

GCML: GROUNDING COMPLEX MOTIONS USING LARGE LANGUAGE MODEL IN 3D SCENES

Anonymous authors

Paper under double-blind review

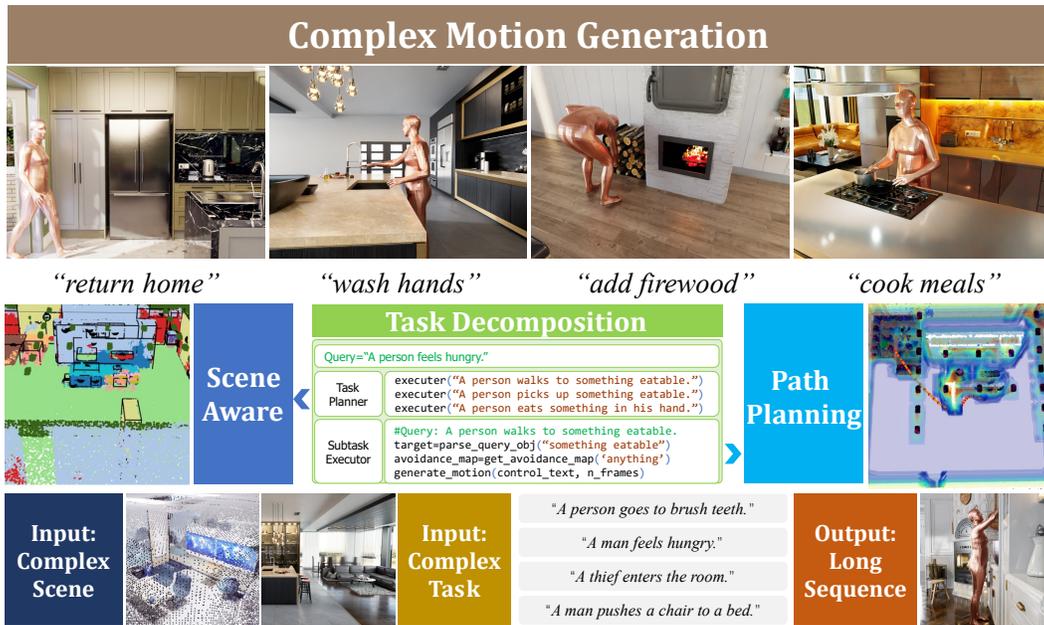


Figure 1: Different from previous methods that primarily focus on generating simple motions, our approach is designed to handle more complex actions. It leverages a Large Language Model for human action inference, task decomposition, and path planning. In combination with a 3D Visual Grounding Model for scene perception, this enables the generation of intricate, extended motions with complex scene and textual input.

ABSTRACT

To solve the problem of generating complex motions, we introduce GCML (Grounding Complex Motions using a Large Language Model). This method supports complex texts and scenes as inputs, such as mopping the floor in a cluttered room. Such everyday actions are challenging for current motion generation models for two main reasons. First, such complex actions are rarely found in existing HSI datasets, which places high demands on the generalization capabilities of current data-driven models. Second, these actions are composed of multiple stages, with considerable variation, making it difficult for models to understand and generate the appropriate motions. Current methods in the HSI field can control the generation of simple actions under multiple constraints, such as walking joyfully toward a door, but they cannot handle the complexity of tasks like the one described above. By incorporating a Large Language Model and a 3D Visual Grounding Model into the HSI domain, our approach can decompose a complex user prompt into a sequence of simpler subtasks and identify interaction targets and obstacles within the scene. Based on these subtask descriptions and spatial control information, the Motion Generation Model generates a sequence of full-body motions, which are then combined into a long motion sequence that aligns with both the user’s input and the scene semantics. Experimental results

054 demonstrate that our method achieves competitive performance for simple ac-
055 tion generation on the HUMANISE dataset and the generalization evaluation set.
056 For complex motion generation, we created a new evaluation set by automati-
057 cally generating possible behaviors of virtual humans in common indoor scenes,
058 where our method significantly outperforms existing approaches. Project Page:
059 <https://anonymous.4open.science/w/GCML-4562/>
060

061 1 INTRODUCTION

062 Research within Human-Scene Interaction (HSI) has advanced considerably in terms of modeling
063 the sophisticated interaction between people and their environments (Zhao et al., 2023); (Zhang
064 et al., 2022). It has also succeeded in synthesizing high-fidelity human motion in various
065 scenes (Hassan et al., 2021); (Zhang & Tang, 2022); (Wang et al., 2022a). However, two major
066 challenges exist:
067
068

069 **(1) Generative models depend heavily on large amounts of high-quality paired data.** The col-
070 lection of HSI datasets and their annotations is a time-consuming and labor-intensive task. Although
071 many efforts have been made (Hassan et al., 2019); (Wang et al., 2022b); (Jiang et al., 2024), chal-
072 lenges like limited diversity in actions, overly simplistic descriptions, and inadequate dataset sizes
073 still exist. On the other hand, existing datasets that only contain human motion, like AMASS (Mah-
074 mood et al., 2019) and HumanML3D (Plappert et al., 2016); (Guo et al., 2022b), are quite substantial
075 in terms of data scale and motion description. Diffusion-based methods trained on these datasets are
076 capable of generating high-quality human motions (Tevet et al., 2023) based on textual descriptions.
077 Recent works (Karunratanakul et al., 2023); (Xie et al., 2023) also enable users to manipulate the
078 style of generated actions via textual prompts and enhance precision using spatial constraints. In-
079 spired by this, we propose tackling scene awareness and motion generation separately. This strategy
080 helps mitigate the shortage of high-quality paired HSI data by maximizing the use of existing human
081 motion and 3D visual grounding datasets.

082 **(2) Only Common and simple motions can be generated.** Due to the lack of diversity in HSI
083 datasets, the majority of current research that generates human motions in scenes tends to focus
084 on producing common and simple movements such as walking, lying, and sitting (Wang et al.,
085 2022b); (Jiang et al., 2024); (Wang et al., 2024). However, industries with a need for human motion
086 generation often require more than these elementary motions. For instance, video game characters
087 should navigate to specified locations and interact with particular targets, while animated movie
088 characters need to perform everyday tasks such as brushing their teeth, cooking, or watering plants.
089 Such complex motions are rare in paired datasets, thus current data-driven methods for human mo-
090 tion generation typically fail to produce them. However, complex motions are often composed of
091 simpler ones, and a practical solution to this challenge is to utilize Large Language Models for task
092 decomposition and reasoning (Huang et al., 2023b); (Lin et al., 2024).

092 To alleviate these limitations, we propose a novel method called Grounding Complex Motions using
093 Large Language Models (GCML), depicted in Figure 1. Our method can generate complex actions
094 like washing hands or cooking meals, accommodating complex scene and textual input as well
095 as producing human motions of long sequences. Furthermore, by incorporating Large Language
096 Models and 3D Visual Grounding Models, our approach creatively generates motions consistent
097 with textual descriptions based on the objects and their layout within a scene. For example, a hungry
098 person in a scene, upon not finding available food, will attempt to open a refrigerator in search of
099 food.

100 Our method operates by taking both text and scene inputs. The text is first processed by the Task
101 Planner, which breaks it down into multiple sub-tasks, each corresponding to a simple motion se-
102 quence. Next, the Sub-task Executor identifies the target objects within the scene for interaction and
103 generates control descriptions for each sub-task with the help of the 3D Visual Grounding Model.
104 These control descriptions and spatial data are sent to the Motion Generation Model, which refines
105 them into whole-body motion sequences for each sub-task. Finally, these sequences are combined
106 into a complete, unified motion.

107 We have thoroughly evaluated GCML on a variety of benchmarks. Experimental results on the HU-
MANISE dataset reveal that even for simple tasks, without relying on HUMANISE training data, the

performance of our method is comparable to the current state-of-the-art data-driven approaches. On the generalization evaluation set proposed by (Wang et al., 2024), our method outperforms afford-motion across most metrics. Additionally, to specifically evaluate GCML’s performance in complex motion generation, we introduced a new evaluation set, the Complex Motion Evaluation Set. Our method was the only one capable of producing satisfactory outcomes on this set.

Our key contributions are as follows:

- We introduce a new task along with a corresponding evaluation set: Complex Motion Generation. The gaming and animation industries increasingly require not just simple actions like walking or sitting, but also the ability to think and interact with surrounding objects like a human. We hope this new task will inspire the development of more methods for generating realistic virtual humans.
- As an attempt to address the challenge of complex human motion generation, we present the GCML framework. By integrating a Large Language Model and a 3D Visual Grounding Model into the HSI domain, our approach circumvents the lack of high-quality paired data, allowing for the generation of motions like brushing teeth and watering plants, which previous methods could not achieve.
- We validated the effectiveness of our proposed method across three datasets. GCML achieves comparable performance in simple motion generation, while in complex motion generation, it consistently outperforms others across all metrics.

2 RELATED WORK

2.1 CONDITIONAL HUMAN MOTION GENERATION

Human-Scene Interaction can be seen as the generation of human motions conditioned on both language descriptions and scene context. A wide range of conditions have been explored for controlled motion generation, such as past motions (Yuan & Kitani, 2020); (Cao et al., 2020); (Xie et al., 2021), music (Tseng et al., 2023); (Li et al., 2021), text (Petrovich et al., 2023); (Guo et al., 2022c); (Tevet et al., 2022); (Chen et al., 2023); (Kim et al., 2023), objects (Ghosh et al., 2023); (Kulkarni et al., 2024); (Xu et al., 2023), and scenes (Huang et al., 2023a); (Wang et al., 2022b). Furthermore, recent work has added spatial constraints to text-guided motion generation (Text2Motion). MDM (Tevet et al., 2023) and priorMDM (Shafir et al., 2023) use motion inbetweening during the diffusion denoising process to replace key control frames, allowing for human motion generation that fits spatial constraints without sacrificing motion quality. GMD (Karunratanakul et al., 2023) introduces a two-stage motion diffusion model to handle sparse control signals, reducing jitter in controlled frames. Omnicontrol (Xie et al., 2023), employing a ControlNet-inspired approach, applies spatial constraints during motion generation and enables control over any key joint.

In the domain of Human Scene Interaction, language, and scene context are the two primary conditioning factors. HUMANISE (Wang et al., 2022b) introduced a comprehensive dataset and developed a cVAE-based model capable of generating motions that respond to both textual descriptions and scene interactions. TRUMANS (Jiang et al., 2024) proposed an even larger human-scene interaction dataset and used autoregressive conditional diffusion to generate HSI motions of any length. AffordMotion (Wang et al., 2024) employed scene affordances as intermediate representations, combining scene embeddings and language-guided motion generation in a two-stage approach. However, these methods are limited by their dependence on training data, generating only the common motion types seen in the datasets. Our work resolves this by separating scene understanding from motion generation, thus mitigating the reliance on specific datasets.

2.2 LARGE LANGUAGE MODELS FOR HUMAN MOTION GENERATION

Large Language Models have seen extensive research in the fields of intelligent agents (Wang et al., 2023); (Lin et al., 2023); (Park et al., 2023) and robotics (Blukis et al., 2020); (Yang et al., 2024); (Liu et al., 2024). The task planning and commonsense reasoning capabilities of pre-trained language models have substantially improved the ability of embodied agents to interpret their environment and tackle complex tasks in these domains.

162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215

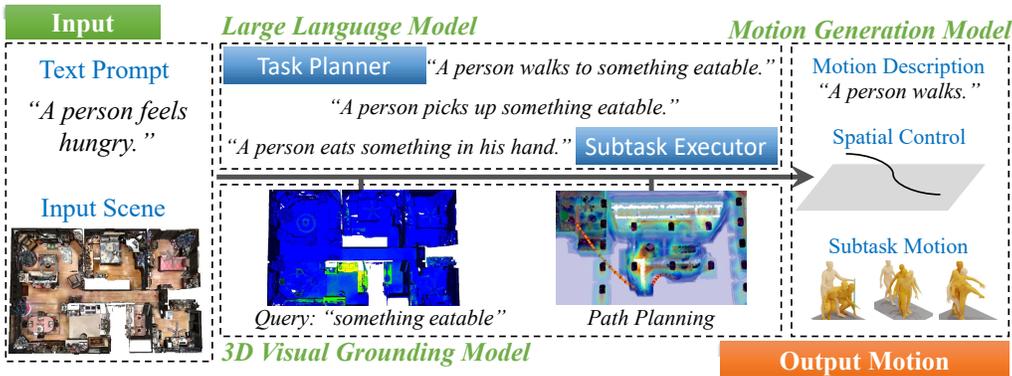


Figure 2: Overview of our method. To generate complex motions that align with text descriptions and scene semantics, GCML first utilizes a Large Language Model to break the task down into a sequence of simpler subtasks. Then, it uses the 3D Visual Grounding Model to extract scene information and generate control data for certain human joints. Finally, the Motion Generation Model produces the full-body motion frames.

In Human Motion Generation, MotionGPT (Jiang et al., 2023) introduced a general motion generator capable of embedding multimodal signals as special input tokens in Large Language Models, allowing it to handle various tasks related to human motion recognition and generation. UniHSI (Xiao et al., 2023) leveraged Large Language Models to decompose language descriptions into a contact chain—a sequence of interactions between human joints and object parts—enabling the continuous generation of motions interacting with multiple objects.

However, both approaches have limitations. MotionGPT cannot perceive the scene, while UniHSI requires fine-grained annotation of scene objects to generate accurate interaction motions. Additionally, UniHSI is unable to generate motions that do not involve contact with fixed objects, such as brushing teeth or eating, limiting its applicability.

2.3 3D VISUAL GROUNDING

3D scene understanding has been extensively researched, particularly in the fields of vision and robotics. Core tasks include 3D object classification (Wu et al., 2015), 3D object detection and localization (Caesar et al., 2020); (Chen et al., 2020), 3D semantic and instance segmentation (Behley et al., 2019); (Liao et al., 2022), and 3D affordance prediction (Deng et al., 2021). A recent approach, OpenScene (Peng et al., 2023), extracts CLIP features for each 3D point, supporting open-vocabulary queries and segmentation based on any text input, making it a versatile tool for scene understanding in our work.

In Human Scene Interaction tasks, few approaches utilize existing 3D scene understanding models to assist in generating motions that align with scene semantics. Instead, data-driven methods typically train models by mapping specific 3D structures to actions, but this introduces two major limitations: the reliance on the quantity and quality of paired data limits performance, and biases in the data may restrict certain interactions. For example, interactions with chairs are typically confined to sitting, ignoring other potential uses. Utilizing existing 3D Visual Grounding models provides distinct advantages in overcoming these issues.

3 METHOD

As illustrated in Figure 2, GCML consists of three components. The Large Language Model is responsible for breaking down complex motion generation tasks into a combination of simpler ones, as well as generating control data for certain human joints in each subtask. The 3D Visual Grounding Model identifies navigable areas and locates the targets of interaction within the scene. The Motion Generation Model produces full-body motion sequences for each subtask based on the control

information provided by the other modules and integrates these sequences into a complete output motion.

This chapter begins by introducing the input and output formats of the task (Section 3.1). It then discusses how the LLM Planner and Subtask Executor convert the text description and scene point cloud into the control data necessary for the motion generation model (Section 3.2). Following this, we explain how the Motion Generation Model synthesizes the control data into a full motion sequence (Section 3.3) and provide further details on the 3D Visual Grounding Model used in our method (Section 3.4). Finally, we present the construction of the Complex Motion Evaluation Set (Section 3.5).

3.1 PROBLEM FORMULATION AND NOTATIONS

The generation of complex motions can be viewed as a recursive process, where complex motions are composed of simpler ones that follow the same input and output structure. Consequently, the task of generating complex motions can be divided into generating multiple simple ones and linking them together. Therefore, a method that can generate complex motions is inherently capable of generating simple motions as well.

Specifically, the input to our task consists of a user text prompt T and a scene S , with the output being a human motion sequence H . The user prompt T is a text description of the motion to be generated, while $S \in \mathbb{R}^{N \times 6}$ represents a point cloud of N points, each with RGB color information. The output motion sequence $\{H_i\}_{i=1}^N$ is a series of human pose parameters over N frames. We use the SMPL-X model (Pavlakos et al., 2019) to represent the human body’s pose and shape. The SMPL-X body mesh $M \in \mathbb{R}^{10475 \times 3}$ is parameterized as $M = F(t, r, \beta, p)$, where $t \in \mathbb{R}^3$ is the global translation, $r \in \mathbb{R}^6$ is the continuous representation of global orientation, $\beta \in \mathbb{R}^{10}$ defines the body shape, and $p \in \mathbb{R}^{J \times 3}$ represents joint rotations in axis-angle format. F is the differentiable linear blend skinning function. In most cases, we do not generate the full set of SMPL-X shape parameters directly; instead, we first generate the 3D world coordinates of 22 key body joints $P \in \mathbb{R}^{22 \times 3}$, which are then used to infer the complete SMPL-X parameters. The process can be summarized by the following formula:

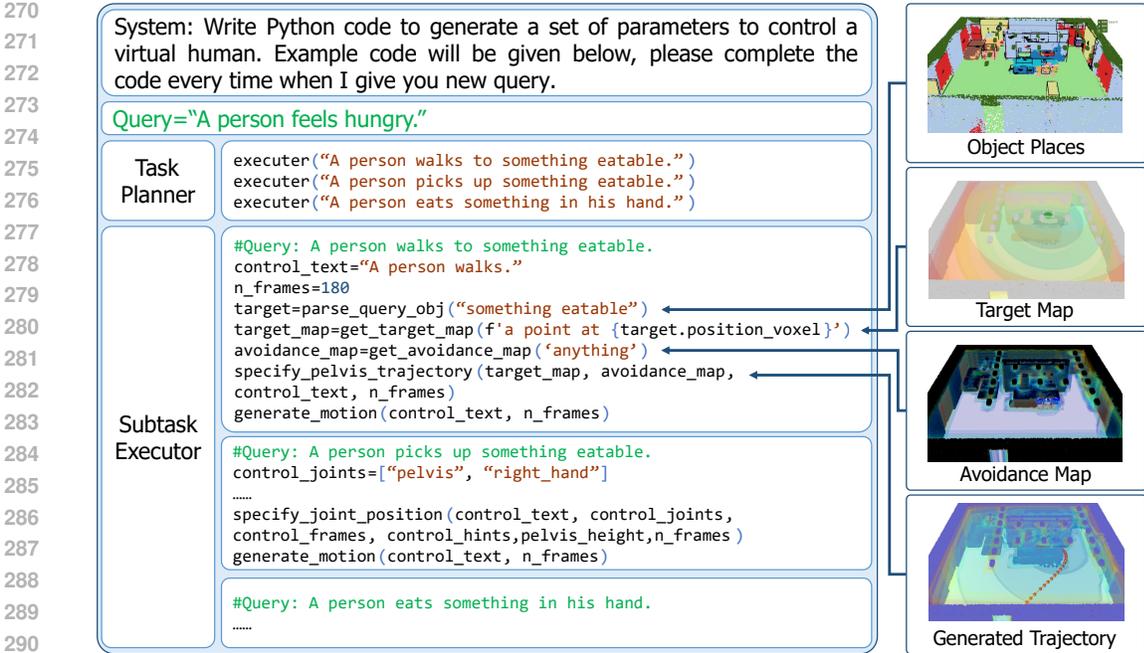
$$T + S \longrightarrow \{P_i\}_{i=1}^{22} \longrightarrow M = F(t, r, \beta, p) \longrightarrow \{H_i\}_{i=1}^N \quad (1)$$

3.2 LLM PLANNER AND SUBTASK EXECUTOR

Similar to VoxPoser (Huang et al., 2023b), our method leverages a Language Model Program (LMP) to use a Large Language Model in generating human motion. Each LMP is responsible for a distinct function—such as breaking down tasks or invoking perception modules—and can call upon other LMPs as needed. The Large Language Model employed in our approach is GPT-4 (Achiam et al., 2023) from OpenAI.

Figure 3 provides an example of the generation process. First, we instruct the Large Language Model to learn from the provided code examples. For each LMP, we offer around 5-10 query samples along with their corresponding responses.

Following the sequence of calls between LMPs, the user’s input motion description is first sent to the Task Planner, where complex or abstract tasks are broken down into simpler ones for the Subtask Executor. These tasks typically involve actions like walking to a location, interacting with an object, or performing a motion while standing still. The Subtask Executor creates human-centric motion descriptions based on the subtask, while object interaction information is provided by the 3D Visual Grounding model. Scene traversability data is directly obtained from the scene’s point cloud. Spatial data is organized using a voxel map, with the Target Map and Avoidance Map values combined to produce a cost map $C \in \mathbb{R}^{100 \times 100 \times 100}$, where lower values correspond to interaction targets and higher values indicate obstacles. Finally, using the following formula, we generate the trajectories of key human joints (such as the pelvis and hand), ensuring they avoid obstacles and reach the interaction target.



292 Figure 3: Example of how the Large Language Model generates human motion based on task descriptions and scene information. The Language Model Program (LMP) leverages provided examples to generate code that calls functions or other LMP instances from the user’s query. This code then uses the perceived scene information to invoke the Motion Generation Model and produce complete motion frames.

$$300 \quad p = \arg \min_{p_i} \sum_{i=1}^n [C(p_i) - w_{\text{inertia}} \cdot \langle p_i - p_{i-1}, d \rangle + w_{z_{\text{penalty}}} \cdot |p_{i,z} - p_{i-1,z}|] \quad (2)$$

303 where p represents the generated trajectory of human keypoints, w_{inertia} is the inertia weight, which helps prevent the trajectory from getting stuck in local minima, $w_{z_{\text{penalty}}}$ is the z-axis offset weight, which ensures the human does not move over obstacles when controlling pelvis. As a result, we obtain control data for the human joints $h \in R^{N \times 22 \times 3}$, along with a human-centric motion description in text form. These data are then passed to the Motion Generation Model to produce a complete human motion sequence.

310 3.3 MOTION GENERATION MODEL

312 Our Motion Generation Model is built upon OmniControl (Xie et al., 2023), a diffusion-based generative model that conditions on both text and spatial keypoint positions. OmniControl was trained on the HumanML3D dataset and supports control over any joints at any time. We adapted it to generate motion sequences of arbitrary length.

316 In the previous steps, we obtained joint control data $h \in R^{N \times 22 \times 3}$. At each step of diffusion, we calculate the L2 distance between the generated joints and the control data h , using the gradient to guide the model in generating sub-motion sequences that align with both the action description T and the spatial control signal h . To ensure smooth transitions between sub-sequences, we use the motion inbetweening method from MDM (Tevet et al., 2023). In each diffusion step, we replace the first frame of the generated motion with the last frame from the previous sequence. We also compute a transformation matrix based on the movement from the first to the last frame of the previously generated sequence. This matrix is applied to the subsequent sequences, merging the sub-task sequences into a coherent long motion sequence.

3.4 3D VISUAL GROUNDING MODEL

In the above generation process, we did not elaborate on how the positions of required objects were obtained from the scene. While many well-established methods exist for detecting and localizing 3D objects (Caesar et al., 2020); (Chen et al., 2020), in the scenarios described in this paper, we sometimes need to detect unconventional objects (e.g., “something eatable” in Figure 3). For this reason, we chose OpenScene (Peng et al., 2023) as our scene perception module due to its support for open vocabulary queries. Once OpenScene assigns object categories to each point in the scene’s point cloud, we use the DBSCAN (Khan et al., 2014) clustering algorithm to filter out noise and pinpoint object instance locations. Additionally, OpenScene’s ability to handle open-vocabulary queries allows us to detect object parts as well. For example, when interacting with a door, we can specifically locate the door handle rather than the entire door, which helps create more realistic motions.

3.5 COMPLEX MOTION EVALUATION SET

In addition to controlling the behavior and style of the virtual human, we are also interested in understanding how they would respond to environmental changes and perform everyday tasks. The former requires the virtual human to simulate human thinking and react appropriately to external stimuli, while the latter involves decomposing actions into subtasks that are easier to execute. These complex actions are difficult to generate, even though they are quite common in everyday life. To evaluate model’s ability to generate these intricate motions, we established the Complex Motion Evaluation Set.

We gathered 16 scenes from ScanNet (Dai et al., 2017) and Replica (Straub et al., 2019) as environments for virtual human activities, annotating these scenes with their offsets from the origin. We then provided the scene information and multi-view RGB images to a vision-language model (VLM). By asking, “What could an advanced virtual human do in this scene?” we generated a series of possible behaviors. After part-of-speech tagging, these behaviors were transformed into HSI descriptions that guide interactions between the virtual human and the environment. Compared to manually designing interactions, this automated approach saves labor and ensures that the generated descriptions are free from personal bias.

The generative model is tasked with producing complex human motions based on the scene mesh and the aforementioned HSI descriptions. Its performance is evaluated using relevant metrics and a human perceptual study. The HUMANISE dataset, Generalization Evaluation Set, and our newly introduced Complex Motion Evaluation Set all use scenes and text as conditional inputs, generating human motions that align with the semantic meanings of both the text and the scenes. Table 1 presents examples cases from each dataset. They show increasing difficulty levels and pose greater challenges for motion generation methods.

4 EXPERIMENTS

We tested our method on the widely-used HUMANISE dataset for generating simple motions. For more complex motions, we evaluated the method on Generalization Evaluation Set, which is specifically designed to assess the generalization ability of models to unseen cases. Additionally, we tested our model’s capacity to generate difficult yet common human motions in our newly introduced Complex Motion Evaluation Set.

4.1 EVALUATION METRICS

Generation Metrics: When evaluating on the Humanise dataset, we followed the evaluation protocols of (Wang et al., 2022b) and (Zhang et al., 2020). Specifically, we used goal distance to measure grounding accuracy, contact to assess the realism of contact between the generated motion and the scene, and non-collision to quantify the proportion of actions that did not collide with the scene. In the afford-motion Generalization Evaluation Set and our Complex Motion Evaluation Set, we applied the same physical metrics but excluded goal distance, as these tasks did not have a clearly defined target object for interaction. Additionally, in alignment with afford-motion evaluation, we utilized the metrics proposed by (Guo et al., 2022a) to assess the quality of the generated motions.

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

Table 1: Example cases of three datasets.

Evaluation Set	Scene	Control Text
HUMANISE		<ol style="list-style-type: none"> 1. Walk to the desk. 2. Walk to the door. 3. Stand up from the chair. 4. Sit on the chair. 5. Lie down on the sofa.
Generalization Evaluation set		<ol style="list-style-type: none"> 1. A person takes a rest on the sofa. 2. A person is shaking hands with someone. 3. A man jumps to the desk like a rabbit. 4. A man dances on the bed happily. 5. Someone sits on the edge of the bed.
Complex Motion Evaluation Set		<p>Human Centric Motions:</p> <ol style="list-style-type: none"> 1. A tired person returns home. 2. A thief enters the room looking for something. 3. An earthquake is coming and a person feels it. <p>Object Maneuver Motions:</p> <ol style="list-style-type: none"> 1. A person throws something into the dustbin. 2. A person moves the desk to the sofa. 3. A person arranges the room for a meeting.

Table 2: Experimental Results on HUMANISE dataset. Bold indicates the best result.

Method	Goal Distance↓	Contact↑	Non-collision↑	Quality score↑	Action score↑
HUMANISE (Wang et al., 2022b)	0.422	0.8406	0.9977	2.25	3.66
Afford-motion (Wang et al., 2024)	0.156	0.9568	0.9970	3.46	4.47
Ours	0.115	0.9526	0.9979	3.85	4.20

These include the Fréchet Inception Distance (FID) for evaluating the naturalness of the generated motions, R-Precision to gauge the alignment between the generated motions and the text prompts, and a diversity metric to measure the variability in the generated actions.

Perceptual Study: We also conducted a human perceptual study to assess the quality of the generated motions and their consistency with the corresponding text and scene. Participants rated the overall motion quality, including naturalness and collision levels, on a 1-5 scale, recorded as the quality score. They also rated how well the generated motions performed the actions specified in the text and interacted with target objects in the scene, recorded as the action score. Higher ratings indicated better alignment with the text and scene. We enlisted 20 evaluators to score 180 generated motion clips and recorded their average scores.

4.2 RESULTS ON HUMANISE DATASET

The HUMANISE dataset is considered the first widely adopted HSI dataset. It aligns motion sequences from the AMASS dataset with 3D scenes from ScanNet (Dai et al., 2017) and employs an automated annotation method to synthesize paired data rich in human-scene interaction information. However, HUMANISE is limited to simple actions like walking, sitting, and lying down, which constrains its applicability. Table 2 shows the quantitative results of the HUMANISE baseline, afford-motion, and our method for generating simple actions. Our method is primarily designed for generating complex motions. As a result, the task planner typically decomposes HUMANISE tasks into two steps: move to the interaction target and perform the interaction. Despite this, our method surpasses others on metrics such as goal distance, non-collision rate, and the quality score

Table 3: **Experimental Results on Afford-Motion’s Generalization Evaluation Set.** “Real” indicates that these data are reference metrics from the HumanML3D test set. “→” indicates metrics that are better when closer to “Real” distribution.

Method	FID↓	R-precision (Top-3)↑	Diversity →	Contact↑	Non-collision↑	Quality score↑	Action score↑
Real	0.000	0.875	9.442	-	-	-	-
Afford-Motion	7.887	0.478	7.935	0.7198	0.9983	2.06	2.63
Ours	8.215	0.687	7.677	0.9520	0.9972	4.03	4.22

Table 4: **Experimental Results on Our Complex Motion Evaluation Set.**

Method	FID↓	R-precision (Top-3)↑	Diversity →	Contact↑	Non-collision↑	Quality score↑	Action score↑
Real	0.000	0.875	9.442	-	-	-	-
Afford-Motion	10.97	0.300	8.087	0.7277	0.9961	2.24	1.72
Ours-text only	43.75	0.251	3.754	0.6955	0.9915	-	-
Ours-w/o planner	12.57	0.362	8.219	0.9104	0.9973	2.09	2.41
Ours	9.31	0.365	7.987	0.8444	0.9979	3.15	3.17

based on human evaluations. Notably, our method excels in goal distance, demonstrating its ability to precisely guide the virtual human to the target and initiate the interaction.

4.3 RESULTS ON GENERALIZATION EVALUATION SET

Researchers in the HSI field have increasingly sought to generate more than just routine actions like walking or sitting. This set contains 16 scenes from ScanNet (Dai et al., 2017), PROX (Hassan et al., 2019), Replica (Straub et al., 2019), and Matterport3D (Chang et al., 2017), along with 80 carefully crafted HSI descriptions. These descriptions often specify multiple aspects of a single action, such as “a person jumps to the desk like a rabbit”, which not only requires the virtual human to jump in a rabbit-like manner but also defines the target location as the desk. Such descriptions are rarely found in existing training datasets, posing significant challenges for motion generation methods in terms of generalization. Table 3 presents the performance metrics of afford-motion and our method on the Generalization Evaluation Set. While our method scores slightly lower than afford-motion on FID and Diversity, it significantly outperforms it on R-precision, indicating better adherence to the text and scene constraints. This is further corroborated by a larger gain in both quality score and action score.

4.4 RESULTS ON COMPLEX MOTION EVALUATION SET

Section 3.5 describes the details of our proposed Complex Motion Evaluation Set. We evaluated the performance of Afford-Motion and GCML on this evaluation set. As shown in Table 4, our method consistently outperforms afford-motion on all metrics in this new test set. In many cases, afford-motion only generates irrelevant or meaningless actions, while our method can break down complex action descriptions into sequences of subtasks and execute them accordingly. Our method performs particularly well in R-precision and contact metrics, indicating that it closely follows textual instructions and exhibits rich interaction with the environment. The perceptual study shows that our method achieves promising results in both overall generation quality and adherence to conditions. Notably, our approach allows precise control of action duration based on the text, supporting the generation of interaction sequences lasting up to several tens of seconds.

4.5 ABLATION STUDY

As shown in Table 4, firstly, we directly pass the text prompts to the motion generation model without using the whole pipeline we proposed. In this case, the generated results were scene agnostic, leading to a significant drop in physical metrics compared to others. Furthermore, a more critical



Figure 4: Comparison of the qualitative results between the Afford-Motion and our method on various datasets. Brighter actions precede darker actions.

issue arose: the motion generation model struggled to comprehend the actual intent behind complex user instructions, resulting in highly distorted human motions in most cases.

Next, we investigated the effect of task decomposition on motion generation outcomes. Here, we skipped the planner stage and directly provided the executor with the text prompts and scene data to produce control commands and spatial information for the motion generation model. The results, shown in the fourth row of Table 4, reveal that the LLM planner plays a crucial role in enhancing motion quality. Without the planner, the LLM Executor generates spatial control data that is challenging for the Motion Generation Model to interpret and follow, thereby diminishing the overall quality of the generated motions.

4.6 QUALITATIVE RESULTS

Figure 4 illustrates the visualization results comparing our method with Afford-Motion across three datasets. The left column shows the generation results for simple actions on the HUMANISE dataset, where both methods produce satisfactory results. However, our method allows for the specification of the character’s initial pose. The middle column presents the outcomes from the Generalization Evaluation Set. In this example, the user prompt indicates that the target should be grasped with the left hand, but Afford-Motion overlooks this instruction during the generation process. The rightmost column presents the generation results from our Complex Motion Evaluation Set. When text prompts do not explicitly specify the desired action, Afford-Motion generates irrelevant or distorted motions, while our method can infer the implicit motion directives. Here, our approach establishes a connection between actions like “avoiding with a lifted foot” and “a sudden mouse”, enabling the generation of coherent character motions based on abstract prompts.

5 CONCLUSION

This paper introduces GCML, a novel method for generating complex human motions guided by textual descriptions within a scene. By utilizing a Large Language Model for task decomposition and subtask execution, and a 3D Visual Grounding Model for scene perception, our method produces complete complex motion frames. We validated the effectiveness of our method across multiple datasets, with experimental results showing that our approach performs well in generating simple human motions. Moreover, on our newly introduced test set for complex human motion generation, our method consistently outperformed existing methods across all evaluation metrics.

REFERENCES

- 540
541
542 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-
543 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical
544 report. *arXiv preprint arXiv:2303.08774*, 2023.
- 545
546 Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and
547 Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In
548 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9297–9307, 2019.
- 549
550 Valts Blukis, Ross A Knepper, and Yoav Artzi. Few-shot object grounding and mapping for natural
551 language robot instruction following. *arXiv preprint arXiv:2011.07384*, 2020.
- 552
553 Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush
554 Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for
555 autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern
556 recognition*, pp. 11621–11631, 2020.
- 557
558 Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term
559 human motion prediction with scene context. In *Computer Vision–ECCV 2020: 16th European
560 Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 387–404. Springer,
561 2020.
- 562
563 Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva,
564 Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor
565 environments. *arXiv preprint arXiv:1709.06158*, 2017.
- 566
567 Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in
568 rgb-d scans using natural language. In *European conference on computer vision*, pp. 202–221.
569 Springer, 2020.
- 570
571 Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your
572 commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on
573 Computer Vision and Pattern Recognition*, pp. 18000–18010, 2023.
- 574
575 Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias
576 Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the
577 IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.
- 578
579 Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A bench-
580 mark for visual object affordance understanding. In *proceedings of the IEEE/CVF conference on
581 computer vision and pattern recognition*, pp. 1778–1787, 2021.
- 582
583 Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek.
584 Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer
585 Graphics Forum*, volume 42, pp. 1–12. Wiley Online Library, 2023.
- 586
587 Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating
588 diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on
589 Computer Vision and Pattern Recognition*, pp. 5152–5161, 2022a.
- 590
591 Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating
592 diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on
593 Computer Vision and Pattern Recognition (CVPR)*, pp. 5152–5161, June 2022b.
- 594
595 Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for
596 the reciprocal generation of 3d human motions and texts. In *European Conference on Computer
597 Vision*, pp. 580–597. Springer, 2022c.
- 598
599 Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D
600 human pose ambiguities with 3D scene constraints. In *International Conference on Computer
601 Vision*, pp. 2282–2292, October 2019. URL <https://prox.is.tue.mpg.de>.

- 594 Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J
595 Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International*
596 *Conference on Computer Vision*, pp. 11374–11384, 2021.
- 597
- 598 Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-
599 Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings*
600 *of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16750–16761,
601 2023a.
- 602 Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer:
603 Composable 3d value maps for robotic manipulation with language models. *arXiv preprint*
604 *arXiv:2307.05973*, 2023b.
- 605
- 606 Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a
607 foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023.
- 608
- 609 Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin
610 Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings*
611 *of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1737–1747, 2024.
- 612 Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided
613 motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF*
614 *International Conference on Computer Vision*, pp. 2151–2162, 2023.
- 615
- 616 Kamran Khan, Saif Ur Rehman, Kamran Aziz, Simon Fong, and Sababady Sarasvady. Dbscan: Past,
617 present and future. In *The fifth international conference on the applications of digital information*
618 *and web technologies (ICADIWT 2014)*, pp. 232–238. IEEE, 2014.
- 619 Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis
620 & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp.
621 8255–8263, 2023.
- 622
- 623 Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and
624 Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis.
625 In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.
626 947–957, 2024.
- 627 Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music condi-
628 tioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Confer-*
629 *ence on Computer Vision*, pp. 13401–13412, 2021.
- 630
- 631 Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene
632 understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45
633 (3):3292–3310, 2022.
- 634 Bill Yuchen Lin, Yicheng Fu, Karina Yang, Faeze Brahman, Shiyu Huang, Chandra Bhagavatula,
635 Prithviraj Ammanabrolu, Yejin Choi, and Xiang Ren. Swiftsage: A generative agent with fast and
636 slow thinking for complex interactive tasks. *Advances in Neural Information Processing Systems*,
637 36, 2024.
- 638
- 639 Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiyue Ping, and Qin Chen. Agentsims: An
640 open-source sandbox for large language model evaluation. *arXiv preprint arXiv:2308.04026*,
641 2023.
- 642 Haokun Liu, Yaonan Zhu, Kenji Kato, Atsushi Tsukahara, Izumi Kondo, Tadayoshi Aoyama, and
643 Yasuhisa Hasegawa. Enhancing the llm-based robot manipulation through human-robot collabo-
644 ration. *arXiv preprint arXiv:2406.14097*, 2024.
- 645
- 646 Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black.
647 AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer*
Vision, pp. 5442–5451, October 2019.

- 648 Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and
649 Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings*
650 *of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.
- 651 Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios
652 Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single
653 image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,
654 pp. 10975–10985, 2019.
- 655 Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas
656 Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of*
657 *the IEEE/CVF conference on computer vision and pattern recognition*, pp. 815–824, 2023.
- 658 Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive
659 3d human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Com-*
660 *puter Vision*, pp. 9488–9497, 2023.
- 661 Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big*
662 *Data*, 4(4):236–252, dec 2016. doi: 10.1089/big.2016.0028. URL [http://dx.doi.org/](http://dx.doi.org/10.1089/big.2016.0028)
663 [10.1089/big.2016.0028](http://dx.doi.org/10.1089/big.2016.0028).
- 664 Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a gener-
665 ative prior. *arXiv preprint arXiv:2303.01418*, 2023.
- 666 Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel,
667 Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor
668 spaces. *arXiv preprint arXiv:1906.05797*, 2019.
- 669 Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Ex-
670 posing human motion generation to clip space. In *European Conference on Computer Vision*, pp.
671 358–374. Springer, 2022.
- 672 Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano.
673 Human motion diffusion model. In *The Eleventh International Conference on Learning Repre-*
674 *sentations*, 2023. URL <https://openreview.net/forum?id=SJ1kSy02jwu>.
- 675 Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music.
676 In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.
677 448–458, 2023.
- 678 Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan,
679 and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models.
680 *arXiv preprint arXiv:2305.16291*, 2023.
- 681 Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and
682 natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE/CVF Conference on*
683 *Computer Vision and Pattern Recognition*, pp. 20460–20469, 2022a.
- 684 Zhan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise:
685 Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information*
686 *Processing Systems*, 35:14959–14971, 2022b.
- 687 Zhan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin
688 Zhu, Wei Liang, and Siyuan Huang. Move as you say interact as you can: Language-guided
689 human motion generation with scene affordance. In *Proceedings of the IEEE/CVF Conference on*
690 *Computer Vision and Pattern Recognition*, pp. 433–444, 2024.
- 691 Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiang
692 Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE*
693 *conference on computer vision and pattern recognition*, pp. 1912–1920, 2015.
- 694 Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiang-
695 miao Pang. Unified human-scene interaction via prompted chain-of-contacts. *arXiv preprint*
696 *arXiv:2309.07918*, 2023.

- 702 Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-
703 based human motion estimation and synthesis from videos. In *Proceedings of the IEEE/CVF*
704 *International Conference on Computer Vision*, pp. 11532–11541, 2021.
- 705
706 Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any
707 joint at any time for human motion generation. *arXiv preprint arXiv:2310.08580*, 2023.
- 708 Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-
709 object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF Interna-*
710 *tional Conference on Computer Vision*, pp. 14928–14940, 2023.
- 711
712 Ziyi Yang, Shreyas S Raman, Ankit Shah, and Stefanie Tellex. Plug in the safety chip: Enforcing
713 constraints for llm-driven robot agents. In *2024 IEEE International Conference on Robotics and*
714 *Automation (ICRA)*, pp. 14435–14442. IEEE, 2024.
- 715 Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In
716 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020,*
717 *Proceedings, Part IX 16*, pp. 346–364. Springer, 2020.
- 718 Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll.
719 Couch: Towards controllable human-chair interactions. In *European Conference on Computer*
720 *Vision*, pp. 518–535. Springer, 2022.
- 721
722 Yan Zhang and Siyu Tang. The wanderings of odysseus in 3d scenes. In *Proceedings of the*
723 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20481–20491, 2022.
- 724
725 Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d
726 people in scenes without people. In *Proceedings of the IEEE/CVF conference on computer vision*
727 *and pattern recognition*, pp. 6194–6204, 2020.
- 728 Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, and Siyu Tang. Synthesizing diverse
729 human motions in 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference*
730 *on Computer Vision*, pp. 14738–14749, 2023.
- 731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755