Synthesizing Diverse Human Motions in 3D Indoor Scenes

Kaifeng Zhao[1] Yan Zhang[1] Shaofei Wang[1] Thabo Beeler[2] Siyu Tang[1]
[1]ETH Zürich [2]Google
{kaifeng.zhao, yan.zhang, shaofei.wang, siyu.tang}@inf.ethz.ch
thabo.beeler@gmail.com

Figure 1: In this work, we propose a method to generate a sequence of natural human-scene interaction events in real-world complex scenes as illustrated here. The human first walks to sit on a stool (yellow to red), then walk to another chair to sit down (red to magenta), and finally walk to and lie on the sofa (magenta to blue).

Abstract

We present a novel method for populating 3D indoor scenes with virtual humans that can navigate in the environment and interact with objects in a realistic manner. Existing approaches rely on high-quality training sequences that contain captured human motions and the 3D scenes they interact with. However, such interaction data are costly, difficult to capture, and can hardly cover the full range of plausible human-scene interactions in complex indoor environments. To address these challenges, we propose a reinforcement learning-based approach that enables virtual humans to navigate in 3D scenes and interact with objects realistically and autonomously, driven by learned motion control policies. The motion control policies employ latent motion action spaces, which correspond to realistic motion primitives and are learned from large-scale motion capture data using a powerful generative motion model. For navigation in a 3D environment, we propose a scene-aware policy with novel state and reward designs for collision avoidance. Combined with navigation mesh-based path-finding algorithms to generate intermediate waypoints, our approach enables the synthesis of diverse human motions navigating in 3D indoor scenes and avoiding obstacles. To generate fine-grained human-object interactions, we carefully curate interaction goal guidance using a marker-based body representation and leverage features based on the signed distance field (SDF) to encode human-scene proximity relations. Our method can synthesize realistic and diverse human-object interactions (e.g., sitting on a chair and then getting up) even for out-of-distribution test scenarios with different object shapes, orientations, starting body positions, and poses. Experimental results demonstrate that our approach outperforms state-of-the-art human-scene interaction synthesis methods in terms of both motion naturalness and diversity. Code, models, and demonstrative video results are available at: https://zkf1997.github.io/DIMOS.

1. Introduction

Simulating how humans interact with environments plays an essential role in many applications, such as generating training data for machine learning algorithms, and simulating autonomous agents in AR/VR and computer games. Although this task is highly related to character animation in computer graphics, most existing character animation methods (e.g. [4, 5, 16]) focus on improving the realism and controllability of character movements. With the traditional character animation workflows, one can produce high-quality animations but can hardly generate autonomous and spontaneous natural human motions interacting with the surroundings in diverse plausible ways as real humans. Previous learning-based interaction synthesis methods [12, 42, 57] require simultaneously capturing human motion and scenes for supervision. However, capturing such training data is costly and challenging, resulting in a notably limited spectrum of human-scene interaction motions and difficulties in handling unseen interaction scenarios. This restriction also results in inferior motion quality of synthesized virtual humans.

To address this problem, we leverage reinforcement learning (RL) [45] to solve our task. By formulating goals as rewards, perception as states, and latent variables of deep generative models as actions, we can synthesize continuous, stochastic, plausible, and spontaneous motions of virtual humans to inhabit the digital world. Although existing RL-based motion synthesis approaches (e.g. [26, 33, 60]) can effectively generate natural motions to achieve goals, their generated virtual humans can only interact with simple scenes, rather than complex environments with functional furniture and diverse objects. For example, GAMMA [60] employs generative motion primitives and a policy network that are generalizable across diverse human body shapes, but it can only synthesize waypoint-reaching locomotions. The trained digital humans are not aware of how to perform actions like sitting on a chair or lying on a sofa, and frequently inter-penetrate with the scene geometry.

To overcome these limitations, we propose a novel framework to learn both scene and interaction-aware motion control policies for synthesizing realistic and diverse human-scene interaction motions. First, in order to improve the physical plausibility of the synthesized human motions, we design a new scene-aware policy to help virtual humans avoid collisions with scene objects. Specifically, we use a 2D occupancy-based local walkability map to incorporate scene information into the locomotion policy. In addition, we add features derived from the signed distance from body markers to the object surface and the gradient direction of the signed distance to encode the proximity between humans and objects for object interaction policies. Second, in order to achieve controllable object interactions, we provide fine-grained guidances based on surface markers [58] of a human body performing the target interactions. Specifically, we use COINS [61] to generate human bodies interacting with scene objects given the interaction semantics, and then use the body markers as the interaction guidance for motion synthesis. Combined with navigation mesh-based path-finding algorithms to generate intermediate waypoints in 3D scenes, virtual humans can autonomously reach target locations in complex environments and mimic target poses in a variety of plausible ways.

We train the policy networks in synthetic scenes consisting of randomized objects to learn generalizable scene-aware locomotion and fine-grained object interactions. With this framework, we investigate how to synthesize diverse in-scene motions consisting of locomotion, sitting, and lying. We empirically evaluate the motion realism and expressiveness of our proposed method, and compare it with state-of-the-art methods. The results show that our approach consistently outperforms the baselines in terms of diversity, physical plausibility, and perceptual scores.

In summary, we aim to let virtual humans inhabit virtual environments, and present these contributions:

1. We propose a reinforcement learning-based framework to generate realistic and diverse motions of virtual humans in complex indoor scenes.

2. We propose to use body surface markers as detailed interaction goals for fine-grained human-object interaction synthesis and leverage COINS [61] to generate articulated 3D human bodies based on interaction semantics to make virtual humans controllable via interaction semantics and fine-grained body poses.

3. We design scene and interaction-aware policies to enable virtual humans to navigate in 3D scenes while avoiding collisions, to interact with scene objects, and to continuously perform sequences of activities in complex scenes.

2. Related Work

Human motion synthesis. Generating high-quality human motion has been widely explored in computer vision and graphics. Motion graph [22] and motion matching [4, 5, 63] generate motion by searching suitable clips in datasets and blending them automatically. Starke *et al*. [41–43] use phase-conditioned neural networks to synthesize character animations and interaction with objects. Zhang *et al*. [58, 60] model motion as sequences of surface markers on the parametric SMPL-X [31] body model, and train autoregressive networks on the large-scale mocap dataset AMASS [29] in order to produce diverse motions of bodies with various shapes. Peng *et al*. [33] use imitation learning with a skill discovery objective to learn a general motion skill space for physically simulated characters. Tang *et al*.

[46] trains a motion manifold model of consecutive frames for real-time motion transition. Recent works [23, 24] propose generative methods to synthesize motions from single or a few example motion sequences. Transformer-based models have been designed to predict or generate stochastic motions conditioned on action categories [35], texts [36, 48], gaze [62], and others. More recently, motion diffusion models [1, 2, 49, 50, 54] achieve appealing performance on motion synthesis conditioned on various control signals and demonstrate flexible motion editing.

Motion control and RL-based motion synthesis. Various motion control methods have been proposed to constrain body movements or guide the body to reach goals. Sampling-based motion control methods [27, 28] generate multiple samples at each step and select the samples that best match the targets. Goal-conditional generation networks [10, 17, 21] are applied for motion control. However, such methods may produce invalid results when the train-test domain gap is large. Optimization-based motion control methods [18, 53] leverage the learned generative motion model as regularization and optimize the motion latent variables to fit the decoded motion to the goals. Motion diffusion models [49, 54] implement control via classifier-free text guidance or gradually reprojecting the generated motion onto the physically plausible space at individual denoising steps. Human motions can be formulated as a Markov decision process, and synthesized and controlled via RL. Imitation learning methods [3, 30, 32, 34] trains policy networks to control humanoids to imitate reference motion and complete given tasks. Peng *et al*. [33] propose a skill discovery objective apart from the imitation objective to learn a latent space for general motion skills and train substream task policies leveraging the general motion space. Follow-up works [20, 47] combine the character controllers with language-based selection or finite state machine to compose more complex movements. Ling *et al*. [26] and Zhang *et al*. [60] leverage the latent variable of the generative motion model as the action, and train policies to generate high-fidelity perpetual motions.

Human-scene interaction synthesis. Synthesizing natural human-scene interactions is an appealing direction in recent years. Object affordance [8, 9, 19, 38] provides cues about how humans can interact with scene objects. Static bodies interacting with scene objects can be generated based on the scene semantics, geometries, and interaction semantics [14, 25, 56, 59, 61]. Wang *et al*. [52] propose to train generative models for human motion in scenes conditioned on scene pointclouds. Wang *et al*. [51] propose a three-stage method to synthesize human motions in scenes, which first puts static bodies in scenes, then generates global paths, and afterwards synthesizes in-between motions. Starke [42] propose an autoregressive neural state machine to synthesize character animations interacting with objects. Hassan *et al*. [12] and Zhang *et al*. [57] propose similar networks as in [42], and incorporate stochastic motion generation and finer-grained motion control modules. Peng *et al*. [33] trains policy to control physically-simulated characters to hit a box. Hassan *et al*. [15] extend the method to generate interactions like carrying a box or sitting on a chair. Zhang *et al*. [55] learns physically simulated tennis players from broadcast videos.

Ours versus others. Our method is most similar to GAMMA [60] and SAMP [12]. GAMMA learns generalizable motion models and policies across human bodies of diverse identities and shapes, without goal-motion paired training data. Despite producing high-fidelity motions, its results are limited to locomotion in the scene, and frequently collide with scene objects. SAMP learns conditional VAEs to produce sitting and lying actions in living environments, with object-motion paired data. Its generated motion has visible artifacts such as foot-ground skating. Our method combines their merits and eliminates their individual disadvantages. We extend the RL-based framework proposed in GAMMA by incorporating fine-grained motion controls (guided by interaction semantics) and scene interaction modules, so as to generate human-scene interactions in complex daily living environments. Compared to SAMP, our produced motion is more diverse, more physically plausible, and can be guided by fine-grained body surface markers. For systematic comparisons, please refer to Section 4.

3. Method

3.1. Preliminaries

SMPL-X [31] and body representation. We use SMPL-X to represent 3D human bodies in our work. Given shape parameter $\boldsymbol{\beta} \in \mathbb{R}^{10}$ and body pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{63}$, SMPL-X produces a posed body mesh with a fixed topology of 10475 vertices. To place the body in a scene, the root location $\boldsymbol{r} \in \mathbb{R}^3$ and the orientation $\boldsymbol{\phi} \in \mathfrak{so}(3)$ w.r.t. the scene coordinates are additionally needed. Since facial expressions and hand gestures are not our focus, we leave their parameters as the default values. In addition, we follow [58, 60] to represent the body in motion by the SSM_67 body surface marker placement. A motion sequence is then formulated as $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$, where N is the length of motion and $\boldsymbol{x}_i \in \mathbb{R}^{67\times 3}$ denotes the body marker 3D locations at frame i. The marker locations are relative to a canonical coordinate frame centered at the body pelvis in the first frame.

GAMMA [60]. GAMMA can synthesize stochastic, perpetual, and realistic goal-reaching actions in 3D scenes. It comprises generative motion primitive models, RL-based control, and tree-based search to implement gradient-free test-time optimization. The motion primitive is formulated by a CVAE model to generate uncertain marker motions for 0.25 seconds into the future given a motion seed, followed by a MLP-based body regressor to yield SMPL-X parameters. Long-term random motion can be generated by running the motion primitive model recursively. The RL-based control is implemented by learning a policy within a simulation area. The actor-critic framework [45] and the PPO algorithm [39] are applied to update the policy network. An additional motion prior term is used to ensure the motion appears natural. During testing, the generated motion primitives are stored in a tree where only the best K primitives at each layer are preserved in order to discard low-quality sampling results.

COINS[61]. COINS generates physically plausible static human-scene interactions with instance-level semantic control. Given the point cloud of an object and action labels, COINS can generate static bodies interacting with the given object based on the specified action, *e.g.*, sitting, lying, or touching. COINS leverages transformer-based generative models trained on a human-scene interaction dataset [13] to first generate a plausible body pelvis for interaction and then the posed body. The generated bodies are further optimized to improve the physical plausibility and to match the predicted action-dependent contact areas with objects. Such generated static bodies capture the characteristics of the human-scene interaction process and can be used as fine-grained interaction guidance.

3.2. RL-based Framework to Inhabit the Virtual

As illustrated in Fig. 2, we propose a motion synthesis framework that enables virtual humans to navigate in complex indoor scenes and interact with various scene objects, *e.g.*, sitting on a chair. Compared to the GAMMA framework [60], our method incorporates scene information into the states to better handle complex human-scene interactions. Also, we use body markers as goals to provide fine-grained guidance on how to drive the body for the target interactions. With modularized path-finding methods and static person-scene interaction generation methods, our framework can synthesize realistic human motions in complex 3D environments. In our work, we use COINS [61] to generate static person-scene interactions from interaction semantics given as 'action-object' pairs. The walking path can be either generated by hand, or by automatic path-finding algorithms like A* [11, 40].

We formulate our motion synthesis tasks with reinforcement learning. At each time step, a virtual human perceives its state s_t in the environment and samples an action a_t from its policy model $\pi(a_t|s_t)$. Based on its motion model, it advances its motion state, and obtains a new perception state s_{t+1}. A reward $r_t = r(s_t, a_t, s_{t+1})$ is calculated, tailored to different tasks.

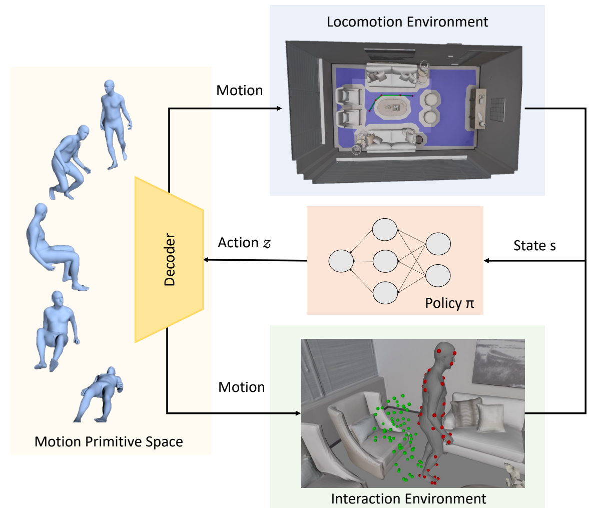

Figure 2: Illustration of our proposed human-scene interaction synthesis framework, which consists of learned motion primitive (actions), alongside locomotion and interaction policies generating latent actions conditioned on scenes and interaction goals. By integrating navigation mesh-based path-finding and static human-scene interaction generation methods, we can synthesize realistic motion sequences for virtual humans with fine-grained controls.

The motion model and the action. We leverage the CVAE-based generative motion primitive [60] as our motion model, and use its latent variables as actions. We train the model conditioned on 1 or 2 past frames using the combination of the SAMP [12] and AMASS [29] motion capture datasets, to learn a latent motion primitive space covering motion skills for human-scene interactions. Each latent variable z in the motion primitive space is regarded as an action and can be decoded to a short clip of motion.

The state. The state is formulated by

$$s_t = (X_s, I, G), \tag{1}$$

where $X_s \in \mathbb{R}^{M \times 67 \times 3}$ is the body markers motion seed that represents a motion history of M frames. I and G denote the person-scene interaction feature and the goal-reaching feature, respectively. The interaction feature and goal-reaching feature vary among the locomotion and object interaction tasks. We introduce the detailed formulation in Sec. 3.3 and 3.4.

The rewards. The rewards evaluate how well the virtual human performs locomotion and fine-grained object interaction tasks. We formulate rewards as

$$r = r_{goal} + r_{contact} + r_{pene}, \tag{2}$$

where r_{goal}, $r_{contact}$, and r_{pene} represent the rewards for goal-reaching, foot-ground contact, and penetration avoidance, respectively. Specifically, the contact reward $r_{contact}$ encourages foot-floor contact and discourages foot skating, and is defined as:

$$r_{contact} = e^{-(|\min x_z|-0.05)_+} \cdot e^{-(\min \|x_{vel}\|_2-0.075)_+}, \tag{3}$$

where F is the set of foot markers, $\boldsymbol{x}_z$ is the height of the markers, $\boldsymbol{x}_{vel}$ is the velocity of the markers, and $(\cdot)_+$ denotes clipping negative values. There are tolerance thresholds of 0.05m for foot-floor distance and 0.075m/s for skating, following GAMMA [60]. The other two reward terms are action-specific and introduced in Sec. 3.3 and 3.4.

Policy network and training. We use the actor-critic algorithm [45] to learn the policy, where a policy network and a value network are trained jointly. The policy network generates a diagonal Gaussian distribution representing the stochastic action distribution given a state, while the value network outputs the value estimation for each state.

Like in GAMMA [60], these two networks are jointly trained by minimizing:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{PPO} + \mathbb{E}[(r_t - V(\boldsymbol{s}_t))^2] \\ + \alpha\text{KL-div}(\pi(\boldsymbol{z}|\boldsymbol{s})||\mathcal{N}(0, I)),\end{aligned} \tag{4}$$

where the first term is the PPO [39] loss, the second updates the value estimation of the critic networks, and the third Kullback–Leibler divergence term regularizes motion in the latent space [60].

Tree sampling for test-time optimization. Given the stochastic nature of our Gaussian policies, sampling motions from the generated action distributions can yield motion primitive results of various qualities. Therefore, we follow [60] to use tree-based sampling during inference to discard motion primitives with inferior goal-reaching and scene interaction scores. Specifically, we sample multiple latent actions at each time step and selectively keep the best K samples, utilizing the same rewards used to train the policies as the selection criteria. This tree-sampling technique yields improved synthesis results of higher quality.

In the following, we elaborate on the design of states and rewards tailored to different actions, namely locomotion and fine-grained object interaction. By combining the learned policies, long-term and coherent motions can be composed by rolling out the initial body and switching between the locomotion and object interaction stages.

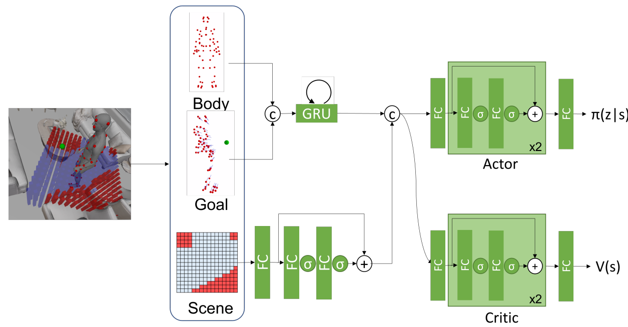

Figure 3: Illustration of the scene-aware locomotion policy network. The locomotion policy state consists of the body markers, the goal-reaching feature of normalized direction vectors from markers to the goal pelvis, and the interaction feature of a 2D binary map indicating the walkability (red: non-walkable area, blue: walkable area) of the local $1.6m \times 1.6m$ square area. The locomotion policy network employs the actor-critic architecture and shares the state encoder.

3.3. Scene-Aware Locomotion Synthesis

Navigating in cluttered scenes here means the human body moving to a target while avoiding collisions with scene objects. Our key idea is to incorporate the walkability information of the surrounding environment into the states and use collision rewards to train the locomotion policy in order to avoid scene collisions. Specifically, we represent the walkability of the environment surrounding the human agent using a 2D binary map $\mathcal{M} \in \{0, 1\}^{16\times16}$ as illustrated in Fig. 3.

The walkability map is defined in the human's local coordinates and covers a $1.6m \times 1.6m$ area centered at the body pelvis and aligned with the body facing orientation. It consists of a 16×16 cell grid where each cell stores a binary value indicating whether this cell is walkable or not. This local walkability map enables the policy to sense surrounding obstacles.

Referring to Eq. 1, the person-scene interaction feature is specified by

$$I = \text{vec}(\mathcal{M}), \tag{5}$$

in which $\text{vec}(\cdot)$ denotes vectorization. The goal-reaching feature is specified by

$$G = (\tilde{\boldsymbol{g}}_p - \boldsymbol{X}_s)_n, \tag{6}$$

where $\boldsymbol{X}_s$ and $\tilde{\boldsymbol{g}}_p \in \mathbb{R}^{M\times67\times3}$ are the body marker seed representing M frames history of motion, and the broadcasted target pelvis location relative to the body-centered canonical coordinate, respectively. $(\tilde{\boldsymbol{g}}_p - \boldsymbol{X}_s)_n$ denotes the normalized vectors pointing from each marker to the goal pelvis.

The rewards contributing to Eq. 2 are defined as

$$r_{pene} = e^{-|\mathcal{M}_0 \cap \mathcal{B}_{xy}(X)|}, \tag{7}$$

where $\mathcal{M}_0$ denotes the non-walkable cells in the walkability map, $\mathcal{B}_{xy}(\cdot)$ denotes the 2D bounding box of the body markers $\boldsymbol{X}$, $\cap$ denotes their intersection, and $|\cdot|$ denotes the number of non-walkable cells overlapping with the human bounding box.

$$r_{goal} = r_{dist} + r_{ori}, \tag{8}$$

$$r_{dist} = 1 - (\|\boldsymbol{p} - \boldsymbol{g}_p\|_2 - 0.05)_+, \tag{9}$$

$$r_{ori} = \frac{\langle \boldsymbol{o}, \boldsymbol{g}_p - \boldsymbol{p} \rangle}{2}, \tag{10}$$

where r_{dist} encourages the body pelvis $\boldsymbol{p}$ to be close to the pelvis goal $\boldsymbol{g}_p$ and r_{ori} encourages the body facing direction $\boldsymbol{o}$ to be aligned with the direction from the current body pelvis $\boldsymbol{p}$ to the pelvis goal $\boldsymbol{g}_p$.

3.4. Fine-grained Object Interaction Synthesis

To synthesize fine-grained human-object interactions, *e.g*. sitting on a chair or lying on a sofa, we use body marker goals as guidance, and model the proximity between the body surface and the scene object in a compact way. The goal marker sets can be generated by static person-scene interaction methods such as [14, 56, 61]. We use COINS [61] to generate the static goal interaction body for its performance and controllability of interaction semantics.

In addition to the marker-based goal guidance, we incorporate the proximity relations between humans and objects into the states. Specifically, we use the signed distance from each marker to the surface of the scene object, as well as the gradient direction of the signed distance to represent the proximity relationship, as illustrated in Fig. 4. Both the signed distance and its gradient direction are calculated using the object's signed distance field (SDF) Ψ_O.

Referring to Eq. 1, the person-scene interaction feature is formulated as

$$I = [\Psi_O(\boldsymbol{X}_s), \nabla\Psi_O(\boldsymbol{X}_s)], \tag{11}$$

in which $\Psi_O \in \mathbb{R}^{M\times 67}$ and $\nabla\Psi_O \in \mathbb{R}^{M\times 201}$ denote the SDF values and the gradient at each marker location in the M frames, respectively, and $[\cdot,\cdot]$ denotes feature concatenation. The goal-reaching feature is formulated as

$$G = [(\tilde{\boldsymbol{g}}_m - \boldsymbol{X}_s)_n, \|\tilde{\boldsymbol{g}}_m - \boldsymbol{X}_s\|_2], \tag{12}$$

in which $\boldsymbol{X}_s \in \mathbb{R}^{M\times 67\times 3}$ denotes the body markers seed, $\tilde{\boldsymbol{g}}_m \in \mathbb{R}^{M\times 67\times 3}$ denotes the broadcasted goal body markers, $(\tilde{\boldsymbol{g}}_m - \boldsymbol{X}_s)_n$ denotes the normalized vector representing the direction from each marker to the corresponding target marker, $\|\tilde{\boldsymbol{g}}_m - X_s\|_2$ denotes the distance from each marker to target marker.

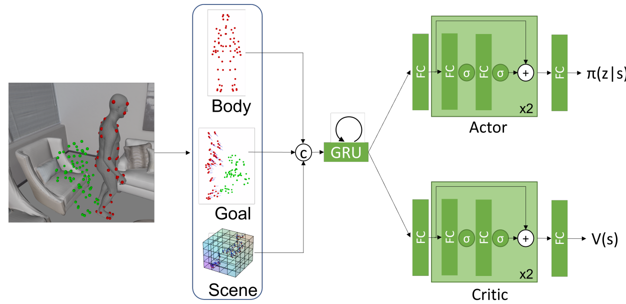

Figure 4: Illustration of the object interaction policy network. The interaction policy state consists of the body markers, the goal-reaching features of both distance and direction from current markers to the goal markers, and the interaction features of the signed distances from each marker to the object surfaces and the signed distance gradient at each marker location. Such interaction features encode the human-object proximity relationship.

The interaction policy is trained using the reward defined in Eq. 2 with the following interaction-specific goal reward and penetration reward:

$$r_{goal} = 1 - (\|\boldsymbol{x} - \boldsymbol{g}_m\|_2 - 0.05)_+, \tag{13}$$

$$r_{pene} = e^{-\frac{1}{|V|}\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{|V|}|(\Psi_O(\boldsymbol{v}_{ti}))_-|}, \tag{14}$$

with $|V|$ being the SMPL-X mesh vertices, T denotes the number of frames in each motion primitive (equals to 10 in our study). The distance reward encourages the final frame body markers $\boldsymbol{x}$ to be close to the goal body markers $\boldsymbol{g}_m$. The penetration reward penalizes all the body vertices within a motion primitive that have negative SDF values. We use body vertices instead of joints because human-object contact happens on the body surface and can be better detected using vertices.

Moreover, we train the interaction policy with a mixture of 'sit/lie down' and 'stand up' tasks. This training scheme enables the human agent to also learn how to stand up and transit from object interaction back to locomotion, which enables the synthesis of a sequence of interaction activities as in Fig. 1.

4. Experiment

Motion Datasets. We combine the large-scale motion capture dataset AMASS [29] with SAMP [12] motion data to train the motion primitive models. Each sequence is first subsampled to 40 FPS and then split into 10 frames and 100 frames clips. Each motion clip is canonicalized using the first frame body. Specifically, we select AMASS sequences annotated with 'sit' or 'lie' in BABEL [37] and all motion data of SAMP to train the motion primitive model. We train

Table 1: Evaluation of the locomotion task. The up/down arrows denote the score is the higher/lower the better and the best results are in boldface.

	time ↓	avg. dist ↓	contact ↑	loco pene ↑
SAMP [12]	5.97	0.14	0.84	0.94
GAMMA [60]	**3.87**	**0.03**	0.94	0.94
Ours	6.43	0.04	**0.99**	**0.95**

separate motion primitive models for locomotion and interactions using task-related data. We observe that extending SAMP motion data with AMASS dataset is the key to making interaction policies work.

Policy Training Environments. To train the scene-aware locomotion policy, we randomly generate synthetic cluttered scenes consisting of random objects from ShapeNet [6]. Random initial and target location pairs are sampled in the walkable areas using navigation meshes to train the locomotion policy. To train the interaction policy, we use the static person-object interaction data from PROX [13] and retargeted to ShapeNet objects. We first randomly sample a furniture and interaction body goal from the retargeted PROX data. Then we sample the initial body with random poses and locations in front of the object to train the interaction policy. We also randomly swap the initial and goal body to learn both 'sit/lie down' and 'stand up' motions. Please refer to Supp. Mat. for more details.

4.1. Locomotion in 3D Scenes

We randomly generate test scenes for locomotion in the same way as the training scenes. The virtual human is instructed to move from the random start point to the random target point while avoiding penetration with scene objects.

Baselines and metrics. We compare our method with SAMP [12] and GAMMA [60] for locomotion. The SAMP results are recorded by running the released Unity demo. The start and termination are manually determined so the reported completion time may be slightly higher than the actual time due to human response time. The evaluation metrics for locomotion include: 1) time from start point to target point or reaching the time limit, measured in seconds. 2) the average distance from the final body to the targets, measured in meters. 3) foot contact score encouraging the lowest feet joints on the floor and discouraging foot skating as defined in Eq. 3. We use body joints instead of markers to calculate the contact score because the marker set annotation for the SAMP body is missing. 4) locomotion penetration score indicating the percentage of body vertices that are inside the walkable areas according to the navigation mesh.

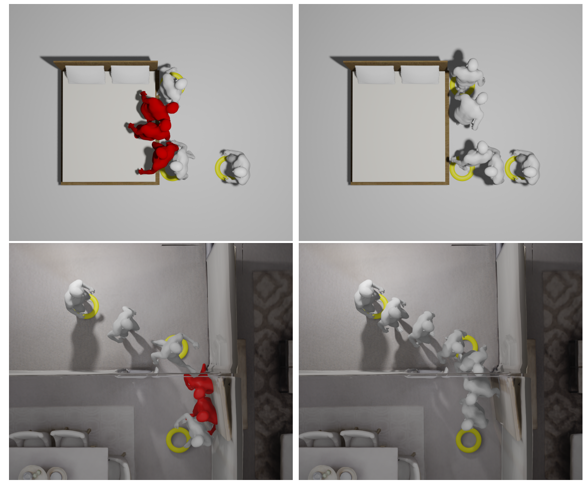

Figure 5: Demonstration of locomotion tasks where GAMMA [60] (left) collides with the obstacles (red bodies) while our scene-aware locomotion policy (right) avoids collision. The yellow circles denote the specified waypoints.

Results. Table 1 shows the empirical evaluation results. Our method achieves a higher contact score (0.99) than both GAMMA (0.94) and SAMP (0.84) which indicates better foot-floor contact and less foot skating. Moreover, our method achieves the highest penetration score which indicates our scene-aware policy can better avoid scene collisions. Fig. 5 shows examples of locomotion tasks where GAMMA collides into the scenes while our scene-aware policy successfully avoids penetration. Compared to GAMMA trained in the same environments, we observe our scene-aware policy learns more conservative behavior with lower moving speed and spent more time (6.43s) walking to the target locations, just like a human afraid of stepping on surrounding traps. All the methods can reach the target location within a reasonable distance.

4.2. Fine-Grained Human-Object Interaction

We evaluate the object interaction task on 10 unseen objects (3 armchairs, 3 straight chairs, 3 sofas, 1 L-sofa) from ShapeNet [6]. We use the object size annotation of ShapeNet and manually filter unrealistic-sized objects. The virtual human is randomly placed in front of the target object and then instructed to perform the interaction, stay for around 2 seconds, and then stand up. We evaluate two interactions of sitting and lying separately.

Baselines and metrics. We compare our method with SAMP [12] for the object interaction task. The evaluation metrics for the interaction tasks are: 1) time of completing the object interaction task. 2) foot contact score as defined in Eq. 3. Note that the foot contact score does not always reflect the motion quality for lying tasks because the foot can

Table 2: Evaluation of the interaction tasks. The up/down arrows denote the score is the higher/lower the better and the best results are in boldface.

	time ↓	contact ↑	pene. mean ↓	pene. max ↓
SAMP [12] sit	8.63	0.89	11.91	45.22
Ours sit	**4.09**	**0.97**	**1.91**	**10.61**
SAMP [12] lie	12.55	0.73	44.77	238.81
Ours lie	**4.20**	**0.78**	**9.90**	**44.61**

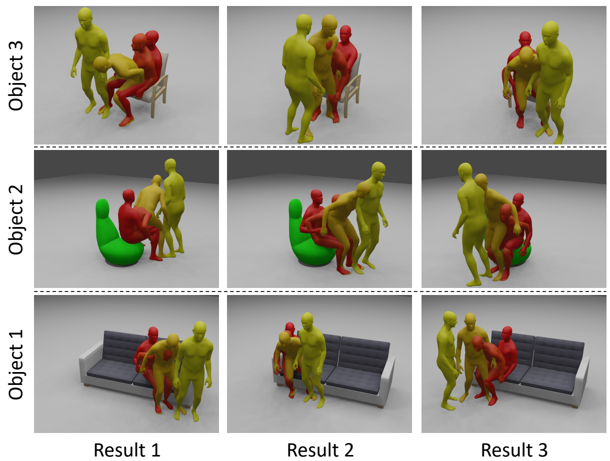

Figure 6: Demonstration of various generated object interactions. Each row shows generated interactions with the same object starting from random initial body locations and orientations. Colors from yellow to red denote time.

often be off the floor during lying. 3) interaction penetration score for each frame is defined as:

$$s_{inter\_pene} = \sum_{v_i \in V} |(\Psi_O(v_i))_-|, \qquad (15)$$

where Ψ_O is the object signed distance field, $(\cdot)_-$ clips all positve distance values to zero, and V is the body vertivces. We show both the average penetration over time and the maximum penetration in one sequence.

Results. Tab. 2 shows the evaluation results. Our method achieves a significantly higher foot contact score, indicating more natural motion. Our method also achieves lower mean and maximum penetration, which means our method generates more physically plausible results. Moreover, our method can complete the interaction tasks much faster than SAMP. This is because SAMP does not generalize well to unseen random body initialization and needs a longer time to start performing interactions. Qualitative results demonstrating the various object interactions generated by our method are shown in Fig. 6. Our object interaction policy generalizes to random initial body locations and orientations, as well as novel objects of various shapes.

Walk to sit on the bus bench   Walk to sit on the bed

Figure 7: Our method can synthesize natural human movements in scene reconstructions that are significantly different from the training scenes. We show the synthesis results in one Paris street [7] (left figure) and one indoor room [13] (right figure). Transparent to solid colors denote time.

Table 3: Evaluation of interaction sequences synthesis. Here ‘contact’ denotes the foot contact score, ‘loco. pene.’ is the percentage of body vertices inside the walkable areas, ‘inter. pene. mean/max’ denotes the mean and maximum penetration with interaction objects, and ‘perceptual’ denotes the ratio of being chosen as perceptually more natural. The best results are in boldface.

	SAMP [12]	Ours
contact ↑	0.87	**0.96**
loco. pene. ↑	0.62	**0.72**
inter. pene. mean↓	15.61	**3.40**
inter. pene. max ↓	101.25	**39.68**
perceptual. ↑	0.15	**0.85**

4.3. Interaction Sequences in 3D Scenes

Humans often continuously perform sequences of interactions with various objects in complex real-world scenes as shown in Fig. 1. Such interaction sequences are combinations of alternating locomotions and fine-grained interactions which we evaluate in Sec. 4.1 and Sec. 4.2 respectively. We conduct the empirical evaluation in real scene scans from Replica [44] and compare our method with SAMP. We select a list of interactable objects in the scene and instruct the virtual human to walk to the objects to perform interactions one by one. The evaluation metrics include the foot contact score, locomotion penetration score, and the mean and maximum interaction penetration score. We also conducted a perceptual study where participants are shown a side-by-side comparison of results from two methods and asked to choose the one perceptually more natural. We report the rate of being chosen as the better.

The evaluation results are shown in Tab. 3. Our method generates high-quality results of human-scene interactions in cluttered environments. Our results achieve higher contact scores and less penetration with the scenes compared

Table 4: Rewards ablation studies results, where '-Contact' and '-Penetration' denote policies trained without the floor contact and penetration avoidance reward, respectively.

	time ↓	contact ↑	pene. mean ↓	pene. max ↓
Ours	4.09	**0.97**	**1.91**	**10.61**
- Contact	**3.75**	0.88	19.49	60.10
- Penetration	4.03	0.95	16.80	45.50

to SAMP. In addition, our method generates perceptually more natural results according to perceptual study.

4.4. Ablation Studies

We perform ablation studies on the foot-ground contact reward and penetration avoidance reward of the interaction policies to substantiate the significance of these rewards. We train two ablation interaction policies using identical networks and environments, differing only in the exclusion of one of these rewards in each case. The quantitative metrics are reported in Tab. 4. The removal of the penetration avoidance reward yields a marked escalation in detected human-scene penetration. The removal of the foot-floor contact reward yields notably both inferior contact and penetration scores, accompanied by observed erratic synthesized motions that either remain suspended in the air or penetrate the floor. These empirical findings underscore the crucial role played by the foot-floor contact and penetration avoidance rewards in the learning of interaction policies.

5. Discussion and Conclusion

Limitations and Future Work. Our current method has various limitations that could be improved in future works. First, our method does not fully resolve penetrations with scene objects and floors since our method only uses rewards to encourage avoiding penetration with scenes, which does not impose hard constraints for penetrations. The combination of our method with physics simulation holds the potential to effectively address and resolve penetration issues. Moreover, the lying motions generated by our method are not as natural as the sitting motions because the motion primitive model fails to learn a comprehensive action space for lying given limited training data. Specifically, the available AMASS motion capture data for lying (167 seconds) is significantly less than the data for sitting (5K seconds). To overcome this issue, we aim to explore more data-efficient learning methods and scalable methods to collect human-scene interaction data. Furthermore, our locomotion policy is now limited to flat-floor scenes due to its reliance on the 2D occupancy-based walkability map. However, to broaden the applicability of our approach to more complex scenes such as uneven outdoor terrains and multi-floor buildings requiring walking upstairs, it will be necessary to replace the walkability map with a more suitable environment sensing mechanism. In addition, our method is restricted to interactions with static scenes. However, in the real world, humans are exposed to dynamic interactions with movable objects and scenes involving autonomous agents including other humans, animals, and vehicles. Extension to dynamic interactions will enable tackling a broader range of intricate and dynamic applications that mirror the complexities inherent in the real world.

Inter-penetrations Unnatural poses

Figure 8: Limitations. Penetrations remain observable in our results (left) since our method only encourages avoiding collision with rewards. Furthermore, the absence of sufficient training data for lying motions impedes the motion primitive model's ability to acquire a comprehensive lying motion space, leading to degraded motion (right).

Conclusion. In this paper, we leverage reinforcement learning to establish a framework to synthesize diverse human motions in indoor scenes, which is stochastic, realistic, and perpetual. The proposed method has large potential to improve many applications such as daily-living activity simulation, synthetic data creation, architecture design, and so on. Compared to existing methods, our method realizes fine-grained control by using body surface keypoints as targets, and achieves autonomous body-scene collision avoidance by incorporating scene information into the states and the rewards. Experiments show that our method effectively enables virtual humans to inhabit the virtual, and outperforms baselines consistently.

Acknowledgements. We sincerely acknowledge the anonymous reviewers for their insightful suggestions. This work was supported by the SNSF project grant 200021 204840 and SDSC PhD fellowship. Lingchen Yang provided valuable suggestions on visualization.

References

[1] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–20, 2023. ISBN: 0730-0301 Publisher: ACM New York, NY, USA. 3

[2] Tenglong Ao, Zeyi Zhang, and Libin Liu. GestureDiffuCLIP: Gesture diffusion model with CLIP latents. *arXiv preprint arXiv:2303.14613*, 2023. 3

[3] Kevin Bergamin, Simon Clavet, Daniel Holden, and James Richard Forbes. DReCon: data-driven responsive control of physics-based characters. *ACM Transactions On Graphics (TOG)*, 38(6):1–11, 2019. Publisher: ACM New York, NY, USA. 3

[4] Michael Buttner. Machine learning for motion synthesis and character control in games. *Proc. of I3D 2019*, 2, 2019. 2

[5] Michael Büttner and Simon Clavet. Motion matching-the road to next gen animation. *Proc. of Nucl. ai*, 2015(1):2, 2015. 2

[6] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, and Hao Su. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 7

[7] Jean-Emmanuel Deschaud, David Duque, Jean Pierre Richa, Santiago Velasco-Forero, Beatriz Marcotegui, and François Goulette. Paris-carla-3d: A real and synthetic outdoor point cloud dataset for challenging tasks in 3d mapping. *Remote Sensing*, 13(22), 2021. 8

[8] Helmut Grabner, Juergen Gall, and Luc Van Gool. What makes a chair a chair? pages 1529–1536. IEEE, 2011. 3

[9] Abhinav Gupta, Scott Satkin, Alexei A Efros, and Martial Hebert. From 3d scene geometry to human workspace. pages 1961–1968. IEEE, 2011. 3

[10] Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura. A recurrent variational autoencoder for human motion synthesis. 2017. 3

[11] Peter E Hart, Nils J Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968. Publisher: IEEE. 4

[12] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. pages 11374–11384, 2021. 2, 3, 4, 6, 7, 8

[13] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3D human pose ambiguities with 3D scene constraints. pages 2282–2292, 2019. 4, 7, 8

[14] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3D scenes by learning human-scene interaction. pages 14708–14718, 2021. 3, 6

[15] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing Physical Character-Scene Interactions, Feb. 2023. arXiv:2302.00883 [cs]. 3

[16] Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa. Learned motion matching. *ACM Transactions on Graphics (TOG)*, 39(4):53–1, 2020. Publisher: ACM New York, NY, USA. 2

[17] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. Publisher: ACM New York, NY, USA. 3

[18] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. Publisher: ACM New York, NY, USA. 3

[19] Ruizhen Hu, Zihao Yan, Jingwen Zhang, Oliver Van Kaick, Ariel Shamir, Hao Zhang, and Hui Huang. Predictive and generative neural networks for object functionality. *arXiv preprint arXiv:2006.15520*, 2020. 3

[20] Jordan Juravsky, Yunrong Guo, Sanja Fidler, and Xue Bin Peng. PADL: Language-Directed Physics-Based Character Control. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 3

[21] Kacper Kania, Marek Kowalski, and Tomasz Trzciński. TrajeVAE: Controllable Human Motion Generation from Trajectories. *arXiv preprint arXiv:2104.00351*, 2021. 3

[22] Lucas Kovar, Michael Gleicher, and Frédéric Pighin. Motion graphs. In *ACM SIGGRAPH 2008 classes*, pages 1–10. 2008. 2

[23] Peizhuo Li, Kfir Aberman, Zihan Zhang, Rana Hanocka, and Olga Sorkine-Hornung. Ganimator: Neural motion synthesis from a single sequence. *ACM Transactions on Graphics (TOG)*, 41(4):1–12, 2022. Publisher: ACM New York, NY, USA. 3

[24] Weiyu Li, Xuelin Chen, Peizhuo Li, Olga Sorkine-Hornung, and Baoquan Chen. Example-based Motion Synthesis via Generative Motion Matching. *arXiv preprint arXiv:2306.00378*, 2023. 3

[25] Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Putting humans in a scene: Learning affordance in 3d indoor environments. pages 12368–12376, 2019. 3

[26] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion vaes. *ACM Transactions on Graphics (TOG)*, 39(4):40–1, 2020. Publisher: ACM New York, NY, USA. 2, 3

[27] Libin Liu, KangKang Yin, and Baining Guo. Improving sampling-based motion control. volume 34, pages 415–423. Wiley Online Library, 2015. Issue: 2. 3

[28] Libin Liu, KangKang Yin, Michiel Van de Panne, Tianjia Shao, and Weiwei Xu. Sampling-based contact-rich motion control. In *ACM SIGGRAPH 2010 papers*, pages 1–10. 2010. 3

[29] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. AMASS: Archive of motion capture as surface shapes. pages 5442–5451, 2019. 2, 4, 6

[30] Josh Merel, Leonard Hasenclever, Alexandre Galashov, Arun Ahuja, Vu Pham, Greg Wayne, Yee Whye Teh, and Nicolas Heess. Neural probabilistic motor primitives for humanoid control. *arXiv preprint arXiv:1811.11711*, 2018. 3

[31] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face,

and body from a single image. pages 10975–10985, 2019. 2, 3

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

[33] Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions On Graphics (TOG)*, 41(4):1–17, 2022. Publisher: ACM New York, NY, USA. 2, 3

[34] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. Sfv: Reinforcement learning of physical skills from videos. *ACM Transactions On Graphics (TOG)*, 37(6):1–14, 2018. Publisher: ACM New York, NY, USA. 3

[35] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. pages 10985–10995, 2021. 3

[36] Mathis Petrovich, Michael J Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. pages 480–497. Springer, 2022. 3

[37] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. BABEL: bodies, action and behavior with English labels. pages 722–731, 2021. 6

[38] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. SceneGrok: Inferring action maps in 3D environments. *ACM transactions on graphics (TOG)*, 33(6):1–10, 2014. Publisher: ACM New York, NY, USA. 3

[39] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 4, 5

[40] Greg Snook. Simplified 3D movement and pathfinding using navigation meshes. *Game programming gems*, 1(1):288–304, 2000. Publisher: Charles River Media Newton Centre, MA, USA. 4

[41] Sebastian Starke, Ian Mason, and Taku Komura. Deepphase: Periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. Publisher: ACM New York, NY, USA. 2

[42] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6):209–1, 2019. 2, 3

[43] Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. Local motion phases for learning multi-contact character movements. *ACM Transactions on Graphics (TOG)*, 39(4):54–1, 2020. Publisher: ACM New York, NY, USA. 2

[44] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, and Shobhit Verma. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 8

[45] Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998. 2, 4, 5

[46] Xiangjun Tang, He Wang, Bo Hu, Xu Gong, Ruifan Yi, Qilong Kou, and Xiaogang Jin. Real-time controllable motion transition for characters. *ACM Transactions on Graphics (TOG)*, 41(4):1–10, 2022. Publisher: ACM New York, NY, USA. 3

[47] Chen Tessler, Yoni Kasten, Yunrong Guo, Shie Mannor, Gal Chechik, and Xue Bin Peng. CALM: Conditional Adversarial Latent Models for Directable Virtual Characters. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–9, 2023. 3

[48] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. pages 358–374. Springer, 2022. 3

[49] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 3

[50] Jonathan Tseng, Rodrigo Castellon, and C Karen Liu. EDGE: Editable Dance Generation From Music. *arXiv preprint arXiv:2211.10658*, 2022. 3

[51] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. pages 20460–20469, 2022. 3

[52] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. pages 9401–9411, 2021. 3

[53] Zhiyong Wang, Jinxiang Chai, and Shihong Xia. Combining recurrent neural networks and adversarial training for human motion synthesis and control. *IEEE transactions on visualization and computer graphics*, 27(1):14–28, 2019. Publisher: IEEE. 3

[54] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. PhysDiff: Physics-Guided Human Motion Diffusion Model. *arXiv preprint arXiv:2212.02500*, 2022. 3

[55] Haotian Zhang, Ye Yuan, Viktor Makoviychuk, Yunrong Guo, Sanja Fidler, Xue Bin Peng, and Kayvon Fatahalian. Learning Physically Simulated Tennis Skills from Broadcast Videos. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. ISBN: 0730-0301 Publisher: ACM New York, NY, USA. 3

[56] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3D environments. pages 642–651. IEEE, 2020. 3, 6

[57] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: towards controllable human-chair interactions. pages 518–535. Springer, 2022. 2, 3

[58] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. pages 3372–3382, 2021. 2, 3

[59] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. pages 6194–6204, 2020. 3

[60] Yan Zhang and Siyu Tang. The wanderings of odysseus in 3D scenes. pages 20481–20491, 2022. 2, 3, 4, 5, 7

[61] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. pages 311–327. Springer, 2022. 2, 3, 4, 6

[62] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao

Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. pages 676–694. Springer, 2022. 3

[63] Fabio Zinno. Ml tutorial day: From motion matching to motion synthesis, and all the hurdles in between. *Proc. of GDC 2019*, 2, 2019. 2