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

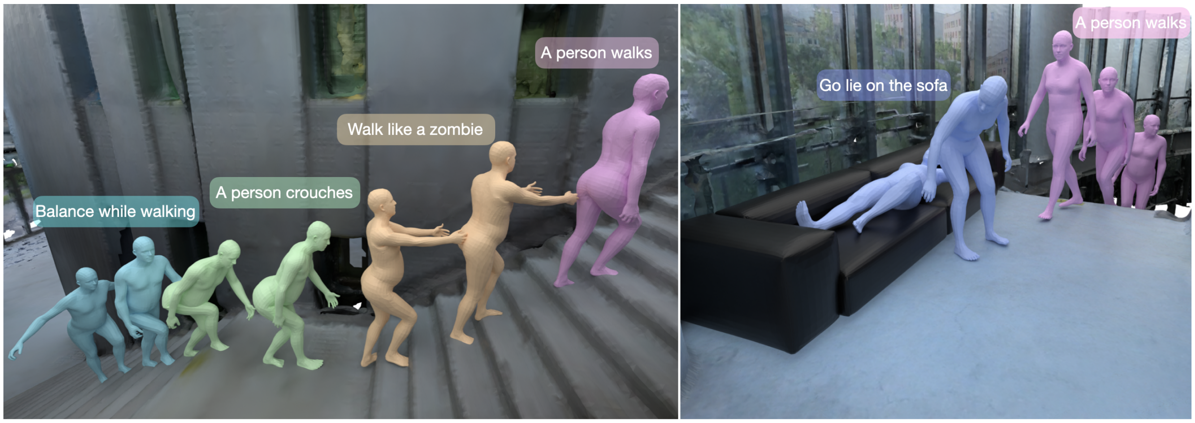


Figure 1: SCENIC is a text-conditioned scene interaction model. It adapts to complex scenes with varying terrains and also supports user-specified semantic control with natural language. Given a 3D scene, our model takes as cues of user-specified trajectory as sub-goals, and text. *We encourage the readers 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 its core is a novel scene-dependent, goal-centric canonicalization that handles high-level goal constraint, and is complemented by an ego-centric distance field that captures local geometric details. This dual representation enables our model to generate physically plausible motion across diverse 3D scenes. By implementing frame-wise text alignment, our system achieves seamless transitions between different motion styles while maintaining scene constraints. Experiments demonstrate 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. Our code, dataset, and models will be released at https://virtualhumans.mpi-inf.mpg.de/scenic/.

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 [29, 31, 53, 91] is crucial for numerous applications ranging from gaming to embodied AI. 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 [61, 69, 74] or motion adaptation to simplified environments [39, 80], they struggle with complex scenarios. Even methods that can adapt to uneven terrain [26, 49, 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 terrain features. Last, traditional approaches require extensive paired motion-scene data, which is expensive to acquire due to tracking difficulties and does not scale well to diverse environments.

Our key insight is that complex scene-aware motion synthesis can be decomposed into hierarchical reasoning levels, similar to how humans approach navigation tasks. At the high level, we synthesize motion in a goal-centric canonical coordinate frame, enabling the model to learn target-reaching behaviors naturally. At a more granular level, we take inspiration from recent 3D generation work [7] that encodes 3D spatial features with 2D planar encoding. We represent detailed scene geometry through a human-centered distance field representation [49, 60]. This efficient representation enables comprehensive reasoning about local scene features, including terrain variations and obstacles. To provide semantic control, we align text and motion on a frame-wise basis, allowing for dynamic instruction changes while ensuring smooth transitions. To address data efficiency, we exploit the compositional nature of human motion, training on short motion segments [29, 31, 53] that can be efficiently augmented by automatically fitting varied terrain surfaces.

With these solutions, we propose the first model which is scene-aware and can be controlled with fine-grained natural language. Experiments demonstrate SCENIC handles complex scene geometry through precise scene-aware adaptation across four real-scene datasets including Replica [64], Matterport3D [8], HPS [21], and LaserHuman [13]. Moreover, SCENIC supports seamless transition between ten distinct motion semantics including "crouching", "climbing", "hopping", "jumping", and "balancing", and can adapt to complicated instruction such as "walking upstairs 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 participants over state-of-the-art alternatives (more details see Table 1).

The key contributions of our work include:

1. We introduce the first unified method for 3D scene-aware human motion synthesis, capable of handling complex terrains like stairs, steps, or slopes, while also enabling fine-grained control through textual prompts.
2. Our novel diffusion model leverages hierarchical scene reasoning, efficiently handles complex 3D environments while maintaining plausibility. Its effectiveness is validated across four diverse real-world datasets.
3. A scalable approach to synthesizing continuous human navigation in 3D scenes, which can be integrated with an object-interaction model, as shown in Figure 1.

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 [11, 15, 25, 35, 48, 50, 61, 69, 84, 85, 89, 92] and comprehensive motion capture datasets like AMASS [51]. 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 [18, 43, 44, 67], joint-level control [32, 70, 74], and style editing [10, 27].

Motion editing through text has evolved along two main paths: in-motion editing for specific body parts [9, 28, 34] and segment-level editing using text prompts. Notably, FlowMDM [2] demonstrated impressive results in seamless transitions between local motion segments. STMC [56] proposed a hybrid method for spatial and temporal motion composition using pre-trained motion models. UniMotion [38] 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.

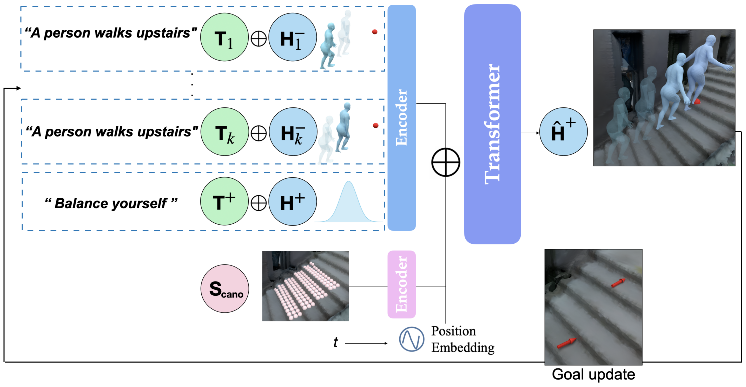


Figure 2. Architecture overview. SCENIC has a 3D scene, a user-defined trajectory, and text prompts, and the past human motion as inputs. The past human motion and the scene encoding first undergo goal-centric canonicalization. The diffusion-based transformer then encodes the aligned text-motion tokens,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

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, 81, 83] spans a wide range, from interactions with large, static objects like chairs and beds [22, 30, 36, 54, 57, 63, 80, 86, 87], to dynamic engagements with moving objects. This includes studies that focus on contact-based object interactions without navigation [16, 55, 73, 75, 76, 78], as well as those that incorporate navigation [39–41, 88]. A parallel line of research leverages reinforcement learning to synthesize interactions [14, 23, 52]. Other studies have concentrated on full body grasps [1, 17, 42, 65, 65, 68] and dexterous hand manipulation [4, 5, 12, 46, 66, 82].

In the context of human-scene interactions, a significant portion of the work is dedicated to generating short-term motion within 3D scenes [6, 71, 72]. PFNN [26] 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, 91]. Others assume uniform height within the scenes [31, 37, 53]. Using reinforcement learning, [49, 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 [80] proposed a two-stage method for collision-free navigation within the scene. TRUMANS [31] unified static and dynamic object interactions, and a recent extension replaced action labels with more versatile text prompts [29], achieving impressive results. However, these models still assume flat terrains or floors. While some concurrent works have demonstrated human motion on stairs [13, 90], they have their limitations. Zhao et al. [90] did not train their model with paired motion-scene data. This lack of scene awareness restricts the model's ability to generalize to complex scene constraints and adapt to changes in terrain surfaces. Moreover, their approach requires the future 3D root position, which is not always available. On the other hand, Cong et al. [13] 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

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 in the goal canonical frame 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 $\mathbf{H}$ Unlike previous motion representation of human motion [19, 31, 69], which requires an additional fitting process to obtain the final animated mesh, our representation can be animated directly. The SMPL model [47] is used to parameterize our human motion. Our motion human $\mathbf{H}$ consists of N frames of the joint rotations in the 6-D continuous form [93] $\boldsymbol{J}_r \in \mathbb{R}^{N\times 22\times 6}$, 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 $\mathbf{S}$ Inspired by [7], 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, 31, 45, 63]. 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 $\mathbf{G}_j$ to our system 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 $\mathbf{T}$ Unlike previous methods that use a single text embedding combined with a timestamp [61, 62, 69], 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 [38], 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. Goal-Centric Canonicalization

One key to our model is the goal-centric canonicalization that ensures robust goal-reaching, while maintaining physical plausibility. This transformation serves two crucial purposes: (1) it simplifies the learning problem by creating a consistent reference frame for motion synthesis, and (2) it enables better generalization across different goal configurations. We transform both our human motion and scene embedding ($\mathbf{H}$ and $\mathbf{S}$) into the coordinate system of the goal so that the model can combine the high-level reasoning of the goal and the fine-grained reasoning of the complex scene geometry. First, under current goal $\mathbf{G}_j$, we apply canonicalization to the motion $\mathbf{H}$ via $\mathbf{H}_{\text{cano}} = \mathcal{T}_{\text{human}}(\mathbf{H}, \mathbf{G}_j)$. Traditional methods [29, 31, 80], which explicitly condition on the goal, can often lead to inaccuracies in reaching the target. Our experiments show that this is accentuated when synthesizing motion on uneven terrain surfaces. Therefore, the model is instead trained to synthesize motion that converges to the origin in the coordinate system defined by the goal. Moreover, the scene embedding is transformed to align with the height of the goal via $\mathbf{S}_{\text{cano}} = \mathcal{T}_{\text{scene}}(\mathbf{S}, \mathbf{G}_j)$. This way, the model does not only implicitly learn to reason about the goal, but also becomes aware of the local scene geometry. Additionally, the current height of the root is encoded.

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 [22] dataset. The interaction diffusion model is trained using the same learning objective as the navigation model.

3.6. Scene-aware Guidance

At test time, diffusion models can be guided to meet specific objectives, alleviating the need for training models with different configurations, and further enhancing the quality of scene interaction. For the sake of readability, we will use the same notation for both estimated and true values in the following discussion. We directly apply the guidance to the clean motion prediction from the model $\mathbf{H}_0$ [24, 32, 39, 80]. At each denoising step, the predicted $\mathbf{H}_0$ is updated with the gradient of an analytic objective function $\mathcal{J}$. This process can be denoted as $\tilde{\mathbf{H}}_0 = \mathbf{H}_0 - \alpha\Delta_{\mathbf{H}_t}\mathcal{J}(\mathbf{H}_0)$, where α controls the strength of the guidance and $\mathbf{H}_t$ is the noisy input motion at diffusion step t. The predicted mean μ is then calculated with the updated motion prediction $\hat{\mathbf{H}}_0$

For navigation, we further introduce a physics plausibility guidance to avoid penetration and encourage realistic contact. By enforcing foot contact when it happens and penalizing the foot penetration with the scene when there is no contact. Formally, the guidance is computed by

$$\mathcal{J}_{\text{phys}} = \boldsymbol{c} \cdot \|\mathbf{J}_{\text{feet}} - \mathbf{h}\|_2 + (1-\boldsymbol{c}) \cdot \mathbb{1}(\mathbf{h} > \mathbf{J}_{\text{feet}}) \cdot \|\mathbf{J}_{\text{feet}} - \mathbf{h}\|_2. \quad (3)$$

Here, we leverage the predicted foot contact label $\boldsymbol{c}$ to enforce accurate foot contact with the scene and to discourage penetration. Furthermore, we denote the predicted foot joint positions as $\mathbf{J}_{feet}$ and the heights of the projected points from the feet as $\mathbf{h}$.

For the interaction model, a collision objective is used to discourage penetrations [39, 80] between humans and objects $\mathcal{J}_{\text{collision}} = \text{SDF}(v)$, where object signed distance field (SDF) is queried by the body vertices v, and the mean penetration distance of the body vertices is minimized. In addition to the object collision guidance, we also incorporate a motion smoothness objective $\mathcal{J}_{\text{smooth}} = \|\mathbf{J}_p^{1:N} - \mathbf{J}_p^{0:N-1}\|_2$

For the navigation model, we set the guidance weight α to 3 for physics guidance and 50 for smoothness guidance. For the interaction model, we utilize weights of 50 for the collision guidance. To ensure smooth generation results, we apply the inference guidance at the final time step of denoising. For a fair comparison with baselines, the inference guidance is not activated for all comparisons.

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, [13, 29] 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 vast database of artificial heightmaps [26], 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 3. To diversify our dataset, we record motion featuring various motion styles across different terrains. Our motion set includes a dataset captured with Inertial Motion Units (IMUs) and the PFNN [26] motion dataset retargeted to the SMPL format. The dataset comprises 15000 sequences, and 1000 sequences are reserved for testing. To augment our data, pose mirroring is performed along the x-axis and for

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↓	Multimodality→	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* [31]	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	**1.57**	**4.51**	**1.38**	**0.0376**	**1.680**	6.354	13.067	**2.671**	**75.6**

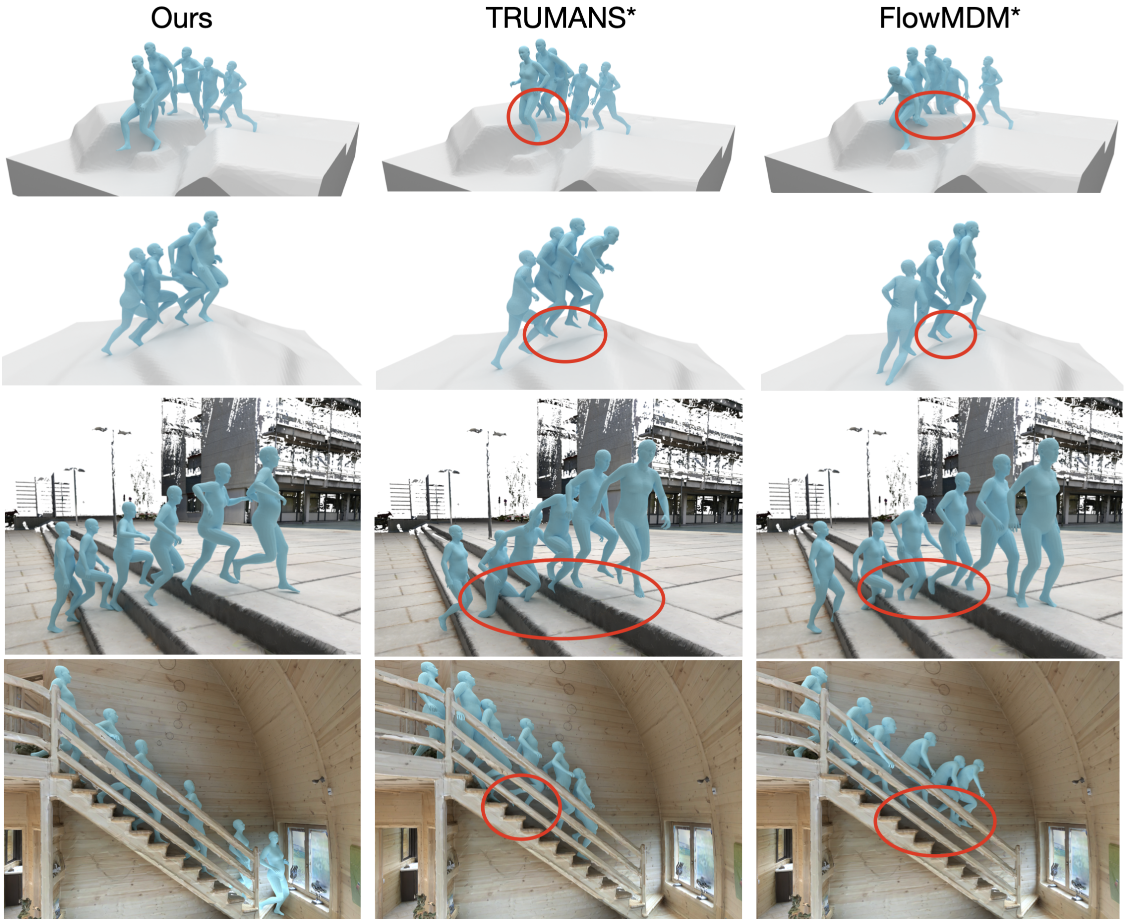


Figure 3. Qualitative comparison with baselines. Results are on the testing set of the SCENIC dataset (top two rows). Without the hierarchical reasoning of the scene, the baseline methods produce more penetration with the legs (first row) and the floating effect (second row). Furthermore, our method generalizes to real-world scene datasets of HPS [21] and MatterPort3D [8] (bottom two rows)

each motion sequence. Three best-fitted terrains are used for training.

Implementation Details All models including the baselines are trained for 400k steps. The navigation models are trained on the SCENIC dataset and the interaction model is trained on our text-annotated SAMP [22]. 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 [31] 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

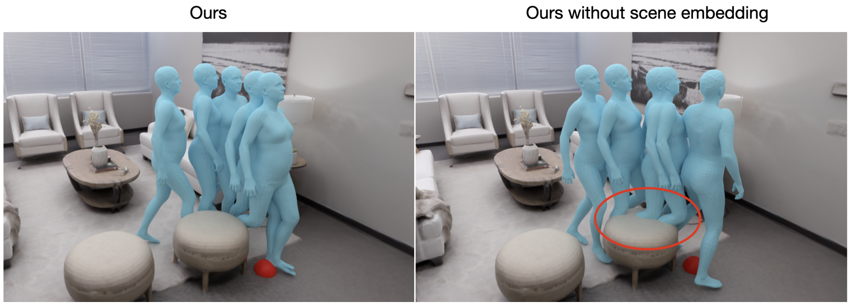


Figure 4. Ablation on the human-centric scene embedding. It is significant in preventing unwanted interactions with cluttered environments.

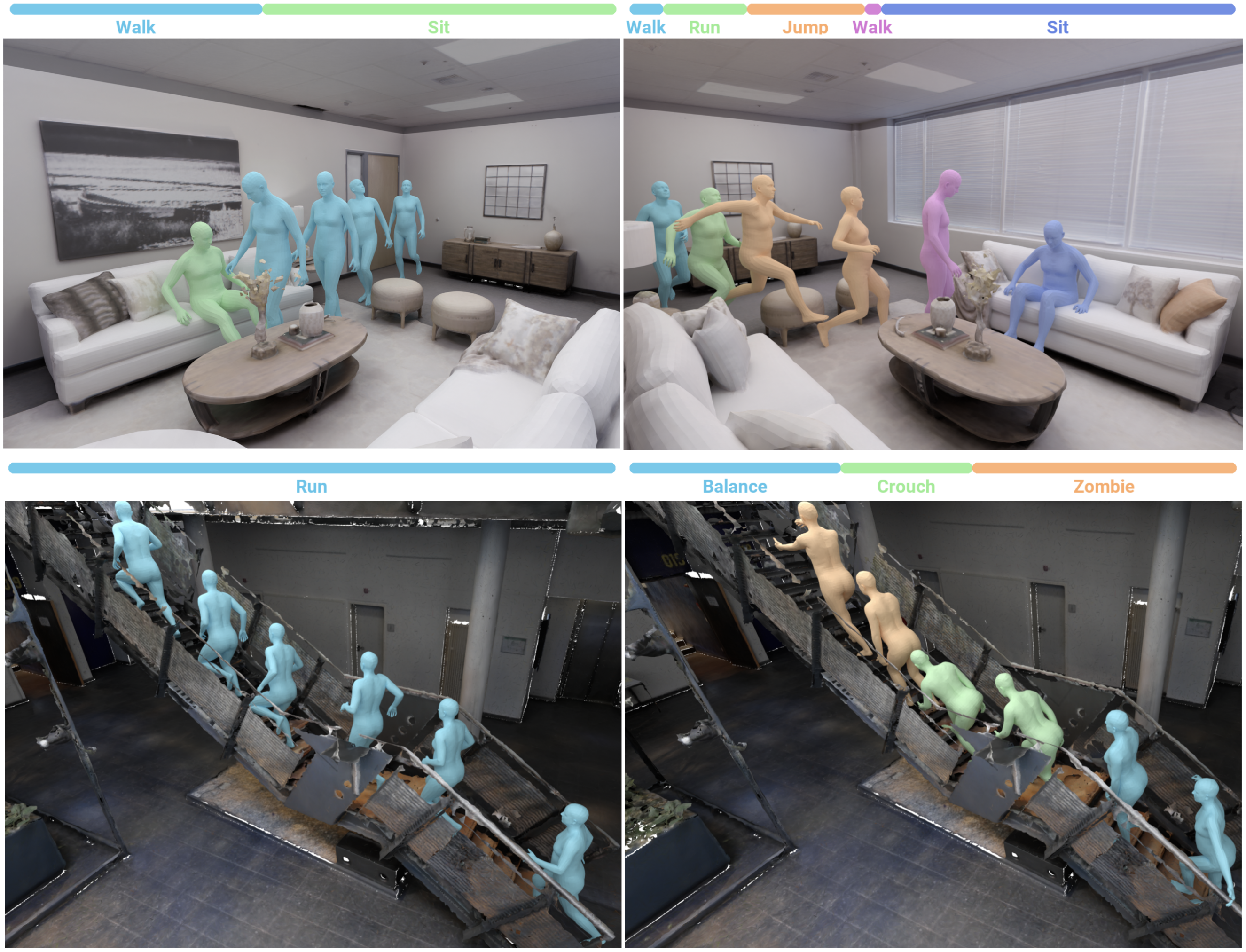


Figure 5. SCENIC generalizes to novel scenes and text instructions, as demonstrated with Replica [64] and HPS [21] scenarios. The model follows instructions like *take a walk*, *sit on the sofa*, and *run up the stairs*, and adapts to more complex commands such as *jump over a stool* while adjusting to scene constraints. In the HPS scene, the model transits between different gait styles, following the text control while adapting to the staircases.

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 baseline is introduced to evaluate the importance of the local scene reasoning by not incorporating the scene embedding.

4.3. Evaluation Metrics

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 [31, 39, 79, 80], 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.
For goal reaching, we evaluate the body-to-goal **positional** (cm) and **rotational offset** (radians). [80, 90].
We follow [20, 31, 39, 69, 80, 90] and evaluate the motion embeddings of an action recognition model [13, 77] trained on the SCENIC dataset with all ten action classes. **Multimodality** measures the alignment between the generated motion and the text instruction. **Frechet Inception Distance** (FID) [20] measures the realism of the motion compared to the ground truth. **Diversity** is computed based on the average pairwise distance between sampled motions.

Human Perceptual Study In addition to the quantitative measures introduced, we also conducted a user study on the realism as well as the controllability of the methods through text. In the user study, we presented animations on the real-world scenes from HPS [21] and Matterport [8] to 24 participants. The participants make three-way comparisons of animations generated by the three methods in shuffled order. We have incomplete responses filtered out. Details of the user study can be found in the supplementary.

4.4. Quantitative Evaluation.

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 (1.57 cm) and contact distance (4.51 cm), outperforming FlowMDM* and TRUMANS*. In goal-reaching, our method exhibits the best performance in both positional accuracy (1.38 cm) and rotational alignment (0.0376 radians). This validates our design choice of goal-centric canonicalization. Regarding motion realism, our approach achieves the lowest FID score (1.680) among all compared methods, being closest to the ground truth. Our method maintains comparable diversity (13.067) and multimodality (6.354) scores close to the ground truth distribution (12.410 and 6.023 respectively). Our model also produces the least foot-skate artifact (2.671 cm). In the user study, 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.

4.5. 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 [8] and HPS [21]. These real-world environments pose more diverse and challenging scenes than those in our training set. Despite these complexities, our method consistently generates physically plausible motions that adhere to scene constraints across these varied terrains. This robust performance again stems from our hierarchical scene reasoning. These results demonstrate that our method not only excels in controlled test scenarios but also effectively adapts to novel, real-world environments. Please refer to our supplementary video for results and comparisons in motion.

4.6. Ablation

The usefulness of our core components of goal-centric canonicalization and human-centric scene embedding are justified through the comparison with the ablative baselines. For goal-reaching capability, it is highlighted in Table 1, where our method (1.38 cm, 0.0376 radians) achieves better performance over the baseline without canonicalization (3.51 cm, 0.0796 radians) validates our design choice of goal-centric canonicalization.

In regards to scene awareness, it is illustrated in Figure 6 that without the scene embedding, the model is more likely to exhibit unwanted penetrations with the cluttered scenes while navigating. With the scene embedding, our model avoids the tea table in the way of reaching the sub-goal. It can navigate while following the sub-goal. The importance of scene awareness is further supported by Table 1, shown in the improvement over our ablation without the scene embedding (2.99 cm penetration) particularly emphasizing the importance of local scene reasoning in preventing body-scene intersections.

4.7. Generalization

SCENIC is capable of generalizing to both novel real-world scenes and text instructions. As shown in Figure 5, SCENIC navigates in Replica [64] and HPS [21] 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 [13].

5. Conclusion

We presented SCENIC, the first diffusion-based motion synthesis model that simultaneously enables text-controlled style editing and adaptation to complex terrains. Our model introduces several key technical innovations, including a goal-centric canonical coordinate frame for long-term navigation and a hierarchical scene reasoning approach that combines high-level goal understanding with fine-grained scene awareness. 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 scene interactions. Directions include incorporating dynamic object manipulation during navigation, such as carrying objects while climbing stairs. Additionally, incorporating collision avoidance mechanisms for dynamic and cluttered environments would benefit real-world applications of virtual social interaction and autonomous driving.

Acknowledgement

A big thank you goes to Hongwei Yi for the useful discussions and exchanging ideas. We appreciate the RVH group members for their useful feedback. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (Emmy Noether Programme, project: Real Virtual Humans), and German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, and Huawei Noah's Ark Lab. Gerard Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. The project was made possible by funding from the Carl Zeiss Foundation.

References

[1] Araújo, J.a.P., Li, J., Vetrivel, K., Agarwal, R., Wu, J., Gopinath, D., Clegg, A.W., Liu, K.: Circle: Capture in rich contextual environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21211–21221 (June 2023) 3

[2] Barquero, G., Escalera, S., Palmero, C.: 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] Bhatnagar, B.L., Xie, X., Petrov, I., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Behave: Dataset and method for tracking human object interactions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2022) 3

[4] Braun, J., Christen, S., Kocabas, M., Aksan, E., Hilliges, O.: Physically plausible full-body hand-object interaction synthesis 3

[5] Braun, J., Christen, S., Kocabas, M., Aksan, E., Hilliges, O.: Physically plausible full-body hand-object interaction synthesis. In: International Conference on 3D Vision (3DV 2024) (2024) 3

[6] Cen, Z., Pi, H., Peng, S., Shen, Z., Yang, M., Shuai, Z., Bao, H., Zhou, X.: Generating human motion in 3d scenes from text descriptions. In: CVPR (2024) 3, 4

[7] Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., Mello, S.D., Gallo, O., Guibas, L., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G.: Efficient geometry-aware 3D generative adversarial networks. In: CVPR (2022) 2, 4

[8] Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. International Conference on 3D Vision (3DV) (2017) 2, 6, 8, 13

[9] Chen, L.H., Dai, W., Ju, X., Lu, S., Zhang, L.: Motionclr: Motion generation and training-free editing via understanding attention mechanisms. arxiv:2410.18977 (2024) 2

[10] Chen, R., Shi, M., Huang, S., Tan, P., Komura, T., Chen, X.: Taming diffusion probabilistic models for character control. In: SIGGRAPH (2024) 2

[11] Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 2

[12] Christen, S., Kocabas, M., Aksan, E., Hwangbo, J., Song, J., Hilliges, O.: D-grasp: Physically plausi-

ble dynamic grasp synthesis for hand-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 3

[13] Cong, P., Wang, Z., Dou, Z., Ren, Y., Yin, W., Cheng, K., Sun, Y., Long, X., Zhu, X., Ma, Y.: Laserhuman: Language-guided scene-aware human motion generation in free environment (2024) 2, 3, 5, 8, 9

[14] Cui, J., Liu, T., Liu, N., Yang, Y., Zhu, Y., Huang, S.: Anyskill: Learning open-vocabulary physical skill for interactive agents. In: Conference on Computer Vision and Pattern Recognition(CVPR), year=2024 3

[15] Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In: Computer Vision and Pattern Recognition (CVPR) (2023) 2

[16] Diller, C., Dai, A.: Cg-hoi: Contact-guided 3d human-object interaction generation (2024) 3

[17] Diomataris, M., Athanasiou, N., Taheri, O., Wang, X., Hilliges, O., Black, M.J.: WANDR: Intention-guided human motion generation. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 3

[18] Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., Slusallek, P.: Remos: 3d motion-conditioned reaction synthesis for two-person interactions. In: European Conference on Computer Vision (ECCV) (2024) 2

[19] 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) (2022) 2, 4

[20] Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020) 8

[21] Guzov, V., Mir, A., Sattler, T., Pons-Moll, G.: 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, 13

[22] Hassan, M., Ceylan, D., Villegas, R., Saito, J., Yang, J., Zhou, Y., Black, M.: Stochastic scene-aware motion prediction. In: Proceedings of the International Conference on Computer Vision 2021 (Oct 2021) 3, 5, 6

[23] Hassan, M., Guo, Y., Wang, T., Black, M.J., Fidler, S., Peng, X.B.: Synthesizing physical character-scene interactions. CoRR **abs/2302.00883** (2023) 3

[24] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv:2204.03458 (2022) 5

[25] Hoang, N.M., Gong, K., Guo, C., Mi, M.B.: Motionmix: Weakly-supervised diffusion for controllable motion generation. In: Thirty-Eighth Conference on Artificial Intelligence, AAAI 2024 2

[26] Holden, D., Komura, T., Saito, J.: Phase-functioned neural networks for character control. ACM Trans. Graph. **36**(4) (2017) 2, 3, 5

[27] Holden, D., Saito, J., Komura, T.: A deep learning framework for character motion synthesis and editing. ACM Trans. Graph. (2016) 2

[28] Huang, Y., Wan, W., Yang, Y., Callison-Burch, C., Yatskar, M., Liu, L.: Como: Controllable motion generation through language guided pose code editing (2024) 2

[29] Jiang, N., He, Z., Li, H., Chen, Y., Huang, S., Zhu, Y.: Autonomous character-scene interaction synthesis from text instruction. In: SIGGRAPH Asia Conference Papers (2024) 2, 3, 4, 5

[30] Jiang, N., Liu, T., Cao, Z., Cui, J., Chen, Y., Wang, H., Zhu, Y., Huang, S.: Full-body articulated human-object interaction. In: ICCV (2023) 3

[31] Jiang, N., Zhang, Z., Li, H., Ma, X., Wang, Z., Chen, Y., Liu, T., Zhu, Y., Huang, S.: Scaling up dynamic human-scene interaction modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1737–1747 (2024) 2, 3, 4, 6, 8

[32] 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 (2023) 2, 5

[33] Kim, J., Kim, J., Na, J., Joo, H.: Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions (2024) 3

[34] Kim, J., Kim, J., Choi, S.: Flame: Free-form language-based motion synthesis & editing. arXiv preprint arXiv:2209.00349 (2022) 2

[35] Kong, H., Gong, K., Lian, D., Mi, M.B., Wang, X.: Priority-centric human motion generation in discrete latent space. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, 2

[36] 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) 3

[37] Lee, J., Joo, H.: Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments. arXiv preprint arXiv:2301.02667 (2023) 3

[38] Li, C., Chibane, J., He, Y., Pearl, N., Geiger, A., Pons-Moll, G.: Unimotion: Unifying 3d human motion synthesis and understanding. arXiv preprint arXiv:2409.15904 (2024) 2, 4

[39] Li, J., Clegg, A., Mottaghi, R., Wu, J., Puig, X., Liu, C.K.: Controllable human-object interaction synthesis 2, 3, 5, 8

[40] Li, J., Clegg, A., Mottaghi, R., Wu, J., Puig, X., Liu, C.K.: Controllable human-object interaction synthesis. In: ECCV (2024)

[41] Li, J., Wu, J., Liu, C.K.: Object motion guided human motion synthesis. ACM Transactions on Graphics **42**(6) (Dec 2023) 3

[42] Li, Q., Wang, J., Loy, C.C., Dai, B.: Task-oriented human-object interactions generation with implicit neural representations. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024 3

[43] Li, S., Gu, T., Yang, Z., Lin, Z., Liu, Z., Ding, H., Yang, L., Loy, C.C.: Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment. In: ICLR (2024) 2

[44] Liang, H., Zhang, W., Li, W., Yu, J., Xu, L.: Intergen: Diffusion-based multi-human motion generation under complex interactions. International Journal of Computer Vision (2024) 2

[45] Liu, X., Hou, H., Yang, Y., Li, Y.L., Lu, C.: Revisit human-scene interaction via space occupancy. arXiv preprint arXiv:2312.02700 (2023) 3, 4

[46] Liu, X., Yi, L.: Geneoh diffusion: Towards generalizable hand-object interaction denoising via denoising diffusion. In: The Twelfth International Conference on Learning Representations (2024) 3

[47] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34** (Oct 2015) 4

[48] Lu, S., Chen, L.H., Zeng, A., Lin, J., Zhang, R., Zhang, L., Shum, H.Y.: Humantomato: Text-aligned whole-body motion generation. arxiv:2310.12978 (2023) 2

[49] Luo, Z., Cao, J., Merel, J., Winkler, A., Huang, J., Kitani, K.M., Xu, W.: Universal humanoid motion representations for physics-based control. In: The Twelfth International Conference on Learning Representations (2024) 2, 3

[50] Ma, S., Cao, Q., Zhang, J., Tao, D.: Contact-aware human motion generation from textual descriptions. arXiv preprint arXiv:2403.15709 (2024) 2

[51] Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: 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

[52] Merel, J., Tunyasuvunakool, S., Ahuja, A., Tassa, Y., Hasenclever, L., Pham, V., Erez, T., Wayne, G., Heess, N.: Catch & carry: reusable neural controllers for vision-guided whole-body tasks. ACM Trans. Graph. (2020) 3

[53] Mir, A., Puig, X., Kanazawa, A., Pons-Moll, G.: Generating continual human motion in diverse 3d scenes. In: International Conference on 3D Vision (3DV) (March 2024) 2, 3

[54] Pan, L., jingbo Wang, Huang, B., Zhang, J., Wang, H., Tang, X., Wang, Y.: Synthesizing physically plausible human motions in 3d scenes (2023) 3

[55] 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) 3

[56] Petrovich, M., Litany, O., Iqbal, U., Black, M.J., Varol, G., Peng, X.B., Rempe, D.: Multi-track timeline control for text-driven 3d human motion generation. In: CVPR Workshop on Human Motion Generation (2024) 2

[57] Pi, H., Peng, S., Yang, M., Zhou, X., Bao, H.: Hierarchical generation of human-object interactions with diffusion probabilistic models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2023) 3

[58] Prokudin, S., Lassner, C., Romero, J.: 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] Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: BABEL: Bodies, action and behavior with english labels. In: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (2021) 2

[60] 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: Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 2, 3

[61] Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. arXiv preprint arXiv:2303.01418 (2023) 2, 4

[62] Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. In: The Twelfth International Conference on Learning Representations (2024) 4

[63] Starke, S., Zhang, H., Komura, T., Saito, J.: Neural state machine for character-scene interactions. ACM Trans. Graph. **38**(6) (2019) 3, 4

[64] Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briales, J., Gillingham, T., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H.M., Nardi, R.D., Goesele, M., Lovegrove, S., Newcombe, R.: The Replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019) 2, 7, 8

[65] 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

[66] Taheri, O., Zhou, Y., Tzionas, D., Zhou, Y., Ceylan, D., Pirk, S., Black, M.J.: Grip: Generating interaction poses conditioned on object and body motion. In: International Conference on 3D Vision (3DV 2024) (2024) 3

[67] Tanaka, M., Fujiwara, K.: Role-aware interaction generation from textual description. In: ICCV (2023) 2

[68] Tendulkar, P., Surís, D., Vondrick, C.: Flex: Full-body grasping without full-body grasps. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 3

[69] Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023) 2, 4, 8

[70] Wan, W., Dou, Z., Komura, T., Wang, W., Jayaraman, D., Liu, L.: Tlcontrol: Trajectory and language control for human motion synthesis. CoRR **abs/2311.17135** (2023) 2

[71] Wang, Z., Chen, Y., Jia, B., Li, P., Zhang, J., Zhang, J., Liu, T., Zhu, Y., Liang, W., Huang, S.: 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

[72] 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) 3

[73] Wu, Q., Shi, Y., Huang, X., Yu, J., Xu, L., Wang, J.: THOR: text to human-object interaction diffusion via relation intervention. CoRR **abs/2403.11208** (2024) 3

[74] Xie, Y., Jampani, V., Zhong, L., Sun, D., Jiang, H.: Omnicontrol: Control any joint at any time for human motion generation. In: The Twelfth International Conference on Learning Representations (2024) 2

[75] Xu, S., Li, Z., Wang, Y.X., Gui, L.Y.: Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In: ICCV (2023) 3

[76] Xu, S., Wang, Z., Wang, Y.X., Gui, L.Y.: Interdreamer: Zero-shot text to 3d dynamic human-object interaction. arXiv preprint arXiv:2403.19652 (2024) 3

[77] Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI (2018) 8

[78] Yang, J., Niu, X., Jiang, N., Zhang, R., Siyuan, H.: F-hoi: Toward fine-grained semantic-aligned 3d human-object interactions. European Conference on Computer Vision (2024) 3

[79] Yi, H., Huang, C.H.P., Tripathi, S., Hering, L., Thies, J., Black, M.J.: MIME: Human-aware 3D scene generation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2023) 8

[80] Yi, H., Thies, J., Black, M.J., Peng, X.B., Rempe, D.: Generating human interaction motions in scenes with text control. arXiv:2404.10685 (2024) 2, 3, 4, 5, 8

[81] Zhang, C., Liu, Y., Xing, R., Tang, B., Yi, L.: Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement. arXiv preprint arXiv:2406.19353 (2024) 3

[82] Zhang, H., Ye, Y., Shiratori, T., Komura, T.: Manipnet: neural manipulation synthesis with a hand-object spatial representation. ACM Trans. Graph. (2021) 3

[83] Zhang, J., Zhang, J., Song, Z., Shi, Z., Zhao, C., Shi, Y., Yu, J., Xu, L., Wang, J.: Hoi-m3: Capture multiple humans and objects interaction within contextual environment. arXiv preprint arXiv:2404.00299 (2024) 3

[84] 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

[85] Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In: 2023 IEEE/CVF International Conference on Computer Vision, ICCV 2

[86] Zhang, W., Dabral, R., Leimkühler, T., Golyanik, V., Habermann, M., Theobalt, C.: Roam: Robust and object-aware motion generation using neural pose descriptors. arXiv preprint arXiv:2308.12969 (2023) 3

[87] 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) (October 2022) 3

[88] Zhang, X., Bhatnagar, B.L., Starke, S., Petrov, I.A., Guzov, V., Dhamo, H., Pérez Pellitero, E., Pons-Moll, G.: Force: Dataset and method for intuitive physics guided human-object interaction. In: Arxiv. vol. 2403.11237 (2024) 3

[89] Zhang, Z., Liu, R., Aberman, K., Hanocka, R.: Tedi: Temporally-entangled diffusion for long-term motion synthesis. In: SIGGRAPH, Technical Papers (2024) 2

[90] Zhao, K., Li, G., Tang, S.: A diffusion-based autoregressive motion model for real-time text-driven motion control. In: Arxiv. vol. abs/2410.05260 (2024) 3, 8

[91] 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

[92] Zhou, W., Dou, Z., Cao, Z., Liao, Z., Wang, J., Wang, W., Liu, Y., Komura, T., Wang, W., Liu, L.: Emdm: Efficient motion diffusion model for fast, high-quality motion generation. arXiv preprint arXiv:2312.02256 (2023) 2

[93] Zhou, Y., Barnes, C., Jingwan, L., Jimei, Y., Hao, L.: On the continuity of rotation representations in neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 4

APPENDIX

1. Details on User Study

Our evaluation encompasses a human perceptual study, which is aimed at assessing both the ability of our methods to satisfy scene constraints and their controllability through text. We utilized animations derived from the HPS [21] and Matterport [8] datasets for this purpose. Each participant was presented with a set of seven questions, as illustrated in Figure 1, requiring them to perform a three-way comparison of animations. These animations were presented in a randomized order to prevent any ordering bias.

The study received 24 complete responses for the final analysis. The results were encouraging, with 75.6% of participants expressing a preference for our model over the baseline alternatives. This strong preference highlights the effectiveness of our method in generating believable human-scene interactions. Notably, our approach significantly reduces floating and penetration artifacts while promoting the generation of realistic contacts.

Overall, our user study validates the effectiveness of our method in creating visually plausible animations that adhere to scene constraints and can be manipulated through text.

2. Dataset

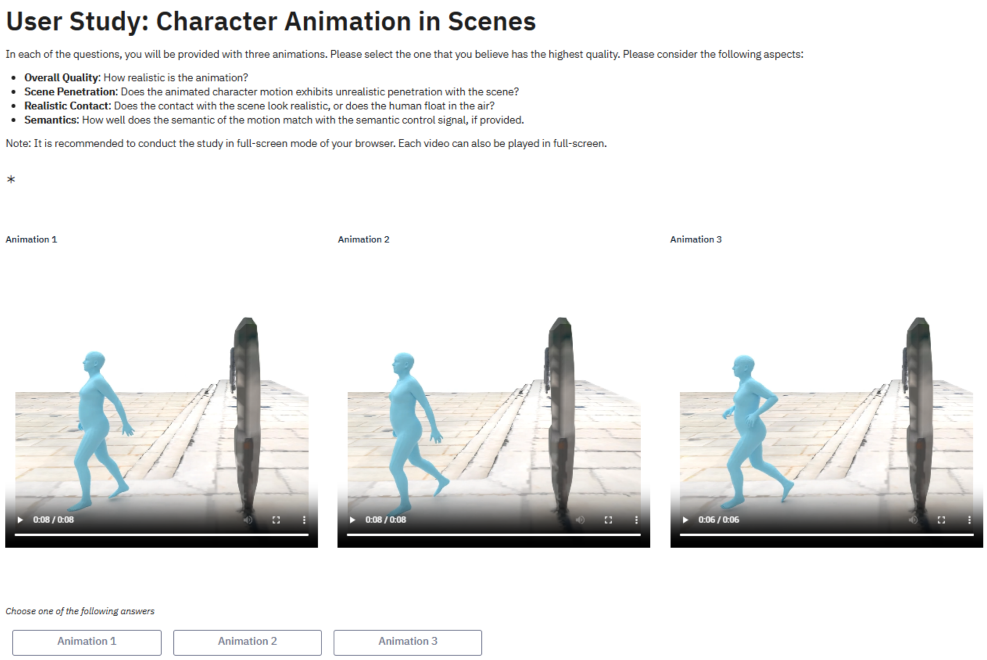
2.1. Terrain Fitting Process

Since capturing simultaneously human motion with scenes that include diverse terrains is expensive and difficult, we leverage a method that fits 2 second motion segments (60 frames) onto a set of 20,000 4x4 meters terrain patches to obtain paired motion-scene data. The terrain patches are sampled at random locations and orientations from large terrain scenes from Source Engine. By leveraging ray-tracing, the full geometric information are encapsulated in the form of heightmaps with a resolution of one pixel per inch. We then construct the patched terrain heightmaps into watertight meshes.

Having sampled the terrain meshes, the motion segments are then fitted in two main stages:

1. **Patch Selection:** Identify the three best-matching terrain patches using a brute-force search that minimizes a comprehensive error function.
2. **Terrain Refinement:** Apply a Radial Basis Function (RBF) mesh editing technique to ensure precise foot placement accuracy.

The error function E_{fit} comprises three key components: E_{contact} ensures foot height matches ground contact point. $E_{\text{penetration}}$ prevents intersection when feet are not in contact with the terrain. E_{jump}: is only activated when the character is jumping, ensuring the height of the terrain is no more than l in distance below the feet.



User Study: Character Animation in Scenes

In each of the questions, you will be provided with three animations. Please select the one that you believe has the highest quality. Please consider the following aspects:

- **Overall Quality**: How realistic is the animation?
- **Scene Penetration**: Does the animated character motion exhibits unrealistic penetration with the scene?
- **Realistic Contact**: Does the contact with the scene look realistic, or does the human float in the air?
- **Semantics**: How well does the semantic of the motion match with the semantic control signal, if provided.

Note: It is recommended to conduct the study in full-screen mode of your browser. Each video can also be played in full-screen.

*

Choose one of the following answers

Figure 6. The layout of our perceptual study for evaluating perceived realism, compliance of scene constraints, and text-based controllability of SCENIC .

$$E_{\text{fit}} = E_{\text{contact}} + E_{\text{penetration}} + E_{\text{jump}} \tag{4}$$

$$E_{\text{contact}} = \sum_{i} \sum_{j \in J} \boldsymbol{c}_j^i (\mathbf{h}_j^i - \mathbf{J}_{\text{feet},j}^i)^2 \tag{5}$$

$$E_{\text{penetration}} = \sum_{i} \sum_{j \in J} (1 - \boldsymbol{c}_j^i) \max(\mathbf{h}_j^i - \mathbf{J}_{\text{feet},j}^i, 0) \tag{6}$$

$$E_{\text{jump}} = \sum_{i} \sum_{j \in J} \mathbb{1}_{\text{jump}}^i (1 - \boldsymbol{c}_j^i) \max((\mathbf{J}_{\text{feet},j}^i - l) - \mathbf{h}_j^i, 0) \tag{7}$$

Here,

- J_{foot}: Set of joint indices (left/right heel and toe)
- c_j^i: Contact label for foot joint j at frame i
- f_j^i: Foot joint height at frame i
- h_j^i: Terrain height under foot joint at frame i
- $\mathbb{1}_{\text{jump}}^i$: Binary indicator for jumping gait
- l: Height threshold (approximately 0.3m)

After computing the fitting error for all terrain patches, we select the 3 patches with the lowest error for further processing. The motion are already well-fitted to the terrains. The further refinement stage involves editing the heightmap to ensure precise foot contact with the ground during contact phases. We use a simplified terrain deformation technique based on Botsch and Kobbelt et al. [**?**], applying a 2D Radial Basis Function (RBF) with a linear kernel to the terrain fit residuals. This approach provides a flexible method for adapting character motion to varied terrain geometries, multiplying the effectiveness of data and enables training generalizable models.

2.2. Dataset Statistics

Our dataset includes ten gait motion styles with annotated text prompts and corresponding terrain scene patches. Table 2 details the dataset's motion style distribution, encompassing various locomotion types from walking and running to more specialized movements like climbing and balancing.

Table 2. Detailed statistics of the SCENIC dataset. The dataset comprises 3 hours of motion (at 30fps), texts annotations, and fitted terrain meshes.

Gait	Minutes	%
Stand	6.88	4.09
Walk	75.95	45.17
Run	50.75	30.18
Crouch	14.06	8.36
Climb	2.30	1.37
Jump	10.01	5.95
Hop	2.54	1.51
Balance	2.69	1.60
Zombie	2.91	1.73
Push	0.07	0.04