SCENIC: Scene-aware Semantic Navigation with Instruction-guided Control

Xiaohan Zhang[1,2], Sebastian Starke[3], Vladimir Guzov[1,2], Zhensong Zhang[4],
Eduardo Pérez-Pellitero[4], Gerard Pons-Moll[1,2]

[1]Tübingen AI Center, University of Tübingen
[2]Max Planck Institute for Informatics, Saarland Informatics Campus
[3]Meta Reality Labs Research
[4]Huawei Noah's Ark Lab

Figure 1: SCENIC is a text-conditioned scene interaction model. It adapts to complex scenes with varying terrains and also supports user-specified semantic control with natural language. Given a 3D scene, our model takes as cues of user-specified trajectory as sub-goals, and text. *We encourage the readers to watch the supplementary video.*

Abstract

Synthesizing natural human motion that adapts to complex environments while allowing creative control remains a fundamental challenge in motion synthesis. Existing models often fall short, either by assuming flat terrain or lacking the ability to control motion semantics through text. To address these limitations, we introduce SCENIC, a diffusion model designed to generate human motion that adapts to dynamic terrains within virtual scenes while enabling semantic control through natural language. The key technical challenge lies in simultaneously reasoning about complex scene geometry while maintaining text control. This requires understanding both high-level navigation goals and fine-grained environmental constraints. The model must ensure physical plausibility and precise navigation across varied terrain, while also preserving user-specified text control, such as "carefully stepping over obstacles" or "walking upstairs like a zombie." Our solution introduces a hierarchical scene reasoning approach. At the core of our method is a novel hierarchical scene reasoning framework. It combines two key components: a motion-scene cross-attention block that aligns the human body's motion features with local scene geometry, enabling precise low-level interactions; and a target point canonicalization module that provides global goal conditioning by normalizing target scene coordinates for high-level guidance. To ensure plausibility and naturalness, we leverage a pre-trained motion diffusion prior and apply scene-constrained diffusion noise optimization during sampling, enabling long-horizon motion generation that respects both scene structure and semantic text input. Experiments demonstrate that our novel diffusion model generates arbitrarily long human motions that both adapt to complex scenes with varying terrain surfaces and respond to textual prompts. Additionally, we show SCENIC can generalize to four real-scene datasets.

1. Introduction

Humans navigate complex environments effortlessly, adapting to varied terrains while performing diverse motions. This fundamental ability to synthesize natural human motion in complex environments [28, 29, 52, 93] is crucial

for numerous applications ranging from gaming to embodied agents. For instance, how can we make virtual characters seamlessly "step over obstacles before sitting" or "walk upstairs like a zombie" (Figure 1). Fundamentally, this requires both scene understanding and semantic control. While recent works have made progress in either text-controlled human motion synthesis [62, 71, 77] or motion adaptation to simplified environments [38, 83], they struggle with complex scenarios. Even methods that can adapt to uneven terrain [25, 48, 60] lack flexible semantic control through natural language. This work bridges this gap by introducing a unified diffusion-based framework that simultaneously handles complex scene geometry and text-based semantic control.

Synthesizing scene-aware semantic motion faces three fundamental challenges. First, the model must generate motion that precisely adapts to complex environment constraints, avoiding penetration, while maintaining natural contact with uneven surfaces, and reaching specific targets. Furthermore, unlike previous approaches that handle either scene geometry or semantic control in isolation, combining both requires sophisticated reasoning about how different motion styles interact with varied geometric features. Last, traditional approaches require extensive paired motion-scene data, which is expensive to acquire due to tracking difficulties and does not scale.

Our key insight is that complex scene-aware motion synthesis can be decomposed into hierarchical reasoning levels, akin to how humans plan and execute navigation tasks. At a high level, we condition motion in the canonicalized of the target scene point, allowing the model to learn goal-directed behaviors while naturally satisfying global constraints such as target elevation or location. At a finer level, we represent detailed scene geometry using a human-centric distance field [48, 60], capturing local affordances around the body. To couple scene perception with motion, we introduce a cross-attention mechanism in the latent space, aligning motion features with spatially structured scene representations. To further enhance realism, we refine motion in the diffusion noise space, ensuring postures conform to fine-grained geometry without sacrificing naturalness, by leveraging the pre-trained motion diffusion prior [30]. Finally, to ensure data efficiency and generalization, we exploit the compositional nature of human motion. Our model is trained on short motion segments [28, 29, 52], which are automatically augmented across a diverse set of terrain surfaces. This enables efficient learning while supporting the synthesis of long-horizon navigation behaviors in complex 3D environments.

With these components, we present the first text-conditioned diffusion model that generates long-term human motion adapted to complex terrain. Experiments demonstrate that SCENIC generalizes to various geometric structures—including uneven terrain, steps, staircases and slopes. We further show that our model supports precise scene adaptation across four real-world 3D scene datasets: Replica [66], Matterport3D [7], HPS [20], and LaserHuman [12] (see Figure 1). Moreover, SCENIC supports seamless transition between ten distinct motion semantics including "crouching", "climbing", "hopping", "jumping", and "balancing", and can adapt to complicated instructions such as "walking up stairs like a zombie". Empirically, our model achieves the best in terms of satisfying the scene and goal constraints, and motion quality. Qualitatively, our model is preferred by *75.6%* of the participants over state-of-the-art alternatives (see Table 1).

The key contributions of our work include:

1. A unified model combining text conditioning with scene-aware motion synthesis, capable of traversing diverse 3D geometries such as terrain, stairs, steps, and slopes.
2. A hierarchical diffusion framework that enables structured scene reasoning via target-based conditioning and local geometry alignment, validated on four real-world datasets.
3. A scene-aware diffusion noise optimization scheme, refining samples with differentiable constraints while preserving naturalness from the pretrained motion prior.

2. Related Work

2.1. Text-guided Motion Diffusion.

Recent years have seen remarkable progress in human motion synthesis, driven by the emergence of diffusion models [10, 14, 23, 34, 47, 49, 62, 71, 87, 88, 92, 96] and comprehensive motion capture datasets like AMASS [50]. The integration of action labels and language descriptions through datasets such as BABEL [59] and HumanML3D [19] has enabled increasingly sophisticated control over generated motions. Recent work has explored various aspects of motion synthesis, including two-person interactions [17, 42, 43, 69], joint-level control [31, 72, 77], and style editing [9, 24].

Motion editing through text has evolved along two main paths: in-motion editing for specific body parts [8, 26, 32] and segment-level editing using text prompts. In particular, FlowMDM [2] demonstrated impressive results in seamless transitions between local motion segments. STMC [55] proposed a hybrid method for spatial and temporal motion composition using pre-trained motion models. UniMotion [37] leveraged per-frame and sequence-level text to enhance motion understanding and control.

While these approaches have advanced the field significantly, they typically assume simplified environments with uniform height and flat terrain. Our work extends these capabilities by incorporating complex scene geometry while

maintaining text-based semantic control.

2.2. Scene-aware Motion Synthesis.

Scene-aware motion synthesis is a comprehensive field that can be broadly classified into two categories: object interaction and scene navigation. Research on human-object interaction [3, 33, 84, 86] spans a wide range, from interactions with large, static objects like chairs and beds [21, 27, 35, 53, 56, 65, 83, 89, 90], to dynamic engagements with moving objects. This includes studies that focus on contact-based object interactions without navigation [15, 54, 75, 78, 79, 81], as well as those that incorporate navigation [38–40, 91]. A parallel line of research leverages reinforcement learning to synthesize interactions [13, 22, 51]. Other studies have concentrated on full body grasps [1, 16, 41, 67, 67, 70] and dexterous hand manipulation [4, 5, 11, 44, 68, 85].

In the context of human-scene interactions, a significant portion of the work is dedicated to generating short-term motion within 3D scenes [6, 73, 74]. PFNN [25] introduced a real-time motion controller that adapts to uneven terrain but requires carefully annotated phase labels and does not enable text-based motion style editing. Some models generate longer-term human motion but often require a full-body target pose as a control signal [45, 93]. Others assume uniform height within the scenes [29, 36, 52]. Using reinforcement learning, [48, 60] propose policies for terrain traversal, however, the motion is not human-like due to the animation of the physical character. Moreover, their synthesis only perform on synthetic terrains with limited complexity.

More recent work incorporates text control into human-scene interaction. TeSMO [83] proposed a two-stage method for collision-free navigation within the scene. TRUMANS [29] unified static and dynamic object interactions, and a recent extension replaced action labels with more versatile text prompts [28], achieving impressive results. However, these models still assume flat terrains or floors. While some concurrent works have demonstrated human motion on stairs [12, 94], they have their limitation of not training on paired motion-scene data. This lack of scene awareness restricts the model's ability to generalize to complex terrain surfaces. Moreover, their approach requires the future 3D root position, which is not always available. On the other hand, Cong et al. [12] did not enable control with the goal location, limiting its controllability and the length of plausible motion sequences it can generate.

Our work addresses these limitations by introducing the first scene-aware motion synthesis model that can adapt to the terrain and is controllable with text-based semantic signals. Our versatile model synthesizes realistic human motion across diverse 3D environments while allowing semantic control over motion style.

3. Method

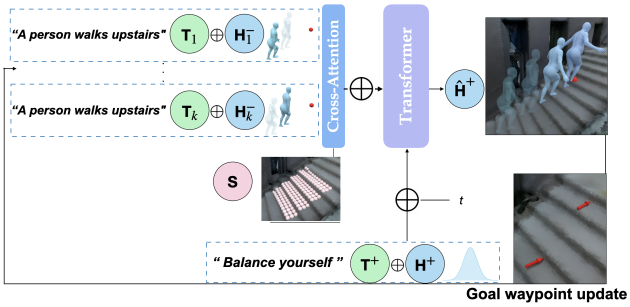

Figure 2. SCENIC has a 3D scene, a set of trajectory waypoints, and text prompts, and the past human motion as inputs. The past human motion and the scene encoding first undergo goal-centric canonicalization. The diffusion-based transformer then encodes the aligned text-motion tokens, scene tokens and a timestamp token to predict the canonicalized future human motion.

Our proposed diffusion model generates arbitrarily long human motions that adapt to complex terrains while allowing semantic control through text prompts. The key insight is decomposing the complex task into hierarchical reasoning levels: high-level movement planning relative to the target scene point and fine-grained scene adaptation through local geometry reasoning.

3.1. Problem Formulation

As illustrated in Figure 2, given a 3D scene, a user-defined trajectory consisting of sub-goals $\{\mathbf{G}_j\}_{j=1}^{M}$, and text prompts $\mathbf{T}$, our model is designed to fulfill both the environmental and textual constraints. It synthesizes motion $\mathbf{H}$ that reaches the goals, adapts to complex scene surfaces, and avoids penetration. Moreover, our motion style can be controlled by user-specified text instructions.

3.2. Data Representations

To synthesize scene-aware semantic motion, our method takes four key representations:

Human Motion H Unlike previous motion representation of human motion [19, 29, 71], which requires an additional fitting process to obtain the final animated mesh, our representation can be animated directly. The SMPL model [46] is used to parameterize our human motion. Our motion human $\mathbf{H}$ consists of $N = 40$ frames of body pose in the 6-D continuous form [97] $\boldsymbol{J}_r \in \mathbb{R}^{N\times 22\times 6}$, and the global root location $\mathbf{J}_{\text{root}} \in \mathbb{R}^{N\times 3}$. The binary foot contact for the heel and toe joints $\boldsymbol{c} \in \mathbb{R}^{N\times 4}$ are also included.

Scene embedding S The scene is encoded by a distance field $\mathbf{S} \in \mathbb{R}^{N\times H\times H}$ centered at the human root joint and its orientation is relative to the Y-rotation of the root. This local

representation enables efficient processing of relevant terrain features while maintaining translation invariance. The embedding is sampled by projecting from the point grid perpendicularly toward the scene. Previous approaches adopt an occupancy representation by encoding the scene with binary values [6, 29, 45, 65]. Instead, our embedding is more efficient and informative for the character to adapt to the terrain. Empirically, we use $H \times H$=144 points that are uniformly sampled from a 1.2×1.2 meter grid.

Goal Representation Each sub-goal $\mathbf{G}_j$ is represented by a target 3D position to be reached on the scene $\boldsymbol{g}_p^j \in \mathbb{R}^3$, and a 2D desired orientation vector represented by $\boldsymbol{g}_r^j \in \mathbb{R}^2$.

Text Control T Unlike previous methods that use a single text embedding combined with a timestamp [62, 63, 71], we employ a different approach. We encode the text on a per-frame basis and treat each frame's text as an individual token within the diffusion transformer. This method of temporal tokenization ensures a precise alignment between the motion and the corresponding text [37], facilitating a seamless transition between different motion styles. The text prompt $\mathbf{T} \in \mathbb{R}^{N \times D}$ is obtained by reducing the dimensionality of the CLIP embeddings using PCA. In our experiments, the CLIP embedding is reduced to $D = 64$ dimensions.

3.3. Hierarchical Scene Reasoning

A key design component of our model is a hierarchical reasoning framework that combines goal-centric canonicalization with cross-modal scene fusion. This allows the model to synthesize long-horizon motion that is both goal-directed and scene-compliant, particularly in challenging environments such as slopes, steps, or cluttered terrains.

Target Scene Point Canonicalization. To simplify learning and enable robust goal-reaching, we transform both the human motion and the local scene context into a canonical coordinate system defined by the current goal $\mathbf{G}_j$. This transformation achieves two critical outcomes: (1) it normalizes diverse spatial configurations into a consistent reference frame, and (2) it encourages the model to focus on relative geometry rather than absolute world positions, which improves generalization across scenes.

Specifically, given a goal $\mathbf{G}_j$, we canonicalize the human motion as: $\mathbf{H}_{\text{cano}} = \mathcal{T}_{\text{human}}(\mathbf{H}, \mathbf{G}_j)$ and the corresponding scene representation as: $\mathbf{S}_{\text{cano}} = \mathcal{T}_{\text{scene}}(\mathbf{S}, \mathbf{G}_j)$. Here, $\mathcal{T}_{\text{human}}$ and $\mathcal{T}_{\text{scene}}$ perform spatial transformations that align both modalities into the goal-centric frame. Rather than explicitly conditioning on the goal vector, which prior work [28, 29, 83] has shown to degrade performance in complex terrain, we instead train the model to predict motion that converges toward the origin in the canonical frame.

Motion-Scene Cross-Attention. To further enhance scene understanding, we introduce a motion-scene cross-attention mechanism that fuses motion features with fine-grained spatial scene context. Operating in the latent space of the diffusion model, this module aligns the canonicalized motion $\mathbf{H}_{\text{cano}}$ with structured scene features $\mathbf{S}_{\text{cano}}$, allowing the model to reason about physical constraints such as ground elevation or obstacles. The resulting fused representation is then passed to downstream self-attention layers, where it interacts with the text prompt embedding. This staged attention mechanism allows the model to first resolve spatial alignment between the body and the scene, before integrating semantic intent.

3.4. Autoregressive Motion Diffusion

The synthesis process seamlessly connects multiple motion segments through an autoregression. As shown in Figure 2, each segment is predicted using the previous one, maintaining continuity while adapting to new goals and terrain features. The model synthesizes scene-aware motion towards the current sub-goal $\mathbf{G}_j$. Once the sub-goal is reached, the goal iterates to $\mathbf{G}_{j+1}$. This way, the model can progressively synthesize arbitrarily long motions that are plausible to the scene. Such an approach not only enables the length of the animation to become unconstrained, but also allows users to control the motion trajectory to avoid obstacles.

Conditional Diffusion Model Each motion segment is generated through a conditional diffusion process, which incorporates a transformer architecture, as depicted in Figure 2. The generation of successive segments is facilitated by using the last k frames of the preceding segment as a seed motion, which then extends to the next segment. We denote the canonicalized motion segment $\mathbf{H}_{\text{cano}}$ defined in Sec 3.3 as a combination of the k frames of seed motion $\mathbf{H}^-$, and the $N-k$ frames of predicted motion $\mathbf{H}^+$. The diffusion process is conditioned on several factors: the scene embeddings $\mathbf{S}$, the text prompt $\mathbf{T}$, and the past seed motion, $\mathbf{H}^-$. Together, these are represented as the condition, $\mathbf{C} = (\mathbf{S}, \mathbf{T}, \mathbf{H}^-)$. In our experiments, we set the values of N and k to 40 and 10, respectively. During the training phase, noise is injected into the future motion, $\mathbf{H}^+$, while the seed motion, $\mathbf{H}^-$, remains unchanged. At each denoising step n, the model learns to reverse the forward diffusion process, with the reverse process defined as

$$p(\mathbf{H}_{n-1}^+ | \mathbf{H}_n^+, \mathbf{C}) := \mathcal{N}(\mathbf{H}_{n-1}^+; \mu(\mathbf{H}_n^+, \mathbf{C}), \Sigma_n), \quad (1)$$

where μ denotes the predicted mean and Σ_n is a fixed variance. Learning the mean can be re-parameterized as learning to predict the clean future motion $\mathbf{H}_0^+$. During training, we also apply an l_2 loss on the predicted joint positions obtained via forward kinematics:

$$\mathcal{L} = \mathbb{E}_{\mathbf{H}_0^+} \|\hat{\mathbf{H}}_0^+ - \mathbf{H}_0^+\|_2 + \lambda \cdot \|\hat{\boldsymbol{J}}_p^{\,+} - \boldsymbol{J}_p^+\|_2. \quad (2)$$

This is crucial for the sharpness of the motion. Here, $\hat{\mathbf{H}}_0^+$ denotes the predicted future motion, while $\hat{\boldsymbol{J}}_p^{\,+}$ denotes the

predicted future joint positions obtained via forward kinematics. The positional loss weight λ is set to be 4.

3.5. Object Interaction

When the human arrives in the vicinity of the target object after the navigation, our method generates full-body motion by interacting with the objects to perform text-controlled sitting and lying. Instead of focusing on the goal and the neighboring scene, the interaction model needs to be aware of the target object geometry. For this reason, we introduce another diffusion model conditioned on an object geometric representation $\mathbf{O} \in \mathbb{R}^{2048}$. The representation comprises the distances from the basis point set (BPS) [58] to the object surface, as well as the distance from the hands and the hip joints to each one of the object voxels. The BPS consists of 512 points uniformly sampled from a sphere of radius 1 meter, centered around the normalized object center. The object is voxelized into an $8 \times 8 \times 8$ grid, and we zero out the distance features for unoccupied voxels. The interaction model employs the same representation for human motion and texts. We train our interaction model on the SAMP [21] dataset. The interaction diffusion model is trained using the same learning objective as the navigation model.

3.6. Scene-Aware Diffusion Noise Optimization

At test time, we apply Diffusion Noise Optimization (DNO) [30, 57, 61, 95] to enforce physics and scene constraints without retraining the model. While a common approach is to apply diffusion guidance [15, 38, 76, 83] which updates the clean prediction at each timestep using gradients of an external objective—this technique has several limitations. Specifically, guidance operates directly in motion space, where small perturbations can easily push samples off the learned motion manifold, often resulting in unrealistic or jerky motions.

To overcome this, we adopt DNO, which instead optimizes the initial Gaussian noise sample $x_T \sim \mathcal{N}(0, I)$, preserving the model's learned motion prior. The final motion is generated by passing the optimized noise through the pretrained diffusion model via an ODE-based DDIM sampler [64]. This setup enables gradients to propagate back through all denoising steps, allowing the noise to be updated to better satisfy scene-level objectives—while staying on the learned manifold of realistic human motion.

Formally, we solve:

$$x_T^* = \arg\min_{x_T} \left[\mathcal{L}(\mathrm{ODE}(G, x_T)) + \mathcal{R}(x_T) \right], \qquad (3)$$

where $\mathcal{L}$ encodes our scene-specific objectives, and $\mathcal{R}(x_T)$ is a regularization term that encourages the noise sample to remain close to the standard normal distribution. G denotes the diffusion model, and ODE represents the DDIM sampler used to produce the final motion.

Scene-Aware Navigation Objectives To promote physically plausible foot contact and prevent ground penetration, we define a contact-aware physics loss:

$$\mathcal{L}_{\text{phys}} = \boldsymbol{c} \cdot \|\mathbf{J}_{\text{feet}} - \mathbf{h}\|_2 + (1 - \boldsymbol{c}) \cdot \mathbb{1}(\mathbf{h} > \mathbf{J}_{\text{feet}}) \cdot \|\mathbf{J}_{\text{feet}} - \mathbf{h}\|_2, \qquad (4)$$

where $\boldsymbol{c}$ is the predicted foot contact label, $\mathbf{J}_{\text{feet}}$ are the 3D foot joint positions, and $\mathbf{h}$ is the ground height at those positions projected from the scene. The first term enforces accurate contact, while the second penalizes interpenetration when contact is not expected.

To reduce jitter and improve motion realism, we apply a smoothness loss $\mathcal{L}_{\text{smooth}} = \left\| \mathbf{J}_p^{1:N} - \mathbf{J}_p^{0:N-1} \right\|_2$, where $\mathbf{J}_p$ denotes the global joint positions over time.

Object Interaction Objectives When interacting with static scene objects (e.g., for sitting or lying), we apply a collision loss $\mathcal{L}_{\text{collision}} = \text{SDF}(\mathbf{v})$, where $\mathbf{v}$ are the body mesh vertices, and SDF is the signed distance field from the object mesh surface. This loss penalizes interpenetration and promotes plausible body-object contact.

Optimization Details We define the final scene objective as: $\mathcal{L} = \lambda_{\text{phys}}\mathcal{L}_{\text{phys}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{collision}}\mathcal{L}_{\text{collision}}$, where the weights are set to $\lambda_{\text{phys}} = 3$, $\lambda_{\text{smooth}} = 50$ for navigation tasks, and $\lambda_{\text{collision}} = 50$ for object interaction. Optimization is performed using the Adam optimizer over 100 of steps. We use the final DDIM trajectory from the optimized noise to synthesize the resulting motion.

4. Experiments

First we introduce our dataset and evaluation metrics. Then we show comparisons of our proposed approach against the baselines. We further conduct a human perceptual study to complement our evaluation and ablation study to verify the effectiveness of our key components.

4.1. Dataset and Implementation Details

The SCENIC Dataset To our knowledge, [12, 28] are the only existing dataset that captures human navigation with scenes and text annotations. However, both its motion style and terrain variation are limited.

To address the scarcity of paired human-scene-text data, we utilize a database of heightmap assets [25], derived from video game environments. This approach allows us to match human motion segments with the most suitable terrain patches, thereby generating paired human and scene data. We divide the motion sequences into clips of 60 frames (2 seconds) each, aligning the human's initial position with the center of the 4×4 meter patches. The terrains with minimized foot contact and penetration error are retrieved, where the error is computed similarly to Equation 4. To diversify our dataset, we record motion featur-

Table 1. Quantitative evaluations against baseline methods, and ablation study on key components and design.

Methods	Scene constraints		Goal reaching		Motion quality				User Study(%)
	Penetration↓	Contact Dist.↓	Pos.↓	Rot.↓	FID↓	Multimodal Dist.→	Diversity→	Foot-skate↓	
Ground Truth	-	-	-	-	0.000	6.023	12.410	-	-
FlowMDM* [2]	4.67	6.94	4.79	0.125	66.485	9.107	17.038	2.949	9.5
TRUMANS* [29]	4.50	6.65	3.38	0.0454	26.533	8.172	14.717	3.329	14.9
Ours no cano.	1.98	5.55	3.51	0.0796	8.021	7.344	13.507	2.710	-
Ours no scene emb.	2.99	5.74	1.57	0.0384	1.924	5.823	12.519	2.678	-
Ours no cross-attn.	2.54	5.61	1.72	0.0392	1.924	5.765	12.540	2.690	-
Ours	**1.58**	**4.56**	**1.40**	**0.0372**	**1.690**	**5.925**	**12.371**	**2.676**	**75.6**

ing various motion styles on different terrains. Our motion set includes a dataset captured with Inertial Motion Units (IMU) and the PFNN [25] motion dataset redirected to the SMPL format. The dataset comprises 15,000 sequences and 1000 sequences are reserved for testing. To augment our data, pose mirroring is performed along the x-axis and for each motion sequence. Three best-fitted terrains are used for training.

Implementation Details All models including the baselines are trained for 400k steps. Navigation models are trained on the SCENIC dataset and the interaction model is trained in our text-annotated SAMP [21]. All models are trained to denoise the input in 100 diffusion steps.

4.2. Baselines

We train all the baselines and perform an ablation study on the SCENIC dataset. We compare our work with state-of-the-art diffusion-based methods. TRUMANS [29] achieves impressive performance for scene interaction, since it does not condition on text prompts, we replace its action encoding with a text encoding. This text variant of TRUMANS is denoted as TRUMANS*. FlowMDM [2] does not consider the surrounding scene, we enhance its scene awareness by additionally incorporating the same occupancy representation that was adopted in the original TRUMANS model.

To justify our key hierarchical scene reasoning, ablation is performed on the goal-centric canonicalization, where instead the motion is canonicalized to the first frame, and the goal is provided explicitly. Another baselines are introduced to evaluate the importance of the local scene reasoning by not incorporating the scene embedding and without the motion-scene cross attention.

4.3. Quantitative Evaluation.

An important aspect of assessing the model is to evaluate how well it satisfies the scene constraint. **Penetration** (cm) measures the average penetration distance for all the human body vertices [29, 38, 82, 83], obtained by querying all body vertices from the computed SDF of the testing scenes. **Contact distance** (cm) evaluates the average distance to the scene when there is contact. For this, we annotate four body vertices - one at the toe and the heel of each foot. From Table 1, our model achieves competitive performance across all evaluation metrics compared to baseline methods. In terms of scene constraints, our approach attains the lowest penetration (1.58 cm) and contact distance (4.56 cm), outperforming FlowMDM* and TRUMANS*. These results collectively demonstrate the effectiveness of our human-centric scene embedding in maintaining physical plausibility of the generated motions.

For goal reaching, we evaluate the body-to-goal **positional** (cm) and **rotational offset** (radians). [83, 94]. In goal-reaching, our method exhibits the best performance in both positional accuracy (1.40 cm) and rotational alignment (0.0372 radians). This validates our design choice of goal-centric canonicalization.

To measure motion quality, we follow previous work [18, 29, 38, 71, 83, 94], using the motion and text embeddings generated from a pre-trained action recognition model [12, 80]. The model is pre-trained on the SCENIC dataset with all ten action classes. To measure the alignment between motion and text, **multimodal distance** measures the average distance between the motion and text embeddings. **Frechet Inception Distance** (FID) [18] measures realism by comparing the motion embedding of the generated and ground-truth sequences. **Diversity** is calculated on the basis of the average pairwise distance between generated motion embeddings. Table 1 shows that our approach achieves the best performance with the lowest FID score (1.690) among all compared methods, being closest to the ground truth. Our method also maintains diversity (12.371) and multimodal distance (5.925) scores closest to the ground truth distribution (12.410 and 6.023 respectively). Our model also produces the least foot-skate artifact (2.676 cm). Note, the "no scene embed" baseline performs better in multimodal distance and diversity metrics due to its absence of terrain constraints. However, this trade-off comes at the cost of physical plausibility, which SCENIC prioritizes.

Human Perceptual Study In addition to the quantitative measures introduced, we also conducted a user study on the realism and controllability of the methods through text. In the user study, we presented animations in real-world scenes from HPS [20] and Matterport [7] to 24 participants. The participants make three-way comparisons of the ani-

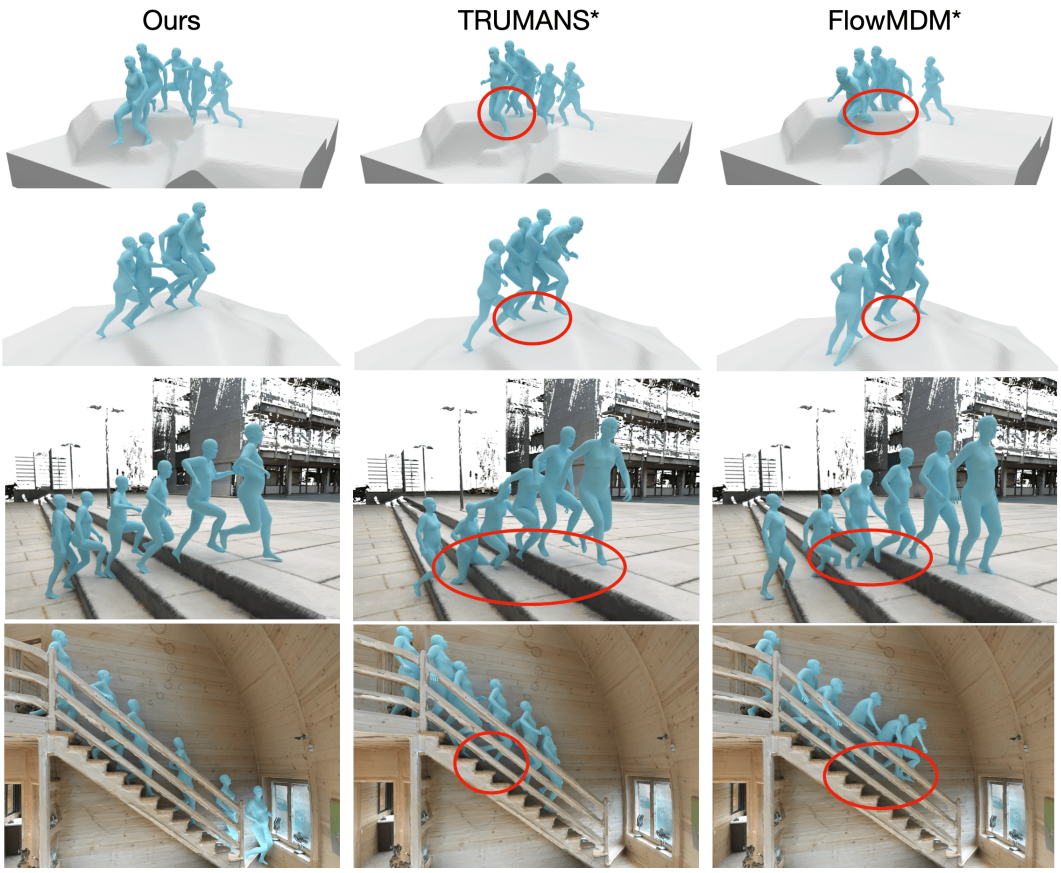

Figure 3. Qualitative comparison with baselines. Results are on the testing set of the SCENIC dataset (top two rows). Without the hierarchical reasoning of the scene, the baseline methods produce more penetration with the legs (first row) and the floating effect (second row). Furthermore, our method generalizes to real-world scene datasets of HPS [20] and MatterPort3D [7] (bottom two rows)

mations generated by the three methods in shuffled order. We have filtered out incomplete responses. Details of the user study can be found in the supplementary. Results show **75.6%** of participants preferred SCENIC over the baselines. This strong preference confirms our method's effectiveness in generating visually plausible human-scene interactions, particularly in reducing floating and penetration artifacts, while generating realistic contacts.

Inference speed SCENIC is more computationally efficient, achieving an inference speed of 0.0203 seconds per data sample vs. 0.0442 seconds for FlowMDM* and 0.0581 seconds for TRUMANS*.

4.4. Qualitative Evaluation

We present qualitative comparisons in Figure 3. The top two rows demonstrate results from the SCENIC dataset's test set, where baseline methods exhibit noticeable artifacts - leg penetration into the ground surface - due to their limited scene understanding. In contrast, our approach, leveraging hierarchical scene reasoning with scene embedding and goal-centric canonicalization, generates motions that maintain proper contact while avoiding both penetration and floating artifacts. The bottom two rows highlight the generalization capabilities of our approach across different scene datasets, namely MatterPort3D [7] and HPS [20]. These real-world environments pose more challenging scenes than those in our training set. Despite these complexities, our method consistently generates physically plausible motions that adhere to scene constraints. Please refer to our supplementary video for results and comparisons in motion.

4.5. Ablation

Goal-centric canonicalization The usefulness of our core components of goal-centric canonicalization and human-centric scene embedding is shown in the comparison with the ablative baselines. Our method (1.40 cm, 0.0372 radians) achieves better performance over the baseline without canonicalization (3.51 cm, 0.0796 radians) validates our design choice of goal-centric canonicalization (Table 1).

Fine-grained scene reasoning As illustrated in Figure 5, without local scene embedding, the model is more likely to exhibit unwanted penetrations with cluttered scenes while

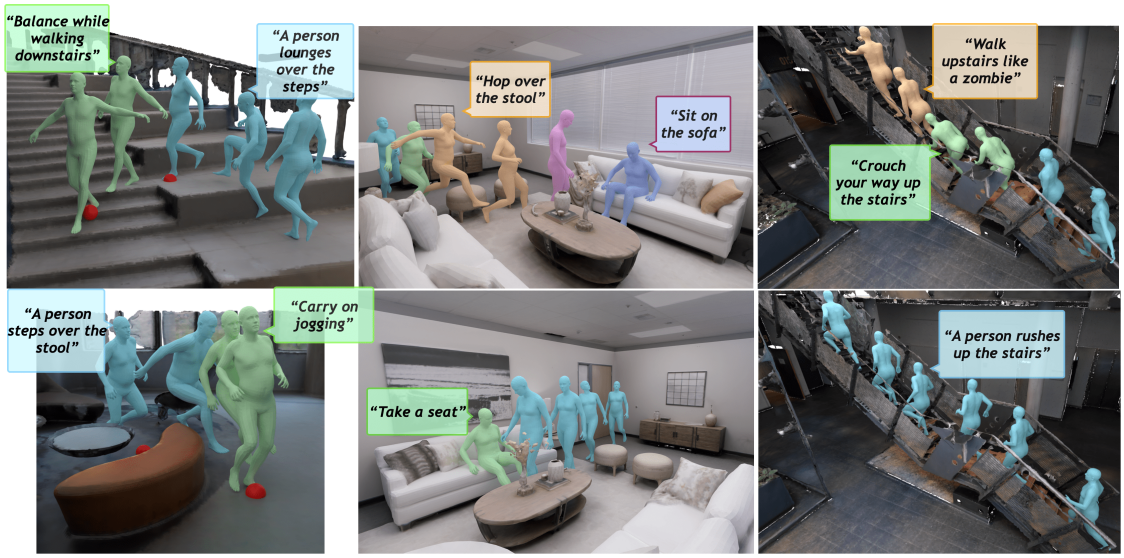

Figure 4. SCENIC generalizes to novel scenes and text instructions, as demonstrated with LaserHuman [12], Replica [66] and HPS [20] scenarios. The model follows instructions like *take a walk*, *sit on the sofa*, and *run up the stairs*, and adapts to more complex commands such as *jump over a stool* while adjusting to scene constraints. In the HPS scene, the model transits between different gait styles, following the text control while adapting to the staircases.

Figure 5. Ablation on the human-centric scene embedding. It prevents unwanted interactions with cluttered environments.

navigating. With the scene embedding, our model avoids the collision in the way of reaching the sub-goal. The importance of local scene reasoning and the motion-scene cross attention is also justified as shown in Table 1.

Motion-text alignment vs. Single-text We compare our per-frame text alignment approach to a baseline that uses a single, global text instruction to describe the target action. Our method demonstrates improved text control by aligning motion generation more closely with the textual description at each frame. This leads to more precise motion-text correspondence, as reflected by a reduction in multimodal distance from 7.88 to 5.925.

4.6. Generalization

Complex real-world environments SCENIC is capable of generalizing to both novel real-world scenes and text instructions. As shown in Figure 4, SCENIC navigates in Replica [66] and HPS [20] The model is firstly instructed to "take a walk" before "sitting on the sofa" (top left) and "running up the stairs" (bottom left). In more complicated scenarios, the model adapts to the scene constraints while following the "jump over a stool" instruction, before "sitting on the sofa" (top right). In the HPS scene, the human transits between various gait styles controlled by text while adapting to the stairs. Similarly in Figure 1, SCENIC is provided a series of text instructions before lying on the sofa in the LaserHuman scene [12].

General Text Prompting As shown in Figure 4, our method handles diverse and semantically rich prompts such as "a person lounges over the steps" or "a person steps over the stool.", while adapting to the scene.

5. Conclusion

We presented SCENIC, the first diffusion-based motion synthesis model that simultaneously enables text control and adaptation to complex terrains. Our model introduces a hierarchical scene reasoning for precise scene adaptation and also a scene-aware diffusion noise optimization scheme. Through extensive experiments across multiple scene datasets, we demonstrated that our approach significantly outperforms existing methods, achieving the best performance in both scene constraint satisfaction and motion quality. User studies further validate our approach, having 75.6% of the participants preferring our method over state-of-the-art methods. In the future, this work can be extended to more complex compositional scene interaction, such as carrying objects while climbing stairs.

References

[1] João Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21211–21221, 2023. 3

[2] German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 6

[3] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[4] Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. 3

[5] Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. In *International Conference on 3D Vision (3DV 2024)*, 2024. 3

[6] Zhi Cen, Huaijin Pi, Sida Peng, Zehong Shen, Minghui Yang, Zhu Shuai, Hujun Bao, and Xiaowei Zhou. Generating human motion in 3d scenes from text descriptions. In *CVPR*, 2024. 3, 4

[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 2, 6, 7

[8] Ling-Hao Chen, Wenxun Dai, Xuan Ju, Shunlin Lu, and Lei Zhang. Motionclr: Motion generation and training-free editing via understanding attention mechanisms. *arxiv:2410.18977*, 2024. 2

[9] Rui Chen, Mingyi Shi, Shaoli Huang, Ping Tan, Taku Komura, and Xuelin Chen. Taming diffusion probabilistic models for character control. In *SIGGRAPH*, 2024. 2

[10] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[11] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[12] Peishan Cong, Ziyi Wang, Zhiyang Dou, Yiming Ren, Wei Yin, Kai Cheng, Yujing Sun, Xiaoxiao Long, Xinge Zhu, and Yuexin Ma. Laserhuman: Language-guided scene-aware human motion generation in free environment, 2024. 2, 3, 5, 6, 8

[13] Jieming Cui, Tengyu Liu, Nian Liu, Yaodong Yang, Yixin Zhu, and Siyuan Huang. Anyskill: Learning open-vocabulary physical skill for interactive agents. In *Conference on Computer Vision and Pattern Recognition(CVPR), year=2024*. 3

[14] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[15] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. 2024. 3, 5

[16] Markos Diomataris, Nikos Athanasiou, Omid Taheri, Xi Wang, Otmar Hilliges, and Michael J. Black. WANDR: Intention-guided human motion generation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[17] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Remos: 3d motion-conditioned reaction synthesis for two-person interactions. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[18] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 6

[19] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[20] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 6, 7, 8

[21] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *Proceedings of the International Conference on Computer Vision 2021*, 2021. 3, 5, 6

[22] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael J. Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. *CoRR*, abs/2302.00883, 2023. 3

[23] Nhat M. Hoang, Kehong Gong, Chuan Guo, and Michael Bi Mi. Motionmix: Weakly-supervised diffusion for controllable motion generation. In *Thirty-Eighth Conference on Artificial Intelligence, AAAI 2024*. 2

[24] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Trans. Graph.*, 2016. 2

[25] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Trans. Graph.*, 36(4), 2017. 2, 3, 5, 6

[26] Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. Como: Controllable motion generation through language guided pose code editing, 2024. 2

[27] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *ICCV*, 2023. 3

[28] Nan Jiang, Zimo He, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia Conference Papers*, 2024. 1, 2, 3, 4, 5

[29] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1737–1747, 2024. 1, 2, 3, 4, 6

[30] Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing diffusion noise can serve as universal motion priors. In *arxiv:2312.11994*, 2023. 2, 5

[31] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[32] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. *arXiv preprint arXiv:2209.00349*, 2022. 2

[33] Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions, 2024. 3

[34] Hanyang Kong, Kehong Gong, Dongze Lian, Michael Bi Mi, and Xinchao Wang. Priority-centric human motion generation in discrete latent space. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023,*. 2

[35] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis, 2023. 3

[36] Jiye Lee and Hanbyul Joo. Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments. *arXiv preprint arXiv:2301.02667*, 2023. 3

[37] Chuqiao Li, Julian Chibane, Yannan He, Naama Pearl, Andreas Geiger, and Gerard Pons-Moll. Unimotion: Unifying 3d human motion synthesis and understanding. *arXiv preprint arXiv:2409.15904*, 2024. 2, 4

[38] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. Controllable human-object interaction synthesis. . 2, 3, 5, 6

[39] Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics*, 42(6), 2023.

[40] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. Controllable human-object interaction synthesis. In *ECCV*, 2024. 3

[41] Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024*, . 3

[42] Siyao Li, Tianpei Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy. Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment. In *ICLR*, 2024. 2

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

[44] Xueyi Liu and Li Yi. Geneoh diffusion: Towards generalizable hand-object interaction denoising via denoising diffusion. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[45] Xinpeng Liu, Haowen Hou, Yanchao Yang, Yong-Lu Li, and Cewu Lu. Revisit human-scene interaction via space occupancy. *arXiv preprint arXiv:2312.02700*, 2023. 3, 4

[46] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34, 2015. 3

[47] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. *arxiv:2310.12978*, 2023. 2

[48] Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris M. Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3

[49] Sihan Ma, Qiong Cao, Jing Zhang, and Dacheng Tao. Contact-aware human motion generation from textual descriptions. *arXiv preprint arXiv:2403.15709*, 2024. 2

[50] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: archive of motion capture as surface shapes. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. 2

[51] Josh Merel, Saran Tunyasuvunakool, Arun Ahuja, Yuval Tassa, Leonard Hasenclever, Vu Pham, Tom Erez, Greg Wayne, and Nicolas Heess. Catch & carry: reusable neural controllers for vision-guided whole-body tasks. *ACM Trans. Graph.*, 2020. 3

[52] Aymen Mir, Xavier Puig, Angjoo Kanazawa, and Gerard Pons-Moll. Generating continual human motion in diverse 3d scenes. In *International Conference on 3D Vision (3DV)*, 2024. 1, 2, 3

[53] Liang Pan, jingbo Wang, Buzhen Huang, Junyu Zhang, Haofan Wang, Xu Tang, and Yangang Wang. Synthesizing physically plausible human motions in 3d scenes, 2023. 3

[54] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553*, 2023. 3

[55] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J. Black, Gül Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation. In *CVPR Workshop on Human Motion Generation*, 2024. 2

[56] Huaijin Pi, Sida Peng, Minghui Yang, Xiaowei Zhou, and Hujun Bao. Hierarchical generation of human-object interactions with diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[57] Huaijin Pi, Zhi Cen, Zhiyang Dou, and Taku Komura. Coda: Coordinated diffusion noise optimization for whole-body manipulation of articulated objects. *arXiv preprint arXiv:2505.21437*, 2025. 5

[58] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 2019. 5

[59] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[60] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[61] Roey Ron, Guy Tevet, Haim Sawdayee, and Amit H. Bermano. Hoidini: Human-object interaction through diffusion noise optimization. *arXiv preprint arXiv::2506.15625*, 2025. 5

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

[63] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *The Twelfth International Conference on Learning Representations*, 2024. 4

[64] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 5

[65] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6), 2019. 3, 4

[66] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2, 8

[67] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[68] Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soren Pirk, and Michael J. Black. Grip: Generating interaction poses conditioned on object and body motion. In *International Conference on 3D Vision (3DV 2024)*, 2024. 3

[69] Mikihiro Tanaka and Kent Fujiwara. Role-aware interaction generation from textual description. In *ICCV*, 2023. 2

[70] Purva Tendulkar, Dídac Surís, and Carl Vondrick. Flex: Full-body grasping without full-body grasps. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[71] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3, 4, 6

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

[73] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[74] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say, interact as you can: Language-guided human motion generation with scene affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[75] Qianyang Wu, Ye Shi, Xiaoshui Huang, Jingyi Yu, Lan Xu, and Jingya Wang. THOR: text to human-object interaction diffusion via relation intervention. *CoRR*, abs/2403.11208, 2024. 3

[76] Zhen Wu, Jiaman Li, Pei Xu, and C Karen Liu. Human-object interaction from human-level instructions. In *IEEE/CVF International Conference on Computer Vision, ICCV*, 2025. 5

[77] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[78] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023. 3

[79] Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. *arXiv preprint arXiv:2403.19652*, 2024. 3

[80] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 6

[81] Jie Yang, Xuesong Niu, Nan Jiang, Ruimao Zhang, and Huang Siyuan. F-hoi: Toward fine-grained semantic-aligned 3d human-object interactions. *European Conference on Computer Vision*, 2024. 3

[82] Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J. Black. MIME: Human-aware 3D scene generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

[83] Hongwei Yi, Justus Thies, Michael J. Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. *arXiv:2404.10685*, 2024. 2, 3, 4, 5, 6

[84] Chengwen Zhang, Yun Liu, Ruofan Xing, Bingda Tang, and Li Yi. Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement. *arXiv preprint arXiv:2406.19353*, 2024. 3

[85] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *ACM Trans. Graph.*, 2021. 3

[86] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-m3: Capture multiple humans and objects interaction within contextual environment. *arXiv preprint arXiv:2404.00299*, 2024. 3

[87] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *2023 IEEE/CVF International Conference on Computer Vision, ICCV*. 2

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

[89] Wanyue Zhang, Rishabh Dabral, Thomas Leimkühler, Vladislav Golyanik, Marc Habermann, and Christian Theobalt. Roam: Robust and object-aware motion generation using neural pose descriptors. *arXiv preprint arXiv:2308.12969*, 2023. 3

[90] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[91] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Ilya A. Petrov, Vladimir Guzov, Helisa Dhamo, Eduardo Pérez Pellitero, and Gerard Pons-Moll. Force: Dataset and method for intuitive physics guided human-object interaction. In *Arxiv*, 2024. 3

[92] Zihan Zhang, Richard Liu, Kfir Aberman, and Rana Hanocka. Tedi: Temporally-entangled diffusion for long-term motion synthesis. In *SIGGRAPH, Technical Papers*, 2024. 2

[93] Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, , and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *International conference on computer vision (ICCV)*, 2023. 1, 3

[94] Kaifeng Zhao, Gen Li, and Siyu Tang. A diffusion-based autoregressive motion model for real-time text-driven motion control. In *Arxiv*, 2024. 3, 6

[95] Kaifeng Zhao, Gen Li, and Siyu Tang. DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. 5

[96] Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast, high-quality motion generation. *arXiv preprint arXiv:2312.02256*, 2023. 2

[97] Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. On the continuity of rotation representations in neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3