

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

Supplementary Material

Paper ID 77

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.

*

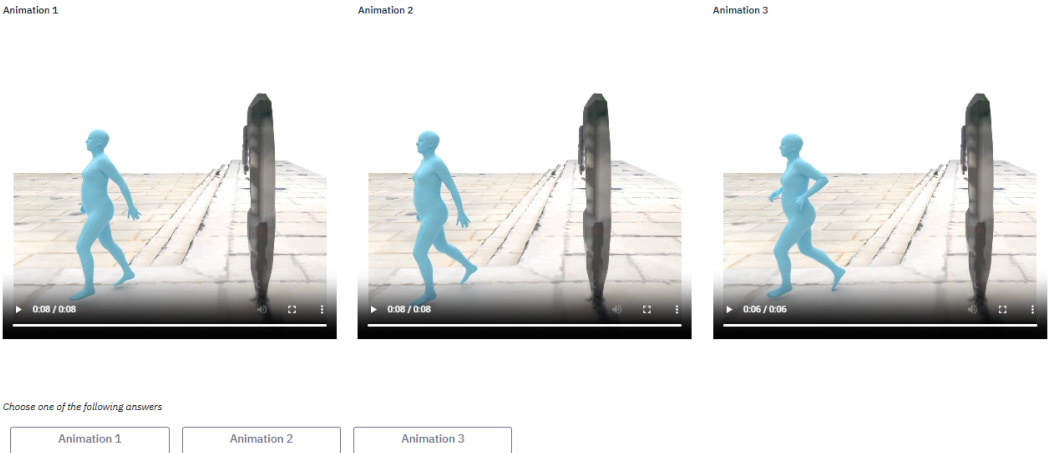


Figure 1: The layout of our perceptual study for evaluating perceived realism, compliance of scene constraints, and text-based controllability of SCENIC .

In this document, we provide details on our qualitative user study (Section 1). Subsequently, we offer explanations on the fitting process and additional statistics of our dataset (Section 2). We encourage readers to refer to our supplementary video for animated qualitative results.

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 [3] and Matterport [2] 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 compar-

ison 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.

$$E_{\text{fit}} = E_{\text{contact}} + E_{\text{penetration}} + E_{\text{jump}} \quad (1)$$

$$E_{\text{contact}} = \sum_i \sum_{j \in J} c_j^i (\mathbf{h}_j^i - \mathbf{J}_{\text{feet},j}^i)^2 \quad (2)$$

$$E_{\text{penetration}} = \sum_i \sum_{j \in J} (1 - c_j^i) \max(\mathbf{h}_j^i - \mathbf{J}_{\text{feet},j}^i, 0) \quad (3)$$

$$E_{\text{jump}} = \sum_i \sum_{j \in J} \mathbb{1}_{\text{jump}}^i (1 - c_j^i) \max((\mathbf{J}_{\text{feet},j}^i - l) - \mathbf{h}_j^i, 0) \quad (4)$$

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. [1], 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 1 details the dataset’s motion style distribution, encompassing various locomotion types from walking and running to more specialized movements like climbing and balancing.

Table 1. 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

References

- [1] Mario Botsch and Leif Kobbelt. Real-time shape editing using radial basis functions. *Comput. Graph. Forum*, 2005. 2
- [2] 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. 1
- [3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1