, C+RNet maintains more consistent and reasonable relative poses between the hand and object during the interaction, showing RefineNet's role in improving pose consistency and reducing artifacts.

The quantitative results are presented in Tab. 1. Since our method focuses on finger-object interaction, the five metrics under the *Interaction* column are particularly important, and we achieve the best performance across all of them.

We further conduct a human perception study to complement our evaluation. We generate 15 sequences using each method and compose 60 pairs, with each pair consisting of our results and the results of a baseline. Each pair was reviewed by 20 Amazon Mechanical Turk (AMT) workers. As shown in Fig. 6, our method is consistently preferred over the baselines by a large margin. Notably, our method achieves even higher preference rates when compared to the real-world data from the FullBodyManipulation [25] dataset, as we generate detailed finger motions that are absent in the original dataset.

7.3. Evaluation of Physics Tracker

The physics tracker can accurately track kinematic motion while ensuring the physical plausibility of the motion. Through physics simulation, artifacts such as foot floating, foot sliding, and human-object penetration are eliminated, as shown in Fig. 7.



(a) Kinematic Motion (b) Physics-Based Motion

Figure 7. Qualitative comparison of kinematic motion and physics-based motion. The physics tracker corrects artifacts such as foot floating and penetration.

Additionally, we evaluate the tracking error of human joints (E_h) and objects (E_o), as shown in Tab. 3. The tracking error is defined as the positional error (cm) between the kinematic motion and the tracked motion. The results were evaluated across 20 interaction sequences, each with an average duration of 30 seconds and 2 objects of varying shapes and types. All sequences were successfully tracked, with relatively small errors, demonstrating high tracking accuracy.

Table 3. Tracking error between the kinematic motion and physics-based motion.

	E_h	E_o
Ours	5.45	4.67

8. Conclusions

In this work, we introduced a system that simulates realistic human-object manipulation in a contextual environment based on human-level language instructions. The framework consists of a high-level LLM planner that infers a target scene layout and execution plan from the given instructions, a low-level motion generator that generates synchronized, full-body, fingers, and object motion, and a physics tracker that imitates the generated kinematic motions and produces physically plausible interaction motions. Our system effectively handles a variety of tasks, showcasing its potential for various real-world applications.

Future Work. Our system currently uses text or image inputs to represent the contextual environment, and could potentially be enhanced with more structured representations like voxel grids as used in [20]. Building on this idea, a promising direction is to create an agent with egocentric vision perception as input to LLM/VLMs for high-level motion planning. This could enable the agent to exhibit more natural behaviors, such as purposeful head orientation for observation and better body positioning to avoid collisions.

Acknowledgement. We thank Wenyang Zhou for data processing. This work is in part supported by the Wu Tsai Human Performance Alliance at Stanford University and the Stanford Institute for Human-Centered AI (HAI).

References

[1] Rio Aguina-Kang, Maxim Gumin, Do Heon Han, Stewart Morris, Seung Jean Yoo, Aditya Ganeshan, R Kenny Jones, Qiuhong Anna Wei, Kailiang Fu, and Daniel Ritchie. Open-universe indoor scene generation using llm program synthesis and uncurated object databases. *arXiv preprint arXiv:2403.09675*, 2024. 2, 3

[2] Joao Pedro Araujo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Deepak Gopinath, Jiajun Wu, Alexander Clegg, and C Karen Liu. Circle: Capture in rich contextual environments. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[3] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2386–2393. IEEE, 2019. 2

[4] Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. *arXiv preprint arXiv:2309.07907*, 2023. 2

[5] Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. *arXiv preprint arXiv:2403.17827*, 2024. 2

[6] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. *arXiv preprint arXiv:2311.16097*, 2023. 2

[7] Yan Ding, Xiaohan Zhang, Chris Paxton, and Shiqi Zhang. Task and motion planning with large language models for object rearrangement. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2086–2092. IEEE, 2023. 3

[8] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[9] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Eurographics*, 2023. 7, 8

[10] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, pages 1–12. Wiley Online Library, 2023. 2

[11] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1471–1481, 2021. 7

[12] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 7

[13] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *International Conference on Computer Vision (ICCV)*, pages 2282–2292, 2019. 2

[14] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *International Conference on Computer Vision (ICCV)*, pages 11354–11364, 2021. 2

[15] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *SIGGRAPH 2023 Conference Papers*, 2023. 2

[16] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023. 2, 3

[17] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv:2311.17842*, 2023. 3

[18] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[19] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11107–11116, 2021. 2

[20] Nan Jiang, Zimo He, Zi Wang, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. Autonomous character-scene interaction synthesis from text instruction. *arXiv preprint arXiv:2410.03187*, 2024. 8

[21] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. *arXiv preprint arXiv:2403.08629*, 2024. 2

[22] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015. 3

[23] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 2

[24] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object

interaction synthesis. *arXiv preprint arXiv:2312.03913*, 2023. 1, 2, 4, 7, 8

[25] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Trans. Graph.*, 42(6), 2023. 2, 7, 8

[26] Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3035–3044, 2024. 7, 8

[27] Ying Li, Jiaxin L Fu, and Nancy S Pollard. Data-driven grasp synthesis using shape matching and task-based pruning. *IEEE Transactions on Visualization and Computer Graphics*, 13(4): 732–747, 2007. 2

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

[29] Libin Liu and Jessica Hodgins. Learning basketball dribbling skills using trajectory optimization and deep reinforcement learning. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 3

[30] Xueyi Liu and Li Yi. Geneoh diffusion: Towards generalizable hand-object interaction denoising via denoising diffusion. *arXiv preprint arXiv:2402.14810*, 2024. 2

[31] Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. Grasping diverse objects with simulated humanoids. *arXiv preprint arXiv:2407.11385*, 2024. 1, 3

[32] Aymen Mir, Xavier Puig, Angjoo Kanazawa, and Gerard Pons-Moll. Generating continual human motion in diverse 3d scenes. *arXiv preprint arXiv:2304.02061*, 2023. 2

[33] OpenAI. Hello gpt-4o, 2024. 6

[34] R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5), 2023. 2

[35] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553*, 2023. 2, 3

[36] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018. 3

[37] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (TOG)*, 40(4):1–20, 2021. 3

[38] Nancy S Pollard and Victor Brian Zordan. Physically based grasping control from example. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 311–318, 2005. 2

[39] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *International Conference on Computer Vision (ICCV)*, pages 4332–4341, 2019. 5

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[41] Alberto Rodriguez, Matthew T Mason, and Steve Ferry. From caging to grasping. *The International Journal of Robotics Research*, 31(7):886–900, 2012. 2

[42] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. 6

[43] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 7

[44] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13263–13273, 2022. 2

[45] Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soren Pirk, and Michael J Black. Grip: Generating interaction poses using latent consistency and spatial cues. *arXiv preprint arXiv:2308.11617*, 2023. 6, 7, 8

[46] Purva Tendulkar, Dídac Surís, and Carl Vondrick. Flex: Full-body grasping without full-body grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21179–21189, 2023. 2

[47] Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. Gpt-4v (ision) for robotics: Multimodal task planning from human demonstration. *arXiv preprint arXiv:2311.12015*, 2023. 3

[48] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9401–9411, 2021. 2

[49] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20460–20469, 2022. 2

[50] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11359–11366. IEEE, 2023. 2, 5

[51] Ruocheng Wang, Pei Xu, Haochen Shi, Elizabeth Schumann, and C Karen Liu. F\" urelise: Capturing and physically synthesizing hand motions of piano performance. *arXiv preprint arXiv:2410.05791*, 2024. 3

[52] Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. Physhoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023. 1, 3

[53] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[54] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan

Huang. Move as you say, interact as you can: Language-guided human motion generation with scene affordance. *arXiv preprint arXiv:2403.18036*, 2024. 2

[55] Qianyang Wu, Ye Shi, Xiaoshui Huang, Jingyi Yu, Lan Xu, and Jingya Wang. Thor: Text to human-object interaction diffusion via relation intervention. *arXiv preprint arXiv:2403.11208*, 2024. 2

[56] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *European Conference on Computer Vision (ECCV)*, pages 257–274, 2022. 2

[57] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 3

[58] Zhaoming Xie, Sebastian Starke, Hung Yu Ling, and Michiel van de Panne. Learning soccer juggling skills with layer-wise mixture-of-experts. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 3

[59] Pei Xu and Ioannis Karamouzas. A gan-like approach for physics-based imitation learning and interactive character control. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 4(3):1–22, 2021. 3

[60] Pei Xu and Ruocheng Wang. Synchronize dual hands for physics-based dexterous guitar playing. *arXiv preprint arXiv:2409.16629*, 2024. 3

[61] Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. *arXiv preprint arXiv:2403.19652*, 2024. 2, 3

[62] Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, et al. Holodeck: Language guided generation of 3d embodied ai environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16227–16237, 2024. 3

[63] Heyuan Yao, Zhenhua Song, Yuyang Zhou, Tenglong Ao, Baoquan Chen, and Libin Liu. Moconvq: Unified physics-based motion control via scalable discrete representations. *ACM Transactions on Graphics (TOG)*, 43(4):1–21, 2024. 3

[64] Yuting Ye and C Karen Liu. Synthesis of detailed hand manipulations using contact sampling. *ACM Transactions on Graphics (ToG)*, 31(4):1–10, 2012. 2

[65] Hongwei Yi, Justus Thies, Michael J Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. *arXiv preprint arXiv:2404.10685*, 2024. 2

[66] Yiming Zeng, Mingdong Wu, Long Yang, Jiyao Zhang, Hao Ding, Hui Cheng, and Hao Dong. Distilling functional rearrangement priors from large models. *arXiv preprint arXiv:2312.01474*, 2023. 3

[67] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics (ToG)*, 40(4):1–14, 2021. 2, 6

[68] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision (ECCV)*, pages 518–535, 2022. 2

[69] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7368–7376, 2024. 3

[70] Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. *arXiv preprint arXiv:2305.12411*, 2023. 2

[71] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 4, 6