Diffusion Implicit Policy for Unpaired Scene-aware Motion Synthesis

Jingyu Gong[1, 2, 3], Chong Zhang[1], Fengqi Liu[4], Ke Fan[4], Qianyu Zhou[5], Xin Tan[1,2], Zhizhong Zhang[1, 3*], 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, China
[4]School of Computer Science, Shanghai Jiao Tong University, Shanghai, China
[5]College of Computer Science and Technology, Jilin University, Jilin, China
{jygong, xtan, zzzhang, yxie}@cs.ecnu.edu.cn, 51265901052@stu.ecnu.edu.cn,
{liufengqi, slipperyfrank}@sjtu.edu.cn, zhouqianyu@jlu.edu.cn

Abstract

Scene-aware motion synthesis has been widely researched recently due to its numerous applications. Prevailing methods rely heavily on paired motion-scene data, while it is difficult to generalize to diverse scenes when trained only on a few specific ones. Thus, we propose a unified framework, termed Diffusion Implicit Policy (DIP), for scene-aware motion synthesis, where paired motion-scene data are no longer necessary. In this paper, we disentangle human-scene interaction from motion synthesis during training, and then introduce an interaction-based implicit policy into motion diffusion during inference. Synthesized motion can be derived through iterative diffusion denoising and implicit policy optimization, thus motion naturalness and interaction plausibility can be maintained simultaneously. For long-term motion synthesis, we introduce motion blending in joint rotation power space. The proposed method is evaluated on synthesized scenes with ShapeNet furniture, and real scenes from PROX and Replica. Results show that our framework presents better motion naturalness and interaction plausibility than cutting-edge methods. This also indicates the feasibility of utilizing the DIP for motion synthesis in more general tasks and versatile scenes.

Code — https://github.com/jingyugong/DIP
Extended version — https://arxiv.org/abs/2412.02261

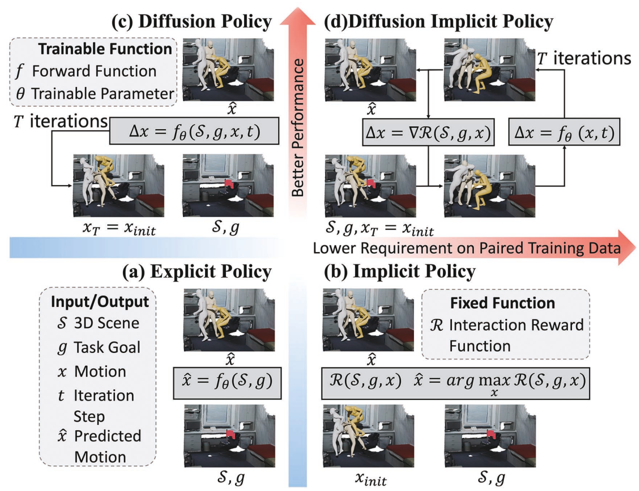

Figure 1: Policy learning frameworks. (a) Explicit policy is trained with paired motion-scene data. (b) Implicit policy optimizes the motion from random initialization. (c) Diffusion policy gradually denoise the motion. (d) Our diffusion implicit policy iteratively denoises and optimizes the motion to ensure motion naturalness, diversity, interaction plausibility without need for any paired motion-scene data.

Introduction

Synthesizing human motion in real 3D scenes has attracted significant attention in recent years (Cao et al. 2020; Wang et al. 2021, 2022a; Zhao et al. 2023; Hu et al. 2024), due to its wide applications in scene simulation, digital human animation, and virtual/augmented reality.

Thanks to learning-based 3D perception (Qi et al. 2017; Zhao et al. 2021; Gong et al. 2021), pioneers (Cao et al. 2020; Wang et al. 2021; Zhao et al. 2023) attempted to synthesize motion in scenes with feasible human-scene interaction. However, in previous works, paired motion-scene data are required to learn scene-aware motion policies. The majority of prevailing methods (Starke et al. 2019; Zhao et al. 2023) learn **Explicit Policies** to directly predict the desired motion based on current states and goals (Fig. 1 (a)). Some of them (Wang et al. 2021, 2022a) utilized a second-stage **Implicit Policy** optimization but sacrifice the motion naturalness for interaction plausibility (Fig. 1 (b)). Recent works (Huang et al. 2023; Wang et al. 2024) utilized **Diffusion Policies** to achieve better performance (Fig. 1 (c)), where massive paired motion-scene data is also necessary.

In fact, human motion data (Mahmood et al. 2019; Lin et al. 2023; Ren et al. 2023) is far more abundant than paired motion-scene data (Hassan et al. 2019; Wang et al. 2022b). Motion synthesis that relies heavily on paired data will inevitably suffer from limited diversity. Meanwhile, the generalization ability is hard to guarantee when trained on limited scenes and applied to various other scenes.

Based on this observation, we propose a unified frame-

[*]Corresponding Author.

work, termed **Diffusion Implicit Policy (DIP)** (Fig. 1 (d)), which disentangles human-scene interaction from motion synthesis during training and then integrate motion denoising with implicit policy optimization during inference. In this way, paired motion-scene data is no longer necessary, and motion naturalness and interaction plausibility can be ensured simultaneously for scene-aware motion synthesis.

In the DIP, a motion diffusion model is employed to make the synthesized motion more and more natural throughout the entire denoising process(Fig. 2 (b)). We equipped the diffusion model with a ControlNet (Zhang, Rao, and Agrawala 2023) branch to provide keyframe joint hints for historical motion and future goals. Following previous works (Tevet et al. 2023; Xie et al. 2024), the diffusion model is designed to predict the original motion at each denoising step and then sample the denoised motion from a normal distribution accordingly. Thus, we can well utilize the stochastic process to pursue plausible human-scene interactions. Specifically, interaction-based reward functions are designed to assess the consistency between motions and scenes. These reward functions are used as implicit policy to optimize the sampling distribution, ensuring that the sampled denoised motion corresponds better to the 3D scenes at each denoising step. As the denoising process can also be treated as optimization for motion naturalness, the entire scene-aware motion synthesis can be framed as an optimization problem to pursue both motion naturalness and interaction plausibility.

To synthesize reasonable motion in 3D scenes, we first train a motion diffusion model conditioned on actions and keyframe joints, which can be derived from motion itself. Furthermore, we design various reward functions to score motion naturalness and interaction plausibility. These rewards will optimize the sample distribution during motion denoising. We choose to adjust the centroid of the distribution in a GAN inversion manner, applying these reward functions to the outputs of the diffusion model at the centroid rather than directly to the centroid itself. In this way, the proposed method can identify a better intermediate noised motion with higher motion naturalness and interaction plausibility in the final synthesized motions.

In addition, for long-term motion synthesis involving multiple tasks, we need to take historical motion as constrain when synthesizing future motion. To maintain continuity between historical and future motions, we employ a time-variant motion blending, where we interpolate the rotation matrix in the power space. Thus far, the proposed framework can synthesize long-term motion in general scenes without any training on paired motion-scene data.

For performance evaluation, we use scenes cluttered with furniture from ShapeNet (Chang et al. 2015) to assess the ability on human-object interaction. We also take PROX (Hassan et al. 2019) and Replica (Straub et al. 2019) to demonstrate the generalization ability in scene-aware motion synthesis. We compared the proposed method with prevailing works based on physical and perceptual scores. Comprehensive experiments support our claims and indicate that the synthesized motion produced by the proposed method demonstrates better performance.

Our main contributions can be summarized as follows: (1) We disentangle human-scene interaction from motion synthesis during training and transform scene-aware motion synthesis into a joint optimization problem, where motion naturalness and interaction plausibility are ensured by iterative diffusion denoising and implicit policy optimization. (2) We propose to adjust the centroid of the sampling distribution during denoising process in a GAN Inversion manner for higher interaction plausibility. (3) We design to generate new motion based on historical constrains via inpainting and blend the motion in the power space of the rotation matrix using time-variant coefficients to synthesize long-term motion for multiple subsequent tasks.

Related Work

Human Scene Interaction. Generating realistic and plausible human-scene interactions has been widely explored (Savva et al. 2014, 2016; Zhang et al. 2020a,b; Zhao et al. 2022). PLACE (Zhang et al. 2020a) modeled the proximity based on the distance between human body and 3D scene during interaction synthesis. POSA (Hassan et al. 2021b) designed a contact feature map for the human body, indicating the contact and semantic information for each vertex in the human mesh. COINS (Zhao et al. 2022) encoded the human body and 3D objects into a shared feature space and synthesizes diverse compositional interactions. Narrator (Xuan et al. 2023) modeled the correlation of 3D scene and text based on a scene graph for interaction generation.

Inspired by the optimization stage in static human-scene interaction, we design interaction-based reward functions as an implicit policy for scene-aware motion synthesis.

Motion Synthesis. Motion synthesis has been studied for a significant period (Clavet et al. 2016; Holden, Komura, and Saito 2017; Starke et al. 2019; Wang, Chai, and Xia 2019; Guo et al. 2020; Petrovich, Black, and Varol 2021; Guo et al. 2022; Wu et al. 2024). TEMOS (Petrovich, Black, and Varol 2022) learned a shared latent space for motion and text alignment. MDM (Tevet et al. 2023) and MotionDiffuse (Zhang et al. 2024) introduced diffusion model into motion synthesis. Subsequent works (Chen et al. 2023; Xie et al. 2023; Zhang et al. 2023a; Dai et al. 2024; Zhang et al. 2023b; Karunratanakul et al. 2023; Xie et al. 2024) further improved the controllability and quality of the generated results.

Thanks to the advancements in motion synthesis, we follow the MDM (Tevet et al. 2023) and extend it to scene-aware motion synthesis, using interaction-based reward functions as an implicit policy.

Scene-Aware Motion Synthesis. Synthesizing human motion in 3D scenes has garnered much attention recently (Hassan et al. 2021a; Wang et al. 2022a; Zhang et al. 2022; Hassan et al. 2023; Mullen et al. 2023; Rempe et al. 2023; Mir et al. 2024; Cen et al. 2024; Jiang et al. 2024; Yi et al. 2024; Liu et al. 2024). Wang $et\ al.$ (Wang et al. 2021) utilized the PointNet (Qi et al. 2017) to provide scene feature and optimized the entire motion after generation. GAMMA (Zhang and Tang 2022) and DIMOS (Zhao et al. 2023) learned a latent space for natural motion, and a policy network was trained over the latent space. SceneDiffuser (Huang et al. 2023) proposed a scene-conditioned diffuser accompanied by a learning-based optimizer and planner. LAMA (Lee and

Joo 2023) introduced a test-time optimization stage for controller network via reinforcement learning to predict the action cues for motion matching (Clavet et al. 2016) and modification. AMDM (Wang et al. 2024) designed a two-stage framework with a scene affordance map as an intermediate representation.

Compared with these methods, we propose to disentangle scene-aware motion synthesis into motion prior learning via diffusion model and implicit policy learning via interaction-based reward functions, and integrate them in a unified framework, termed Diffusion Implicit Policy.

Method

Preliminary

Motion Representation. For human motion, we take the SMPL-X model (Pavlakos et al. 2019) to represent the pose at each frame. Here, we mainly consider the global orientation represented in axis-angle $\theta_{global} \in \mathbb{R}^3$, joint rotation in axis-angle $\theta_{j=1:21} \in \mathbb{R}^{63}$ and the translation $\tau \in \mathbb{R}^3$. Accordingly, for each frame s, the human pose can be defined as $P_s = \{\theta_{s,global}, \theta_{s,j=1:21}, \tau_s\} \in \mathbb{R}^{69}$, and the synthesized motion consisting of consecutive poses can be annotated as $\hat{P} = \{\hat{P}_s\}_{s=1:S}$. The body shape $\beta \in \mathbb{R}^{10}$ and hand pose $\theta_h \in \mathbb{R}^{24}$ are always keep the same as initial human body for simplicity. The first K joints $J = J_{1:K} \in \mathbb{R}^{K\times 3}$ and body mesh with V vertices $M(\tau, \theta_{global}, \beta, \theta_j, \theta_h) \in \mathbb{R}^{V\times 3}$ are taken as auxiliary representation for human pose.

Motion Diffusion Model. Motion diffusion is modeled as a noising process which gradually add noise to the original motion with S frames $x_0 = \{P_s\}_{s=1:S}$

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

where $\mathcal{N}$ is a normal distribution and $\alpha_{t=1:T}$ are hyper-parameters. The distribution of final noised motion x_T will approximate to $\mathcal{N}(0, I)$. Like MDM (Tevet et al. 2023), we train a diffusion model to predict the original motion directly

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

where t is the time step and a is the action label. For motion synthesis, the denoising procedure can be formulated as:

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

where $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$, $\beta_t = 1-\alpha_t$, $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$, and $\mu_t(\hat{x}_0^{\phi}, x_t) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{x}_0^{\phi} + \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t$. Thanks to the diffusion model ϕ, we can easily adjust $P(x_{t-1})$ via optimizing $\phi(\mu_t, t-1, a)$ to pursue higher interaction plausibility in the final synthesized motion at each denoising step.

Overview

In this paper, we attempt to synthesize human motion in 3D scenes given a sequence of interaction sub-tasks (Fig. 2 (a)).

We can first decompose the command into a list of sub-task (interaction behavior and object pair) via current LLMs (Touvron et al. 2023). For each sub-task, the human may go to some place or interact with the objects in the scene (*e.g.* , sitting on the chair).

Given a sub-task, we will first locate the goal position using COINS (Zhao et al. 2022), and then fetch reward functions for implicit policy according to current action.

We train a diffusion model conditioned on human action and keyframe joints to synthesize human motion with easy control. Later, we model the interaction-based reward functions and take them to optimize the sampling distribution of denoised motion at each denoising step. The motion prior from diffusion model and implicit policy from reward function are integrated together to synthesize scene-aware motion with desired interactions (as shown in Fig. 2 (b)).

Given historical motion, the synthesized motion should be consistent with it. Thus, we design to derive the long-term motion via a motion blending where translation are interpolated linearly and rotation are blended in the matrix power space. By now, we can synthesize long-term motion in 3D scenes when all sub-tasks are completed.

Conditional Diffusion Model

We take $x_0 \in \mathbb{R}^{S\times 69}$ consisting of human translation, orientation, and joint rotations in S frames as motion representation. To simplify the formulation of reward functions in our method, we train a diffusion model ϕ (Fig. 3 (a)) to predict the original motion $\hat{x}_0^{\phi}$ for each noised x_t as shown in Eq. 2.

For any original motion x_0, we take the local coordinate of motion to reduce representation redundancy. Specifically, we translate and horizontally rotate the motion based on the human pose in the first frame. In the transformed motion, the orientation of the human body in the first frame lie in the yz-plane ($y \geq 0$), and the pelvis is positioned at the origin. After motion synthesis, the motion in scenes can be derived via simple translation and horizontal rotation.

To better maintain consistency with historical motion and achieve future goals, we also take a ControlNet branch (Zhang, Rao, and Agrawala 2023) to provide the hint of keyframe human joints (available from motion data for training) that need to be controlled (Fig. 3 (b)). We choose to take the joint positions rather than joint rotation as the external hint for easier understanding of the space information. Concretely, we will first calculate the joints' positions according to the original human motion, and then randomly select one from these joints in a few frames as the hint and others are padded with zeros, thus the input of ControlNet branch take the form of $\mathcal{J} = \{J_{s,k}\}_{s=1:S,k=1:K}$ where non-zero values provide the controlled joint position hints.

For training the ControlNet branch, all its parameters are randomly initialized, and the link layers are initialized to zero to maintain the motion synthesis capability of the main branch. Meanwhile, all parameters in original motion diffusion model ϕ are loaded from a pre-trained model. The controlled diffusion model φ is also supervised to predict the origin motion $x_0 = \{P_s\}_{s=1:S}$. After fine-tuning, the controlled motion diffusion model can be formulated as

$$\hat{x}_0^{\varphi} = \varphi(x_t, t, a, \mathcal{J}). \tag{4}$$

For simplicity, the action a and joint hint $\mathcal{J}$ are termed as the condition c. Later, the controlled diffusion model φ are taken as motion prior to ensure motion naturalness.

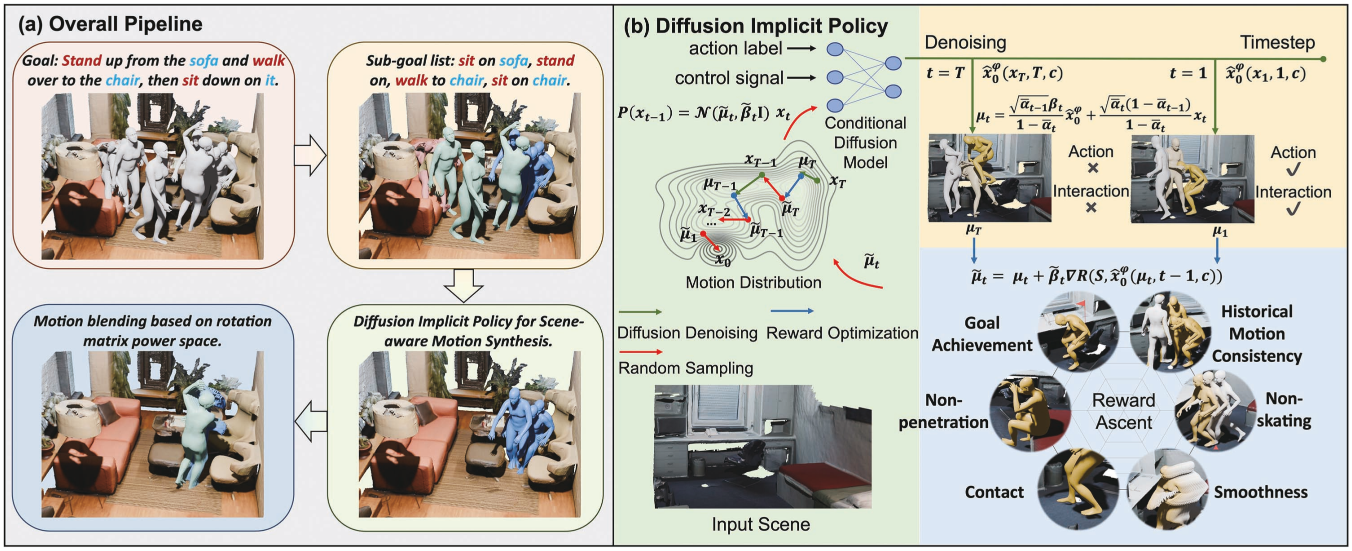

Figure 2: (a) indicates the overall pipeline. Any feasible command will be decomposed into sub-tasks with action-object pairs. Then, we will synthesize future motion according to current sub-task. Last, the synthesized motions will be fused into the historical motion to obtain the final long-term motion. (b) presents the Diffusion Implicit Policy (DIP). In each iteration, the denoising step will make the synthesized motion appear more **natural**), and the implicit policy optimization will endow the motion with **plausible** interaction. The random sampling step can help the framework synthesize motion with **diverse** styles.

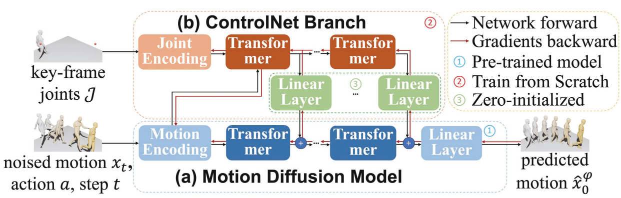

Figure 3: Illustration of conditional diffusion model. A diffusion model is first trained conditioned on action, and then a ControlNet branch is taken to provide keyframe joints' hint.

To better control the poses at specific frames, especially for the historical motion constrains, we maintain the fixed poses in an inpainting manner when $t > T_{inpaint}$. The predicted original motion $\hat{x}_0^{\varphi}$ will be updated via inpainting that can be formulated as $\hat{x}_0^{\varphi} = m \cdot x_m + (1 - m) \cdot \hat{x}_0^{\varphi}$, where x_m indicates the poses that should be maintained and $m \in \mathbb{R}^{S \times 69}$ represents the inpainting mask.

Thus far, the diffusion model can provide motion prior with explicit joint control for later DIP.

Reward Functions for Implicit Policy

Furthermore, we introduce a set of differentiable reward functions that score the motions within scenes and serve as an implicit policy to enhance overall performance (as shown in Fig. 2 (b) blue part). Here, we primarily focus on three key aspects in the design of reward functions: motion continuity, goal achievement, and interaction plausibility.

To promote motion continuity, we introduce a Historical Motion Consistency reward $\mathcal{R}_{his}$ to ensure continuous transition between historical and synthesized motions, alongside a Smoothness reward to prevent abrupt changes with excessive acceleration $\mathcal{R}_{acc}$. Additionally, we implement a Goal Achievement reward $\mathcal{R}_{goal}$ that encourage the human to move closer to the target. We also incorporate a Contact reward $\mathcal{R}_{cont}$ and a Non-Penetration reward $\mathcal{R}_{pene}$ to facilitate plausible human-scene interactions, as well as a Non-skating reward $\mathcal{R}_{skt}$ to anchor the vertices of contact. The total reward function for implicit policy take the form of

$$\begin{aligned} \mathcal{R}_{ip} = \lambda_{his}\mathcal{R}_{his} + \lambda_{acc}\mathcal{R}_{acc} + \lambda_{goal}\mathcal{R}_{goal} \\ + \lambda_{cont}\mathcal{R}_{cont} + \lambda_{pene}\mathcal{R}_{pene} + \lambda_{skt}\mathcal{R}_{skt} \end{aligned} \quad (5)$$

where $\lambda_{(\cdot)}$ are a series of hyper-parameters.

Diffusion Implicit Policy

By now, we utilize a diffusion model with ControlNet branch φ to predict $\hat{x}_0^{\varphi}$. Thus, similar to Eq. 3, we can sample the x_{t-1} according to x_t and $\hat{x}_0^{\varphi}$

$$P(x_{t-1}|\hat{x}_0^{\varphi}, x_t) = \mathcal{N}(\mu_t(\hat{x}_0^{\varphi}, x_t), \tilde{\beta}_t \mathbf{I}). \quad (6)$$

Such denoising process (Fig. 2 (b) yellow part) can also be considered as an optimization problem based on a motion naturalness reward function $\mathcal{R}_{nat}$ which can be defined implicitly by its gradient $\nabla \mathcal{R}_{nat}(x_t) = \mu_t - x_t$. Meanwhile, the denoising process is also accompanied by a stochastic disturbance item (Fig. 2 (b) green part) following $\mathcal{N}(0, \tilde{\beta}_t \mathbf{I})$.

Thus, the stochastic item can be well utilized to search for motion with higher interaction plausibility. We design the interaction-based implicit policy (Fig. 2 (b) blue part) to partly play the role of such disturbance, and scene-aware motion synthesis can be treated as a joint optimization problem which both maximize the motion naturalness and interaction plausibility in Diffusion Implicit Policy

$$\hat{x}_0 = arg \max_x \mathcal{R}_{dip}(x), \quad (7)$$

$$\mathcal{R}_{dip}(x) = \mathcal{R}_{nat}(x) + \mathcal{R}_{ip}(\hat{x}_0^{\varphi}(x)). \tag{8}$$

In order to synthesize human motion that can maximize the total reward, we integrate the implicit policy optimization into each denoising step where motion naturalness and interaction plausibility can be enhanced iteratively.

In Eq. 6, x_{t-1} is sampled from $\mathcal{N}(\mu_t, \tilde{\beta}_t \mathbf{I})$, thus we can adjust μ_t (*i.e.* the mean value of x_{t-1}) based on implicit policy and more suitable x_{t-1} can be sampled accordingly. It is noteworthy, we need the final synthesized human motion x_0 to be consistent with the scene and achieve high reward. Thus, we propose to optimize μ_t through $\hat{x}_0^{\varphi}(\mu_t, t-1, c)$ rather than μ_t itself. Here, we take $t-1$ as denoising step because μ_t is the mean value of the denoised x_{t-1}.

Thanks to the motion representation that we take, the reward functions are fully differentiable. Meanwhile, we find that optimizing $\hat{x}_0^{\varphi}(\mu_t, t-1, c)$ shows better performance than directly modifying μ_t itself as previous work (Xie et al. 2024). That is because direct optimization over μ_t does not ensure motion continuity. On the other side, optimizing μ_t via $\hat{x}_0^{\varphi}(\mu_t, t-1, c)$ can help search a better distribution with higher interaction reward for the final synthesized motion x_0, and μ_t can be be adjusted as a whole in a GAN Inversion manner (given a Generator $x = G(z)$, adjust the latent code z via loss $\mathcal{L}(x)$ for more desired output) (Bau et al. 2019a,b). Similarly, given the diffusion model $\hat{x}_0^{\varphi}(\mu_t, t-1, c)$, we optimize μ_t via $\mathcal{R}(\hat{x}_0^{\varphi})$ according the following formulation

$$\tilde{\mu}_t = \mu_t + \tilde{\beta}_t \cdot \nabla \mathcal{R}_{ip}(\mathcal{S}, \hat{x}_0^{\varphi}(\mu_t, t-1, c)), \tag{9}$$

where $\mathcal{S}$ indicate the 3D scene information, including scene semantics, SDF, and floor height. Further, $\tilde{\mu}_t$ is taken as the mean of distribution to sample x_{t-1}.

Multi-Task Motion Synthesis

As for the command for multi-task motion synthesis, such as "The person first sits on the bed, then goes to the corner of the room, and finally sits on the chair.", we need to infer future motion "Sit on the chair" while maintaining continuity with previously synthesized motions "The person first sits on the bed and then goes to the corner of the room".

For any previous synthesized motion $\tilde{P}_{1:\tilde{S}}$, we will select the latest $H = min(\tilde{S}, H_{max})$ frames as external historical constrains for future motion synthesis. Explicitly, we extract those pelvis joints $\{\tilde{J}_{s,pelvis}\}_{s=\tilde{S}-H+1:\tilde{S}}$ as trajectory hints for conditional motion diffusion. In addition, we take the body skeleton joints from the historical frames to form $\mathcal{J} = \{\tilde{J}_s\}_{s=\tilde{S}-H+1:\tilde{S}}$, and use them for pose constraints in the implicit policy optimization.

After a new round of motion synthesis, we obtain the generated motion $\{\hat{P}_s\}_{s=1:S}$. For more natural motion transition in the overlapping H frames, we take a time-variant motion blending. Different from direct linear interpolation in the pose representation like priorMDM (Shafir et al. 2024). We only utilize linear interpolation for the human translation $\check{\tau}_{1:H}$. As for human orientation and joints' rotations, we take the axis-angle representation $\bar{\theta}_{rot,s} = \tilde{\theta}_{\tilde{S}-H+s}$ and $\hat{\theta}_{rot,s}$ and convert them to the rotation matrix $\bar{M}_{rot,s}$ and $\hat{M}_{rot,s}$.

	time ↓	avg. dist ↓	contact ↑	loco pene ↑
SAMP	5.97	0.14	0.84	0.94
GAMMA	3.87	**0.03**	0.94	0.94
DIMOS	6.43	0.04	**0.99**	**0.95**
Ours	**3.35**	**0.03**	0.91	**0.95**

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.

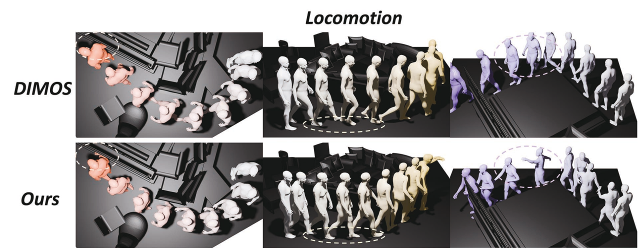

Figure 4: Visual results given by DIMOS and ours for locomotion task. The dashed circles indicate lower penetration, less skating and higher diversity in the synthesized motion.

The blending occur in the power space of rotation matrix which can be formulated as:

$$\check{M}_{rot,s} = (\hat{M}_{rot,s} \bar{M}_{rot,s}^{-1})^{\gamma} \bar{M}_{rot,s} \tag{10}$$

where $\gamma = (1 : H)/(H+1)$, and $\check{\theta}_{1:H}$ is converted from $\check{M}_{rot,1:H}$. Thus, the blended pose take the form of $\check{P}_{1:H} = \{\check{\theta}_{s,global}, \check{\theta}_{s,j}, \check{\tau}_s\}_{s=1:H}$. The newly updated long-term motion is derived as

$$\tilde{P}_{1:\tilde{S}+S-H} = \{\tilde{P}_{1:\tilde{S}-H}, \check{P}_{1:H}, \hat{P}_{H+1:S}\}. \tag{11}$$

The entire motion can be synthesized in an iterative manner until all sub-tasks are completed (Fig. 2 (a)).

Experiments

Datasets

Motion Datasets. Here, we use captured motion data with action/description labels from AMASS (Mahmood et al. 2019) to train our model. Babel (Punnakkal et al. 2021) provided action labels and the start/end frames for several subsets of AMASS. HumanML3D (Guo et al. 2022) provided additional sentence annotations and start/end frames for more motion data in AMASS.

Scene Datasets. We evaluate the performance in both synthesized scenes and real scanned scenes. Following DIMOS (Zhao et al. 2023), we use randomly generated scenes consisting of furniture from ShapeNet (Chang et al. 2015) to validate the performance on atomic locomotion and human-scene interaction. As for real scanned scenes from PROX (Hassan et al. 2019) and Replica (Straub et al. 2019), we take them to evaluate the performance of the pipeline in synthesizing long-term motions within scenes that involve multiple tasks. All experiments are conducted using the same model and pipeline, thus indicating the generalization ability of our method.

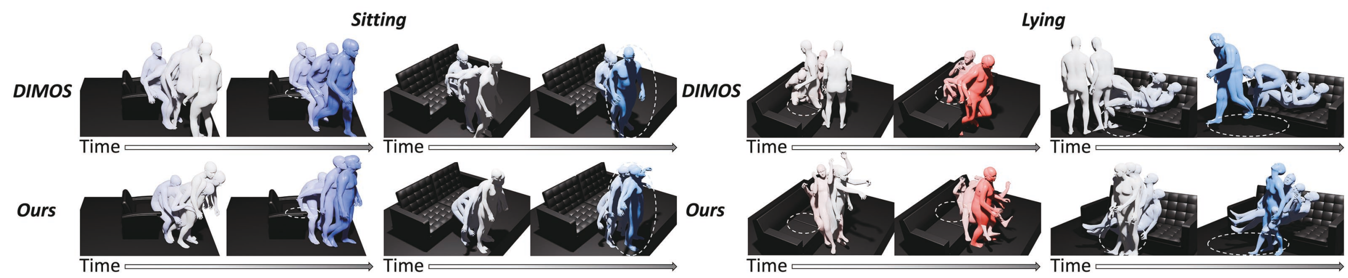

Figure 5: Visual results of synthesized motions given by DIMOS and our method for sitting and lying. The dashed circles indicates obvious advantages over DIMOS in less collision (col. 1,3), higher diversity (col. 2) and better foot contact (col. 4).

Scene Navigation

We take the generated scenes from DIMOS (Zhao et al. 2023) for testing, where scenes are cluttered with furniture from ShapeNet (Chang et al. 2015). In this experimental setting, the human need to walk from the starting point to the target point and avoid collisions with the furniture in scenes.
Metrics. We also evaluate the performance of synthesized motion for locomotion from four aspects as DIMOS (Zhao et al. 2023), where (1) finish time measured in seconds, (2) average horizontal distance from final human body to target point measured in meters, (3) foot joint contact score

$$s_{cont} = e^{-(|min j_z|-0.05)_+} \cdot e^{-(min||j_{vel}||_2-0.075)_+}, \quad (12)$$

and (4) penetration score indicating the percentage of body vertices within the walkable area are taken as our metrics.
Results. We compare our method with SAMP (Hassan et al. 2021a), GAMMA (Zhang and Tang 2022), and DIMOS (Zhao et al. 2023) for locomotion in 3D scenes and report the results in Tab. 1. The results indicate our method can achieve the shortest finish time (3.35s), closest distance (0.03m), and lowest penetration (0.95). It's noteworthy, the locomotion speed of the proposed method is much similar to that of real human than other methods. As can be seen, the performance on the contact score is inferior. We believe that is because our method focuses more on foot vertex contact, whereas the contact score calculation is based on foot joints.

We visualize a few examples of DIMOS and the proposed for locomotion task in Fig. 4. As shown, the synthesized motion for locomotion demonstrates lower scene penetration, less skating and higher motion diversity. That is attributes to the implicit policy and the stochastic sampling procedure during denoising.

Scene Object Interaction

We take scenes with furniture from ShapeNet (Chang et al. 2015) to evaluate the performance on scene object interaction, and 10 objects (3 armchairs, 3 straight chairs, 3 sofas and 1 L-sofa) are chosen for atomic interaction as previous work (Zhao et al. 2023). For each scene, the human is initialized to stand in front of the interaction object and guided to interact with it and finally return to original position.
Metrics. We take 4 metrics to evaluate the performance, including (1) the time to finish the task measured in seconds, (2) the foot contact score mentioned in Eq. 12, (3) mean human mesh vertex penetration $\sum_{v \in M} |(f_{SDF}(v))_-|$ over time, and (4) maximum penetration over time.
Results. We compare the proposed method with prevailing methods (Hassan et al. 2021a; Zhao et al. 2023) on human-scene interaction. The interaction tasks for sitting and lying are evaluate separately, and the results are reported in Tab. 2.

	time ↓	pene. mean ↓	pene. max ↓
SAMP sit	8.63	11.91	45.22
DIMOS sit	4.09	1.91	10.61
Ours sit	**3.71**	**1.86**	**7.13**
SAMP lie	12.55	44.77	238.81
DIMOS lie	4.20	9.90	44.61
Ours lie	**3.55**	**9.80**	**30.8**

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.

Fig. 5 visualizes the synthesized human-scene interaction given by DIMOS and our method. The visual results indicate our method presents more plausible interaction with higher human-scene contact and lower human-scene collision.

Long-term Motion Synthesis

For long-term motion synthesis in 3D scenes where multiple tasks are completed consecutively, objects with feasible interaction in scenes are randomly selected. We utilize COINS (Zhao et al. 2022) to sample the static interactions with these objects as the goals. All compared methods use the same initial state and task goals to synthesize long-term motions, ensuring a fair comparison.
Metrics. To comprehensively evaluate the synthesized motions, we conduct a user study where results given by different methods are directly rated by participants. We totally generate 20 motion for each method, 10 for scenes from PROX, and 10 for scenes from Replica. We present motions synthesized by different methods in the same scene to participants simultaneously, and ask the them to rate the results on a scale of 1 to 5 (in increments of 0.5) from four perspectives: motion naturalness, diversity, interaction plausibility and overall performance.
Results. Finally, $1,200$ ratings from 15 participants are collected for each method (60 results per sample from 4 as-

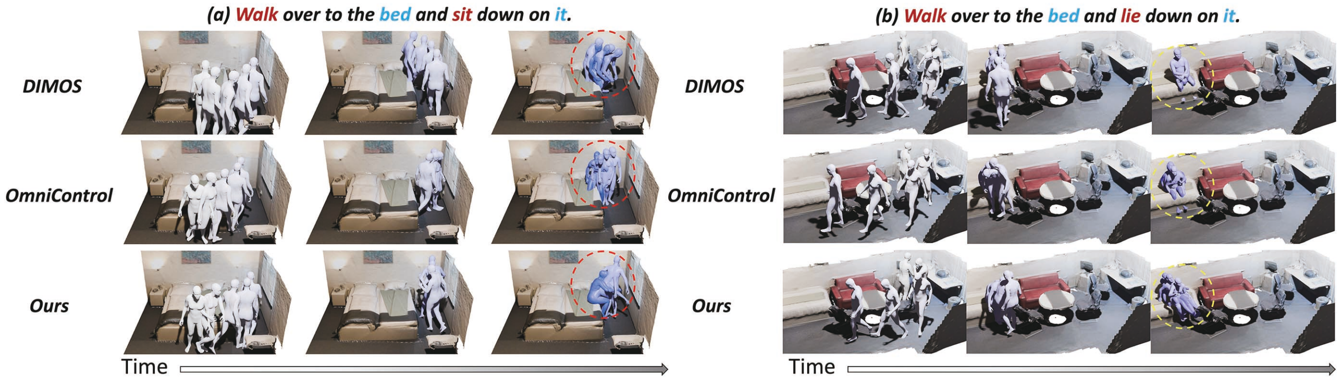

Figure 6: Visual comparison of different methods on motion synthesis in 3D scenes from PROX.

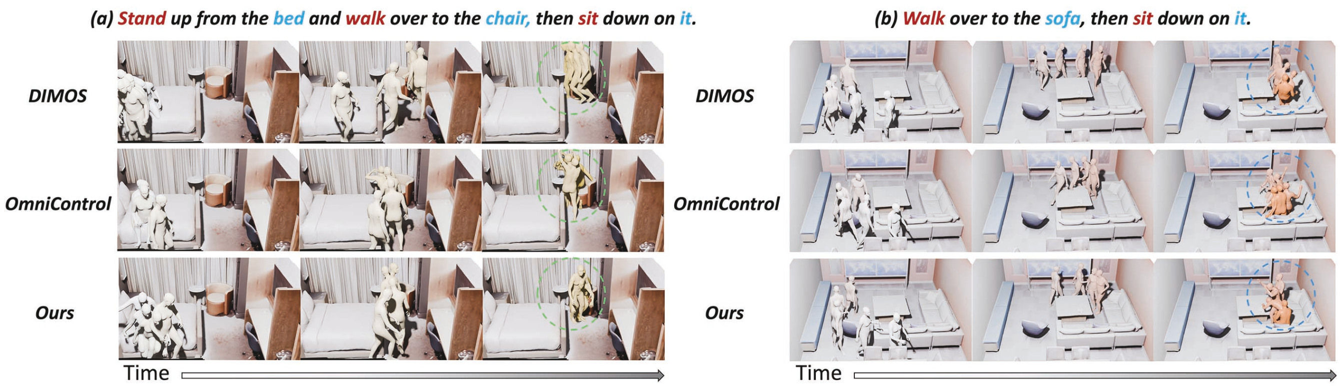

Figure 7: Visual results of synthesized motions given by compared methods in Replica scenes.

pects) and we report the scores in Tab. 3. It can be seen, the proposed method achieve the best performance in motion diversity, interaction plausibility and overall performance even without paired motion-scene data for training. For motion naturalness, our proposed method is on par with DIMOS, as no specific designs are taken to improve motion naturalness.

	Naturalness ↑	Diversity ↑	Plausibility ↑	Overall ↑
DIMOS	2.72/**3.09**	3.00/3.21	2.55/2.85	2.86/3.11
OmniControl	2.66/2.67	2.83/3.07	2.26/2.5	2.61/2.69
Ours	**2.81**/3.03	**3.34/3.36**	**3.15/3.17**	**3.17/3.26**

Table 3: Comparison between competitive methods based on user study. Users are asked to give scores (ranging from 1 to 5, ↑) according to motion naturalness, diversity, interaction plausibility, and overall performance. Results on PROX-/Replica are reported on the left/right respectively.

We present the results of motion synthesis in scenes from PROX and Replica dataset respectively in Fig. 6 and Fig. 7. It can be seen that the proposed Diffusion Implicit Policy performs better in terms of motion naturalness, interaction plausibility and motion diversity. This is due to the integration of motion denoising, implicit policy optimization, and random sampling within a unified framework.

Conclusion

In this paper, we propose a unified framework, termed Diffusion Implicit Policy, for motion synthesis in 3D scenes, where paired motion-scene data is no longer needed for training. In this framework, interaction is disentangled from motion learning during training. Motion prior from diffusion model and implicit policy from reward functions are integrated together to iteratively optimize the motion from random noise, pursuing motion naturalness, diversity and interaction plausibility simultaneously. We utilize joint hints and inpainting to ensure that keyframe poses remain consistent with the historical motion. We adjust the motion distribution centroid in a GAN Inversion manner to achieve better interaction plausibility while maintaining motion continuity. We introduce motion blending in the power space of the joint rotation matrix to ensure smooth transitions between multiple tasks for long-term motion synthesis. Comprehensive experiments on generated scenes with ShapeNet furniture, and scenes from PROX and Replica indicate the effectiveness and generalization capability. Such a promising solution can also encourages future works to learn scene-aware motion synthesis from unpaired motion and scene data.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 62222602, 62302167, U23A20343, 62476090, 62502178, 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

Bau, D.; Strobelt, H.; Peebles, W.; Wulff, J.; Zhou, B.; Zhu, J.-Y.; and Torralba, A. 2019a. Semantic photo manipulation with a generative image prior. *ACM TOG*, 38(4): 1–11.

Bau, D.; Zhu, J.-Y.; Wulff, J.; Peebles, W.; Strobelt, H.; Zhou, B.; and Torralba, A. 2019b. Seeing what a gan cannot generate. In *ICCV*, 4502–4511.

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

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.

Guo, C.; Zuo, X.; Wang, S.; Zou, S.; Sun, Q.; Deng, A.; Gong, M.; and Cheng, L. 2020. Action2motion: Conditioned generation of 3d human motions. In *ACM MM*, 2021–2029.

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.

Holden, D.; Komura, T.; and Saito, J. 2017. Phase-functioned neural networks for character control. *ACM TOG*, 36(4): 1–13.

Hu, Z.; Yin, Z.; Haeufle, D.; Schmitt, S.; and Bulling, A. 2024. Hoimotion: Forecasting human motion during human-object interactions using egocentric 3d object bounding boxes. *IEEE TVCG*.

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.

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

Karunratanakul, K.; Preechakul, K.; Suwajanakorn, S.; and Tang, S. 2023. GMD: Controllable Human Motion Synthesis via Guided Diffusion Models. *arXiv preprint arXiv:2305.12577*.

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

Lin, J.; Zeng, A.; Lu, S.; Cai, Y.; Zhang, R.; Wang, H.; and Zhang, L. 2023. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *arXiv preprint arXiv:2307.00818*.

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

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.

Mir, A.; Puig, X.; Kanazawa, A.; and Pons-Moll, G. 2024. Generating continual human motion in diverse 3d scenes. In *International Conference on 3D Vision (3DV)*, 903–913.

Mullen, J. F.; Kothandaraman, D.; Bera, A.; and Manocha, D. 2023. Placing Human Animations into 3D Scenes by Learning Interaction-and Geometry-Driven Keyframes. In *WACV*, 300–310.

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. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 652–660.

Rempe, D.; Luo, Z.; Bin Peng, X.; Yuan, Y.; Kitani, K.; Kreis, K.; Fidler, S.; and Litany, O. 2023. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *CVPR*, 13756–13766.

Ren, Y.; Zhao, C.; He, Y.; Cong, P.; Liang, H.; Yu, J.; Xu, L.; and Ma, Y. 2023. LiDAR-aid Inertial Poser: Large-scale Human Motion Capture by Sparse Inertial and LiDAR Sensors. *IEEE TVCG*, 29(5): 2337–2347.

Savva, M.; Chang, A. X.; Hanrahan, P.; Fisher, M.; and Nießner, M. 2014. SceneGrok: Inferring action maps in 3D environments. *ACM TOG*, 33(6): 1–10.

Savva, M.; Chang, A. X.; Hanrahan, P.; Fisher, M.; and Nießner, M. 2016. Pigraphs: learning interaction snapshots from observations. *ACM TOG*, 35(4): 1–12.

Shafir, Y.; Tevet, G.; Kapon, R.; and Bermano, A. H. 2024. Human Motion Diffusion as a Generative Prior. In *ICLR*.

Starke, S.; Zhang, H.; Komura, T.; and Saito, J. 2019. Neural state machine for character-scene interactions. *ACM TOG*, 38(6): 178.

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*.

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*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.; Chai, J.; and Xia, S. 2019. Combining recurrent neural networks and adversarial training for human motion synthesis and control. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 27(1): 14–28.

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.

Wu, Z.; Wang, Q.; Zheng, X.; Ye, J.; Yang, P.; Wang, Y.; and Wang, Y. 2024. Doodle Your Motion: Sketch-Guided Human Motion Generation. *IEEE TVCG*.

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*.

Xie, Z.; Wu, Y.; Gao, X.; Sun, Z.; Yang, W.; and Liang, X. 2023. Towards detailed text-to-motion synthesis via basic-to-advanced hierarchical diffusion model. *arXiv preprint arXiv:2312.10960*.

Xuan, H.; Li, X.; Zhang, J.; Zhang, H.; Liu, Y.; and Li, K. 2023. Narrator: Towards Natural Control of Human-Scene Interaction Generation via Relationship Reasoning. In *ICCV*, 22268–22278.

Yi, H.; Thies, J.; Black, M. J.; Peng, X. B.; and Rempe, D. 2024. Generating human interaction motions in scenes with text control. In *ECCV*, 246–263. Springer.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*, 3836–3847.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2024. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE TPAMI*.

Zhang, M.; Guo, X.; Pan, L.; Cai, Z.; Hong, F.; Li, H.; Yang, L.; and Liu, Z. 2023a. ReMoDiffuse: Retrieval-Augmented Motion Diffusion Model. *arXiv preprint arXiv:2304.01116*.

Zhang, M.; Li, H.; Cai, Z.; Ren, J.; Yang, L.; and Liu, Z. 2023b. FineMoGen: Fine-Grained Spatio-Temporal Motion Generation and Editing. *arXiv preprint arXiv:2312.15004*.

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, X.; Bhatnagar, B. L.; Starke, S.; Guzov, V.; and Pons-Moll, G. 2022. Couch: Towards controllable human-chair interactions. In *ECCV*, 518–535. Springer.

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.

Zhao, H.; Jiang, L.; Jia, J.; Torr, P. H.; and Koltun, V. 2021. Point transformer. In *ICCV*, 16259–16268.

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.