SemGeoMo: Dynamic Contextual Human Motion Generation with Semantic and Geometric Guidance

Peishan Cong[1,2,*], **Ziyi Wang**[1,*], **Yuexin Ma**[1,†], **Xiangyu Yue**[2,†]
[1] ShanghaiTech University [2] The Chinese University of Hong Kong
{congpsh,wangzy17,mayuexin}@shanghaitech.edu.cn

Abstract

Generating reasonable and high-quality human interactive motions in a given dynamic environment is crucial for understanding, modeling, transferring, and applying human behaviors to both virtual and physical robots. In this paper, we introduce an effective method, SemGeoMo, for dynamic contextual human motion generation, which fully leverages the text-affordance-joint multi-level semantic and geometric guidance in the generation process, improving the semantic rationality and geometric correctness of generative motions. Our method achieves state-of-the-art performance on three datasets and demonstrates superior generalization capability for diverse interaction scenarios. The project page and code can be found at https://4dvlab.github.io/project_page/semgeomo/.

1. Introduction

Dynamic contextual human motion generation [19, 21, 22] aims to generate human interaction motions that are both commonsense and geometrically accurate, adapting seamlessly to real dynamic environments. Its core lies in constructing an interaction-oriented world model for humans, enabling reasonable adaptation to changes of interactive objects or people. This interaction-oriented world model can support a wide range of applications, including human-robot interaction, closed-loop simulators, intelligent sports coaching, and immersive VR/AR gaming experiences.

As the importance of interaction becomes increasingly recognized, some studies have evolved from text-driven human motion generation [4, 6, 16, 31, 32, 49] to text-driven joint generation of human-object or human-human interactions [2, 22, 23, 37, 42]. However, generating motions jointly for both the human and the interactive target creates an excessively large search space, often leading to sub-optimal generation results. Additionally, the lack of fine-grained control over the generated data hinders the creation of personalized interactive motions, limiting its applicability to real-world scenarios such as humanoid operation and human-robot interaction.

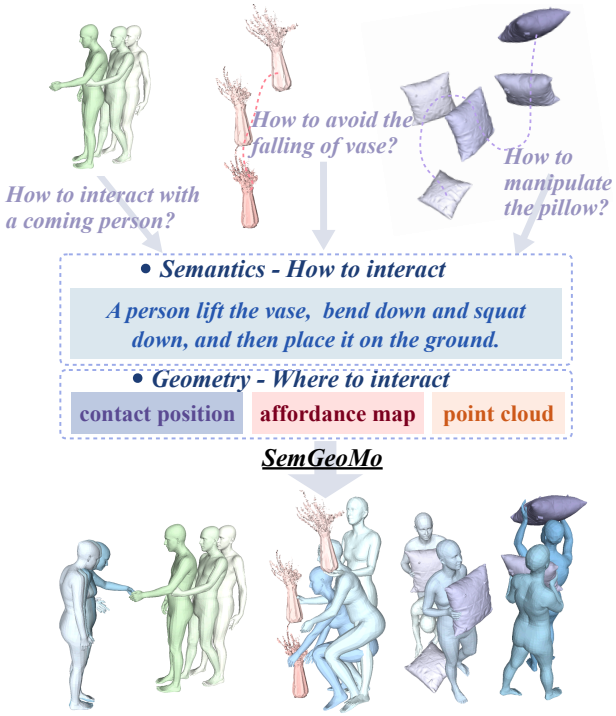

Figure 1. Given sequential point clouds of interactive targets, SemGeoMo generates realistic and high-quality human interactive motions along with corresponding textual descriptions. By leveraging both semantic and geometric guidance, our method ensures the semantic coherence and geometric accuracy of the generated results.

Recently, contextual human motion generation has garnered increasing attention for its ability to produce controllable interactions based on specified scenario conditions. However, many works [3, 14, 35, 38, 40] only focus on generating coarse-grained human trajectories and motions in static environments with fixed furniture or layouts

*Equal contribution. † Corresponding author. This work was supported by NSFC (No.62206173, No. 62306261), Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI), MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (KLIP-HuMaCo), and the Shun Hing Institute of Advanced Engineering (SHIAE) No. 8115074.

rather than the fine-grained details of interactions, limiting adaptability to dynamic conditions. A few recent studies [5, 22, 43] have begun to explore dynamic contextual human motion generation. However, these approaches have notable limitations: 1) they lack textual guidance, which undermines the semantic coherence of interactions and limits the generalization capability of the approach, and 2) they fail to incorporate fine-grained geometric representations, resulting in insufficient constraints on the geometric accuracy of generated interaction motions.

In this work, we propose a novel dynamic contextual human motion generation method, named **SemGeoMo**, illustrated in Fig.1, which could generate reasonable and high-quality interactive motions by comprehensively integrating semantic information from textual descriptions with hierarchical geometric features extracted from interactive objects. The first challenge is *how to construct the semantic guidance*. Given that large language models (LLMs) possess general knowledge and can provide rich information on the attributes of interacting objects as well as guidance for the interaction process, we introduce an automated interaction text annotator. By leveraging careful prompt design and fine-tuning of the model [30, 33], our **LLM Annotator** eliminates the need for manual text labeling, offering strong support for the semantic coherence and generalization of generated human interaction behaviors. The second challenge is *how to construct the geometric guidance*. To ensure the geometric accuracy of interactive motions, such as avoiding geometric penetration and ensuring appropriate contact between human and object, we propose a two-stage framework that decouples contact geometry generation from interactive motion generation. In the first stage, **SemGeo Hierarchical Guidance Generation**, a diffusion model generates affordance-level and joint-level interaction cues guided by semantic information to capture both coarse and precise geometric positioning. In the second stage, **SemGeo-guided Motion Generation**, these cues are effectively utilized to guide the generation of detailed human motions, improving both semantic plausibility and geometric accuracy. It is worth noting that our model simultaneously generates human motions and language descriptions at varying levels of granularity, which not only improves the quality of the generated motions but also enhances the interpretability and comprehensibility of the interactions.

Extensive experiments demonstrate that we achieve state-of-the-art performance on three human-object interaction datasets. Moreover, we demonstrate the generalization capability of our method in more challenging scenarios, including interactions with unseen objects, human-human interactions, and interactions with deformable objects. To summarize, our work makes the following contributions:

- We propose a novel method that generates responsive human motions and corresponding textual descriptions based on observed dynamic interactive targets.
- We introduce an automated text annotator that interprets reactions during interactions, reducing the burden of manual labeling and enhancing the generalization capability.
- Our method fully utilizes multi-level semantic and geometric guidance—including text, affordance, and joint-level cues—throughout the generation process, improving both the semantic rationality and geometric accuracy of interactive motion.
- Our method generates high-quality human motions and achieves state-of-the-art performance across three benchmarks and an unseen dataset.

2. Related Work

2.1. Text-guided Human Motion Generation

Text-conditioned human motion generation [6, 7, 15, 31, 32, 34, 49, 50] have made significant progress with the rise of diffusion models [11, 12]. With the emergence of human-object interaction datasets [2, 22, 47, 51] and human-human interaction datasets [23, 42], some works have begun to explore text-driven interaction motion generation. Several studies [8, 27, 39, 44] dive into jointly generating a sequence of human and object poses based on textual conditions. HOI-Diff [27] emphasizes the significance of affordance information and Thor [39] refines object rotation during each inverse diffusion step for human-object interaction. Furthermore, CHOIS [21] further integrates the 2D waypoints with object geometry loss during sampling process. Other studies [23, 42] focus on human-human interaction, where two human motions are jointly generated based on language descriptions. However, the joint generation of human motions and interactive targets creates a large optimization search space during training, leading to lower quality in the generated motions. Moreover, conditioning solely on text lacks fine-grained control over the generation process, limiting its applicability in areas such as robotic operations, human-robot interaction, and AR/VR immersive experiences.

2.2. Contextual Human Motion Generation

Contextual human motion generation explores settings that are more applicable to real-world scenarios, where human motion is influenced by and interacts with the given environment. Several works [18, 20] focus on interactions with seating furniture [17] and others [3, 14, 35, 38] generate natural human motions in 3D indoor scenes [1, 9, 19, 36, 40, 45]. SceneDiff [14] utilizes point clouds as conditions to generate feasible interactions. AffordMotion [38] provides a two-stage framework that employs a scene affordance map as an intermediary. However, these works are limited to static environments, where interacted objects are often restricted to fixed furniture like beds and chairs, and

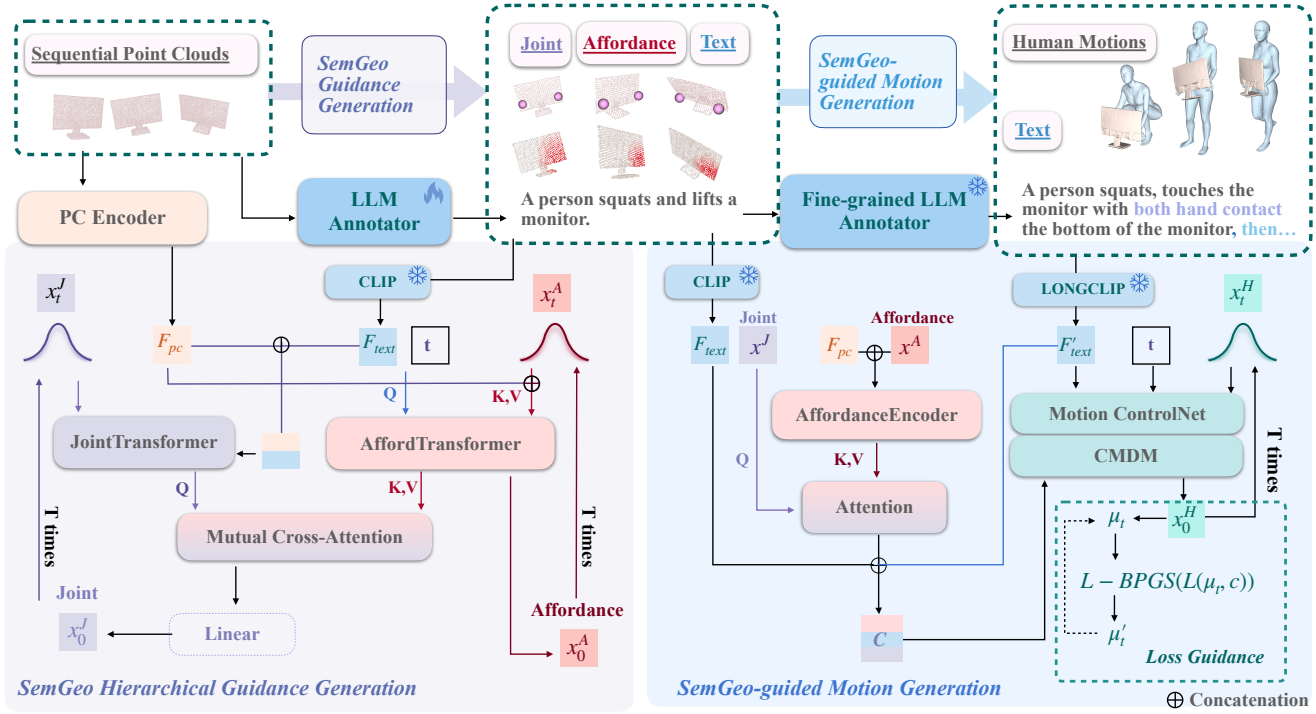

Figure 2. The pipeline of our two-stage framework. LLM Annotator provides the semantic guidance. SemGeo Hierarchical Guidance Generation takes textual information and sequential point cloud as condition and generate affordance-level and joint-level guidance. Then SemGeo-guided Motion Generation utlizes semantic and geometric information to generate responsive human motion.

the category of motions is limited to sitting, lying, and walking. They focus on the trajectory and the goal state, rather than interacting with dynamic, ever-changing targets.

Following works [5, 22, 43] introduce dynamic, interactive targets, such as movable objects or other people. ReGenNet [43] generates human reactions conditioned on given human motion in SMPL [25] representation and the action condition. However, SMPL representation requires additional processing on raw sensor data and is not suitable for all dynamic targets such as objects. A more flexible point cloud-based representation is used in [5] for interacting scenes or objects. However, it lacks a carefully designed feature modeling method for geometric and temporal information, resulting in suboptimal performance. OMOMO [22] presents a framework for generating human behaviors based on object motions, utilizing conditional diffusion to generate hand joint positions as extra guidance. While this work lacks the integration of textual information, which can provide important semantic guidance. By contrast, our method fully leverages textual reasoning and incorporates hierarchical semantic and geometric features to enhance contextual motion generation.

3. Method

Our goal is to generate responsive human motions conditioned solely on the interactive target, represented as a 4D sequential point cloud. To achieve this, we introduce the LLM Annotator to provide semantic features, which are then processed in the following two stages: SemGeo Hierarchical Guidance Generation and SemGeo-guided Motion Generation. In the first stage, multi-level geometric information is generated in a coarse-to-fine manner. These generated guidance signals are then used in the second stage to guide the motion generation process, enhancing both semantic plausibility and geometric accuracy. The overall architecture is shown in Fig. 2.

3.1. Data Representation

Motion Representations. We denote the human motion as $\mathbf{x}^h \in \mathbb{R}^{L\times D}$, where L represents the number of frames in the sequence and D is the feature dimension, which includes the pelvis velocity, local joint positions, rotations and velocities of other joints in the pelvis space, as well as binary foot-ground contact labels, following the representation in MDM [32].

Point Cloud Representations. To extract geometric features, point clouds are well-suited for capturing the dynamic changes of objects, providing a unified input modality for representing deformable or non-rigid interactive targets without the need for additional preprocessing. The sequential point cloud is down-sampled to $N = 1024$ points per frame, denoted as $\mathbf{P} \in \mathbb{R}^{L\times N\times 3}$. We then adopt Basis Point Set [28] (BPS) representation, following the approach in [22], to encode the object geometry. The process begins by uniform sampling the basis points from a unit ball with

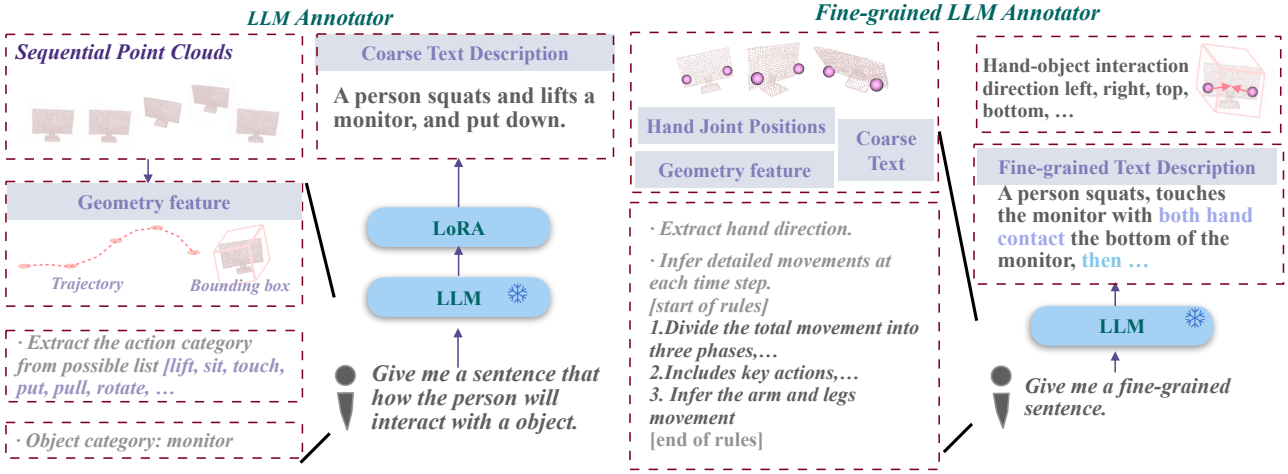

Figure 3. LLM Annotator pipeline. It first takes a sequential point cloud to infer a coarse text description. Then uses joint positions, the coarse text, and geometric features to generate a fine-grained sentence with a designed prompt.

a radius of 1 meter, followed by calculating the minimum Euclidean distance from the basis points to their nearest neighbors. Finally, these distances are concatenated with the center position of the original point cloud. The resulting point cloud representation for each frame is denoted as $\mathbf{p} \in \mathbb{R}^{N\times 3+3}$. An MLP is then applied to project the BPS representation into a lower-dimensional space, yielding the sequential geometry features $\mathbf{F_{pc}} \in \mathbb{R}^{L\times 256}$.

Affordance Map Representations. The affordance map serves as an intermediate representation, providing crucial geometric information regarding which parts of the target are most likely to come into contact during the interaction. For each frame, we calculate the ℓ_2 distance between each point and the skeleton joints, resulting in a per-frame distance map $d \in \mathbb{R}^{L\times N\times J}$, where J is the number of skeleton joints. We then transform this distance field into a normalized distance map, Affordance, denoted as $\mathbf{A}$, which encodes the spatial relationship between the target and the interactive human. The affordance map is computed as:

$$\textbf{Affordance}(n,j) = \exp\left(-\frac{1}{2}\frac{\mathbf{d}(n,j)}{\sigma^2}\right), \qquad (1)$$

where σ is the normalizing factor.

3.2. LLM Annotator

Textual descriptions provide essential semantic information. Given the interactive targets, envisioning how to interact with them is crucial for generating realistic human motion. Previous work [31, 32, 49] on text-to-motion has also validated the effectiveness of feature mapping from textual descriptions to motion generation.

To obtain textual guidance, we employ a Large Language Model (LLM) as an annotator in the initial stage to generate a coarse motion description. The whole pipeline is shown in Fig. 3. From the sequential point cloud, we extract the bounding box of the interactive target along with its movement trajectory. Utilizing a predefined list of actions and categories, we derive initial language-based guidance. Specifically, we utilize the pre-trained LLaMA model [33], enhanced with LoRA [13] finetuning on given text-interaction pairs from [22].

More detailed, context-aware cues can capture the dynamic changes in contact, guiding the model to generate more precise body movements. Leveraging the strong reasoning capabilities of LLMs, we design a coarse-to-fine automated language guidance annotation system. After generating initial textual guidance and predicting hand joint positions, we determine sequential contact information by calculating distances between predicted joint positions and interactive targets, inferring where the body should make contact with each part of the targets. The LLM then organizes this contact information into language descriptors (*e.g.*, "the left-hand contacts the lower left of the box"). Simultaneously, we let the LLM divide the interaction into three steps, with the textual descriptions reflecting changes in contact at each step, with also deducing finer body movements of the arms and legs.

Thus, the generated textual descriptions assist in the motion generation task by providing semantic information. On the other hand, our entire pipeline also enables reasoning capabilities, allowing for a more comprehensive understanding and generation of human motion.

3.3. SemGeo Hierarchical Guidance Generation

The joint position $\mathbf{J_h} \in \mathcal{R}^{L\times J\times 3}$ provides precise spatial information, and the affordance map $\mathbf{A}$ offers coarse geometric clues. These two types of features are crucial for modeling interactions in a coarse-to-fine manner, thus we introduce a conditional diffusion model with dual-branch transformer to jointly generate contact information with capturing their mutual influence in the first stage.

We utilize CLIP[29] as text encoder to obtain the text feature F_{text}. At each step in the diffusion process, the model take the text feature F_{text}, point cloud feature F_{pc}, and noisy signals x_t^J and x_t^A as input and predict clean x_0^J and x_0^A. For JointTransformer, the inputs are concatenated together and then fed into the multi-head self-

attention blocks followed with position-wise feedforward layer. For AffordanceTransformer, inspired by [38], the affordance map is more closely related to the point cloud geometry. Therefore, we encode the point cloud feature F_{pc} along with x_t^A, which act as the key and value in attention module. The concatenation of the language feature and diffusion step embeddings serves as the query. After passing through the cross-attention mechanism and multiple self-attention blocks, the refined point features are obtained. To enhance the coarse to fine interactions and further refine the joint position, we introduce a mutual cross-attention mechanism. This take the output from Jointtransformer as query, output from AffordanceTransformer as key and value, which updates the hand joint position.

The conditional diffusion model learns the reverse diffusion process to generate clean data from a Gaussian noise x_t over T consecutive denoising steps. Specifically, we use c to represent the conditions, and the reverse diffusion process is modeled as:

$$p_\theta\left(x^{t-1} \mid x^t, \boldsymbol{c}\right) := \mathcal{N}\left(x^{t-1}; \boldsymbol{\mu}_\theta\left(x^t, t, \boldsymbol{c}\right), \sigma_t^2 I\right). \quad (2)$$

Finally, our model directly estimates the input signal. The training process optimizes from the reconstruction loss:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}^0, t}\left\|\hat{x}_\theta\left(x^t, t, \boldsymbol{c}\right) - x^0\right\|_1. \quad (3)$$

3.4. SemGeo-guided Motion Generation

In the second stage, a denoising network architecture is used to generate full-body motions based on the predicted joint positions $\mathbf{J}_h'$, affordance map $\mathbf{A}'$, and text descriptions.

SemGeo Condition Module To effectively process and integrate these diverse features, we design SemGeo Condition Module to encode the combined features. After obtaining both coarse and fine-grained text descriptions, we use a text encoder to extract the semantic information. Since the language generated by the Fine-grained LLM Annotator contains phase-specific details and longer descriptions, we adopt LONGCLIP [46], which can better capture and represent fine-grained attributes without the length limitations inherent in CLIP. The extracted feature is F'_{text}.

To extract features from the affordance map, we first concatenate the point cloud feature F_{pc} with the affordance map $\mathbf{A}'$, and then pass the combined input through a 3-layer MLP. The temporal transformer is applied to extract latent features F over time, which helps to capture both spatial and temporal dependencies. The operation is formalized as:

$$F = \text{TemporalTransformer}(MLP(F_{pc} \oplus \mathbf{A}'), \quad (4)$$

where $\oplus$ denotes concatenation.

We apply mutual cross-attention to extract the mutual features between joint positions and the affordance map. The joint feature, after being mapped to a higher dimensional space with MLP, serves as the query. The latent feature from the previous step is used as key and value in the attention mechanism. This interaction between joint positions and the affordance map enables the model to capture the relationship between joint movements and object geometry, improving motion prediction:

$$F_{fuison} = \text{CrossAttention}(MLP(\mathbf{J}_h')_q, F_k, F_v). \quad (5)$$

The final condition c is the concatenation result of F_{text}, F'_{text}, F_{fuison}.

This comprehensive feature representation ensures that the model has a rich, multi-dimensional understanding, capturing both coarse-to-fine semantic information from the text and spatial-temporal geometric information from the joint positions and affordances.

Motion ControlNet Inspired by [37, 41], we introduce Motion ControlNet to generate high-fidelity motions conditioned on c. With MDM frozen during training, each transformer encoder layer in ControlNet [48] is linked to its MDM counterpart via a zero-initialized linear layer.

Loss Guidance To refine our generated interactions, we employ joint guidance and foot guidance during sampling with classifier guidance. The joint-based guidance loss aligns the generated global joint positions $\mathbf{J}_{\text{pred}}$ from the second stage with the target control joint positions $\mathbf{J}'_{\mathbf{h}}$ from the first stage, ensuring consistency in the generated motion across both stages. Constraints are applied to the joints that are in contact with objects when the distance between the joint and the nearest point on the object is below a predefined threshold τ. Specifically, the mask is defined as $\text{Mask} = \text{Dis}(\mathbf{J}_{\text{h}}, V) \leq \tau$. The joint guidance function is then defined as:

$$L_{\text{joint}} = \frac{1}{J}\sum_{i=1}^{L} |\mathbf{J}_{\text{pred}_i} - \mathbf{J}'_{\text{h}_i}|_2 \cdot \text{Mask}_i. \quad (6)$$

The foot-stability loss L_{foot} is designed to ensure that the foot stays near the ground and penalize sudden changes in velocity to eliminate foot sliding:

$$L_{\text{foot}} = \frac{1}{L}\sum_{i=1}^{L}\left((y_i - h_g)^2 + \alpha \mathbf{M}_c(v_i^2) + \beta \mathbf{M}_c(a_i^2)\right), \quad (7)$$

where $y_i = \min(h_{l,i}, h_{r,i})$ is the height of the lower foot for each frame and h_g is the empirical values indicating contact with the ground. $v_i = \|p_{i+1} - p_i\|$ is the foot velocity at frame i, calculated from the foot position p and $a_i = \|v_{i+1} - v_i\|$ is the foot acceleration. $\mathbf{M}_c$ represents the mask for contact with the floor. α and β are hyperparameters. Inspired by [37], we employ L-BFGS for several iterations at each denoising step to update the posterior mean.

4. Experiment

4.1. Datasets

FullBodyManipulation [22] takes a total duration of approximately 10 hours. It provides, paired object and human

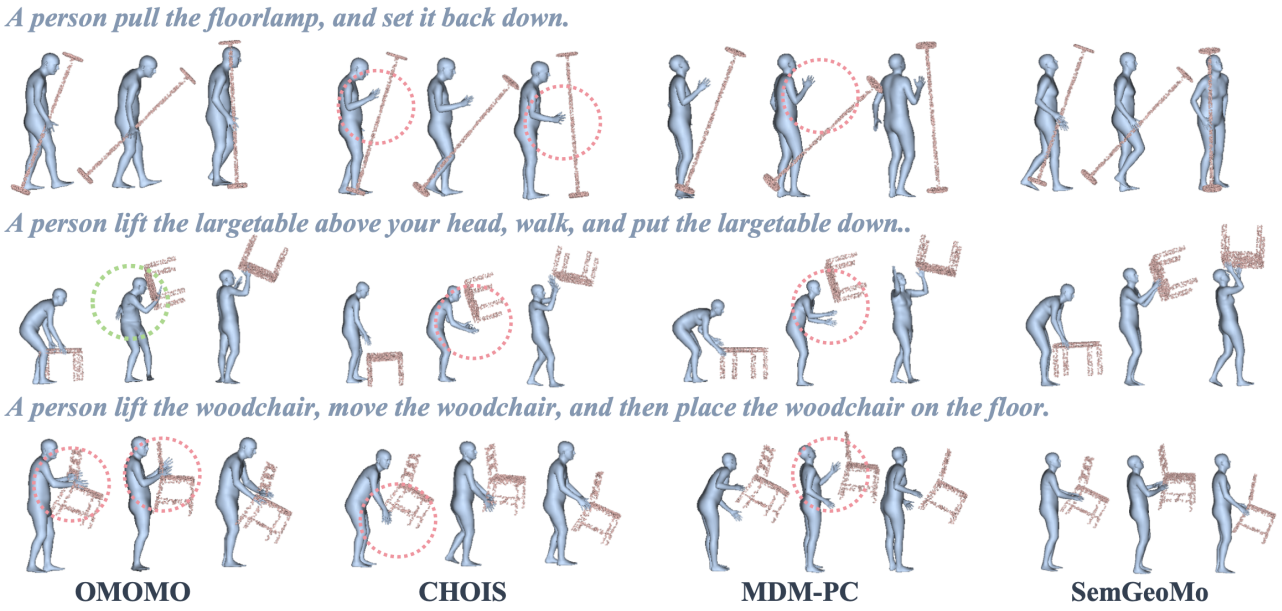

Figure 4. Qualitative results on the FullBodyManipulation dataset. We circle areas of low contact performance in pink and instances of contorted motion in green.

Table 1. Human motion generation result on FullBodyManipulation.

		HandJPE↓	MPJPE ↓	C_{prec} ↑	C_{rec} ↑	C_{acc} ↑	c%↑	F1 ↑	FID ↓	R-score ↑	Diversity ↑	FS ↓
w/o text	SceneDiff [14]	95.38	19.84	0.64	0.19	0.45	0.18	0.27	1.64	0.59	9.86	0.38
	OMOMO [22]	33.18	18.06	0.77	0.71	0.74	0.61	0.75	1.98	0.38	8.99	0.50
w GT text	MDM-PC [32]	51.35	18.25	0.71	0.41	0.62	0.33	0.49	**0.65**	0.57	9.52	0.51
	CHOIS [21]	31.68	17.12	0.76	0.58	0.61	0.52	0.59	2.27	0.49	6.04	0.47
	AffordMotion [38]	98.66	25.34	0.45	0.14	0.31	0.13	0.16	4.71	0.45	8.15	**0.43**
	SemGeoMo	**27.84**	**16.62**	**0.84**	**0.74**	**0.85**	**0.66**	**0.77**	1.17	**0.66**	**10.15**	0.57
w Gen text	SemGeoMo	30.35	17.98	0.82	0.74	0.82	0.66	0.74	1.05	0.64	9.78	0.47

motion, including interactions of 17 subjects with 15 different objects with text descriptions. We follow the official train/test split for evaluation.

BEHAVE [2] consists of the interactions of 8 subjects with 20 different objects. We follow the official train/test split provided by BEHAVE. [27] provides the text annotations while we observe that these annotations merely combine action labels and categories of motion, offering phase-specific details. We provide a revised version of the annotations with phased interactions that more clearly describe the motions.

IMHD^2 [51] and **HoDome** [47] are challenging 3D human-object interaction datasets for motion capture, covering interactions between 10 objects and 10 subjects, and 23 diverse objects with 10 subjects, respectively. We annotate the text descriptions using our LLM annotation pipeline to facilitate further text-driven human-object interaction studies and enable comparisons with other text-guided methods.

4.2. Evaluation Metrics

We mainly follow the metrics in OMOMO [22] and MDM [32]. **HandJPE** and **MPJPE** represent mean hand joint position errors, and mean per-joint position errors, computed using the Euclidean distance between the predicted and ground truth in centimeters (cm). For measuring the interaction quality, we employ contact metrics including precision ($\mathbf{C}_{prec}$), recall ($\mathbf{C}_{recall}$), accuracy ($\mathbf{C}_{acc}$) and **F1 score** following [22]. The contact percentage ($c\%$) is the proportion of frames where contact is detected. **FID** measures the distance of the generated motion distribution to the ground truth distribution in latent space. **R-score** measures the text and motion matching accuracy and **Diversity** measures the generation diversity. **FS** represents foot sliding metric and is computed following [10].

4.3. Results

Baselines OMOMO [22] is the only work which is aligned with our setting. In addition, we adapt several related works, including SceneDiff [14], MDM-PC [32], AffordMotion [38], and CHOIS [21], to fit our problem setting. SceneDiff [14] utilizes a diffusion model conditioned on static scenes, so we modified the conditional model to accommodate sequential point cloud inputs. We modified the text-to-motion generation work MDM as MDM-PC, incorporating our sequential point cloud representation. CHOIS and AffordMotion are conditioned on both textual descriptions and scenarios. CHOIS requires a sequence of object

Table 2. Human motion generation result on Behave.

		HandJPE↓	MPJPE ↓	C_{prec} ↑	C_{rec} ↑	C_{acc} ↑	c%↑	F1 ↑	FID ↓	R-score ↑	Diversity ↑	FS ↓
w/o text	SceneDiff [14]	51.58	18.25	0.73	0.38	0.40	0.32	0.47	1.69	0.13	5.32	0.33
	OMOMO [22]	45.35	21.56	0.71	0.60	0.61	0.60	0.62	1.94	0.14	5.11	0.42
w Gen text	MDM-PC [32]	35.41	18.61	0.73	0.48	0.51	0.47	0.57	1.52	0.10	5.45	**0.32**
	CHOIS [21]	36.75	18.17	0.72	0.41	0.43	0.41	0.51	2.26	0.13	5.02	0.46
	AffordMotion [38]	55.65	19.16	0.72	0.23	0.28	0.25	0.32	1.92	0.13	4.38	0.51
	SemGeoMo	**27.91**	**16.22**	**0.84**	**0.67**	**0.67**	**0.66**	**0.74**	**1.47**	**0.15**	**5.64**	0.52

Table 3. Human motion generation result on IMHD^2.

		HandJPE↓	MPJPE ↓	C_{prec} ↑	C_{rec} ↑	C_{acc} ↑	c%↑	F1 ↑	FID ↓	R-score ↑	Diversity ↑	FS ↓
w/o text	SceneDiff [14]	82.01	25.08	0.45	0.18	0.21	0.16	0.22	1.89	0.15	5.22	0.57
	OMOMO [22]	39.40	23.36	0.58	0.39	0.43	0.41	0.42	2.09	**0.16**	4.56	0.55
w Gen text	MDM-PC [32]	63.57	23.81	0.48	0.24	0.28	0.30	0.24	1.73	0.14	5.23	0.59
	CHOIS [21]	44.92	25.09	0.56	0.31	0.35	0.31	0.32	2.67	0.11	4.45	0.53
	AffordMotion [38]	75.32	24.37	0.55	0.16	0.22	0.16	0.21	2.88	0.12	4.48	**0.42**
	SemGeoMo	**35.43**	**20.85**	**0.72**	**0.49**	**0.51**	**0.49**	**0.52**	**1.64**	0.14	**5.35**	0.49

Table 4. Human motion generation result on HoDome.

		HandJPE↓	MPJPE ↓	C_{prec} ↑	C_{rec} ↑	C_{acc} ↑	c%↑	F1 ↑	FID ↓	R-score ↑	Diversity ↑	FS ↓
w/o text	SceneDiff [14]	107.50	29.53	0.21	0.11	0.16	0.13	0.14	4.88	0.08	4.72	0.60
	OMOMO [22]	86.12	27.07	0.42	0.23	0.31	0.22	0.25	5.47	0.10	4.93	0.31
w Gen text	MDM-PC [32]	95.91	26.29	0.25	0.13	0.18	0.14	0.15	**3.64**	0.13	4.89	0.67
	CHOIS [21]	76.74	**24.17**	0.55	0.12	0.26	0.11	0.18	6.12	0.11	4.14	**0.28**
	AffordMotion [38]	94.24	29.31	0.49	0.11	0.15	0.10	0.13	5.29	0.11	5.14	0.45
	SemGeoMo	**44.22**	24.28	**0.78**	**0.47**	**0.47**	**0.45**	**0.54**	4.29	**0.13**	**5.22**	0.35

states in 2D waypoints, which we modify this part into 3D states. AffordMotion predicts the affordance map and subsequently generates motion in scenes, we adopted this approach for our interactive target setting. Note that for the aforementioned works, since FullBodyManipulation provide the textual description, we provided ground-truth (GT) text as input, and compared the variant of our model under the same conditions with ground-truth text as well.

Results on the FullBodyManipulation Dataset. The results on the FullBodyManipulation dataset are illustrated in Tab. 1, and we provide a visualization comparison in Fig. 4. SceneDiff, MDM-PC, and CHOIS directly use the original point cloud as a condition in a single stage, the dynamic nature of the point cloud makes it challenging to infer low-level contact information, resulting in a significant drop in contact metrics. Even though MDM incorporates text guidance and improves performance on FID, the contact accuracy remains poor. CHOIS applies a human-object loss to refine contact, but such improvement in performance is limited. OMOMO predicts joint positions as the first-stage output, but without textual guidance, it performs poorly on FID and R-score, struggling to capture realistic interaction motions. Distortions and abnormal movements may occur, as shown in Fig. 4 circled in green. AffordMotion uses the affordance map for geometric guidance, while this approach lacks sufficient detail for fine-grained human-object interactions. In contrast, we leverage both semantic and geometric information by generating textual descriptions and predicting coarse-to-fine contact information, resulting in superior performance. Compared to the variant, which utilizes ground-truth text information as a condition, our model achieves comparable performance, demonstrating the efficiency and accuracy of the language description generated by our method.

Results on the Behave and IMHD^2 Dataset. We further conduct experiments on the BEHAVE and IMHD^2 datasets to validate our model's performance in Tab. 2 and Tab. 3. Notably, the IMHD^2 dataset includes more challenging interactions, such as sports activities, which make precise contact and accurate motion generation more difficult. However, our proposed method still outperforms other approaches in these scenarios.

Results on the HoDome Dataset. To test our model's ability to generate new scenarios, we further conduct the experiment on totally unseen objects with directly sampling on HoDome [47]. The result is illustrated in Tab. 4, our proposed method outperforms other methods. The visualization results are illustrated in Fig. 5.

4.4. Ablation Studies

Ablation studies on the impact of semantic and geometric information. To verify the effectiveness of our multi-level guidance, we conduct ablation studies to assess the impact of each component in Tab. 5. The multi-level geometric guidance and semantic information enhance human motion generation, reflected in the improvement of the contact and FID&R-score indicators, respectively. We also compare our designed SemGeo conditional module with cross-

Table 5. Ablation studies on the impact of semantic and geometric information with our model design.

	HandJPE↓	MPJPE ↓	C_{prec} ↑	C_{rec} ↑	C_{acc} ↑	c%↑	F1 ↑	FID ↓	R-score ↑	Diversity ↑	FS ↓
w/o affordanc map	29.89	18.85	0.80	0.71	0.81	0.63	0.73	2.21	0.62	9.84	0.58
w/o joint	31.23	20.60	0.76	0.70	0.78	0.61	0.72	3.52	0.51	8.15	0.74
w/o attention	29.18	19.47	0.81	0.66	0.78	0.59	0.70	9.27	0.44	6.84	0.70
w/o text	30.36	24.44	0.78	0.62	0.72	0.54	0.64	1.78	0.41	9.40	0.61
w/o fine-grained text	27.84	**16.62**	0.84	0.74	0.85	0.66	0.77	1.17	0.66	**10.15**	0.57
Full	**27.97**	17.01	**0.84**	**0.75**	**0.86**	**0.67**	**0.77**	**1.03**	**0.68**	10.05	**0.46**
Full-GT	6.44	13.61	0.88	0.82	0.93	0.70	0.84	0.97	0.70	10.48	0.58

Table 6. Ablation study on guidance loss.

L_{Joint}	L_{FS}	HandJPE↓	MPJPE ↓	C_{prec} ↑	C_{rec} ↑	C_{acc} ↑	c%↑	F1 ↑	FID ↓	R-score ↑	Diversity ↑	FS ↓
✓		27.84	16.97	0.83	0.75	0.86	0.67	0.75	1.15	0.65	10.13	0.61
	✓	85.28	29.75	0.62	0.25	0.35	0.24	0.22	0.93	0.66	10.05	0.31
✓	✓	27.84	16.62	0.84	0.74	0.85	0.66	0.77	1.17	0.66	10.15	0.57

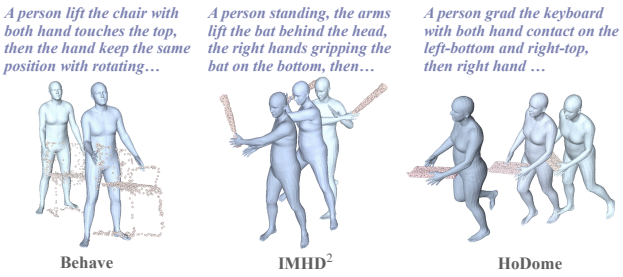

Figure 5. Visulization on more datasets.

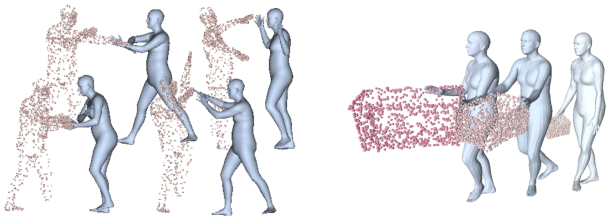

Figure 6. Extension on other scenarios. (a) is the interaction with humans and (b) is the interaction with objects of varying sizes.

Table 7. Text generation result on FullBodyManipulation.

BLEU-4	ROUGE -1	ROUGE-2	ROUGE-L
87.21	90.91	83.79	89.39

attention. The results show that cross-attention significantly improves performance by effectively sharing information between coarse-to-fine features. During implementation, we predict the hand joint and compare with generation process with ground truth position (denoted as Full-GT), it is worth noting that when the hand is fixed to the exact ground truth, this may reduce the FS score due to some foot sliding in the generated motion. Our results are comparable to the Full-GT setting, indicating that our joint and affordance predictions are sufficiently accurate to provide effective guidance for generating plausible results.

Analysis on text annotation. To verify the correctness of our generated text, we calculate BLEU [26] and ROUGE [24] scores, compared with the textual information provided by FullBodyManipulation in Tab. 7. These metrics illustrate the accuracy of our reasoning results. As shown in Tab. 1, we achieve performance comparable to using ground truth text, further demonstrating the effectiveness of our textual descriptions. Additionally, the results on both coarse and fine-grained text highlight the effectiveness of incorporating multi-level semantic information.

Ablation studies on loss guidance. We conduct experiments on the loss guidance described in Tab. 6. The joint loss plays a key role in improving contact performance, while the foot loss enhances the feasibility of the generated results. These improvements are reflected in the enhancement of FID and FS metrics.

4.5. Extension

We further evaluate our model on more challenging scenarios, including complex human-human interactions and shape-varying object manipulations, as Fig. 6 shows. The varying size of objects are simulated to assess how well our model generates responsive and adaptive interactions. The results demonstrate the feasibility of our approach, with generating realistic human responsive motions.

5. Conclusion

In this work, we introduce SemGeoMo, a novel method for generating responsive human motions and corresponding textual descriptions based on dynamic interactive targets. We design an automated text annotator to provide semantic information. By integrating text-affordance-joint semantic and geometric guidance, SemGeoMo ensures the semantic coherence of the generated text and the geometric precision of the corresponding motion. Our method achieves state-of-the-art performance on three benchmarks and demonstrates generalization abilities on an unseen dataset.

References

[1] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21211–21221, 2023. 2

[2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 1, 2, 6

[3] Zhi Cen, Huaijin Pi, Sida Peng, Zehong Shen, Minghui Yang, Shuai Zhu, Hujun Bao, and Xiaowei Zhou. Generating human motion in 3d scenes from text descriptions. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 1855–1866, 2024. 1, 2

[4] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF CVPR*, pages 18000–18010, 2023. 1

[5] Peishan Cong, Ziyi Wang, Zhiyang Dou, Yiming Ren, Wei Yin, Kai Cheng, Yujing Sun, Xiaoxiao Long, Xinge Zhu, and Yuexin Ma. Laserhuman: Language-guided scene-aware human motion generation in free environment. *arXiv preprint arXiv:2403.13307*, 2024. 2, 3

[6] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9760–9770, 2023. 1, 2

[7] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9760–9770, 2023. 2

[8] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19888–19901, 2024. 2

[9] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019. 2

[10] Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. Nemf: Neural motion fields for kinematic animation. *Advances in Neural Information Processing Systems*, 35:4244–4256, 2022. 6

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 4

[14] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16750–16761, 2023. 1, 2, 6, 7

[15] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023. 2

[16] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Y, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural IPS*, 36, 2024. 1

[17] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Chairs: Towards full-body articulated human-object interaction. *arXiv preprint arXiv:2212.10621*, 3, 2022. 2

[18] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9365–9376, 2023. 2

[19] 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. 1, 2

[20] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 947–957, 2024. 2

[21] 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, 6, 7

[22] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 1, 2, 3, 4, 5, 6, 7

[23] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, pages 1–21, 2024. 1, 2

[24] C Lin. Recall-oriented understudy for gisting evaluation (rouge). *Retrieved August*, 20:2005, 2005. 8

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 3

[26] K Papinesi. Bleu: A method for automatic evaluation of machine translation. In *Proc. 40th Actual Meeting of the Association for Computational Linguistics (ACL), 2002*, pages 311–318, 2002. 8

[27] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthe-

sis of 3d human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553*, 2023. 2, 6

[28] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4332–4341, 2019. 3

[29] 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. 4

[30] Konstantinos I Roumeliotis and Nikolaos D Tselikas. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6):192, 2023. 2

[31] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 1, 2, 4

[32] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2, 3, 4, 6, 7

[33] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2, 4

[34] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. *arXiv preprint arXiv:2311.17135*, 2023. 2

[35] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9401–9411, 2021. 1, 2

[36] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems*, 35:14959–14971, 2022. 2

[37] Zhenzhi Wang, Jingbo Wang, Dahua Lin, and Bo Dai. Intercontrol: Generate human motion interactions by controlling every joint. *arXiv preprint arXiv:2311.15864*, 2023. 1, 5

[38] 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. In *Proceedings of the IEEE/CVF CVPR*, pages 433–444, 2024. 1, 2, 5, 6, 7

[39] 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

[40] 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. *arXiv preprint arXiv:2309.07918*, 2023. 1, 2

[41] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. *arXiv preprint arXiv:2310.08580*, 2023. 5

[42] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22260–22271, 2024. 1, 2

[43] Liang Xu, Yizhou Zhou, Yichao Yan, Xin Jin, Wenhan Zhu, Fengyun Rao, Xiaokang Yang, and Wenjun Zeng. Regennet: Towards human action-reaction synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1759–1769, 2024. 2, 3

[44] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14928–14940, 2023. 2

[45] Hongwei Yi, Justus Thies, Michael J Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. In *ECCV*, pages 246–263. Springer, 2025. 2

[46] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *ECCV*, pages 310–325. Springer, 2025. 5

[47] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8834–8845, 2023. 2, 6, 7

[48] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 5

[49] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 1, 2, 4

[50] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[51] Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I'm hoi: Inertia-aware monocular capture of 3d human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 729–741, 2024. 2, 6