Human Motion Synthesis in 3D Scenes via Unified Scene Semantic Occupancy

Jingyu Gong[1,2,3], Kunkun Tong[1], Zhuoran Chen[1], Chuanhan Yuan[4], Mingang Chen[3], Zhizhong Zhang[1, 3], Xin Tan[1,2*], Yuan Xie[1,2]

[1]School of Computer Science and Technology, East China Normal University, Shanghai, China
[2]Chongqing Key Laboratory of Precision Optics, Chongqing Institute of East China Normal University, Chongqing, China
[3]Shanghai Key Laboratory of Computer Software Evaluating and Testing, Shanghai Development Center of Computer Software Technology, Shanghai, China
[4]College of Computer Science, Chongqing University, Chongqing, China
{jygong, zzzhang, xtan, yxie}@cs.ecnu.edu.cn, {51265901010, 10235102518}@stu.ecnu.edu.cn, 20241401019@stu.cqu.edu.cn, cmg@sscenter.sh.cn

Abstract

Human motion synthesis in 3D scenes relies heavily on scene comprehension, while current methods focus mainly on scene structure but ignore the semantic understanding. In this paper, we propose a human motion synthesis framework that take an unified Scene Semantic Occupancy (SSO) for scene representation, termed SSOMotion. We design a bi-directional tri-plane decomposition to derive a compact version of the SSO, and scene semantics are mapped to an unified feature space via CLIP encoding and shared linear dimensionality reduction. Such strategy can derive the fine-grained scene semantic structures while significantly reduce redundant computations. We further take these scene hints and movement direction derived from instructions for motion control via frame-wise scene query. Extensive experiments and ablation studies conducted on cluttered scenes using ShapeNet furniture, as well as scanned scenes from PROX and Replica datasets, demonstrate its cutting-edge performance while validating its effectiveness and generalization ability.

Code — https://github.com/jingyugong/SSOMotion
Extended version — https://arxiv.org/abs/2511.07819

Introduction

Understanding and simulating human behaviors in real 3D scenes has attracted lots of attention due to its wide application in robotics, games, and AR/VR (Zhao et al. 2023; Huang et al. 2023; Wang et al. 2024; Hwang et al. 2025).

Pioneers attempted to emulate human behavior and then synthesize human motion in given scenarios (Starke et al. 2019; Cao et al. 2020; Wang et al. 2021, 2022a). Thanks to the advances in 3D scene understanding (Qi et al. 2017a,b; Gong et al. 2021a,b), recent works can perceive the whole scene and provide guidance for scene-aware human motion synthesis (Wang et al. 2021, 2022a; Tang et al. 2024). More recent works paid much attention on exploiting fine-grained scene structures, and grid sensors were widely utilized for its simplicity (Lim, Jeong, and Kim 2023; Lee and Joo 2023; Zhao et al. 2023; Jiang et al. 2024b; Liu et al. 2024; Cen et al. 2024). However, the scene semantic information, which is highly correlated with human behavior, cannot be effectively formulated and leveraged in motion synthesis.

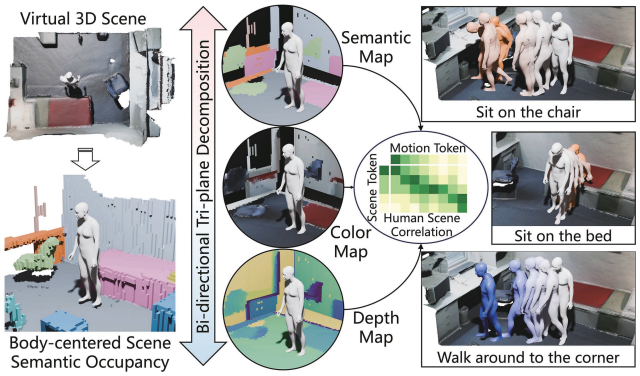

Figure 1: Illustration of human motion synthesis within 3D Scene Semantic Occupancy (SSO). We decompose the SSO into bi-directional tri-plane as a unified scene representation. Human scene correlation is modeled as control signal for instruction-aware motion synthesis in 3D scenes.

This motivates us to explore a general representation comprising both scene semantics and structures, and finally we find that Scene Semantic Occupancy (SSO) can fulfill our requirements (Cao and De Charette 2022; Tong et al. 2023; Ma et al. 2024; Chen et al. 2025). Nevertheless, directly performing feature extraction on the SSO is highly resource-consuming. In addition, semantic category definitions are dataset-dependent in current scene comprehension, making it challenging to achieve cross-dataset generalization.

Observing the inherent sparsity in scene occupancy, we propose to employ bi-directional tri-plane decomposition on it. We will first take grid sensors surrounding the human to obtain localized SSO, and then project the localized occupancy along the $\pm xyz$ axes to obtain semantic occupancy maps (Fig. 1 left). These semantic maps encapsulate body-centered semantic, depth, and color maps from diverse perspectives. Such low-dimension SSO can provide comprehensive semantic structure hints while substantially decreases computational demands.

*Corresponding Author.

To take a unified semantic representation for the SSO, we take the CLIP text encoder (Radford et al. 2021) for category feature embedding. Due to the repetitiveness of category information in semantic map and the high-dimension of CLIP feature, we decide to first take a shared mapping layer for semantic feature dimensionality reduction. Later, the compressed semantic features will be distributed into the semantic map for further scene comprehension. Thus far, we can derive the scene semantic structures for scene-aware motion synthesis with low computation overhead.

These scene hints will further guide the human to achieve the goal following the instructions (Fig. 1 right). We choose to translate the instructions into target poses or positions via human population (Hassan et al. 2021b; Zhao et al. 2022). Later, we model the correlation between 3D scene and goal-aware human motion via a frame-wise scene query, and then employ it for motion control in 3D scenes.

For performance evaluation, we have conducted experiments on cluttered scenes with ShapeNet (Chang et al. 2015) furniture for primitive task evaluation. In addition, we follow the instructions to synthesize human motions in 3D scenes from PROX (Hassan et al. 2019) and Replica (Straub et al. 2019) to show the cross-dataset performance. Extensive experiments and comprehensive ablation studies demonstrate the effectiveness and universality of the proposed method.

Related Works

Human Motion Synthesis. Motion synthesis has been researched for a long time (Clavet et al. 2016; Starke et al. 2019). As the technical development of motion capture and generative model, frontier researchers have continuously pursued generating more natural and realistic human motions (Tevet et al. 2023; Amballa, Akkinapalli, and Muralikrishnan 2025). ACTOR (Petrovich, Black, and Varol 2021) and TEMOS (Petrovich, Black, and Varol 2022) utilized a latent feature space for motion sequence, thus reducing accumulative error in motion synthesis. T2M (Guo et al. 2022) estimated the motion length according to the text before final motion synthesis. MDM (Tevet et al. 2023) and MLD (Chen et al. 2023) introduced diffusion model into motion synthesis and achieved performance boost. OmniControl (Xie et al. 2024) utilized the joint hints to control the synthesized motion. MotionLCM (Dai et al. 2024) and MotionMamba (Zhang et al. 2024) further improved the speed and efficiency for real-time application. Motion Anything (Zhang et al. 2025) generated human motion according signals like text and music.

Thanks to the advancements in human motion synthesis, we can extend to scene-aware motion synthesis with easier control and higher motion naturalness.

Scene-aware Motion Synthesis. Scene-aware Motion Synthesis has attracted lots of attention recently due to its wide applications (Cao et al. 2020; Hassan et al. 2021a, 2023; Gong et al. 2024; Wang et al. 2024; Jiang et al. 2024a; Gong et al. 2026). GAMMA (Zhang and Tang 2022) and DIMOS (Zhao et al. 2023) first learned a latent space for plausible motion and later trained policy models to control motion synthesis. SceneDiffuser (Huang et al. 2023) designed a scene-based diffuser along with learning-based optimizer and planner to achieve the goal in 3D scenes. AMDM (Wang et al. 2024) took scene affordance map as intermediate representation for human-scene interaction. SceneMI (Hwang et al. 2025) utilized motion inbetweening with better keyframe control in scene-aware motion synthesis.

However, scene spatial semantics which is highly related to human motion is usually ignored by previous works. Thus, we decide to exploit the correlation between human behavior and scene semantics in this paper.

Body-centered Scene Perception. Scene perception has played a vital role in scene-aware motion synthesis. Pioneer (Wang et al. 2021) in scene-aware motion synthesis utilized PointNet (Qi et al. 2017a) to extract scene feature. However, human motion is more related to body-centered local scene. PLACE (Zhang et al. 2020a) calculated the distance between scene and human body to model the scene proximity. PSI (Zhang et al. 2020b) took scene semantics into consideration during static pose generation in 3D scenes. SAMP (Hassan et al. 2021a) and DIMOS (Zhao et al. 2023) designed virtual grid sensor to recognize the surrounded obstacles and avoid collision. Recent works (Liu et al. 2024; Hwang et al. 2025) attempted to utilize space occupancy to represent the scene and provided scene perception for human motion synthesis.

Different from previous works, we introduce unified scene semantic occupany with bi-directional tri-plane decomposition as our scene representation. It can embed both scene structure and semantics in a lightweight way, and can generalize well to various scenes.

Method

Preliminary

Human Motion. Human motion can be treated as a sequence of body poses. Following previous works (Wang et al. 2021; Zhao et al. 2023), we decide to take the parametric human model SMPL-X (Pavlakos et al. 2019) to represent the human pose. Within SMPL-X parameters, we mainly consider the global translation $\tau \in \mathbb{R}^3$, global orientation in axis-angle $\theta_g \in \mathbb{R}^3$, and body joint rotation in axis-angle $\theta_p = \theta_{j\in 1:21} \in \mathbb{R}^{63}$. Thus, the human pose can be represented by $P = \{\tau, \theta_g, \theta_{1:21}\} \in \mathbb{R}^{63}$. The human shape β and hand pose θ_h remain invariant throughout the motion. The human mesh $\mathcal{M}$ and skeleton joints $\mathcal{J}$ can be directly derive from parameters aforementioned by SMPL-X human model $\mathcal{M}, \mathcal{J} = \mathrm{SMPLX}(\tau, \theta_g, \beta, \theta_p, \theta_h)$.

Scene Semantic Occupancy. We decide to take scene semantic occupancy (Cao and De Charette 2022) to represent a scene. As the occupied voxels in 3D scenes are sparse, we take a compact scene occupancy representation $\mathcal{S} \in \mathbb{R}^{N\times 8}$. Each element $\mathcal{S}_i$ in $\mathcal{S}$ indicate an occupied voxel, consisting of xyz coordinate, $rgba$ color and semantic label s.

Motion Diffusion Model. For simplicity, we annotate any human motion as $x_0 = \{P_s\}_{s\in 1:S}$. The motion diffusion procedure will gradually add noise to the original motion

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}), \quad (1)$$

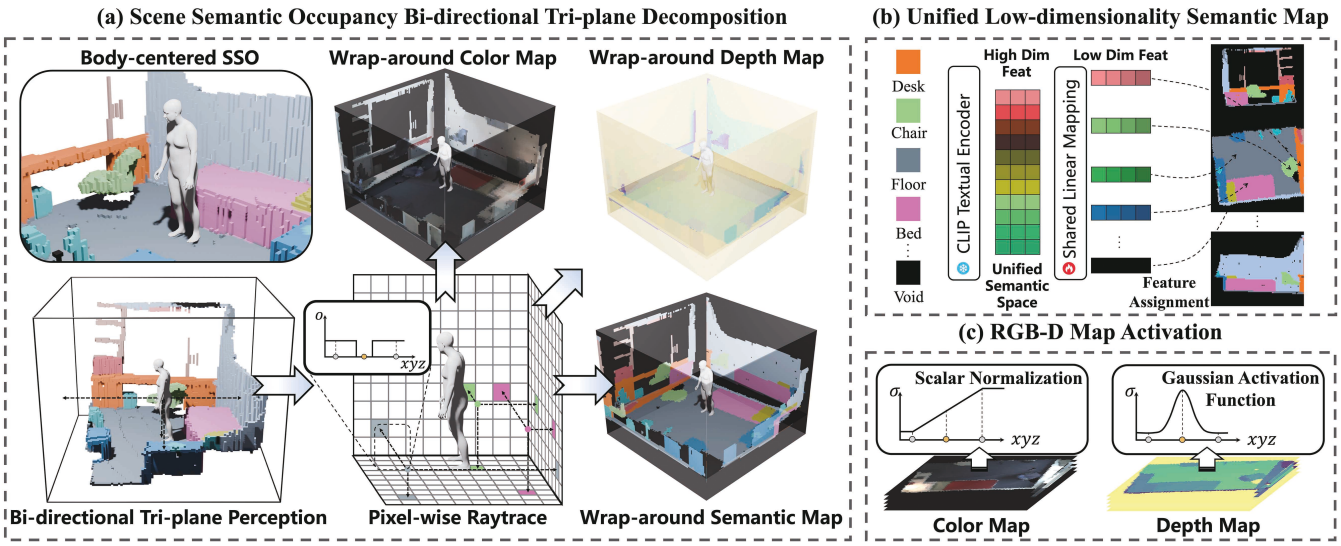

Figure 2: Pipeline of the Scene Semantic Occupancy perception (SSO). (a) presents the Bi-directional Tri-plane Decomposition of the SSO, where scene color, semantics and depth are perceived in body-centered coordinate. In (b), we map the semantic labels into a unified semantic space via the CLIP textual encoder and a shared linear layer. Then, the unified low-dimension semantic features will be scattered into the semantic map. (c) indicates the normalization functions for distance and color space.

where $\mathcal{N}$ indicates a normal distribution and $\alpha_{t\in 1:T}$ are hyper-parameters. Finally, x_T will approximate to $\mathcal{N}(\mathbf{0}, \mathbf{I})$. In the reverse procedure, we will gradually denoise the human motion via $P(x_{t-1}|x_t)$.

In our implementation, we supervise the network to directly predict the original motion given x_t at any time step t following MDM (Tevet et al. 2023)

$$\hat{x}_0^\phi = \phi(x_t, t, c), \tag{2}$$

where c combines the action label a and masked joints $\mathcal{J}_{1:S}$. During inference, we denoise the motion according to

$$P(x_{t-1}|x_t) = \mathcal{N}(\mu_t(\hat{x}_0^\phi(x_t), x_t), \tilde{\beta}_t \mathbf{I}). \tag{3}$$

Here, $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$, $\beta_t = 1-\alpha_t$, $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$, and $\mu_t(\hat{x}_0^\phi, x_t) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{x}_0^\phi + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t$.

Overview

Problem Definition. This work attempts to synthesize human motion $\{P_s\}_{s=1:S}$ in 3D scenes $\mathcal{S}$ according to the textual instructions. The instruction contains a sequence of interaction sub-tasks (action and object pairs $\{(a_i, o_i)\}_{i=1:I}$). For each sub-task, the human may go somewhere or interact with the objects (*e.g.* stools or tables) within scenes.

Pipeline. Given current task instruction (a, o), we will synthesize future motion based on surrounding scene $\mathcal{S}_l$ and historical motion, as shown in Fig. 2 (a). Specifically, given any historical motion $\tilde{P}_{1:\tilde{S}}$, we decide to take the last $H = \min(\tilde{S}, H_{max})$ frames as our historical motion, where $\tilde{P}_{\tilde{S}-H+1}$ is treated as current human pose P_1, and $P_{1:H} = \tilde{P}_{\tilde{S}-H+1:\tilde{S}}$ will be utilize to control the motion synthesis. Specifically, we take the human joints $J_{1:H} = \tilde{J}_{\tilde{S}-H+1:\tilde{S}}$ as historical motion hint f_h. For body-centered scene perception, we take the localized scene semantic occupancy $\mathcal{S}_l$, and then decompose it into bi-directional tri-plane representation, as shown in Fig. 2 (b). Then, the bi-directional tri-plane semantic occupancy will be encoded into scene feature f_s as shown in Fig. 2 (c). We further design a motion controller (Fig. 3) to synthesize future motion $\hat{P}_{1:S}$ based on historical motion f_h, current scene f_s and future interaction goal (a, o). Finally, the synthesized motion $\hat{P}_{1:S}$ will be updated to the historical motion $\tilde{S}_{1:\tilde{S}}$ with motion blending for overlapping H frames.

Semantic Occupancy Guidance

Bi-directional Tri-plane Decomposition. Given the global scene in a compact representation $\mathcal{S}$, we will first convert it into a sparse SSO $\mathcal{O}_g \in \mathbb{R}^{H_g \times W_g \times D_g \times 4}$, consisting of rgb and semantic information.

As shown in Fig. 2 (a), we will perceive the body-centered local scene based on current human status. Given any historical human motion $\tilde{P}_{1:\tilde{S}}$, we fetch the last H frames $\tilde{P}_{\tilde{S}-H+1:\tilde{S}}$ as guidance for future scene-aware motion synthesis. We will take $\tilde{P}_{\tilde{S}-H+1}$ as the initial pose to perceive surrounding environments. Here, we can obtain human translation $\tau^{\mathcal{J}}$ and horizontal orientation $\theta_z^{\mathcal{J}}$ according to human skeleton joints $\tilde{\mathcal{J}}_{\tilde{S}-H+1}$. After that, we will arrange $H \times W \times D$ semantic occupancy sensors distributed evenly around human body $\mathcal{I}_l \in \mathbb{R}^{H\times W\times D\times 3}$. These sensors will be transformed into world coordinate through

$$\mathcal{I}_g = R(\theta_z^{\mathcal{J}})\mathcal{I}_l + \tau^{\mathcal{J}}, \tag{4}$$

where $R(\cdot)$ indicates the rotation matrix. Later, these sensors will be taken to perceive the semantic occupancy. The

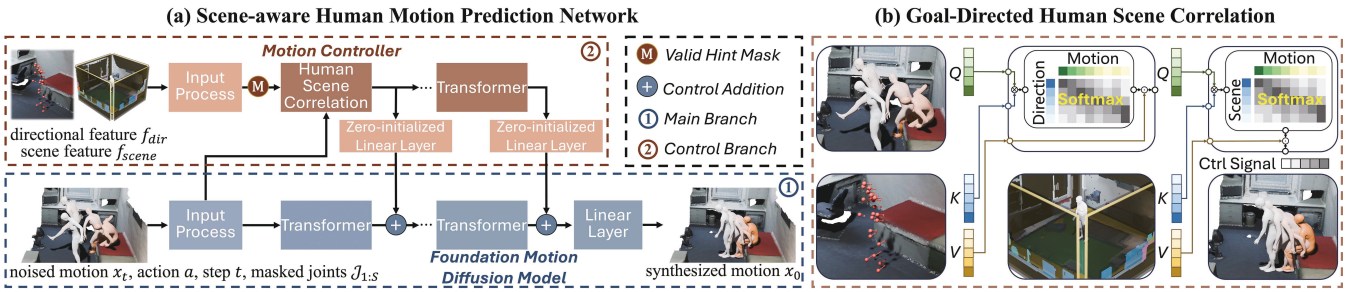

Figure 3: Overview of (a) the network for instruction-aware human motion synthesis in 3D scenes and (b) the motion controller based on Goal-directed Human Scene Correlation.

local semantic occupancy can be represented by $\mathcal{O}_l \in \mathbb{R}^{H\times W\times D\times 4}$.

Last, the local semantic occupancy will be decomposed into bi-directional tri-plane. Specifically, $\mathcal{O}_l$ will be rendered along $\pm xyz$ axes, and obtain the semantic occupancy maps $\mathcal{O}_{yz}, \mathcal{O}_{zy} \in \mathbb{R}^{H\times D\times 5}$, $\mathcal{O}_{zx}, \mathcal{O}_{xz} = \mathbb{R}^{W\times D\times 5}$, $\mathcal{O}_{xy}, \mathcal{O}_{yx} \in \mathbb{R}^{H\times W\times 5}$. Each semantic occupancy map $\mathcal{O}_{ij}$ contains color $\mathcal{O}^c_{ij}$, depth $\mathcal{O}^d_{ij}$ and semantic label $\mathcal{O}^s_{ij}$ information. It's noteworthy, $\mathcal{O}_{yx}$ is not utilized in body-centered scene comprehension as ceiling won't influence human behavior in most cases.

Unified Semantic Occupancy Representation. As the semantic category sets are different for different 3D scene datasets, we need to take a unified semantic space for various category rather than utilizing one-hot embedding. It means we need to convert the semantic map $\mathcal{O}^s_{ij}$ into an unified semantic feature map $\mathcal{O}^{sf}_{ij} \in \mathbb{R}^{D_1\times D_2\times C}$, where D_1, D_2 indicates map shape, and C indicates semantic feature dimension.

In this work, we choose to take the CLIP text encoder (Radford et al. 2021) to embed the semantic category (Fig. 2 (b)). Due to the repetitive nature of semantics across most regions, performing feature extraction on high-dimensional CLIP semantic features $f_{clip} \in \mathbb{R}^{C_h}$ will result in significant computational redundancy. Thus, rather than directly transfer the semantic map into feature map, we choose to first employ a shared linear layer to reduce the dimensionality of all semantic category features. Then, the reduced-dimensional CLIP features $\tilde{f}_{clip} \in \mathbb{R}^{C_l}$ will be distributed into the semantic map, obtaining the unified semantic feature map $\mathcal{O}^{sf}_{ij} \in \mathbb{R}^{D_1\times D_2\times C_l}$. Based upon this, we can easily extract scene semantic feature f_{sem} via a naive image encoder.

For bi-directional tri-plane depth map $\mathcal{O}^d_{ij}$, as shown in Fig. 2 (c), we choose to focus more on nearby objects, as scene geometry in close proximity exhibits stronger correlation with human behaviors. Thus, we take a gaussian kernel as activation function to obtain the activated depth map

$$\mathcal{O}^{da}_{ij} = \frac{1}{\sigma\sqrt{2\pi}} e^{-(\mathcal{O}^d_{ij})^2/(2\sigma^2)}, \tag{5}$$

where nearby objects can be easily perceived. The scene geometry feature f_{geo} can be directly extracted from it.

As for color map $\mathcal{O}^c_{ij}$, we simply normalize it to the range $[0, 1]$, and derive the scene texture feature f_{tex}.

By now, the perceived scene feature f_{scene} will combine the semantic, geometry, and texture feature

$$f_{scene} = f_{sem} \oplus f_{geo} \oplus f_{tex}, \tag{6}$$

which will be utilized for motion control in 3D scenes.

Motion Controller

Action Intention Cues. Given the 3D scene, we need to follow the instruction to synthesize future motions (as shown in Fig. 3 (a)). Besides the noised motion x_t, step t, and mask joints $\mathcal{J}$, we take the action a encoded by a learnable codebook as our input.

Meanwhile, we need to follow the instruction to navigate to the target position or interact with the target object o. For locomotion, we will calculate the direction from current human pelvis to target position $\mathbf{d} = o - J_{1,0} \in \mathbb{R}^3$. The direction will be repeated for K times where K indicates the number of human joints. As for interaction, we will sample target human pose with skeleton joints $\bar{J}$, then we will calculate the direction for all joints $\mathbf{d} = \bar{J} - J_1 \in \mathbb{R}^{K\times 3}$. In addition, we clip the norm of $\mathbf{d}$ to a normalized range

$$\mathbf{d}_n = \frac{\min(||\mathbf{d}||, 1) + \epsilon}{||\mathbf{d}|| + \epsilon}\mathbf{d}, \tag{7}$$

which provide directional hint f_{dir} for motion control.

Goal-Directed Human Scene Correlation. For motion control, we mainly consider sub-task goal given by textual instructions and scene constrains. Here, we first model the correlation between human motion and goal human status (as shown in Fig. 3 (b)). Specifically, we map directional hint feature f_{dir} into L_d tokens f^t_{dir}, which further provide the directional hint keys $K_{d,i} = W^k_{d,i} f^t_{dir}$ and values $V_{d,i} = W^v_{d,i} f^t_{dir}$. We further design to convert f_{mot} into frame-wise queries $Q_{d,i} = W^q_{d,i} f_{mot}$. The goal-directed human motion feature can be derived by

$$f^g_{mot} = \underset{i}{\text{concat}}(\text{softmax}(\frac{Q_{d,i}K^T_{d,i}}{\sqrt{d_d}})V_{d,i})W_{o,d}, \tag{8}$$

where W^k_{dir}, W^v_{dir}, W^q_{dir}, and $W_{o,d}$ are learnable parameters. Furthermore, we map scene feature f_{scene} into scene-related keys $K_{s,j}$ and values $V_{s,j}$, and model goal-directed

human scene correlation via

$$f_{mot}^{gs} = \underset{j}{\text{concat}}(\text{softmax}(\frac{Q_{s,j}K_{s,j}^{T}}{\sqrt{d_s}})V_{s,j})W_{o,s}, \quad (9)$$

where $Q_{s,j}$ are frame-wise human motion queries given by f_{mot}^{g}. The goal-directed human scene correlation f_{mot}^{gs} will be fed into the control branch.

Human Motion Control. The motion control branch is built proportionally to the main branch. Meanwhile, we employ multiple zero-initialized linear layers to inject control signal from the control branch into the main branch. As aforementioned, f_{mot}^{gs} is fed into the control branch during training to ensure that the synthesized motion comply with the instruction requirements and scene constraints.

Training and Inference

Due to the scarcity of motion-in-scene datasets relative to motion capture datasets, SSOMotion adopts a decoupled training strategy in which the base diffusion model and control branch are trained separately on different datasets.

Fundamental Model Training. We follow the MDM (Tevet et al. 2023) to train our base diffusion model. At each denoising step, the model predicts the clean motion sequence, and the training objective is to minimize the reconstruction loss between the predicted and ground-truth sequences. To enable conditional generation, the model is trained with action labels and masked human joints as conditioning inputs. These forms of guidance are commonly available in standard motion capture datasets that do not include explicit scene context, making them well-suited for training the scene-agnostic base model.

Control Branch Training. We train a separate control branch on a motion-in-scene dataset to incorporate scene awareness. Each scene is represented by the SSO to capture the semantic layout of the environment. During training, the parameters of the base diffusion and control branch are jointly optimized. The control branch receives the SSO representation along with the human movement direction as input and produces control signals that modulate the denoising process of the diffusion model.

Motion Synthesis. Our model supports multi-task motion synthesis in 3D scenes, including locomotion and different types of human-scene interactions. Given the textual instruction, we will decompose it in action-object pair (a, o). The target object or position will be fed into the control branch along with the decomposed SSO to derive the control signal. Meanwhile, the main branch will iteratively denoised the motion under the guidance of initial human status, action label, and control signal. In addition, we take advantage of the DIP (Gong et al. 2026) to introduce scene constraints during denoising. For long-term motion synthesis, our model can generate the future motion (consisting of S frames) according to the historical motion constrains, with H overlapping frames to ensure continuity in motion transition. The historical motion is represented in the form of masked subsequent human skeleton joints, which serve as the input to the main branch. The historical motion is updated by blending it with the synthesized motion. The motion synthesis process continues until all tasks are completed. Thus, our approach allows the model to produce coherent and goal-directed motion with indefinite length.

Experiments

Datasets

Motion Datasets. In this work, we use AMASS (Mahmood et al. 2019) dataset for base diffusion model training . Babel (Punnakkal et al. 2021) provided action labels and the initial/final frames for several subsets of the AMASS dataset and HumanML3D (Guo et al. 2022) supply additional sentence annotations and initial/final frames for more motion data in AMASS. All motions are downsampled to 30 FPS and segmented into 30-frame clips with a 5-frame stride. Each motion clip is transformed based on the human pose in the first frame, aligning the initial pose to the origin with the body orientation facing the positive y-axis. These processed motion clips, along with each corresponding action labels, are then used to train the base motion diffusion model.

Scene-aware Motion Datasets. We use the HUMANISE dataset (Wang et al. 2022b) for motion controller training, which aligns motion data from AMASS with scenes from ScanNet (Dai et al. 2017). We voxelize the dense scene point cloud with a voxel size of 4cm to generate scene occupancy and apply a nearest-neighbor approach to propagate semantic labels from sparse annotations. The final scene semantic occupancy is stored in an compact version where coordinate–label pairs are used for all non-empty voxels.

Scene Datasets. We evaluate the proposed method in two widely adopted dataset, namely PROX (Hassan et al. 2019) and Replica (Straub et al. 2019). By benchmark on the two dataset, we assess our model's long-term motion synthesis across diverse tasks. The experiments share the same motion diffusion model and pipeline, ensuring the method's generalization capability.

Scene Navigation

We take the cluttered scenes constructed by DIMOS (Zhao et al. 2023) for locomotion evaluation, where furniture from ShapeNet (Chang et al. 2015) is randomly placed.

Metrics. Following the DIMOS (Zhao et al. 2023) metrics for scene navigation evaluation, we adopt four criteria: average distance from the final body position to the target point, foot contact score indicating the degree of ground contact, locomotion penetration score representing the percentage of body vertices within the walkable area, and the task completion time measured in seconds.

Results. We report the results of motion synthesis for locomotion in Tab. 1. The results demonstrate the proposed method can achieve cutting-edge performance in destination distance (0.02m), non-penetration score (0.95), and execution time (3.60s). The proposed method has inferior performance in foot contact score as we focus more on contact status of foot vertices rather than the inner human joints.

	avg. dist ↓	ft. cont ↑	loco. pene ↑	time ↓
SAMP	0.14	0.84	0.94	5.97
GAMMA	0.03	0.94	0.94	3.87
DIMOS	0.04	**0.99**	**0.95**	6.43
OmniControl	0.04	0.76	0.93	**2.43**
Ours	**0.02**	0.90	**0.95**	3.60

Table 1: Evaluation of motion synthesis on locomotion task. The up/down arrows (↑/↓) indicate higher/lower is better. Metrics with best performance are annotated in boldface.

We present the visual results in Fig. 4 where the proposed method shows higher motion diversity and less collision with scenes.

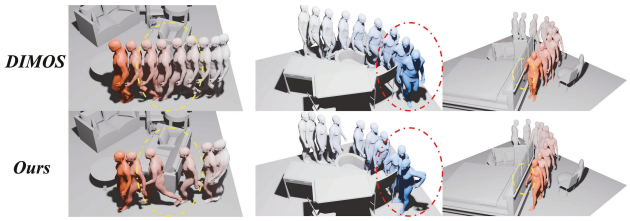

Figure 4: Visual comparison of locomotion synthesis between DIMOS and the proposed method.

Human-Scene Interaction

Following previous work (Zhao et al. 2023), we take the scenes consisting of specific 10 furniture from the ShapeNet (Chang et al. 2015) for interaction evaluation.

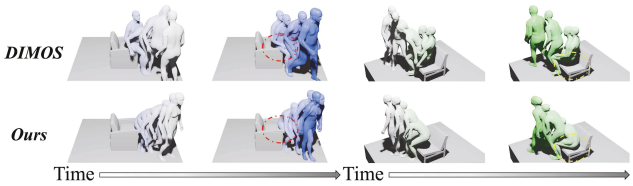

Figure 5: Visual results given by DIMOS and the proposed method for sitting action.

Metrics. We take three metrics to evaluate the performance of human-scene interaction, including task completion time, mean human mesh vertex penetration with scene, and maximum penetration observed over time.

Results. We report the results of motion synthesis for scene interaction in Tab. 2. It can be seen, the synthesized motion given by the proposed method can execute the instruction more rapidly with minimum scene penetration.

Long-term Motion Synthesis

We attempt to synthesize long-term human motion in 3D scenes following consequent textual instructions. We utilize the COINS (Zhao et al. 2022) to populate the human in target position following textual instruction, which further guide goal-directed motion synthesis.

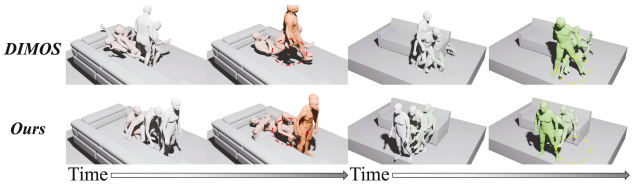

Figure 6: Visualization of lying motions synthesized by DIMOS and the proposed method.

	action	time ↓	pene. mean ↓	pene. max ↓
SAMP	sit	8.63	11.91	45.22
DIMOS	sit	4.09	1.91	10.61
Ours	sit	**3.45**	**1.83**	**6.78**
SAMP	lie	12.55	44.77	238.81
DIMOS	lie	4.20	9.90	44.61
Ours	lie	**3.69**	**5.74**	**40.48**

Table 2: Evaluation of motion synthesis on interaction tasks. The up/down arrows (↑/↓) indicate higher/lower is better. The best results are shown in boldface.

Metrics. We conducted a comprehensive user study to obtain a more accurate evaluation of long-term motion synthesis performance. Participants were asked to give scores to the generated motions without prior knowledge of the corresponding method. Each participant assessed four metrics on all generated motion samples, including naturalness, diversity, plausibility, and goal achievement.

Results. Finally, we collect $4,080$ ratings from 17 participants and report the results in Fig. 7. It can be observed that our method achieves optimal performance across all aspects.

We also present the visual differences between competitive methods in Fig. 8 and Fig. 9 for motion synthesis in scenes from the PROX and Replica dataset.

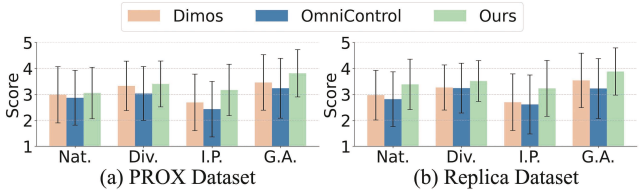

Figure 7: (a) and (b) present the scores of competitive methods across four aspects (motion naturalness, motion diversity, interaction plausibility and goal achievement) in the user study. Higher scores indicate better performance.

Ablation Study

In this part, we mainly analyze the computational cost of scene comprehension and human scene correlation in scene-aware human motion synthesis.

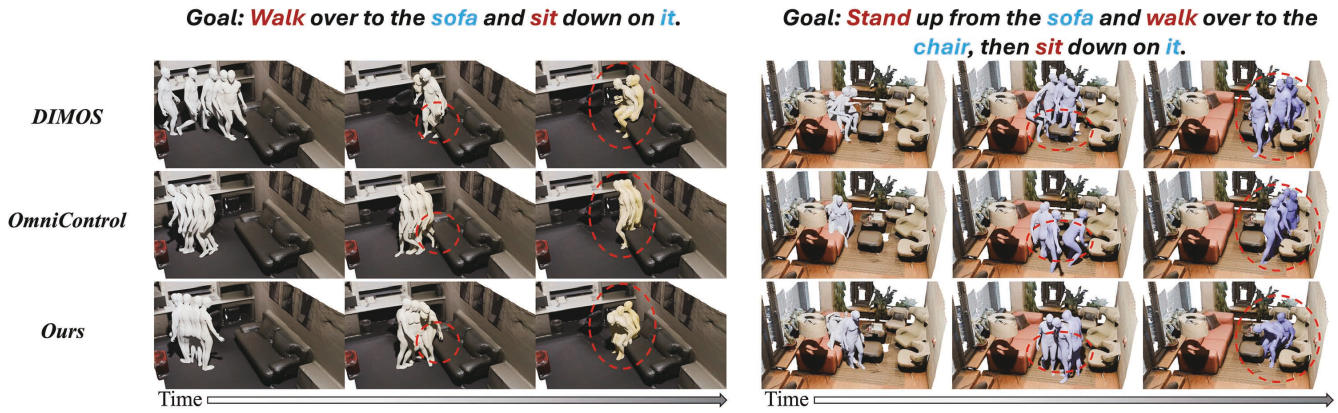

Figure 8: Visual comparison of different methods for instruction-based motion synthesis in 3D scenes from PROX dataset.

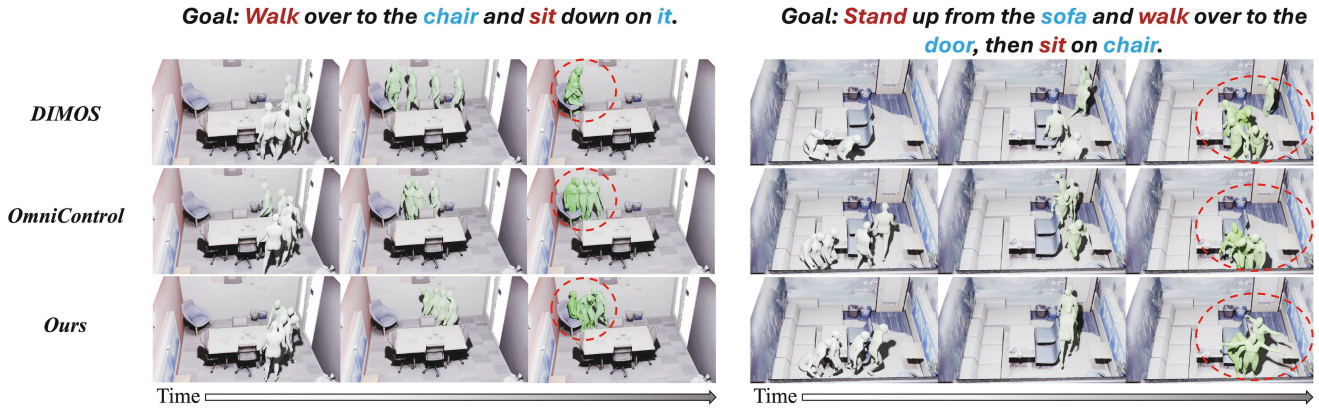

Figure 9: Visualization of synthesized human motion given by different method in 3D scenes from the Replica dataset.

BT Decomp.	Dim. Red.	Comp. Cost (GFLOPs)
		20783.78
✓		421.73
	✓	21.21
✓	✓	0.49

Table 3: Computational Cost of the Scene Comprehension and Motion Scene Correlation for only 1 sample.

Computational Cost. In the proposed method, we decompose the scene semantic occupancy into bi-directional tri-plane RGB-DS map for scene comprehension and motion control. Before semantic feature map construction, we introduce a shared linear layer for dimensionality reduction of semantic feature from CLIP textual encoder.

We report the computational cost (batch size equals 1) of (1) Neither Bi-directional Tri-plane Decomposition (BT Decomp.) nor Dimensionality Reduction (Dim. Red.), (2) only "BT Decomp.", (3) only "Dim. Red.", and (4) both "BT Decomp." and "Dim. Red" in Tab. 3. It can be seen, such design can significantly reduce computational cost while preserve the scene semantic occupancy information. Without the SSO Bi-directional Tri-plane Decomposition and semantic feature Dimensionality Reduction, employing high-resolution SSO for scene comprehension in motion synthesis is computationally prohibitive.

Conclusion

In this paper, we present SSOMotion, an effective framework for human motion synthesis in 3D scenes. We adopt a novel approach to integrate semantic and structural scene understanding by leveraging the SSO. Through a bi-directional tri-plane decomposition of the SSO and applying a dimensionality reduction on CLIP-based semantic encoding, our model is capable to efficiently capture fine-grained scene semantics while minimizing the computational overhead. We then validate the effectiveness of the proposed SSOMotion through extensive experiments on synthesized scenes from DIMOS with ShapeNet furniture, and cluttered scenes from PROX and Replica datasets.

Despite its effectiveness, SSOMotion currently has certain limitations and future work should focus on optimizing the efficiency and extending the framework to support real-time deployment in dynamic environments.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 62176092, 62222602, 62302167, U23A20343, 62476090, 62502159), Natural Science Foundation of Shanghai (Grant No. 25ZR1402135), Shanghai Sailing Program (Grant No. 23YF1410500), Young Elite Scientists Sponsorship Program by CAST (Grant No. YESS20240780), the Chenguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission (Grant No. 23CGA34), Natural Science Foundation of Chongqing (Grant No. CSTB2023NSCQ-JQX0007,CSTB2023NSCQ-MSX0137, CSTB2025NSCQ-GPX0445), Open Project Program of the State Key Laboratory of CAD&CG (Grant No. A2501), Zhejiang University, Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science-MOE, ECNU.

References

Amballa, A.; Akkinapalli, G.; and Muralikrishnan, V. 2025. LS-GAN: Human Motion Synthesis with Latent-space GANs. In *WACV*, 326–335.

Cao, A.-Q.; and De Charette, R. 2022. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, 3991–4001.

Cao, Z.; Gao, H.; Mangalam, K.; Cai, Q.-Z.; Vo, M.; and Malik, J. 2020. Long-term human motion prediction with scene context. In *ECCV*, 387–404. Springer.

Cen, Z.; Pi, H.; Peng, S.; Shen, Z.; Yang, M.; Zhu, S.; Bao, H.; and Zhou, X. 2024. Generating Human Motion in 3D Scenes from Text Descriptions. In *CVPR*, 1855–1866.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Chen, H.; Yuan, L.; Sun, T.; Gong, J.; Tan, X.; Zhang, Z.; and Xie, Y. 2025. YouTube-Occ: Learning Indoor 3D Semantic Occupancy Prediction from YouTube Videos. *arXiv preprint arXiv:2506.18266*.

Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; and Yu, G. 2023. Executing your Commands via Motion Diffusion in Latent Space. In *CVPR*, 18000–18010.

Clavet, S.; et al. 2016. Motion matching and the road to next-gen animation. In *Proc. of GDC*, volume 2, 4.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 5828–5839.

Dai, W.; Chen, L.-H.; Wang, J.; Liu, J.; Dai, B.; and Tang, Y. 2024. Motionlcm: Real-time controllable motion generation via latent consistency model. *arXiv preprint arXiv:2404.19759*.

Gong, J.; Wang, M.; Liu, W.; Qian, C.; Zhang, Z.; Xie, Y.; and Ma, L. 2024. Demos: dynamic environment motion synthesis in 3D scenes via local spherical-bev perception. *arXiv preprint arXiv:2403.01740*.

Gong, J.; Xu, J.; Tan, X.; Song, H.; Qu, Y.; Xie, Y.; and Ma, L. 2021a. Omni-supervised point cloud segmentation via gradual receptive field component reasoning. In *CVPR*, 11673–11682.

Gong, J.; Xu, J.; Tan, X.; Zhou, J.; Qu, Y.; Xie, Y.; and Ma, L. 2021b. Boundary-aware geometric encoding for semantic segmentation of point clouds. In *AAAI*, volume 35, 1424–1432.

Gong, J.; Zhang, C.; Liu, F.; Fan, K.; Zhou, Q.; Tan, X.; Zhang, Z.; and Xie, Y. 2026. Diffusion Implicit Policy for Unpaired Scene-aware Motion Synthesis. In *AAAI*.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022. Generating diverse and natural 3d human motions from text. In *CVPR*, 5152–5161.

Hassan, M.; Ceylan, D.; Villegas, R.; Saito, J.; Yang, J.; Zhou, Y.; and Black, M. J. 2021a. Stochastic scene-aware motion prediction. In *ICCV*, 11374–11384.

Hassan, M.; Choutas, V.; Tzionas, D.; and Black, M. J. 2019. Resolving 3D human pose ambiguities with 3D scene constraints. In *ICCV*, 2282–2292.

Hassan, M.; Ghosh, P.; Tesch, J.; Tzionas, D.; and Black, M. J. 2021b. Populating 3D scenes by learning human-scene interaction. In *CVPR*, 14708–14718.

Hassan, M.; Guo, Y.; Wang, T.; Black, M.; Fidler, S.; and Peng, X. B. 2023. Synthesizing physical character-scene interactions. In *ACM SIGGRAPH*, 1–9.

Huang, S.; Wang, Z.; Li, P.; Jia, B.; Liu, T.; Zhu, Y.; Liang, W.; and Zhu, S.-C. 2023. Diffusion-based generation, optimization, and planning in 3d scenes. In *CVPR*, 16750–16761.

Hwang, I.; Zhou, B.; Kim, Y. M.; Wang, J.; and Guo, C. 2025. SceneMI: Motion In-betweening for Modeling Human-Scene Interactions. *arXiv preprint arXiv:2503.16289*.

Jiang, N.; He, Z.; Wang, Z.; Li, H.; Chen, Y.; Huang, S.; and Zhu, Y. 2024a. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia 2024 Conference Papers*, 1–11.

Jiang, N.; Zhang, Z.; Li, H.; Ma, X.; Wang, Z.; Chen, Y.; Liu, T.; Zhu, Y.; and Huang, S. 2024b. Scaling up dynamic human-scene interaction modeling. In *CVPR*, 1737–1747.

Lee, J.; and Joo, H. 2023. Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments. In *ICCV*, 9663–9674.

Lim, D.; Jeong, C.; and Kim, Y. M. 2023. Mammos: Mapping multiple human motion with scene understanding and natural interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4278–4287.

Liu, X.; Hou, H.; Yang, Y.; Li, Y.-L.; and Lu, C. 2024. Revisit Human-Scene Interaction via Space Occupancy. In *ECCV*. Springer.

Ma, Q.; Tan, X.; Qu, Y.; Ma, L.; Zhang, Z.; and Xie, Y. 2024. Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction. In *CVPR*, 19936–19945.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 5442–5451.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 10975–10985.

Petrovich, M.; Black, M. J.; and Varol, G. 2021. Action-Conditioned 3D Human Motion Synthesis with Transformer VAE. In *ICCV*.

Petrovich, M.; Black, M. J.; and Varol, G. 2022. TEMOS: Generating diverse human motions from textual descriptions. In *ECCV*, 480–497. Springer.

Punnakkal, A. R.; Chandrasekaran, A.; Athanasiou, N.; Quiros-Ramirez, A.; and Black, M. J. 2021. BABEL: Bodies, action and behavior with english labels. In *CVPR*, 722–731.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 652–660.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 30.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Starke, S.; Zhang, H.; Komura, T.; and Saito, J. 2019. Neural state machine for character-scene interactions. *ACM TOG*, 38(6): 209–1.

Straub, J.; Whelan, T.; Ma, L.; Chen, Y.; Wijmans, E.; Green, S.; Engel, J. J.; Mur-Artal, R.; Ren, C.; Verma, S.; et al. 2019. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*.

Tang, J.; Wang, J.; Ji, K.; Xu, L.; Yu, J.; and Shi, Y. 2024. A unified diffusion framework for scene-aware human motion estimation from sparse signals. In *CVPR*, 21251–21262.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; and Bermano, A. H. 2023. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*.

Tong, W.; Sima, C.; Wang, T.; Chen, L.; Wu, S.; Deng, H.; Gu, Y.; Lu, L.; Luo, P.; Lin, D.; et al. 2023. Scene as occupancy. In *ICCV*, 8406–8415.

Wang, J.; Rong, Y.; Liu, J.; Yan, S.; Lin, D.; and Dai, B. 2022a. Towards Diverse and Natural Scene-aware 3D Human Motion Synthesis. In *CVPR*, 20460–20469.

Wang, J.; Xu, H.; Xu, J.; Liu, S.; and Wang, X. 2021. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *CVPR*, 9401–9411.

Wang, Z.; Chen, Y.; Jia, B.; Li, P.; Zhang, J.; Zhang, J.; Liu, T.; Zhu, Y.; Liang, W.; and Huang, S. 2024. Move as You Say Interact as You Can: Language-guided Human Motion Generation with Scene Affordance. In *CVPR*, 433–444.

Wang, Z.; Chen, Y.; Liu, T.; Zhu, Y.; Liang, W.; and Huang, S. 2022b. Humanise: Language-conditioned human motion generation in 3d scenes. *NeurIPS*, 35: 14959–14971.

Xie, Y.; Jampani, V.; Zhong, L.; Sun, D.; and Jiang, H. 2024. OmniControl: Control Any Joint at Any Time for Human Motion Generation. In *ICLR*.

Zhang, S.; Zhang, Y.; Ma, Q.; Black, M. J.; and Tang, S. 2020a. PLACE: Proximity learning of articulation and contact in 3D environments. In *International Conference on 3D Vision (3DV)*, 642–651. IEEE.

Zhang, Y.; Hassan, M.; Neumann, H.; Black, M. J.; and Tang, S. 2020b. Generating 3d people in scenes without people. In *CVPR*, 6194–6204.

Zhang, Y.; and Tang, S. 2022. The wanderings of odysseus in 3d scenes. In *CVPR*, 20481–20491.

Zhang, Z.; Liu, A.; Reid, I.; Hartley, R.; Zhuang, B.; and Tang, H. 2024. Motion mamba: Efficient and long sequence motion generation. In *ECCV*, 265–282. Springer.

Zhang, Z.; Wang, Y.; Mao, W.; Li, D.; Zhao, R.; Wu, B.; Song, Z.; Zhuang, B.; Reid, I.; and Hartley, R. 2025. Motion Anything: Any to Motion Generation. *arXiv preprint arXiv:2503.06955*.

Zhao, K.; Wang, S.; Zhang, Y.; Beeler, T.; and Tang, S. 2022. Compositional human-scene interaction synthesis with semantic control. In *ECCV*, 311–327. Springer.

Zhao, K.; Zhang, Y.; Wang, S.; Beeler, T.; and Tang, S. 2023. Synthesizing diverse human motions in 3d indoor scenes. In *ICCV*, 14738–14749.