Generating Human Interaction Motions in Scenes with Text Control

Hongwei Yi[1,2], Justus Thies[2,3], Michael J. Black[2],
Xue Bin Peng[1,4], and Davis Rempe[1]

[1] NVIDIA
[2] Max Planck Institute for Intelligent Systems, Tubingen, Germany
[3] Technical University of Darmstadt
[4] Simon Fraser University

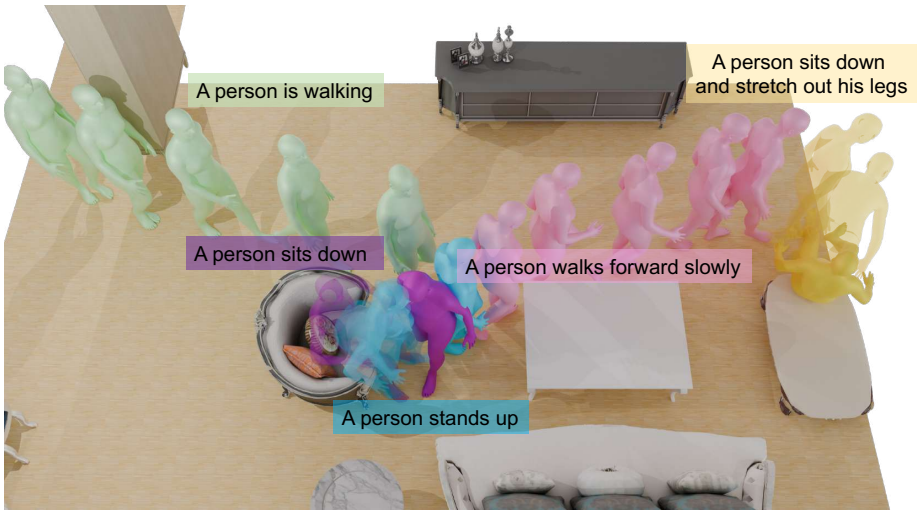

Fig. 1: We present TeSMo, a method for generating diverse and plausible human-scene interactions from text input. Given a 3D scene, TeSMo generates scene-aware motions, such as walking in free space and sitting on a chair. Our model can be easily controlled using textual descriptions, start positions, and goal positions.

Abstract. We present TeSMo, a text-controlled scene-aware motion generation method based on denoising diffusion models. Previous text-to-motion methods focus on characters in isolation without considering scenes due to the limited availability of datasets that include motion, text descriptions, and interactive scenes. Our approach begins with pre-training a scene-agnostic text-to-motion diffusion model, emphasizing goal-reaching constraints on large-scale motion-capture datasets. We then enhance this model with a scene-aware component, fine-tuned using data augmented with detailed scene information, including ground plane and object shapes. To facilitate training, we embed annotated navigation and interaction motions within scenes. The proposed method produces

realistic and diverse human-object interactions, such as navigation and sitting, in different scenes with various object shapes, orientations, initial body positions, and poses. Extensive experiments demonstrate that our approach surpasses prior techniques in terms of the plausibility of human-scene interactions and the realism and variety of the generated motions. Code and data are available at https://research.nvidia.com/labs/toronto-ai/tesmo.

Keywords: Scene-Aware Human Motion Generation · Text-to-Motion

1 Introduction

Generating realistic human movements that can interact with 3D scenes is crucial for many applications, ranging from gaming to embodied AI. For example, character animators for games and films need to author motions that successfully navigate through cluttered scenes and realistically interact with target objects, while still maintaining artistic control over the style of the movement. One natural way to control style is through text, e.g., "skip happily to the chair and sit down". Recently, diffusion models have shown remarkable capabilities in generating human motion from user inputs. Text prompts [36, 47] let users control style, while methods incorporating spatial constraints enable more fine-grained control, such as specifying desired joint positions and trajectories [18, 33, 41]. However, these works have predominantly focused on characters in isolation, without considering environmental context or object interactions.

In this work, we aim to incorporate scene-awareness into user-controllable human motion generation models. However, learning to generate motions involving scene interactions is challenging, even without text prompts. Unlike large-scale motion capture datasets that depict humans in isolation [26], datasets with paired examples of 3D human motion and scene/object geometry are limited. Prior work uses small paired datasets without text annotations to train VAEs [9, 34, 49] or diffusion models [16, 30] that generate human scene interactions with limited scope and diversity. Reinforcement learning methods are able to learn interaction motions from limited supervision [11, 22, 52], and can generate behaviors that are not present in the training motion dataset. However, designing reward functions that lead to natural movements for a diverse range of interactions is difficult and tedious.

To address these challenges, we introduce a method for Text-conditioned Scene-aware Motion generation, called TeSMo. As shown in Fig. 1, our method generates realistic motions that navigate around obstacles and interact with objects, while being conditioned on a text prompt to enable stylistic diversity. Our key idea is to combine the power of general, but scene-agnostic, text-to-motion diffusion models with paired human-scene data that captures realistic interactions. First, we pre-train a text-conditioned diffusion model [36] on a diverse motion dataset with no objects (e.g., HumanML3D [7]), allowing it to learn a realistic motion prior and the correlation with text. We then fine-tune the

model with an augmented scene-aware component that takes scene information as input, thereby refining motion outputs to be consistent with the environment.

Given a target object with which to interact and a text prompt describing the desired motion, we decompose the problem of generating a suitable motion in a scene into two components, *navigation* (e.g., approaching a chair while avoiding obstacles) and *interaction* (e.g., sitting on the chair). Both stages leverage diffusion models that are pre-trained on scene-agnostic data, then fine-tuned with an added scene-aware branch. The *navigation* model generates a pelvis trajectory that reaches a goal pose near the interaction object. During fine-tuning, the scene-aware branch takes, as input, a top-down 2D floor map of the scene and is trained on our new dataset containing locomotion sequences [26] in 3D indoor rooms [6]. The generated pelvis trajectory is then lifted to a full-body motion using motion in-painting [33]. Next, the *interaction* model generates a full-body motion conditioned on a goal pelvis pose and a detailed 3D representation of the target object. To further improve generalization to novel objects, the model is fine-tuned using augmented data that re-targets interactions [9] to a variety of object shapes while maintaining realistic human-object contacts.

Experiments demonstrate that our navigation approach outperforms prior work in terms of goal reaching and obstacle avoidance, while producing full-body motions on par with scene-agnostic diffusion models [18, 41]. Meanwhile, our interaction model generates motions with fewer object penetrations than the state-of-the-art approach [52], being preferred 71.9% of the time in a perceptual study. The central contribution of this work includes: (**1**) a novel approach to enable scene-aware and text-conditioned motion generation by fine-tuning an augmented model on top of a pre-trained text-to-motion diffusion model, (**2**) a method, TeSMo, that leverages this approach for navigation and interaction components to generate high-quality motions in a scene from text, (**3**) data augmentation strategies for placing navigation and interaction motions with text annotations realistically in scenes to enable scene-aware fine-tuning.

2 Related Work

2.1 Scene-aware Motion Generation

Motion synthesis in computer graphics has a rich history, encompassing areas such as locomotion [1, 19, 23, 50], human-scene/object interaction [21, 35], and dynamic object interaction [3,24,25]. We refer readers to an extensive survey [53] for an overview and focus on scene-aware motion generation in this section.

A particular challenge in modeling scene-aware motion is the lack of paired, high-quality human-scene datasets. One line of work [37,38] employs a two-stage method that first predicts the root path, followed by the full-body motion based on the scene and predicted path. However, these methods suffer from low-quality motion generation, attributed to the noise in the training datasets captured from monocular RGB-D videos [10]. Neural State Machine (NSM) [34] proposes the use of phase labeling [15] and local expert networks [5, 17, 45] to generate high-quality object interactions, such as sitting and carrying, after training on a small

human-object mocap dataset. Nonetheless, it struggles with recognizing walkable regions in 3D scenes, often failing to avoid obstacles. Therefore, later work in this vein requires using the A* algorithm for collision-free path planning [9]. These and related approaches [48,49] are moreover limited by the diversity of the small human-scene interaction datasets with no text annotations.

Various approaches ameliorate the data issue by creating synthetic data with captured [42, 43] or generated [20] motions placed in scenes heuristically. HUMANISE [39] does this for text-conditioned scene interactions, but rely entirely on short synthetic sequences for training, where the realism is limited by the data generation heuristics used. The reinforcement learning (RL) approach DIMOS [52] learns autoregressive policies to reach goal poses in a scene without requiring paired human-scene data for training, but still relies on A* and is constrained by the accuracy of goal pose generation [51]. RL with physical simulators [2, 11, 28, 40] has been used to produce physically plausible movements but faces challenges in generalizing across varied scenes and objects.

Unlike most prior works, our approach is text-conditioned and leverages a mix of both scene-agnostic and paired human-scene data. Pre-training is done with a diverse scene-agnostic dataset, while scene-aware fine-tuning uses motion data with scene context. For training, we adopt both synthetic data creation with real motions and data augmentation of real-world human-object interactions [9].

2.2 Diffusion-Based Motion Generation

Recently, diffusion models have demonstrated the ability to generate high-quality human motions, especially when conditioned on a text prompt [29,36,47]. In addition to text, several diffusion models add spatial controllability. Some works [33, 36] adopt image inpainting techniques to incorporate dense trajectories of spatial joint constraints into generated motions. OmniControl [41] and GMD [18] allow control with sparse signals and a pre-defined root path, respectively.

A few diffusion works handle interactions with objects or scenes. TRACE [32] generates 2D trajectories for pedestrians based on a rasterized street map. SceneDiffuser [16] conditions generation on a full scanned scene point cloud, but motion quality is limited due to noisy training data [10]. Another approach [30] tackles single-object interactions through hierarchical generation of milestone poses followed by dense motion, but it lacks text control. A concurrent line of work enables text conditioning for single-object interactions [4,27], but they focus on humans manipulating dynamic objects rather than interactions in full scenes.

We leverage a pre-trained text-to-motion diffusion model [36] and a fine-tuned scene-aware branch to enable both text controllability and scene-awareness with diffusion. We break motion generation into navigation and interaction with static objects by conditioning on 2D floor maps and 3D geometry, respectively, and create specialized human-scene data to enable diversity and quality.

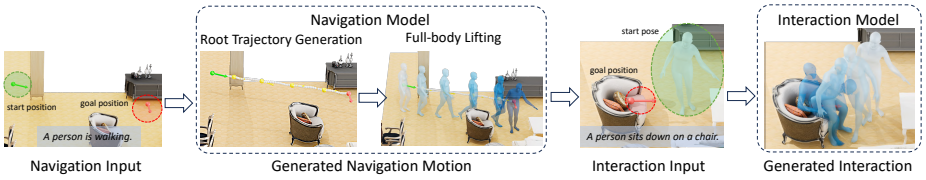

Fig. 2: Pipeline overview: given the start position (green arrow), goal position (red arrow), 3D scene, and text description, the navigation root trajectory is first generated and then the full-body motion is completed through in-painting. Subsequently, the interaction is generated from a start pose (i.e., the end pose from navigation), goal position, and the target object, enabling the generation of object-specific motion.

3 Text-Conditioned Scene-Aware Motion Generation

3.1 Overview

Given a 3D scene and a target interaction object, our goal is to generate a plausible human-scene interaction, where the motion style can be controlled by a user-specified text prompt. Our approach decomposes this task into two components, *navigation* and *interaction*, as illustrated in Fig. 2. Both components are diffusion models that leverage a fine-tuning routine to enable scene-awareness without losing user controllability, as introduced in Sec. 3.2. To interact with an object, the character must first navigate to a location in the scene near the object, which is easily calculated heuristically or specified by the user, if desired. As described in Sec. 3.3, we design a hierarchical *navigation* model, which generates a root trajectory starting from an initial location that moves to the goal location while navigating around obstacles in the scene. The generated root trajectory is then lifted into a full-body motion using in-painting techniques [33, 41]. Since the navigation model gets close to the object in the first stage, for generating the actual object *interaction*, we can focus on scenarios where the character is already near the object. This allows a one-stage motion generation model that directly predicts the full-body motion from the starting pose (*i.e.*, the last pose of navigation), a goal pelvis pose, and the object (as detailed in Section 3.4).

3.2 Background: Controllable Human Motion Diffusion Models

Motion diffusion models. Diffusion models have been successfully used to generate both top-down trajectories [32] and full-body motions [36, 47]. These models generate motions by iteratively denoising a temporal sequence of N poses (e.g., root positions or full-body joint positions/angles) $\mathbf{x} = [\mathbf{x}^1, \ldots, \mathbf{x}^N]$. During training, the model learns to reverse a forward diffusion process, which starts from a clean motion $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, sampled from the training data, and after T diffusion steps is approximately Gaussian $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Then at each step t of motion denoising, the reverse process is defined as:

$$p_\phi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c}) = \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\phi(\mathbf{x}_t, \mathbf{c}, t), \beta_t \mathbf{I}\right) \quad (1)$$

where $\mathbf{c}$ is some conditioning signal (e.g., a text prompt), and β_t depends on a pre-defined variance schedule. The denoising model $\boldsymbol{\mu}_\phi$ with parameters ϕ predicts the denoised motion $\hat{\mathbf{x}}_0$ from a noisy input motion $\mathbf{x}_t$ [13]. The model is trained by sampling a motion $\mathbf{x}_0$ from the dataset, adding random noise, and supervising the denoiser with a reconstruction loss $\|\mathbf{x}_0 - \hat{\mathbf{x}}_0\|^2$.

Augmented controllability. In the image domain, general pre-trained diffusion models are specialized for new tasks using an augmented ControlNet [46] branch, which takes in a new conditioning signal and is fine-tuned on top of the frozen base diffusion model. OmniControl [41] adopts this idea to the human motion domain. For motion diffusion models with a transformer encoder architecture, they propose an augmented transformer branch that takes in kinematic joint constraints (e.g., pelvis or other joint positions) and at each layer connects back to the base model through a linear layer that is initialized to all zeros.

As described in Secs. 3.3 and 3.4, our key insight is to use an augmented control branch to enable scene awareness. We first train a strong scene-agnostic motion diffusion model to generate realistic motion from a text prompt, and then fine-tune an augmented branch that takes scene information as input (e.g., a 2D floor map or 3D geometry). This new branch adapts generated motion to be scene-compliant, while still maintaining realism and text controllability.

Test-time guidance. At test time, diffusion models can be controlled to meet specific objectives through guidance. We directly apply guidance to the clean motion prediction from the model $\hat{\mathbf{x}}_0$ [14, 32]. At each denoising step, the predicted $\hat{\mathbf{x}}_0$ is perturbed with the gradient of an analytic objective function $\mathcal{J}$ as $\tilde{\mathbf{x}}_0 = \hat{\mathbf{x}}_0 - \alpha \nabla_{\mathbf{x}_t} \mathcal{J}(\hat{\mathbf{x}}_0)$ where α controls the strength of the guidance and $\mathbf{x}_t$ is the noisy input motion at step t. The predicted mean $\boldsymbol{\mu}_\phi$ is then calculated with the updated motion prediction $\tilde{\mathbf{x}}_0$ as in [14, 32]. As detailed later, we define guidance objectives for avoiding collisions and reaching goals.

3.3 Navigation Motion Generation

The goal of the navigation stage is for the character to reach a goal location near the target object using realistic locomotion behaviors that can be controlled by the user via text. We design a hierarchical method that first generates a dense root trajectory with a diffusion model, then leverages a powerful in-painting model [33] to generate a full-body motion for the predicted trajectory. This approach facilitates accurate goal-reaching with the root-only model, while allowing diverse text control through the in-painting model.

Root trajectory generation. Our root trajectory diffusion model, shown in Fig. 3(a), operates on motions where each pose is specified by $\mathbf{x}^n = [x, y, z, \cos\theta, \sin\theta]_n$, with (x, y, z) being the pelvis position and θ the pelvis rotation, both of which are represented in the coordinate frame of the *first* pose in the sequence. The model is conditioned on a text prompt along with starting and ending goal positions and orientations. In contrast to the representation from prior work [7],

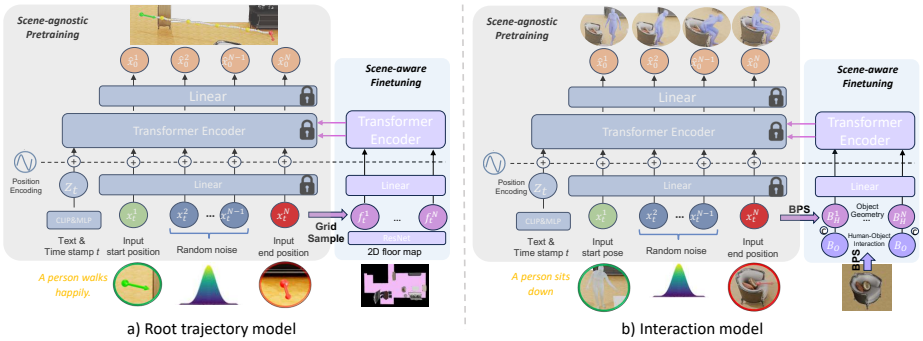

Fig. 3: Network architecture of the (**a**) root trajectory model and (**b**) interaction motion model. Initially, the base transformer encoder is trained on scene-agnostic motion data using start pose, target pose, and text as input. Subsequently, a scene-aware component is fine-tuned, which incorporates the 2D floor map (a) or 3D object (b).

which uses relative pelvis velocity and rotation, our representation using absolute coordinates facilitates constraining the outputs of the model with goal poses. Inspired by motion in-painting models [33,36], given a start pose $\mathbf{s}$ and end goal pose $\mathbf{g}$, at each denoising step, we mask out the input $\mathbf{x}_t$ such that $\mathbf{x}_t^1 = \mathbf{s}$ and $\mathbf{x}_t^N = \mathbf{g}$, thereby providing clean goal poses directly to the model. To achieve this, a binary mask $\mathbf{m} = \left[\mathbf{m}^1, \ldots, \mathbf{m}^N\right]$ with the same dimensionality as $\mathbf{x}_t$ is defined, where $\mathbf{m}^1$ and $\mathbf{m}^N$ are a vector of 1's and all other $\mathbf{m}^n$ are 0's. During training, overwriting occurs with $\tilde{\mathbf{x}}_t = \mathbf{m} * \mathbf{x}_0 + (\mathbf{1} - \mathbf{m}) * \mathbf{x}_t$ where $*$ indicates element-wise multiplication and $\mathbf{x}_0$ is a ground truth root trajectory. We then concatenate the mask with the overwritten motion $[\tilde{\mathbf{x}}_t; \mathbf{m}]$ and use this as input to the model to indicate which frames have been overwritten.

At test time, goal-reaching is improved using a guidance objective $\mathcal{J}_g = (\hat{\mathbf{x}}_0^N - \mathbf{g})^2$ that measures the error between the end pelvis position and orientation of the predicted clean trajectory $\hat{\mathbf{x}}_0^N$ and the final goal pose.

Incorporating scene representation. The model as described so far is trained on a locomotion subset of the HumanML3D dataset [7] to enable generating realistic, text-conditioned root trajectories. However, it will be entirely unaware of the given 3D scene. To take the scene into account and avoid degenerating the text-following and goal-reaching performance, we augment the base diffusion model with a control branch that takes a representation of the scene as input. This scene-aware branch is a separate transformer encoder that is fine-tuned on top of the frozen base model. As input, we extract the walkable regions from the 3D geometry of the scene and project them to a bird's-eye view, yielding a 2D floor map $\mathcal{M}$. Following [32], a Resnet-18 [12] encodes the map $\mathcal{M}$ as feature grid, and at denoising step t, each 2D projected pelvis position $(x, z) \in \boldsymbol{x}_t^n$ is queried in the feature grid $\mathcal{M}$ to get the corresponding feature $\mathbf{f}_t^n$. The resulting sequence of features $\mathbf{f}_t = \left[\mathbf{f}_t^1, \ldots, \mathbf{f}_t^N\right]$, along with the text prompt and noisy motion $\mathbf{x}_t$, become the input to the separated transformer branch.

At test time, a collision guidance objective further encourages scene compliance. This is defined as $\mathcal{J}_c = \text{SDF}(\hat{\mathbf{x}}_0, \mathcal{M})$ where SDF calculates the 2D transform distance map from the 2D floor map, then queries the 2D distance value at each time step of the root trajectory. Positive distances, indicating pelvis positions outside the walkable region, are averaged to get the final loss.

Scene-aware training and data. To train the scene-aware branch, it is important to have a dataset featuring realistic motions navigating through scenes with corresponding text prompts. For this purpose, we create the **Loco-3D-FRONT** dataset by integrating locomotion sequences from HumanML3D into diverse 3D environments from 3D-FRONT [6]. Each motion is placed within a different scene with randomized initial translation and orientation, following the methodology outlined in [43], as depicted in Fig. 4(a). Additionally, we apply left-right mirroring to both the motion and its interactive 3D scenes to augment the dataset [7]. This results in a dataset of approximately 9,500 walking motions, each motion accompanied by textual descriptions and 10 plausible 3D scenes on average, resulting in 95k locomotion-scene training pairs.

Added control with trajectory blending. Our root trajectory diffusion model generates scene-aware motions and, unlike many prior works [9,52], does not require a navigation mesh to compute A* [8] paths to follow. However, a user may want a character to take the shortest path to an object by following the A* path, or to control the general shape of the path by drawing a 2D route themselves. To enable this, we propose to fuse an input 2D trajectory $\mathbf{p} \in \mathbb{R}^{N\times 2}$ with our model's predicted clean trajectory at every denoising step. At step t, we extract the 2D (x, z) components $\hat{\mathbf{p}}_0$ from the predicted root trajectory $\hat{\mathbf{x}}_0$ and interpolate them with the input trajectory $\tilde{\mathbf{p}}_0 = s * \hat{\mathbf{p}}_0 + (1 - s) * \mathbf{p}$ where s is the blending scale that controls how closely the generated trajectory matches the input. We then overwrite the 2D components of $\hat{\mathbf{x}}_0$ with $\tilde{\mathbf{p}}_0$ and continue denoising. This blending procedure ensures outputs roughly follow the desired path but still maintain realism inherent to the trained diffusion model.

Lifting to full-body poses. To lift the generated pelvis trajectory to a full-body motion, we leverage the existing text-to-motion in-painting method PriorMDM [33], which takes a dense 2D root trajectory as input. By using this strong model that is pre-trained for text-to-motion, we can effectively generate natural and scene-aware full-body motion, while offering diverse stylistic control through text.

3.4 Object-Driven Interaction Motion Generation

After navigation, the character has reached a location near the target object and next should execute a desired interaction motion. Due to the fine-grained relationship between the body and object geometry during interactions, we propose a single diffusion model to directly generate full-body motion, unlike the two-stage navigation approach from Sec. 3.3.

Interaction motion generation. The interaction motion model operates on a sequence of full-body poses and is shown in Fig. 3(b). Our pose representation extends that of HumanML3D [7] to add the absolute pelvis position and heading $(x, y, z, \cos\theta, \sin\theta)$, similar to our navigation model. Each pose in the motion is $\mathbf{x}^n = \left[x,\ y,\ z,\ \sin\theta,\ \cos\theta,\ \dot{r}^a,\ \dot{r}^x,\ \dot{r}^z,\ r^y,\ \mathbf{j}^p,\ \mathbf{j}^v,\ \mathbf{j}^r,\ \mathbf{c}^f\right]_n \in \mathbb{R}^{268}$ with $\dot{r}^a$ root angular velocity, $(\dot{r}^x, \dot{r}^z)$ root linear velocity, r^y root height, $\mathbf{c}^f$ foot contacts, and $\mathbf{j}^p$, $\mathbf{j}^v$, $\mathbf{j}^r$ the local joint positions, velocities, and rotations, respectively.

The model is conditioned on a text prompt along with a starting full-body pose (*i.e.*, the final pose of the navigation stage) and a final goal pelvis position and orientation. The goal pelvis pose can usually be computed heuristically, but may also be provided by the user or predicted by another network [9]. The same masking procedure described in Sec. 3.3 is used to pass the start and end goals as input to the model. At test time, we also use the same goal-reaching guidance to improve the accuracy of hitting the final pelvis pose.

Object representation. The base interaction diffusion model is first trained on a dataset of interaction motions from HumanML3D and SAMP [9] without any objects, which helps develop a strong prior on interaction movements driven by text prompts. Similar to navigation, we then augment the base model with a new object-aware transformer encoder and fine-tune this encoder separately.

For the input to this branch at each denoising step t, we leverage Basis Point Sets (BPS) [31] to calculate two key features: object geometry and the human-object relationship. First, a sphere with a radius of 1.0m is defined around the object's center, and 1024 points are randomly sampled inside this sphere to form the BPS. The distance between each point in the BPS and the object's surface is then calculated, capturing the object's geometric features and stored as $\mathbf{B}_O \in \mathbb{R}^{1024}$. Next, for each body pose $\mathbf{x}_t^n$ at timestep n in the noisy input sequence, we calculate the minimum distance from each BPS point to any body joint, giving $\mathbf{B}^n \in \mathbb{R}^{1024}$. The resulting sequence of features $\mathbf{B}_H = \left[\mathbf{B}^1, \ldots, \mathbf{B}^N\right]$ represents the human-object relationship throughout the entire motion. Finally, the object and human-object interaction features are concatenated with the original pose representation at each timestep $[\mathbf{x}_t^n;\ \mathbf{B}^n;\ \mathbf{B}_O]$ and fed to an MLP to generate a merged representation, which serves as the input to the scene-aware branch.

At test time, a collision objective is used to discourage penetrations between human and object. This is very similar to the collision loss described in Sec. 3.3, but the SDF volume is computed for the 3D object and body vertices that are inside the object are penalized. Please see the supplementary material for details.

Scene-aware training and data. To train the scene-aware branch, we utilize the SAMP dataset [9], which captures motions and objects simultaneously. Specifically, we focus on "sitting" and "stand-up" interactions extracted from 80 sitting motion sequences in the SAMP dataset involving chairs of varying heights, as shown in Fig. 4(b). To diversify the object geometry, we randomly select objects from 3D-FRONT [6] to match the contact vertices on human poses in the original SAMP motion sequences. This matching is achieved using the contact loss and collision loss techniques outlined in MOVER [44].

Fig. 4: (**a**) Loco-3D-FRONT contains locomotion placed in 3D-FRONT [6] scenes without collisions. (**b**) We augment SAMP [9] by randomly selecting chairs from 3D-FRONT to match the motions and annotating a text description for each sub-sequence.

The original SAMP motions are often lengthy (∼100 sec) and lack paired textual descriptions. For instance, a "sit" motion sequence involves walking to an object, sitting down, standing up, and moving away. To effectively learn individual skills, we extract sub-sequences containing specific interactions that begin or end with a sitting pose, such as "walk then sit", "stand up then sit", "stand up from sitting", and "walk from sitting." Furthermore, we annotate textual descriptions for each sub-sequence, which often incorporate the style of sitting poses, such as "a person walks and sits down on a chair while crossing their arms." Applying left-right data augmentation to motion and objects results in approximately 200 sub-sequences for each motion sequence, each paired with corresponding text descriptions and featuring various objects.

4 Experimental Evaluation

4.1 Evaluation Data and Metrics

Navigation. Navigation performance is assessed using the test set of Loco-3D-FRONT, comprising roughly 1000 sequences. Our metrics evaluate the generated root trajectory and the full-body motion after in-painting separately. For the root trajectory, we measure goal-reaching accuracy for the 2D (horizontal xz) root **position** (m), **orientation** (rad), and root **height** (m). The **collision ratio** evaluates the consistency of root motions with the environment. For the full-body motion after in-painting, we use **FID**, **R-precision**, and **diversity** from prior work [7], along with **foot skating** ratio [18], which evaluates the physical plausibility of motion-ground interaction. Please see the Sup. Mat. for details.

Interactions. To evaluate full-body human-object interactions, we use the established test split of the SAMP dataset [9], which contains motions related to sitting. Same as navigation, we analyze goal-reaching accuracy through position, orientation, and height errors. Furthermore, we assess physical plausibility by computing average **penetration values** and **penetration ratios** between the generated motion and interaction objects. We also perform a **user study** to compare methods,. Please see more details in the Sup. Mat.

Table 1: Evaluation of navigation motion generation on the Loco-3D-FRONT test set. (**Left**) For generated pelvis trajectories, our approach achieves the best goal-reaching accuracy with low collision rate. (**Right**) After in-painting the full-body motion, our method maintains diverse and realistic motion that aligns with the given text prompt, competitive with diffusion-based scene-agnostic GMD and OmniControl.

	Root trajectory evaluation				**Full-body motion evaluation**			
	Goal-reaching error ↓							
Method	Pos.	Orient.	Height	Collision ↓	FID ↓	R-precision ↑	Diversity ↑	Foot skating ↓
Ground Truth	-	-	-	-	0.010	0.672	7.553	0.000
GMD [18]	0.374	1.231	-	-	**13.160**	0.114	4.488	0.181
OmniContol [41]	1.226	1.018	1.159	-	22.930	**0.458**	**7.128**	0.094
TRACE [32]	0.205	0.152	0.010	0.055	22.669	0.144	6.501	0.058
Ours (1-stage train)	0.197	0.132	0.013	**0.028**	22.372	0.152	6.347	0.062
Ours	**0.169**	**0.119**	**0.008**	0.031	20.465	0.376	6.415	**0.056**

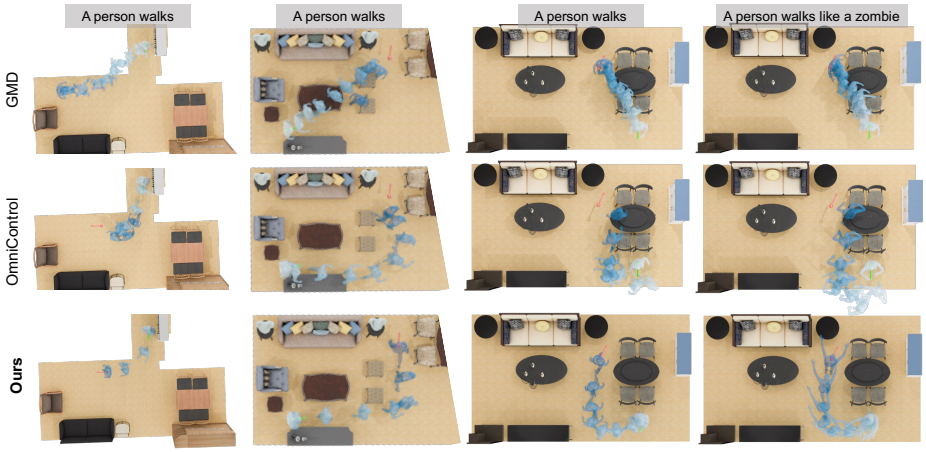

Fig. 5: Navigation generation performance. The start pose is the green arrow, and the goal pose is the red arrow. Our method more accurately reaches the goal and avoids obstacles while style is controlled by a text prompt.

4.2 Comparisons

Navigation. We conduct a comparative analysis of our method with previous scene-aware and scene-agnostic motion generation approaches in Tab. 1. Every method is conditioned on a text prompt along with start/end goal poses. The TRACE baseline and our method TeSMo also receive the 2D floor map as input.

We first compare to GMD [18] and OmniControl [41], previous scene-agnostic text-to-motion diffusion models trained on HumanML3D to follow a diverse range of kinematic motion constraints. GMD utilizes the horizontal pelvis positions (x, z) of both the start and end goals to generate a dense root trajectory and subsequently the full-body motion. OmniControl takes as input the horizontal pelvis positions (x, z) along with the height y to directly generate full-body motion in a single stage. Our navigation model achieves better goal-reaching

Table 2: Evaluation of human-object interaction motion generation on SAMP [9] sitting test set. Compared to DIMOS, our approach reaches the goal pose more accurately and exhibits fewer object penetrations, resulting in higher human preference.

Method	Goal-reaching error ↓			Object penetration ↓		User study preference ↑
	Pos.	Height	Orient.	Value	Ratio	
DIMOS [52]	0.2020	0.1283	0.4731	0.0193	0.1076	29.1%
Ours	**0.1445**	**0.0120**	**0.2410**	**0.0043**	**0.0611**	**71.9%**

accuracy, e.g., 16.9 cm for root position, since it is trained specifically for the goal-reaching locomotion task. More importantly, the right half of Tab. 1 shows that our method's full-body motion after in-painting is comparable in realism, text-following, and diversity, while achieving the best foot skating results. This demonstrates our approach adds scene-awareness to locomotion generation without compromising realism or text control.

To justify our two-branch model architecture, we adapt TRACE [32], a recent root trajectory generation model designed to take a 2D map of the environment as input. The adapted TRACE architecture is very similar to our model in Fig. 3(a), but instead of using a separate scene-aware branch, the base transformer directly takes the encoded 2D floor map features as input. This results in a single-branch architecture that must be trained from scratch, as opposed to our two-branch fine-tuning approach. Tab. 1 reveals that our method generates more plausible root trajectories with fewer collisions and more accurate goal-reaching. We also see that training our full two-branch architecture from scratch (*1-stage train* in Tab. 1), instead of using pre-training then fine-tuning, degrades both goal reaching and final full-body motion after in-painting.

A qualitative comparison of generated motions in different rooms is shown in Fig. 5. GMD tends to generate simple walking-straight trajectories. OmniControl and GMD do not reach the goal pose accurately and ignore the surroundings, leading to collisions with the environment. Our method TeSMo is able to generate diverse locomotion styles controlled by text in various scenes, achieving superior goal-reaching accuracy compared to other methods.

Interaction. Tab. 2 compares our approach to DIMOS [52], a state-of-the-art method to generate interactions trained with reinforcement learning. DIMOS requires a full-body final goal pose as input to the policy, unlike our approach which uses just the pelvis pose. Despite this, DIMOS struggles to reach the goal accurately, likely due to error accumulation during autoregressive rollout. Our method showcases fewer instances of interpenetration with interaction objects and the user study reveals a distinct preference for motions generated by our approach (preferred 71.9%) over those produced by DIMOS. Fig. 6 compares the approaches qualitatively, where we see that more accurate goal-reaching reduces floating or penetrating the chair during sitting. Moreover, the interactions generated by DIMOS lack diversity, and cannot be conditioned on text.

Table 3: Test-time guidance evaluation. Adding guidance to reach goal poses and avoid collisions during inference improves performance. Lower is better for all metrics.

Guidance		Navigation		Interaction		
Goal Reach	Collision	Goal Pos.	Collision	Goal Pos.	Pen. Val.	Pen. Ratio
✗	✗	0.1568	0.0294	0.1445	0.0043	0.0611
✓	✗	**0.118**	0.0342	0.1453	0.0050	0.0554
✗	✓	0.1550	0.0013	0.1407	**0.0040**	**0.0414**
✓	✓	0.1241	**0.0012**	**0.1404**	0.0045	0.0494

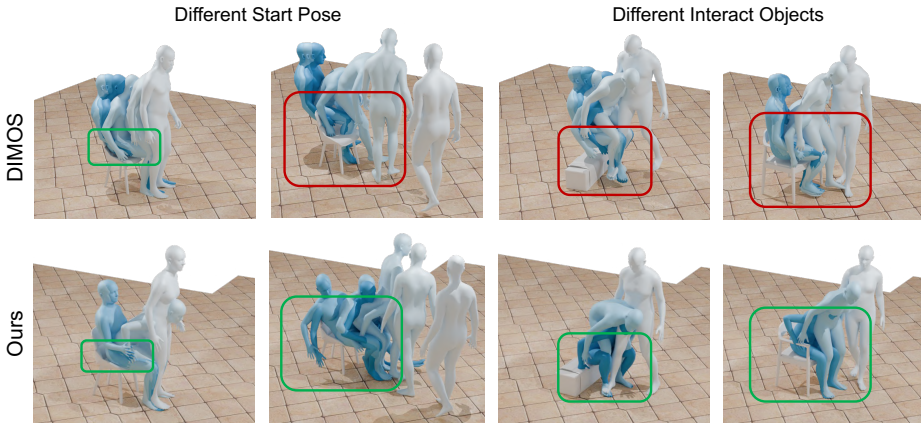

Fig. 6: Compared with DIMOS [52], our method generates more realistic human-object interactions with reduced floating and interpenetrations.

4.3 Analysis of Capabilities

In Fig. 1, our method carries out a sequence of actions, enabling traversal and interaction with multiple objects within a scene. Fig. 7 demonstrates additional key capabilities. In the top section, our method is controlled through a variety of text prompts. For interactions in particular, diverse text descriptions disambiguate between actions like sitting or standing up, and allow stylizing the sitting motion, e.g., with crossed arms. In the middle section, we enable user control over trajectories by adhering to a predefined A* path. By adjusting the blend scale, users can adjust how closely the generated trajectory follows A*. At the bottom of Fig. 7, we harness guidance at test time to encourage motions to reach the goal while avoiding collisions and penetrations. As shown in Tab. 3, combining guidance losses gives improved results both for navigation and interactions.

5 Discussion & Future Work

We introduced TeSMo, a novel method for text-controlled scene-aware motion generation. While our navigation model ensures accurate goal-reaching and text-to-motion control, the two-stage process can disconnect the generated pelvis

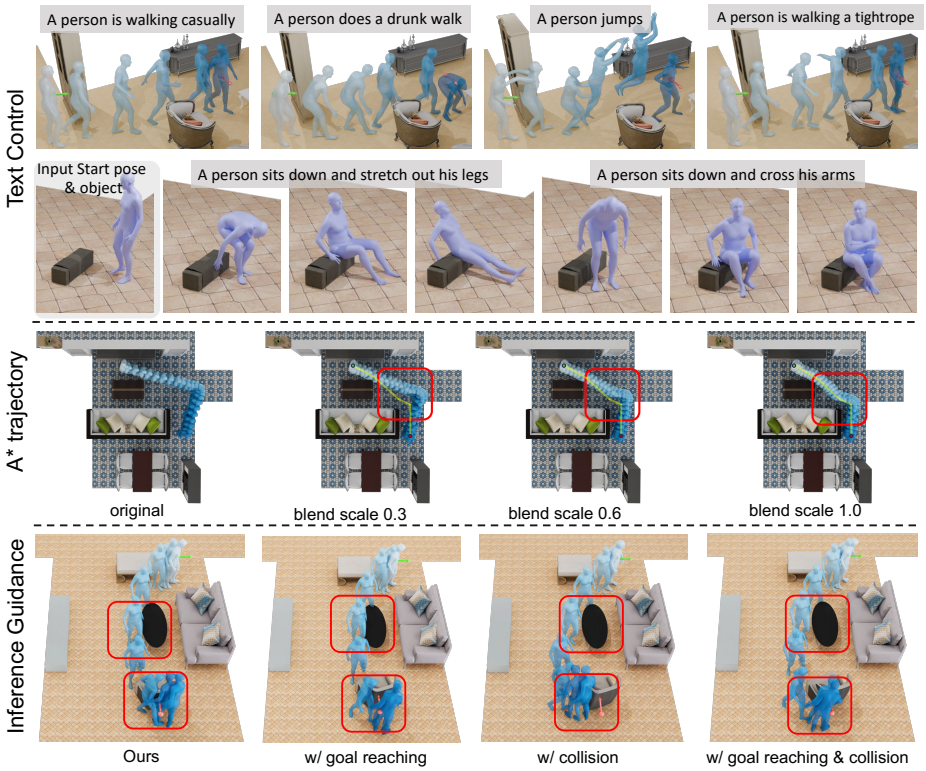

Fig. 7: TeSMo capabilities. (**Top**) Diverse text control; (**Middle**) Following A* path with adherence controlled by the blend scale; (**Bottom**) Test-time guidance encourages locomotion to reach the goal accurately without colliding with the environment.

trajectory from the in-painted full-body poses. Exploring one-stage models for simultaneous pelvis and pose generation could streamline the process. Additionally, our 2D floor map approach limits handling intricate interactions, like stepping over a small stool.

Our approach aims at controllability, letting users specify text prompts or goal objects and locations. It may integrate with recent pipelines [40] that employ LLM planners to specify a sequence of actions and contact information that could be used to guide our motion generation. Looking ahead, we aim to broaden the spectrum of actions modeled by the system, to encompass activities such as lying down and touching. Furthermore, enabling interactions with dynamic objects will allow for more interactive and realistic scenarios.

Acknowledgments. Thanks Y. Huang and Y. Liu for technical support. Thanks M. Petrovich and N. Athanasiou for the fruitful discussion about text-to-motion synthesis. Thanks T. Niewiadomski, T. McConnell, and T. Alexiadis for running the user study.

Disclosure. https://files.is.tue.mpg.de/black/CoI_ECCV_2024.txt

References

1. Agrawal, S., van de Panne, M.: Task-based locomotion. ACM Transactions on Graphics **35**(4), 1–11 (Jul 2016). https://doi.org/10.1145/2897824.2925893, http://dx.doi.org/10.1145/2897824.2925893 3
2. Chao, Y.W., Yang, J., Chen, W., Deng, J.: Learning to sit: Synthesizing human-chair interactions via hierarchical control. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 5887–5895 (2021) 4
3. Corona, E., Pumarola, A., Alenyà, G., Moreno-Noguer, F.: Context-aware human motion prediction. Cornell University - arXiv,Cornell University - arXiv (Apr 2019) 3
4. Diller, C., Dai, A.: Cg-hoi: Contact-guided 3d human-object interaction generation (2024) 4
5. Eigen, D., Ranzato, M., Sutskever, I.: Learning factored representations in a deep mixture of experts. arXiv: Learning,arXiv: Learning (Dec 2013) 3
6. Fu, H., Cai, B., Gao, L., Zhang, L.X., Wang, J., Li, C., Zeng, Q., Sun, C., Jia, R., Zhao, B., et al.: 3d-front: 3d furnished rooms with layouts and semantics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10933–10942 (2021) 3, 8, 9, 10
7. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (June 2022) 2, 6, 7, 8, 9, 10
8. Hart, P.E., Nilsson, N.J., Raphael, B.: A formal basis for the heuristic determination of minimum cost paths. IEEE transactions on Systems Science and Cybernetics **4**(2), 100–107 (1968) 8
9. Hassan, M., Ceylan, D., Villegas, R., Saito, J., Yang, J., Zhou, Y., Black, M.J.: Stochastic scene-aware motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11374–11384 (October 2021) 2, 3, 4, 8, 9, 10, 12
10. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3d human pose ambiguities with 3d scene constraints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019) 3, 4
11. Hassan, M., Guo, Y., Wang, T., Black, M., Fidler, S., Peng, X.B.: Synthesizing physical character-scene interactions. In: ACM SIGGRAPH 2023 Conference Proceedings. SIGGRAPH '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3588432.3591525, https://doi.org/10.1145/3588432.3591525 2, 4
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015) 7
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf 6
14. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv preprint (2022) 6
15. Holden, D., Komura, T., Saito, J.: Phase-functioned neural networks for character control. ACM Transactions on Graphics p. 1–13 (Aug 2017). https://doi.org/10.1145/3072959.3073663, http://dx.doi.org/10.1145/3072959.3073663 3

16. Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., Zhu, S.C.: Diffusion-based Generation, Optimization, and Planning in 3D Scenes. arXiv e-prints arXiv:2301.06015 (Jan 2023). https://doi.org/10.48550/arXiv.2301.06015 2, 4
17. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural Computation p. 79–87 (Feb 1991). https://doi.org/10.1162/neco.1991.3.1.79, http://dx.doi.org/10.1162/neco.1991.3.1.79 3
18. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2151–2162 (2023) 2, 3, 4, 10, 11
19. Kovar, L., Gleicher, M., Pighin, F.: Motion graphs. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2. pp. 723–732 (2023) 3
20. Kulkarni, N., Rempe, D., Genova, K., Kundu, A., Johnson, J., Fouhey, D., Guibas, L.: Nifty: Neural object interaction fields for guided human motion synthesis (2023) 4
21. Lee, J., Chai, J., Reitsma, P.S.A., Hodgins, J.K., Pollard, N.S.: Interactive control of avatars animated with human motion data. In: Proceedings of the 29th annual conference on Computer graphics and interactive techniques (Jul 2002). https://doi.org/10.1145/566570.566607, http://dx.doi.org/10.1145/566570.566607 3
22. Lee, J., Joo, H.: Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments. In: International Conference on Computer Vision (ICCV) (2023) 2
23. Lee, K.H., Choi, M.G., Lee, J.: Motion patches. ACM Transactions on Graphics p. 898–906 (Jul 2006). https://doi.org/10.1145/1141911.1141972, http://dx.doi.org/10.1145/1141911.1141972 3
24. Li, J., Clegg, A., Mottaghi, R., Wu, J., Puig, X., Liu, C.K.: Controllable human-object interaction synthesis. arXiv preprint arXiv:2312.03913 (2023) 3
25. Li, J., Wu, J., Liu, C.K.: Object motion guided human motion synthesis. ACM Transactions on Graphics (TOG) **42**(6), 1–11 (2023) 3
26. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision. pp. 5442–5451 (Oct 2019) 2, 3
27. Peng, X., Xie, Y., Wu, Z., Jampani, V., Sun, D., Jiang, H.: Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. arXiv preprint arXiv:2312.06553 (2023) 4
28. Peng, X.B., Guo, Y., Halper, L., Levine, S., Fidler, S.: Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. ACM Trans. Graph. **41**(4) (Jul 2022) 4
29. Petrovich, M., Litany, O., Iqbal, U., Black, M.J., Varol, G., Peng, X.B., Rempe, D.: STMC: Multi-track timeline control for text-driven 3d human motion generation. arXiv:2401.08559 (2024) 4
30. Pi, H., Peng, S., Yang, M., Zhou, X., Bao, H.: Hierarchical generation of human-object interactions with diffusion probabilistic models. In: ICCV. pp. 15061–15073 (October 2023) 2, 4
31. Prokudin, S., Lassner, C., Romero, J.: Efficient learning on point clouds with basis point sets. In: International Conference on Computer Vision (ICCV). pp. 4332–4341 (2019) 9

32. Rempe, D., Luo, Z., Peng, X.B., Yuan, Y., Kitani, K., Kreis, K., Fidler, S., Litany, O.: Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In: CVPR (2023) 4, 5, 6, 7, 11, 12
33. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. arXiv preprint arXiv:2303.01418 (2023) 2, 3, 4, 5, 6, 7, 8
34. Starke, S., Zhang, H., Komura, T., Saito, J.: Neural state machine for character-scene interactions. ACM Trans. Graph. **38**(6) (nov 2019). https://doi.org/10.1145/3355089.3356505, https://doi.org/10.1145/3355089.3356505 2, 3
35. Taheri, O., Choutas, V., Black, M.J., Tzionas, D.: GOAL: Generating 4D whole-body motion for hand-object grasping. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2022), https://goal.is.tue.mpg.de 3
36. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Bermano, A.H., Cohen-Or, D.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022) 2, 4, 5, 7
37. Wang, J., Xu, H., Xu, J., Liu, S., Wang, X.: Synthesizing long-term 3d human motion and interaction in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9401–9411 (2021) 3
38. Wang, J., Yan, S., Dai, B., Lin, D.: Scene-aware generative network for human motion synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12206–12215 (2021) 3
39. Wang, Z., Chen, Y., Liu, T., Zhu, Y., Liang, W., Huang, S.: Humanise: Language-conditioned human motion generation in 3d scenes. In: Advances in Neural Information Processing Systems (NeurIPS) (2022) 4
40. Xiao, Z., Wang, T., Wang, J., Cao, J., Dai, B., Lin, D., Pang, J.: Unified human-scene interaction via prompted chain-of-contacts. In: International Conference on Learning Representations (ICLR) (2024) 4, 14
41. Xie, Y., Jampani, V., Zhong, L., Sun, D., Jiang, H.: Omnicontrol: Control any joint at any time for human motion generation. arXiv preprint arXiv:2310.08580 (2023) 2, 3, 4, 5, 6, 11
42. Ye, S., Wang, Y., Li, J., Park, D., Liu, C.K., Xu, H., Wu, J.: Scene synthesis from human motion. In: SIGGRAPH Asia 2022 Conference Papers. SA '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3550469.3555426, https://doi.org/10.1145/3550469.3555426 4
43. Yi, H., Huang, C.H.P., Tripathi, S., Hering, L., Thies, J., Black, M.J.: MIME: Human-aware 3D scene generation. In: Computer Vision and Pattern Recognition (CVPR) (June 2023) 4, 8
44. Yi, H., Huang, C.H.P., Tzionas, D., Kocabas, M., Hassan, M., Tang, S., Thies, J., Black, M.J.: Human-aware object placement for visual environment reconstruction. In: Computer Vision and Pattern Recognition (CVPR). pp. 3959–3970 (Jun 2022) 9
45. Yuksel, S.E., Wilson, J.N., Gader, P.D.: Twenty years of mixture of experts. IEEE Transactions on Neural Networks and Learning Systems p. 1177–1193 (Aug 2012). https://doi.org/10.1109/tnnls.2012.2200299, http://dx.doi.org/10.1109/tnnls.2012.2200299 3
46. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023) 6
47. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022) 2, 4, 5

48. Zhang, W., Dabral, R., Leimkühler, T., Golyanik, V., Habermann, M., Theobalt, C.: Roam: Robust and object-aware motion generation using neural pose descriptors. International Conference on 3D Vision (3DV) (2024) 4
49. Zhang, X., Bhatnagar, B.L., Starke, S., Guzov, V., Pons-Moll, G.: Couch: Towards controllable human-chair interactions. In: European Conference on Computer Vision (ECCV). Springer (October 2022) 2, 4
50. Zhang, Y., Tang, S.: The wanderings of odysseus in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20481–20491 (2022) 3
51. Zhao, K., Wang, S., Zhang, Y., Beeler, T., , Tang, S.: Compositional human-scene interaction synthesis with semantic control. In: European conference on computer vision (ECCV) (2022) 4
52. Zhao, K., Zhang, Y., Wang, S., Beeler, T., , Tang, S.: Synthesizing diverse human motions in 3d indoor scenes. In: International conference on computer vision (ICCV) (2023) 2, 3, 4, 8, 12, 13
53. Zhu, W., Ma, X., Ro, D., Ci, H., Zhang, J., Shi, J., Gao, F., Tian, Q., Wang, Y.: Human motion generation: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 3