Synthesizing Diverse Human Motions in 3D Indoor Scenes

Appendix

A. Policy Training Environment Details

A.1. Locomotion Environments

We train the scene-aware locomotion policy using random synthetic scenes to learn generalizable locomotion skills of moving from the initial location to the goal location while avoiding collision with scenes. The initial and goal locations used for training are waypoints of collision-free paths sampled in the synthetic scenes.

Each synthetic scene has a random scene size, consists of random numbers and categories of objects sampled from ShapeNet [1], and has a random scene layout. The synthetic scenes are generated using the following steps:

- Sample the initial scene shape as a rectangle with edges ranging from 2 meters to 7 meters.
- Randomly sample furniture objects constituting the scene from ShapeNet. Specifically, we sample objects from chairs, beds, sofas, desks, and tables. We limit the number of objects belonging to categories that normally have large sizes (e.g. beds) to avoid the scenes being fully occupied, leaving no space for human movements. We use the real object size annotation of ShapNet and transform the object model to make the z-axis point up.
- Randomly rotate and translate the objects in the scene to obtain random scene layouts.
- Expand the scene boundary so that every object keeps a reasonable distance from the boundary and humans can potentially walk by.

After synthetic scene generation, we calculate the corresponding navigation mesh as described in Sec. B, and randomly sample pairs of collision-free initial-goal locations in the walkable areas. We first randomly sample two initial and goal locations on the navigation mesh. Then we use navigation mesh-based pathfinding to generate a sequence of waypoints that constitute a collision-free path. Each pair of consecutive waypoints are used as one initial-goal location pairs to train the locomotion policy. One sample synthetic scene and waypoints for training the locomotion policy are shown in Fig. S1.

We train the locomotion policy using the synthetic scenes and corresponding initial-goal location pairs. The locomotion policy is trained to move from the initial location to the goal location while avoiding penetration with scene objects. We further randomize the initial body pose and orientation to make the policy generalize to various initial body configurations.

A.2. Object Interaction Environments

We train the human-object interaction policy to reach the fine-grained body marker goals that perform the specified interaction. The static goal human-scene interaction data is the prerequisite for training the interaction policy, which we obtain from the PROX [3] dataset using the following steps:

- We obtain the static human-object interaction estimation from PROX recordings, which consist of SMPL-X body estimation from LEMO [9], and object mesh from the instance segmentation and annotation from COINS [11]. Specifically, we use the static human-scene interactions annotated as 'sit on' and 'lie on' according to COINS.
- To improve object diversity and augment the interaction data from PROX, we retarget the static interaction data to random ShapeNet [1] objects similar to the data augmentation in [6]. For each static human-object interaction data from PROX, we randomly sample an object from ShapeNet and fit it to the original PROX object by optimizing scale, rotation, and translation. After fitting, we replace the original PROX object with the fitted object. Then we augment the interaction data by applying slight scaling and rotation augmentation to the fitted object, and the corresponding human bodies are updated using contact points and relative vectors similar to [6].

With the object retargeting and augmentation, we obtain goal static human-object interaction data with increased diversity compared to the original PROX dataset. When training the object interaction policy, we randomly sample one frame of static interaction to retrieve the interaction object mesh and fine-grained goal body markers for the training environment setup. We

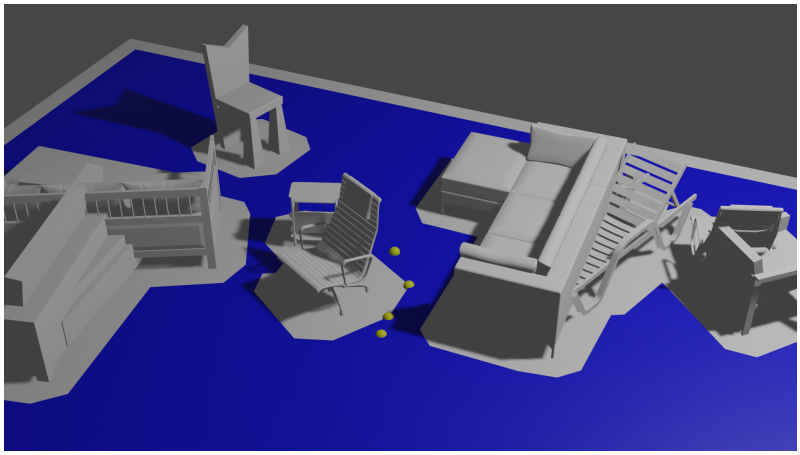
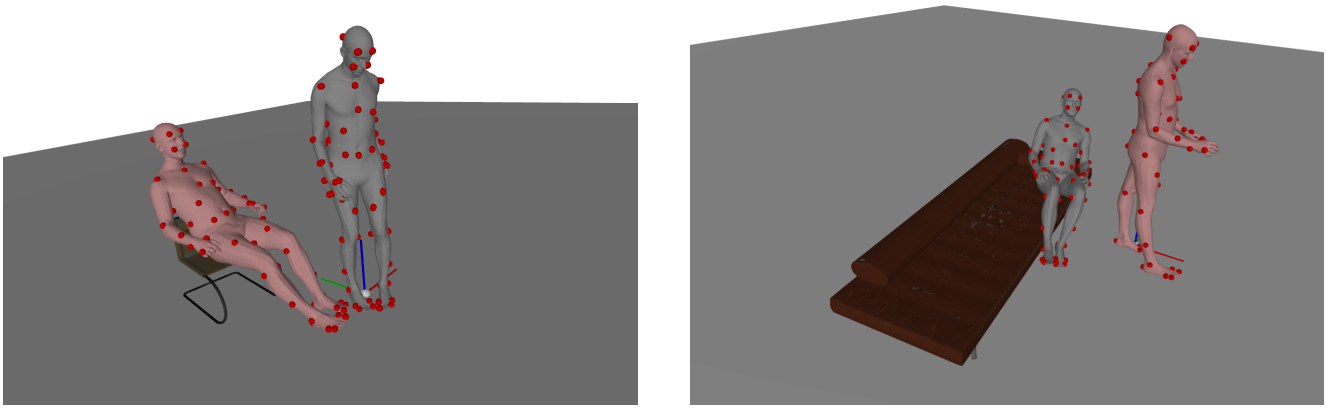


Figure S1: Illustration of a synthetic scene and sampled waypoints used to train the locomotion policy. The corresponding navigation mesh is visualized as blue, denoting the walkable areas for humans. The yellow spheres denote pair-wise collision-free waypoints found by navigation mesh-based path finding and are used to train the locomotion policy.



(a) 'Sit down' training environment

(b) 'Stand up' training environment

Figure S2: Demonstration of two example training environments (left: sit down, right: stand up) for the fine-grained human-object interaction policy. Each training scene consists of an interaction object, an initial body (gray), and a goal body (pink). The object interaction policy is trained to reach the goal interaction body while avoiding collision with the interaction object and the floor. The red spheres denote the body markers.

randomly sample the initial body location and pose in front of the interaction object. Furthermore, we randomly swap the initial body and goal body with a probability of 0.25, in order to also learn 'stand up' behaviors in addition to 'sit/ lie down' behaviors. Two example training scenes of sitting down and standing up are demonstrated in Fig. S2.

B. Implementation Details

Goal static interaction synthesis using COINS. In this paper, we use a modified version of COINS[11], incorporating slightly improved object generalization, for the synthesis of goal static human-scene interactions. COINS synthesizes static human-scene interactions conditioned on interaction semantics and object geometries. However, the original COINS models are trained on the PROX [3] dataset which contains a very limited number of object models. This restricted training object diversity constrains the object generalization capabilities of COINS models. Empirical observations reveal that generation quality deteriorates when applied to objects with domain gaps, such as CAD models from ShapeNet [1] and in-the-wild scans from ScanNet [2]. We observed that the object generalization failures are mainly due to the pelvis generation stage.

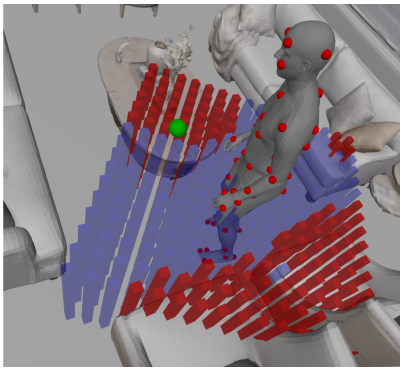


Figure S3: Illustration of the human-centric walkability map. The walkability map is a 2D occupancy map indicating which areas surrounding the human are walkable (blue cells) or occupied by obstacles (red cells). The walkability map is dynamically updated at each step according to the human body's location and orientation.

Therefore, we annotated the pelvis frame data of sitting and lying interactions on a subset of ShapeNet objects that covers more diverse objects than contained in the PROX dataset. We retrain the PelvisNet of COINS with the annotated pelvis frame data and keep the BodyNet and other sampling algorithms untouched. The goal static interactions are generated and filtered in advance of the interaction motion synthesis.

Navigation mesh and path planning. We use the open-sourced pynavmesh [4] library for navigation mesh creation and collision-free pathfinding. We implemented utility code to adapt the library to general scenes with the z-axis pointing up. We enforce the generated navigation mesh only containing triangle faces. For each scene, we create two versions of navigation meshes with different agent radii. The navigation mesh created with a larger agent radius of 0.2 is used for collision-free pathfinding, and the other one created with a smaller radius of 0.02 gives a tight fitting of the unoccupied areas and is used for the local walkability map calculation, as described in the next paragraph.

Walkability map. The walkability map is implemented as a 2D occupancy map centered at the human pelvis and aligns with the body's forward orientation. We leverage a 16x16 walkability map covering 1.6 meters by 1.6 meters square area. Each cell of the map has a binary value indicating whether this cell is walkable or occupied by obstacles. At each time step, we dynamically update this human-centric local walkability map. We first sample the 256 cells in the human-centric local coordinates frames, then transform the cell center to the scene coordinates, and evaluate each cell occupancy using the tightly fitted navigation mesh by querying whether the centroid is inside any triangle faces of the navigation mesh. One example walkability map is shown in Fig. S3.

SDF-based features. We leverage the mesh-to-sdf [5] library to calculate the marker-object signed distance and gradient features, which are used by the fine-grained human-object interaction policy. For each interaction object, we precompute a 128x128x128 SDF grid and a corresponding gradient grid. At each test step of interaction synthesis, we calculate the body marker-object distance and gradient features by evaluating the SDF and gradient grids at the current marker locations using grid sampling with trilinear interpolation. We utilize this grid sampling-based SDF feature calculation to achieve a balance between accuracy and computational efficiency.

C. Comparison to More Related Works

Here we discuss and compare with more related works on synthesizing human motions in 3D scenes. [7, 8] share a similar multi-stage motion synthesis framework of first placing anchor bodies in scenes and then generating in-between trajectories and poses, which requires the pre-specification of the total number of frames. In contrast, our method offers greater flexibility by not constraining the number of frames to generate in advance. Since [7] doesn't provide source code, we add the quantitative locomotion comparison with [8] in their test scenes as shown in Tab. S1. Our method can generate locomotion results with significantly more natural foot-floor contact and less scene collision. The generated human motion

results from [8] exhibit artifacts like obvious jittering and foot skating, as evident in the video qualitative comparison in our project website.

Table S1: Quantitative comparison with [8] on locomotion.

	time ↓	avg. dist ↓	contact ↑	loco pene ↑
Wang etc. [8]	4.00	**0.03**	0.92	0.86
Ours	**3.09**	0.04	**0.99**	**0.95**

COUCH [10] autoregressively synthesize human motions sitting to chairs satisfying given hand-chair contact constraints. However, COUCH can not handle complex scenes nor generate standing-up motion, and COUCH repeats generating deterministic motion. Nevertheless, we quantitatively compare with COUCH on the task of sitting down using their test chairs. The quantitative metrics in Tab. S2 show that our method can generate sitting interaction results with faster interaction completion, more natural foot contact, and less human-object penetration. We refer to our project website for the qualitative comparison of video results.

Table S2: Quantitative comparison with [10] on sitting interactions.

	time ↓	contact ↑	pene. mean ↓	pene. max ↓
COUCH [10]	6.47	0.91	5.24	14.14
Ours	**3.25**	**0.97**	**1.50**	**5.89**

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, and Hao Su. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1, 2

[2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 2

[3] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3D human pose ambiguities with 3D scene constraints. pages 2282–2292, 2019. 1, 2

[4] Shekn Itrch. pynavmesh: Python implementation of path finding algorithm in navigation meshes. 3

[5] Marian Kleineberg. mesh-to-sdf: Calculate signed distance fields for arbitrary meshes. 3

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

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

[8] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. pages 9401–9411, 2021. 3, 4

[9] Siwei Zhang, Yan Zhang, Federica Bogo, Pollefeys Marc, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *International Conference on Computer Vision (ICCV)*, Oct. 2021. 1

[10] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: towards controllable human-chair interactions. pages 518–535. Springer, 2022. 4

[11] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. pages 311–327. Springer, 2022. 1, 2