Synthesizing Long-Term 3D Human Motion and Interaction in 3D Scenes

Jiashun Wang
UC San Diego

Huazhe Xu
UC Berkeley

Jingwei Xu
Shanghai Jiao Tong University

Sifei Liu
NVIDIA

Xiaolong Wang
UC San Diego

Figure 1: Two examples of our generated long-term 3D motion in the 3D scene. *Left*: A human walks around the furniture in the room; *Right*: A human walks in the hallway and then sits down on the bench.

Abstract

Synthesizing 3D human motion plays an important role in many graphics applications as well as understanding human activity. While many efforts have been made on generating realistic and natural human motion, most approaches neglect the importance of modeling human-scene interactions and affordance. On the other hand, affordance reasoning (e.g., standing on the floor or sitting on the chair) has mainly been studied with static human pose and gestures, and it has rarely been addressed with human motion. In this paper, we propose to bridge human motion synthesis and scene affordance reasoning. We present a hierarchical generative framework to synthesize long-term 3D human motion conditioning on the 3D scene structure. Building on this framework, we further enforce multiple geometry constraints between the human mesh and scene point clouds via optimization to improve realistic synthesis. Our experiments show significant improvements over previous approaches on generating natural and physically plausible human motion in a scene.[1]

1. Introduction

Capturing and synthesizing realistic human motion in 3D scenes has played an essential role in various applications in virtual reality, video game animations, and human-robot interactions. As shown in Fig. 1, given the 3D scenes, our goal is to generate long-term human motion and interaction in the scene, such as walking around the room avoiding collision with the furniture (left), as well as walking through the hall way, turning around and then sitting down (right). To achieve this, there are two main challenges on: (i) generating realistic motion in long-term; (ii) modeling human-scene interaction and affordance.

Recent works have made substantial efforts on human motion synthesis, which generates visually appealing and natural pose sequences using optimization-based statistical models [64, 5], or deep neural networks [23, 22, 65]. However, while focusing on realistic motion, these works rarely address the interactions between the human and the scene. On the other hand, a line of researches on 3D scene affordance [61, 34, 67] have studied on the "opportunities for interactions" [13] in the scene. For example, Wang *et al.* [61] propose to learn to predict human skeletons from an empty scene by training with a large-scale sitcom dataset. While focusing on the scene context, these approaches are only able to generate a single static human pose.

In this paper, we intend to bridge human motion synthesis and affordance learning. We consider a novel problem setting: *Given the start and the end positions far away in a 3D scene, synthesize the human motion moving in between.* To generate long-term motion in the scene, instead

[1]Project page: https://jiashunwang.github.io/Long-term-Motion-in-3D-Scenes

of synthesizing a long route of poses at one time, we introduce a 2-level hierarchical framework: (i) we first set several sub-goal positions between the start and the end locations. We predict the human pose for each sub-goal, start, and end positions, conditioning on the 3D scene context. (ii) we synthesize the short-term human motion between every two sub-goals, using the predicted poses on the sub-goals as well as the 3D scene as inputs. The short-term motion will be then connected together for the final long-term motion synthesis.

We model the interaction between human and the scene in both stages of our framework. Instead of using human skeletons, we emphasize that we adopt the differentiable SMPL-X [44] model for representing both the shape and the pose of the human, which allows more flexible geometry constraints and more realistic modeling of contacts. Specifically, in the first stage, given a single sub-goal in a 3D scene, we utilize a Conditional Variational Autoencoder (CVAE) [53] to generate the SMPL-X parameters. In the second stage for short-term motion generation, we use a bi-directional LSTM [20] which takes the start-end human SMPL-X representations and 3D scene representation as inputs, and generates a sequence of human bodies represented by SMPL-X. Besides training the deep models in a data-driven manner for motion synthesis, we also perform explicit geometry reasoning between the human mesh and 3D scene point clouds by optimization. Our optimization approach considers both the naturalness of motion and the physical collisions with the environment. By unifying both learning based and optimization based techniques, we are able to synthesize realistic human motion in long-term.

We perform our experiments on both the PROX [17] and the MP3D [7] 3D environments. By considering 3D scene affordance and structural constraints in motion synthesis, we qualitatively show realistic and physically plausible human motion generation results. We also show large advantage quantitatively against state-of-the-art motion and human pose generation approaches, using multiple metrics and human evaluation.

Our contributions in this paper include: (i) A hierarchical learning framework for motion synthesis considering both the realism of the motion and the affordance of the scene; (ii) An optimization process to explicitly improve the synthesized human poses; (iii) state-of-the-art motion synthesis results on various 3D environments.

2. Related work

Affordance learning. Learning scene affordance has captured a lot attention in recent years [11, 62, 56, 42, 9, 27, 10, 15, 31, 34, 52, 71, 70, 41, 51, 8, 67, 66, 6]. One paradigm to study the scene affordance is to understand how to put a human skeleton in a scene [56, 34, 67, 66, 56, 42]. For example, Tan *et al.* [56] predict where to put a human in a given image and search a person that fits the scene from a database. Li *et al.* [34] introduce a 3D pose generative model to predict physically feasible human poses in a given scene. Recently, due to more refined needs and the development of 3d human representations [38, 49, 44], more researches start to study how to place a human body shape in a 3d scene instead of skeleton. For example, Zhang *et al.* [67] use a two-stage CVAE [53] to generate plausible 3D human bodies that are posed naturally in 3D scene. However, most studies only focus on the single static body generation and they have barely addressed the problem of generating physically plausible motion in the scene.

Human dynamics prediction. Our work is related to the research in modeling and predicting the human dynamics. Both research in trajectory prediction [18, 57, 30, 2, 21, 14, 50, 3, 39, 54, 6] and pose prediction [59, 68, 60, 26, 24, 33, 12, 19, 40, 45] are raising a lot of attention in recent years. Instead of isolating the environments from prediction, researchers have studied on using 3D information or bird-eye view image to predict future human dynamics [14, 50, 3, 39, 54]. To study on how the surroundings would influence the human dynamics, Helbing *et al.* [18] use physical forces to model social-scene interaction for pedestrian dynamics. Cao *et al.* [6] consider scene context when predicting the goal, path and poses of human movement given the scene image and past 2D pose histories as inputs. While these studies make great contribution to the human dynamics and skeleton-based pose prediction, they pay little attention on the physical and geometric interaction between the human and the scene for realistic generation. On the contrary, our paper focuses on synthesizing natural human pose and shape when interacting with the scene. We also want to emphasize we are not performing future prediction since our goal is given in the task.

Motion synthesis. The main focus of our work is on motion synthesis, which has been long standing problem in computer graphics and vision [5, 32, 43, 55, 23, 35, 46, 47, 65]. For example, Kovar *et al.* [32] introduce a novel data structure called a registration curve that expands the class of motions that can be successfully blended without manual input. Pavllo *et al.* [46] represent rotations with quaternions and uses a sequentially network and a novel loss function to perform forward kinematics on a skeleton to penalize absolute position errors instead of angle errors. There are also studies [58, 63, 16, 37] focusing on synthesizing intermediate states between the given key frames. But these works can only synthesize a transient motion with a small position change. Xu *et al.* [65] proposes a hierarchical way to generate long-term motion by using a memory bank to retrieve short-term motion references and a bi-directional interpolation to connect the short-term motions. However, most of the previous studies in motion synthesis lack consideration of the surroundings. This may cause problems when apply-

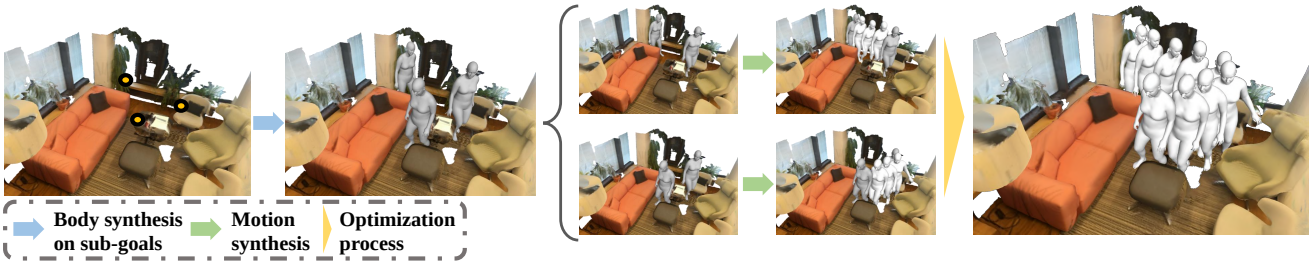


Figure 2: **Framework of our long-term motion synthesis.** We first generate the sub-goal bodies with the given $\{\beta, t, r\}$ as inputs. Sub-goal bodies are in gray color. Then we divide it into several short-term start/end pairs and synthesize short-term motion each. Finally we use an optimization process to connect all these short-term motion to a long-term motion. Generated motion is in white.

ing to the real world environment where the opportunities for interaction are rich. In this work, we enable long-term motion generation in complex scenes. To better model the interaction, we use the representation of human mesh instead of human skeleton during motion synthesis.

3. Method

3.1. Overview

Representation. We denote scene mesh as $S = (v^s, f^s)$ where v^s representes the vertices and f^s represents the faces. Instead of using skeleton-based representation, we use modified sequential SMPL-X parameters [44] to represent human bodies. Concretely, we define $M_i = \mathcal{M}(t_i, r_i, \beta, p_i, h_i)$, where $t \in \mathbb{R}^3$ is the translation, 6d continuous rotation $r \in \mathbb{R}^6$ [69] is used to replace the original global orientation in SMPL-X, $\beta \in \mathbb{R}^{10}$ is the body shape, $p \in \mathbb{R}^{32}$ represents the body pose [44] and $h \in \mathbb{R}^{24}$ represents the hand pose.

Problem definition. In this paper, We design a two-level hierarchical framework to generate long-term 3D motion in the 3D scene. Our method takes the inputs of the start, end and sub-goal positions and orientations. The sub-goals divide the higher-level long trajectory into lower-level short paths. We first generate the human bodies on these given positions with given shape β, using a Conditional Variational Autoencoder (CVAE) [53]. We give β to control the body shape of the motion with more diversity. Given the human body estimation results on each sub-goal, we aim to generate plausible motion between every two sub-goals, which will then be connected together to a long-term motion.

Concretely, we propose a motion synthesis network to generate short-term motion. Given the start body mesh $M_0 = \mathcal{M}(t_0, r_0, \beta, p_0, h_0)$, the end body mesh $M_k = \mathcal{M}(t_k, r_k, \beta, p_k, h_k)$ estimated by the CVAE and the scene mesh S as inputs, the motion synthesis network will generate a motion sequence $M_{1:k-1}$ between the start and end bodies. Multiple short-term motion sequence $M_{1:k-1}$ will be connected together to form a long-term motion. We assume the length of each short-term motion is $k+1$ steps, where we set $k = 61$ (2 seconds in a frame rate of 30fps) in our experiments. Finally, we adopt a geometric optimization process to further enforce realistic and physically plausible synthesis.

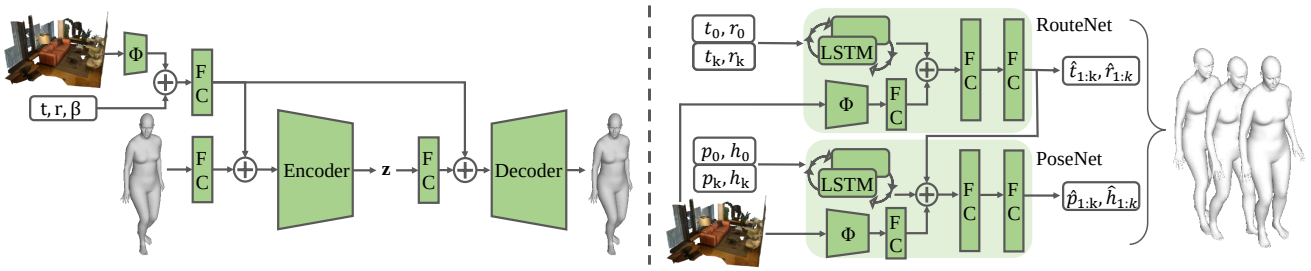
The benefits of our two-level generation approach lie in two folds: (i) First generating the poses on the sub-goals, and then generating the motion in between allows the short-term motion easier to connect, since the ending pose of one sequence will be same as the starting pose of the next sequence; (ii) By sampling the latent variables in CVAE, we can generate diverse human poses on the sub-goals, which leads to diverse motion in the long-term trajectory.

3.2. Static Human Body Synthesis on Sub-Goals

We propose to use a Conditional Variational Autoencoder (CVAE) for generating bodies on each sub-goal, given the inputs of $\{\beta, t, r\}$ presented in the sub-goal and the scene point cloud v^s, as shown in Fig. 3 (a). During training time, we first extract the feature for the scene point cloud as $F^s = \Phi(v^s)$ using a PointNet Φ [48]. We concatenate it with the shape, location and orientation inputs $\{\beta, t, r\}$ and forward them to a fully connected layer to obtain the integrated feature F^{hs}. This feature is the conditional feature for the CVAE. The CVAE is presented with an encoder and a decoder as follows.

Encoder. We forward the human body parameters M_0 to two residual blocks containing two fully connected layers each. The output is then concatenated with F^{hs}, followed by two fully connect layers to predict the mean $\mu \in \mathbb{R}^{32}$ and variance $\sigma^2 \in \mathbb{R}^{32}$ of a Gaussian distribution $Q(z|\mu, \sigma^2)$. We sample the latent code z from this distribution as one of the decoder inputs.

Decoder. We concatenate the latent code z with the conditional feature F^{hs} as the input for the decoder, which is another two residual blocks containing two fully connected layers each. The output of the decoder predicts the desired human body parameters, which is trained computing the reconstruction loss against the ground-truth human mesh M_0. Following [29], the training objective also includes maxi-



(a) Human Body Synthesis with CVAE

(b) Motion Synthesis Networks

Figure 3: **Network architectures.** (a) shows the architecture of our static human body synthesis network. (b) is the architecture of our motion synthesis networks. $\bigoplus$ means concatenation.

mizing the KL-Divergence between $Q(z|\mu, \sigma^2)$ and a standard Gaussian distribution $\mathcal{N}(0, I^2)$.

During inference, only the decoder is adopted. We sample the latent code from $\mathcal{N}(0, I^2)$ and concatenate it with F^{hs} as inputs. By sampling different z, we can generate different human bodies in the same sub-goal position.

Implementation details for PointNet. We modify the PointNet architecture for affordance prediction. We remove the transform architecture to better fit our task of generating a body mesh in a specified location, since applying transformation to the scene may cause the mismatch of human and scene coordinates. We pretrain the PointNet using the S3DIS [4] dataset with a segmentation task. We then take the encoder of the PointNet as Φ and the output is a 256-dimension feature for point cloud representation.

To train the CVAE for static human body synthesis, we adopt the standard reconstruction loss and the KL divergence loss proposed in [53], following by the contact and collision constraints proposed in [67].

3.3. Motion Synthesis Framework

We propose two sequentially connected networks in the motion synthesis framework: A RouteNet $\mathcal{R}$ for predicting the locations and orientations of the route between two sub-goals, and a PoseNet $\mathcal{P}$ for predicting the human body pose on each locations of the route.

As shown in the top row of Fig. 3 (b), the RouteNet $\mathcal{R}$ takes the start $\{t_0, r_0\}$ and the end $\{t_k, r_k\}$ locations and orientations, as well as the scene point cloud v_s as inputs, and generates the route in between as,

$$\widehat{t}_{1:k-1}, \widehat{r}_{1:k-1} = \mathcal{R}(t_0, r_0, t_k, r_k, v^s) \tag{1}$$

where $\{\widehat{t}_{1:k-1}, \widehat{r}_{1:k-1}\}$ represents the route locations and orientations from step 1 to $k-1$. Concretely, to extract the feature for of the scene, we utilize a PointNet given v_s as inputs. Note that the PointNet feature extractor is not shared with the CVAE in the previous section. We denote the point cloud feature as F^{rs}. To connect the start and end locations, we utilize a bi-directional LSTM [20] which takes the start $\{t_0, r_0\}$ and the end $\{t_k, r_k\}$ as inputs and outputs the features for each time step in between. We concatenate the features from all time steps as well as the point cloud feature F^{rs} together. The output is then forwarded to two fully connected layers for predicting the route $\{\widehat{t}_{1:k-1}, \widehat{r}_{1:k-1}\}$.

We illustrate the PoseNet $\mathcal{P}$ as the bottom row of Fig. 3 (b). It takes the start pose $\{p_0, h_0\}$, the end pose $\{p_k, h_k\}$, point cloud v^s and the predicted route $\{\widehat{t}_{1:k-1}, \widehat{r}_{1:k-1}\}$ from the RouteNet as inputs. The outputs of the PoseNet are the pose parameters for the input route as,

$$\widehat{p}_{1:k-1}, \widehat{h}_{1:k-1} = \mathcal{P}(p_0, h_0, p_k, h_k, \widehat{t}_{1:k-1}, \widehat{r}_{1:k-1}, v^s) \tag{2}$$

where $\{\widehat{p}_{1:k-1}, \widehat{h}_{1:k-1}\}$ are the body pose and hand pose parameters from step 1 to $k-1$. Similar to the RouteNet, we use another PointNet to extract the feature for v^s as F^{ps}. The start and end pose parameters $\{p_0, h_0\}$ and $\{p_k, h_k\}$ are also forwarded to another bi-directional LSTM, which predicts the pose features between the two locations. We concatenate the pose features from step 1 to $k-1$ and F^{ps} together. We then forward the feature to two fully connected layers for generating the a sequence of pose parameters $\{\widehat{p}_{1:k-1}, \widehat{h}_{1:k-1}\}$. Finally, we combine the predicted SMPL-X [44] parameters $\{t_{1:k}, r_{1:k}, p_{1:k}, h_{1:k}\}$ and the given β to generate the mesh sequence $M_{1:k}$.

We compute the training losses for both the RouteNet and the PoseNet with the L_1 distance between the predictions and the ground-truth location and pose parameters. Specifically, we define the RouteNet loss $\mathcal{L}_{route}$ and the PoseNet loss $\mathcal{L}_{pose}$ as,

$$\mathcal{L}_{route} = \lambda_t \sum_{i=1}^{k-1} |\widehat{t}_i - t_i| + \lambda_r \sum_{i=1}^{k-1} |\widehat{r}_i - r_i| \tag{3}$$

$$\mathcal{L}_{pose} = \lambda_p \sum_{i=1}^{k-1} |\widehat{p}_i - p_i| + \lambda_h \sum_{i=1}^{k-1} |\widehat{h}_i - h_i| \tag{4}$$

where λ_t, λ_r, λ_p and λ_h are constant coefficients to balance the loss. During training, we train the RouteNet first and then fix its weight for training the PoseNet.

3.4. Optimization

We perform optimization with geometric and physics constraints to improve the generation results from the mo-

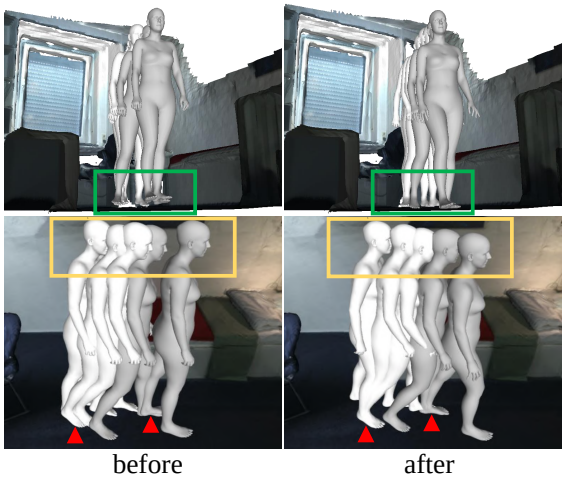


Figure 4: An example of human motion before and after optimization in two views. The green box shows that after optimization, the motion sequence is contacting with the floor. The red arrow points at the pivot foot. White human bodies are from the first sub-sequence, gray bodies are from the second, it can be seen that in each sub-sequence, the pivot foot is more stable after optimization. Yellow box shows the effectiveness of vertices smoothing.

tion synthesis networks and help connect the short-term motions to a long-term motion at the same time. We will introduce three types of constraints in the following, including the constraint on the foot, contact constraints, and the smoothness of the motion. For simplicity, we will take the generated mesh $M_{1:k-1}$ as an example for explaining our approach. While in our experiments, we apply the optimization approach on the whole long-term sequence.

Inputs and variables.

The scene point cloud is provided to calculate the contact and we have the signed distance field [17] of the scene mesh surface to calculate the collision. All the SMPL-X parameters are given and our goal is to optimize the translation t, global orientation r, body pose p and also hand pose h.

Foot location constraints. While most previous works [36, 65, 1] assume the floor is flat and also have information on which foot should be fixed on the floor at each time step, our approach does not make any of such assumptions. This allows the generated motion to be more diverse and natural *e.g.* sitting down or jumping on the bed. However, we still need to constrain the stableness of the foot motion.

As the human is moving in the room, in each moment, we assume one foot is stable and the other is moving. Thus we aim to separate the motion into multiple sub-sequences, where each sub-sequence has the same foot stable and this foot switches between the sub-sequences. We get the sub-sequences utilizing the nearer foot between two frames from networks' outputs: nearer one should be the stable one. We compute the average location of the stable foot in each sub-sequences, and we encourage the foot to be close to this average location since it should stand still. We use the ℓ_2 distance for computing this error as E_{foot} for adjusting the foot location.

Environment constraints. We also consider the physical plausibility between the human and the environment during optimization. The constraint design here is motivated by [67, 17]. On one hand, we constrain the human mesh to avoid collision with the scene; On the other hand, we also encourage the human mesh to get close to the scene for physical support.

For collision constraint, we utilize the negative signed distance field of the scene $\Psi_s^-(\cdot)$, where we constrain the human mesh from intersecting with the scene 3D surfaces, we can represent the collision error as,

$$E_{col} = \sum_{i=1}^{k-1} \mathbb{E}(|\Psi_s^-(M_i)|), \tag{5}$$

where $|\Psi_s^-(M_i)|$ computes the negative signed distance values of the body vertices, $\mathbb{E}$ represents an average function. The goal is to minimize the loss so that the the body stays on the positive level-set of the signed distance field.

We also encourage the human body to contact the scene. Our goal is to minimize E_{cont} as below,

$$E_{cont} = \sum_{i=1}^{k-1} \sum_{v' \in v_i^c} \min_{v'' \in v^s} \rho(|v' - v''|) \tag{6}$$

where v_i^c denotes a set of the predefined [17] body vertices which are encouraged to contact with the scene vertices. The Geman-McClure error function $\rho(\cdot)$ is used to down-weight the scene vertices which are far from the human. **Motion smoothness.** We encourage the human mesh nearby in time to be smooth. To achieve this, we define the smoothness constraint to be minimizing the L_2 distance between the mesh vertices as,

$$E_{smooth} = \sum_{i=1}^{k-2} \left\| v^{M_i} - v^{M_{i+1}} \right\|_2 \tag{7}$$

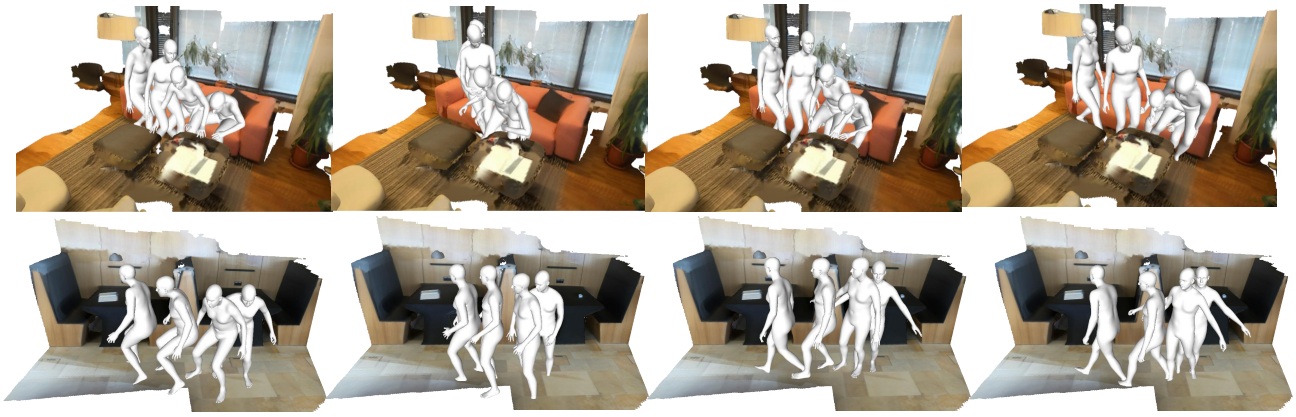
where v^{M_i} is the vertices of body M_i, and we perform this constraint between every two consecutive steps.

During optimization, we combine all the three error terms together as $\lambda_{foot}E_{foot} + \lambda_{col}E_{col} + \lambda_{cont}E_{cont} + \lambda_{smooth}E_{smooth}$. We perform gradient descent with Adam [28] to optimize the predicted mesh parameters directly. We show the effects of our optimization in Fig. 4.

4. Experiments

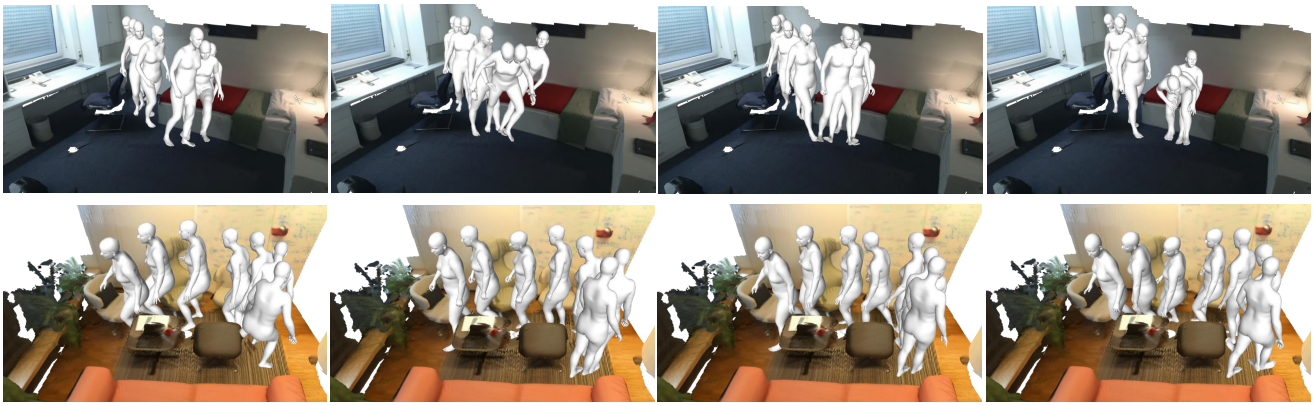
4.1. Implementation details

We use Adam [28] as the optimizer for our networks and also optimization process. To train our CVAE generation network, we use 0.001 as the learning rate (lr) with a batch



(a) Route+CVAE (b) HSNMS (c) Ours (d) Pseudo-ground truth

Figure 5: Comparisons on generating 2-second motion given the same inputs: (a) the result of Route+CVAE [67]; (b) the result of HSNMS [65]; (c) our result; (d) the pseudo-ground truth. We show that our results have more natural motions and a more consistent sitting down motion. (b) and (d) have shown cases of foot inside the ground. The poses in (a) are not natural.



(a) Route+CVAE (b) HSNMS (c) Ours (d) Pseudo-ground truth

Figure 6: Comparisons on generating 4-second motion given the same inputs over different methods: While our method shows natural and physically plausible motions, (b) has the problem in generating suitable poses in the first row, and the motion in (a) does not look natural.

size of 16 and training epochs of 40. For RouteNet the lr is 0.001. We train 20 epochs with a batch size of 32. We train PoseNet with 0.001 lr and a batch size of 16 with 20 epochs. We set $\lambda_t = 1$, $\lambda_r = 1$, $\lambda_p = 1$ and λ_h is chosen as 0.1 to have better empirical results. Our short-term motion lasts for 2 seconds and in 30 fps, which means $k = 61$ in our experiments. Please see more implementation details in Appendix 2.

4.2. Datasets

We use the PROX dataset [17] for training where the ground-truth SMPL-X[44] parameters of the human motion are generated by a fitting algorithm, which we denote as **pseudo-ground truth (p-gt)**. For training CVAE body synthesis network, we sample 4.3k bodies from 8 scenes as the training data. For training motion synthesis networks, we sample 10k two-second frames whose distance between start and end larger than 0.5 meters in 8 scenes as the training data. For evaluation, we first test our short-term motion synthesis networks using 3k sequences in 4 unseen scenes using standard metrics introduced in the next section. We then perform human evaluation for which we generate 50 motion sequences each last for 2s, 4s and 6s. We also investigate the generalization ability of the models trained with the PROX dataset by directly testing them on the MP3D dataset [7], using the same metrics.

4.3. Evaluation Metrics

Reconstruction error metrics. We introduce the standard metrics for evaluating the short-term motion: Given the p-gt start and end human bodies, we aim to generate motion in between close to the p-gt. We first compare the reconstruction error by computing the ℓ_1 distance (reported in $\times 100$) on translation t, global orientation r, pose parameters p (of SMPL-X parameters) between the predictions and the p-gt. We also use MPJPE [25] and Mean Per Vertices Position Error (MPVPE) both in millimeters to measure the mean distance between the predictions and the p-gt. In addition, start/end sides of the generated short-term motion sequence should be close to the input start/end bodies: We

method	scene	transl	orientation	pose	MPJPE	MPVPE
RouteNet(t)+PoseNet	×	7.54	11.65	46.45	212.5	201.5
RouteNet(t)+PoseNet	✓	7.31	11.82	46.73	213.1	202.1
RouteNet(t+r)+PoseNet	×	7.58	**9.38**	44.16	201.6	190.5
RouteNet(t+r)+PoseNet	✓	**6.91**	9.71	44.89	195.8	184.9
BaseNet	×	9.33	10.61	44.09	235.8	227.1
BaseNet	✓	9.02	10.68	43.69	226.2	217.6
Route+CVAE [67]	✓	8.47	10.03	59.16	294.2	278.2
HSNMS [65]	×	10.01	13.72	63.41	293.5	275.2
Ours w/o opt	✓	**6.91**	9.71	**41.17**	**191.6**	**180.9**
Route+CVAE [67] w/ opt	✓	9.60	9.97	61.65	311.8	293.8
HSNMS [65] w/ opt	✓	10.39	14.32	69.62	285.8	270.0
Ours	✓	**8.06**	**9.53**	**46.68**	**219.1**	**205.4**

Table 1: **Results of accuracy comparison.** This table shows the accuracy comparison with ablation methods and baselines. Best results are shown in boldface. ✓ in scene means this method has scene information and × means not. w/ opt means including the optimization process and w/o means not.

method	non-collision		contact	
	PROX	MP3D	PROX	MP3D
Route+CVAE	98.57	95.19	88.68	91.36
HSNMS [65]	96.88	95.29	99.33	90.12
Ours w/o opt	97.53	95.33	99.32	92.59
Route+CVAE [67] w/ opt	99.88	99.51	99.17	95.06
HSNMS [65] w/ opt	99.88	**99.57**	99.22	95.53
Ours	**99.91**	99.30	**99.35**	**97.53**
p-gt	98.08	-	99.98	-

Table 2: **Results of the in-environment evaluation.** We use the non-collision score and modified contact score to evaluate how reasonable the generated motion is in the given scene. Best results are shown in boldface. It can be seen our method performs nearly the best in different scenes and our optimization process can help to improve the environmental adaptability of the baselines.

calculate the ℓ_2 distance between the input and generated start/end bodies. We refer this metric as **neighbour v2v distance** (v2v is short for vertice to vertice). Reducing this error pushes the end of one short motion sequence close to the start of next short motion, which leads to better connections between two short motions for long-term motion synthesis.

Naturalness metrics. We perform naturalness evaluation for both short and long-term motion. We perform naturalness evaluation in two aspects: in-environment evaluation and human evaluation. For in-environment evaluation, we use the non-collision score [67] and a modified contact score. We use Amazon Mechanical Turk(AMT) for human evaluation. For each task in AMT, we give a group of comparison given the same inputs and ask users to score from 1 to 5 and higher score means more natural. Please see more details on human evaluation metric in Appendix 4.

4.4. Baselines

We compare our approach with state-of-the-art baselines as well as the ablative variants of our method. Note the previous motion synthesis works lack environment information and contact insensitive representation. For affordance prediction, previous approaches focus more on static or one-frame body generation. Thus, instead of directly comparing to previous approaches, we improve them to create stronger baselines. All baselines and our method use the modified SMPL-X representation. We introduce each baseline as following.

Route+CVAE. Zhang et al. [67] propose to use a CVAE to synthesize a single body given an image. However, this approach does not consider the motion information. Thus we provide the route using our RouteNet for CVAE to generate a sequence of human bodies. Specifically, we provide $\{t, r\}$ from RouteNet and adding scene condition v_s, body shape β, translation t and rotation r to CVAE. We call this baseline Route+CVAE. We apply optimization to it to improve the continuity.

HSNMS [65] is a state-of-the-art motion synthesis method combining both advantage of data-driven and neural network. For short-term motion, it searches the most similar motion sequence from a memory bank. Since this method lacks consideration of environment, we also apply our optimization process to improve this approach.

Ablative baselines. We compare with ablative baselines focusing on accuracy. We try to prove that PoseNet would perform better when having the predicted route as input so we design RouteNet+PoseNet which means the PoseNet separately generates the pose without given predicted route. We want to explore whether the rotation should be predicted in the route or pose network. RouteNet(t) means it only predicts t and RouteNet(t+r) means it predicts t and r; other parameters are generated in PoseNet. We also try an end-to-end neural network to predict route and pose jointly at one time and we name this network **BaseNet**. We also explore the effectiveness of scene information, thus for each ablative baseline there is a version without scene information as inputs.

4.5. Evaluation on reconstruction error

Comparison with baselines. As shown in Tab. 1, our method performs better in the accuracy of synthesis both in locations and poses than the baselines. Though the optimization process would slightly increase the error, it would largely improve the naturalness performance (Fig. 4). It can be seen our method has the lowest neighbour v2v distance from Tab. 4, which is easier to connect to long-term motion.

Comparison with ablative baselines. As shown in Tab. 1, we prove that predicted route as inputs would improve the performance of PoseNet and it is more accurate if we use RouteNet to predict t and r. According to the results of BaseNet, directly using one network is not as good as our architecture. Finally, all the results with scene information would improve the accuracy of synthesis.

method	2 seconds		4 seconds		6 seconds	
	PROX	MP3D	PROX	MP3D	PROX	MP3D
Route+CVAE [67] w/ opt	2.23±1.10	3.06±1.07	2.75±0.98	2.94±0.99	2.83±1.06	3.19±0.95
HSNMS [65] w/ opt	2.90±1.09	3.21±0.97	3.26±1.02	3.14±0.99	3.22±1.08	3.13±1.05
Ours	**3.39±1.11**	**3.33±0.90**	**3.55±0.88**	**3.22±1.05**	**3.65±0.96**	**3.30±0.96**
p-gt	3.60±1.10	-	3.76±0.97	-	3.92±0.91	-

Table 3: **Results of the human evaluation.** We show human evaluation for motions last for 2 seconds, 4 seconds and 6 seconds. We provide the average human evaluated score(1-5) w.r.t. the average ± the standard deviation. Best results are shown in boldface.

method	neighbour v2v distance	
	w/ opt	w/o opt
Route+CVAE [67]	21.78	10.59
HSNMS [65]	21.05	10.84
Ours	**8.25**	**7.31**
p-gt	3.59	-

Table 4: **Neighbour v2v distance with and without optimization.** Best results are shown in boldface.

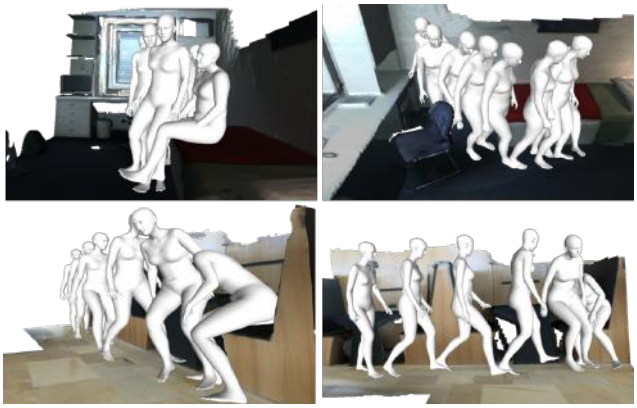


Figure 7: Two pairs of results in two views in PROX [17] scenes. Both rows show a human walking and then sitting down.

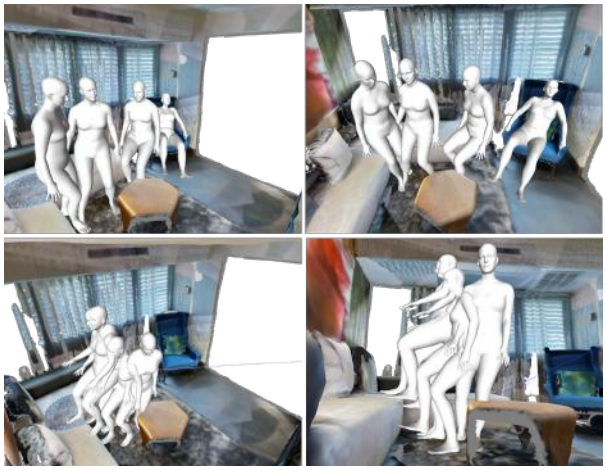


Figure 8: Two pairs of our results in two views in scene from MP3D [7]. The first row shows a human standing up and walking and the second row shows a human jumping on the sofa.

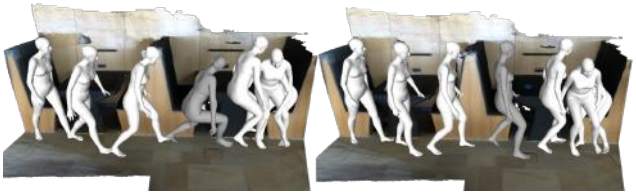


Figure 9: We show a pair of example with different generated sub goal bodies as the inputs.

4.6. Evaluation on naturalness

We firstly show qualitative naturalness comparison for 2-second motion in Fig. 5, and 4-second motion in Fig. 6. Our method can produce motions close to the real record, with less unreasonable motion and better human-scene interaction. We provide two pairs of our results in two views in Fig. 7. We also do in-environment naturalness evaluation between these methods. As shown in Tab. 2, our method has nearly the best performance in both contact and avoiding collision. Not only that, our optimization process can largely improve baselines' performance on generating physically plausible motion.

For human evaluation, we compare our approach with the other two baselines and the p-gt from PROX [17] dataset. For fair comparison, the start, end and sub-goal bodies are from PROX [17]. Since the p-gt is generated by fitting, it would also have implausible motion sequences. As shown in Tab. 3, our method has the highest score compared to the other two baselines for 2s, 4s and 6s motion and the scores are close to the p-gt, which shows our methods can generate realistic and natural motions. Through the experiments, we prove that compared to previous affordance learning method, our sequentially networks has a better performance in motion continuity. Previous motion-based method which heavily relies on existing data, may have problems in handling the environment.

Generalization on MP3D [7] dataset. We evaluate naturalness performance with our trained model directly on MP3D. We randomly sample the input shape, positions and orientations for the start/end and sub-goals and generate plausible bodies using our body synthesis model. We apply our motion synthesis networks to the input bodies and then optimize the whole motion sequence. The input start/end and sub-goal bodies for baselines are the same. We show in-environment evaluation in Tab. 2 and human evaluation in Tab. 3. Since the positions and orientations are randomly selected, there would exist some challenging cases, such as jumping on the sofa. We also provide our results shown in two views in Fig. 8. Generally, our method can generate the most realistic results with best environment adaptability.

4.7. Diversity

One advantage of our hierarchical framework is we can control the shape, positions, gestures of the sub-goal bodies (including start/end), which makes our long-term mo-

tion with more diversity. We show an example in Fig. 9, the gray color body is the generated sub-goal body. Since we can generate different sub-goal bodies, the motion sequence would look differently as what we want, *e.g*. jumping on the sofa. This would enhance the interaction between human and scene and generate motions with more diversity.

5. Conclusion

In this paper, we propose a novel hierarchical generative framework to synthesize long-term human motion in the 3D scene. We generate the sub-goal bodies and short-term motions with two deep models both designed with human-scene interaction. We further design an optimization-based method to improve realistic synthesis and connect the short-term motions to a long-term motion. Compared with other baselines, our framework can synthesize more natural and physically plausible long-term motion in the 3D scene.

References

[1] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *arXiv preprint arXiv:2005.05751*, 2020. 5

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016. 2

[3] Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. Socially-aware large-scale crowd forecasting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2203–2210, 2014. 2

[4] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1534–1543, 2016. 4

[5] Matthew Brand and Aaron Hertzmann. Style machines. In *SIGGRAPH*, pages 183–192, 2000. 1, 2

[6] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. *arXiv preprint arXiv:2007.03672*, 2020. 2, 12

[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. *arXiv preprint arXiv:1709.06158*, 2017. 2, 6, 8, 12

[8] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8648–8657, 2019. 2

[9] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 975–983, 2018. 2

[10] Vincent Delaitre, David F Fouhey, Ivan Laptev, Josef Sivic, Abhinav Gupta, and Alexei A Efros. Scene semantics from long-term observation of people. In *European conference on computer vision*, pages 284–298. Springer, 2012. 2

[11] David F Fouhey, Vincent Delaitre, Abhinav Gupta, Alexei A Efros, Ivan Laptev, and Josef Sivic. People watching: Human actions as a cue for single view geometry. *International journal of computer vision*, 110(3):259–274, 2014. 2

[12] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015. 2

[13] James J. Gibson. The ecological approach to visual perception. *Houghton Mifflin*, 1979. 1

[14] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018. 2

[15] Abhinav Gupta, Scott Satkin, Alexei A Efros, and Martial Hebert. From 3d scene geometry to human workspace. In *CVPR 2011*, pages 1961–1968. IEEE, 2011. 2

[16] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020. 2

[17] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2282–2292, 2019. 2, 5, 6, 8, 12

[18] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. 2

[19] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7134–7143, 2019. 2

[20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 2, 4

[21] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[22] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Trans. Graph.*, 36(4):42:1–42:13, 2017. 1

[23] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Trans. Graph.*, 35(4):138:1–138:11, 2016. 1, 2

[24] Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce. Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH Asia 2015 Technical Briefs*, pages 1–4. 2015. 2

[25] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 6, 12

[26] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 5308–5317, 2016. 2

[27] Vladimir G Kim, Siddhartha Chaudhuri, Leonidas Guibas, and Thomas Funkhouser. Shape2pose: Human-centric shape analysis. *ACM Transactions on Graphics (TOG)*, 33(4):1–12, 2014. 2

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5, 12

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[30] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *European Conference on Computer Vision*, pages 201–214. Springer, 2012. 2

[31] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013. 2

[32] Lucas Kovar and Michael Gleicher. Flexible automatic motion blending with registration curves. In *SIGGRAPH/Eurographics Symposium*, pages 214–224, 2003. 2

[33] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5226–5234, 2018. 2

[34] Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Putting humans in a scene: Learning affordance in 3d indoor environments. In *CVPR*, 2019. 1, 2, 12

[35] Yan Li, Tianshu Wang, and Heung-Yeung Shum. Motion texture: a two-level statistical model for character motion synthesis. *ACM Trans. Graph.*, 21(3):465–472, 2002. 2

[36] Zimo Li, Yi Zhou, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. *arXiv preprint arXiv:1707.05363*, 2017. 5

[37] C Karen Liu and Zoran Popović. Synthesis of complex dynamic character motion from simple animations. *ACM Transactions on Graphics (TOG)*, 21(3):408–416, 2002. 2

[38] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2

[39] Osama Makansi, Eddy Ilg, Ozgun Cicek, and Thomas Brox. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7144–7153, 2019. 2

[40] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9489–9497, 2019. 2

[41] Aron Monszpart, Paul Guerrero, Duygu Ceylan, Ersin Yumer, and Niloy J Mitra. imapper: interaction-guided scene mapping from monocular videos. *ACM Transactions on Graphics (TOG)*, 38(4):1–15, 2019. 2

[42] Xi Ouyang, Yu Cheng, Yifan Jiang, Chun-Liang Li, and Pan Zhou. Pedestrian-synthesis-gan: Generating pedestrian data in real scene and beyond. *arXiv preprint arXiv:1804.02047*, 2018. 2

[43] Sang Il Park, Hyun Joon Shin, and Sung Yong Shin. On-line locomotion generation based on motion blending. In *SIGGRAPH/Eurographics Symposium*, pages 105–111, 2002. 2

[44] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 2, 3, 4, 6

[45] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019. 2

[46] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. *arXiv preprint arXiv:1805.06485*, 2018. 2

[47] Vladimir Pavlovic, James M Rehg, and John MacCormick. Learning switching linear models of human motion. *Advances in neural information processing systems*, 13:981–987, 2000. 2

[48] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 3

[49] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6):245, 2017. 2

[50] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1349–1358, 2019. 2

[51] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. Scenegrok: Inferring action maps in 3d environments. *ACM transactions on graphics (TOG)*, 33(6):1–10, 2014. 2

[52] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. Pigraphs: learning interaction snapshots from observations. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016. 2

[53] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491, 2015. 2, 3, 4

[54] Lei Tai, Jingwei Zhang, Ming Liu, and Wolfram Burgard. Socially compliant navigation through raw depth inputs with generative adversarial imitation learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1111–1117. IEEE, 2018. 2

[55] Charlie Irawan Tan and Wen-Kai Tai. Characteristics preserving racer animation: a data-driven race path synthesis in formation space. *Journal of Visualization and Computer Animation*, 23(3-4):215–223, 2012. 2

[56] Fuwen Tan, Crispin Bernier, Benjamin Cohen, Vicente Ordonez, and Connelly Barnes. Where and who? automatic semantic-aware person composition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1519–1528. IEEE, 2018. 2

[57] Adrien Treuille, Seth Cooper, and Zoran Popović. Continuum crowds. *ACM Transactions on Graphics (TOG)*, 25(3):1160–1168, 2006. 2

[58] Raquel Urtasun, David J. Fleet, Andreas Geiger, Jovan Popovic, Trevor Darrell, and Neil D. Lawrence. Topologically-constrained latent variable models. In *ICML*, pages 1080–1087, 2008. 2

[59] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. *arXiv preprint arXiv:1704.05831*, 2017. 2

[60] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *Proceedings of the IEEE international conference on computer vision*, pages 3332–3341, 2017. 2

[61] Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *CVPR*, 2017. 1

[62] Zhe Wang, Liyan Chen, Shaurya Rathore, Daeyun Shin, and Charless Fowlkes. Geometric pose affordance: 3d human pose with scene constraints. *arXiv preprint arXiv:1905.07718*, 2019. 2

[63] Guiyu Xia, Huaijiang Sun, Qingshan Liu, and Renlong Hang. Learning-based sphere nonlinear interpolation for motion synthesis. *IEEE Trans. Industrial Informatics*, 15(5):2927–2937, 2019. 2

[64] Shihong Xia, Congyi Wang, Jinxiang Chai, and Jessica K. Hodgins. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Trans. Graph.*, 34(4):119:1–119:10, 2015. 1

[65] Jingwei Xu, Huazhe Xu, Bingbing Ni, Xiaokang Yang, Xiaolong Wang, and Trevor Darrell. Hierarchical style-based networks for motion synthesis. In *ECCV*, 2020. 1, 2, 5, 6, 7, 8

[66] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. Place: Proximity learning of articulation and contact in 3d environments. *arXiv preprint*, 2020. 2

[67] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6194–6204, 2020. 1, 2, 4, 5, 6, 7, 8, 12

[68] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris Metaxas. Learning to forecast and refine residual motion for image-to-video generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 387–403, 2018. 2

[69] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 3

[70] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *European conference on computer vision*, pages 408–424. Springer, 2014. 2

[71] Yixin Zhu, Chenfanfu Jiang, Yibiao Zhao, Demetri Terzopoulos, and Song-Chun Zhu. Inferring forces and learning human utilities from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3823–3833, 2016. 2

Appendix

We provide more details about datasets, optimization implementation, discussion of other methods, and naturalness evaluation in the appendix.

1. Datasets

We use 'MPH112', 'MPH11', 'MPH8', 'N0Sofa', 'N3Library', 'N3Office', 'BasementSittingBooth' and 'Werkraum' in PROX[17] as training scenes and we use 'MPH16', 'MPH1Library', 'N0SittingBooth', 'N3OpenArea' in PROX[17] and the family room, living room and bedroom of '17DRP5sb8fy' in MP3D[7] dataset as testing scenes.

For the training data of sub-goal body synthesis network, we down-sample the original motion sequences and use the static body every 0.33 seconds. For the training data of motion synthesis networks, we first sample the start and end bodies which has a duration of 2 seconds and the Euclidean distance between them is larger than 0.5 meters. We use the motion in between these start/end pairs as our motion training data.

2. Implementation details

To better balance the environmental constraints and plausibility of motion, we perform our optimization in two stages. In the first stage, we enhance the optimization for environment constrains and motion smoothness and set $\lambda_{foot} = 0$, $\lambda_{col} = 1$, $\lambda_{cont} = 1$ and $\lambda_{smooth} = 0.25$. In the second stage, we want to improve the motion plausibility and set $\lambda_{foot} = 1$, $\lambda_{col} = 1$, $\lambda_{cont} = 1$ and $\lambda_{smooth} = 0.25$.

3. Discussion of other methods

We also try to create a baseline inspired by CVAE interpolation for motion synthesis. Since our setting is to give the start and end bodies to generate motion in between, we first perform gradient descent with Adam [28] to fit two latent z of the start and end bodies. After we get the latent z of the start/end, we can use interpolation to get the sequence in between. However, this method may only be applied to a few cases. For motion with a certain distance, this method is more like average interpolation rather than following the law of human motion. As shown in Fig. 10, CVAE interpolation can not generate a complete human motion.

Another related work is [6], which uses past 1 second motion to predict future 2 seconds motion using skeleton and rgb images to represent human and scenes. Their motion is in 10fps and ours is 30fps. Their paper's w/ gt destination setting is the most similar to ours. They report the path error and MPJPE [25] error in PROX [17] which can be compared to us. Their path error is from 19.3 to 23.7, and the average of ours is 8.06. For MPJPE in millimeters, ours is 219.1 while their method is 237. Considering their weakness in generating dense motion sequence, large requirements of past sequence and different representation setting, we do not compare the other aspects.

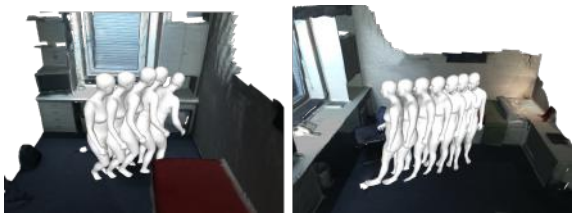


Figure 10: Two examples of CVAE interpolation results. Left example shows that if the start and end bodies we give are without legs walking motion, the result of interpolation is more like standing but being pushed forward. The right example shows that if there is legs changing motion but with a certain distance, no matter how far the motion is, the interpolation result will finish this in one step, thus the whole motion is full of foot skating.

4. Naturalness evaluation

We provide more details about modified contact score, human evaluation and show more qualitative results in this section.

Modified contact score. Since our task is a motion synthesis task, we set a threshold of 0.01 of the signed distance value and if it is smaller than 0.01, we take it as contact, unlike 0 in [67].

Human evaluation details. Different from [34] giving two examples once and asking user to compare which is better and [67] giving just one example to score from 1 to 5, we give 4 examples (two baselines, ours and pseudo-ground truth) once with the same start, end, sub-goal body inputs and ask users to score from 1 (strongly not natural) to 5 (strongly natural) each. The advantage of this is we can ensure that for the same motion, people who scored are the same, which is fairer for the comparison. Each task will be scored by 3 users and we calculate the average score.

More qualitative results. We provide more qualitative results of our generated sub-goal (start/end) bodies and generated motion in between in different scenes in Fig. 11 and Fig. 12. It can be seen that our method can synthesize different kinds of challenging long-term motion such as walking, sitting down, jumping on the bed and lying down in different scenes. Furthermore, we provide examples of randomly sampled body shape β in Fig. 13 and also examples of randomly sampled latent variables for sub-goal bodies in Fig. 14. It can be seen that our method can synthesize diversified motion with different body shape and different motion style.

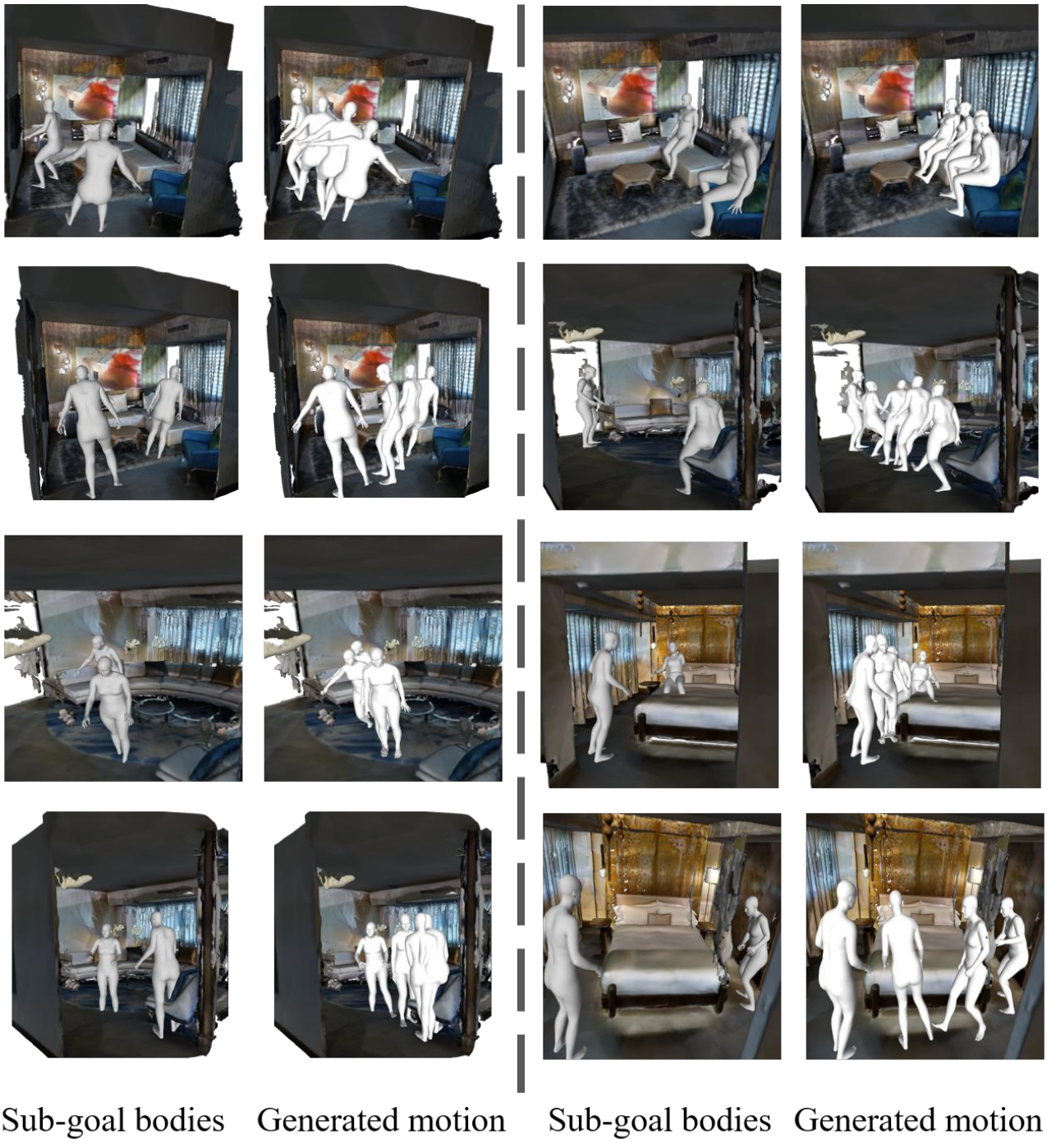


Figure 11: **Our results.** We show the generated sub-goal bodies and motion between in sub-goal bodies.

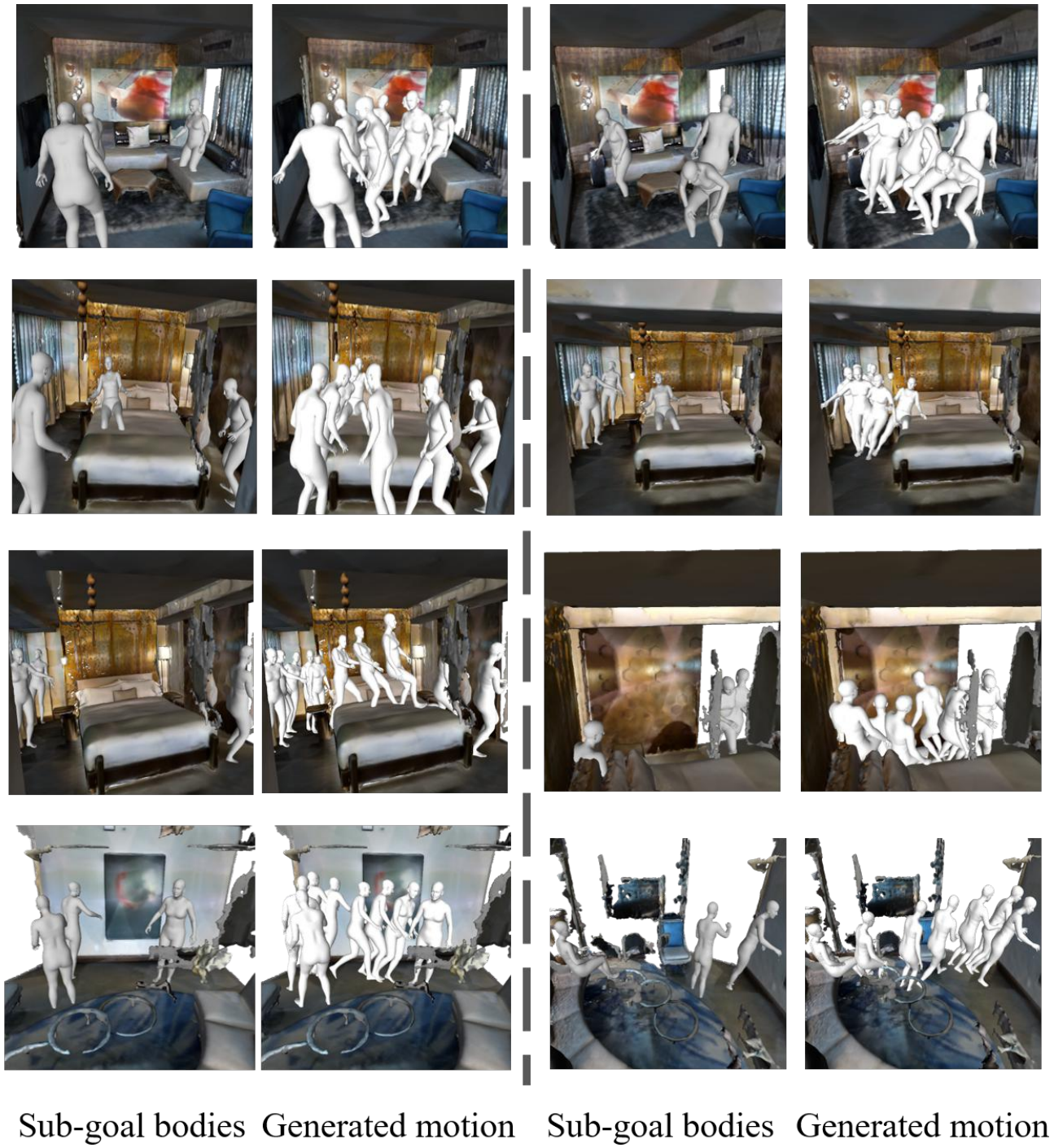


Figure 12: **Our results.** We show the generated sub-goal bodies and motion between in sub-goal bodies.

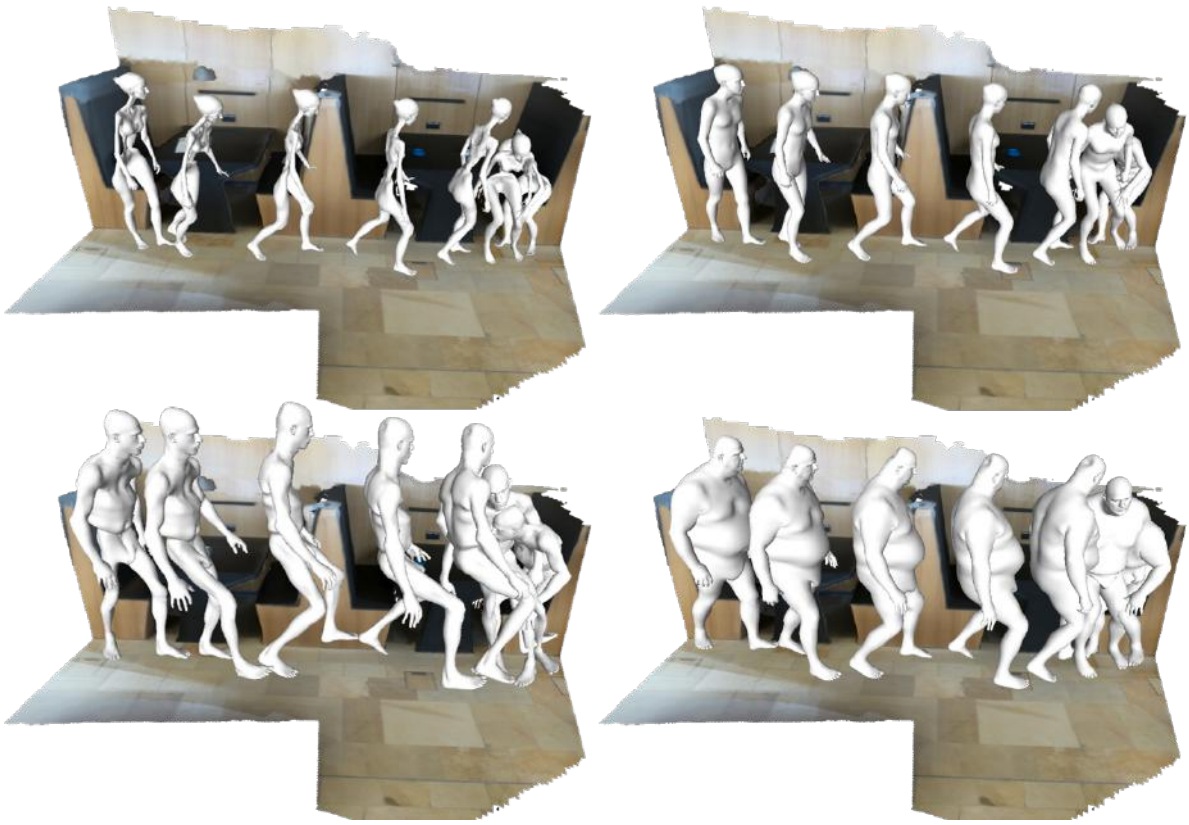


Figure 13: Four examples of diversified motion with different body shape β.

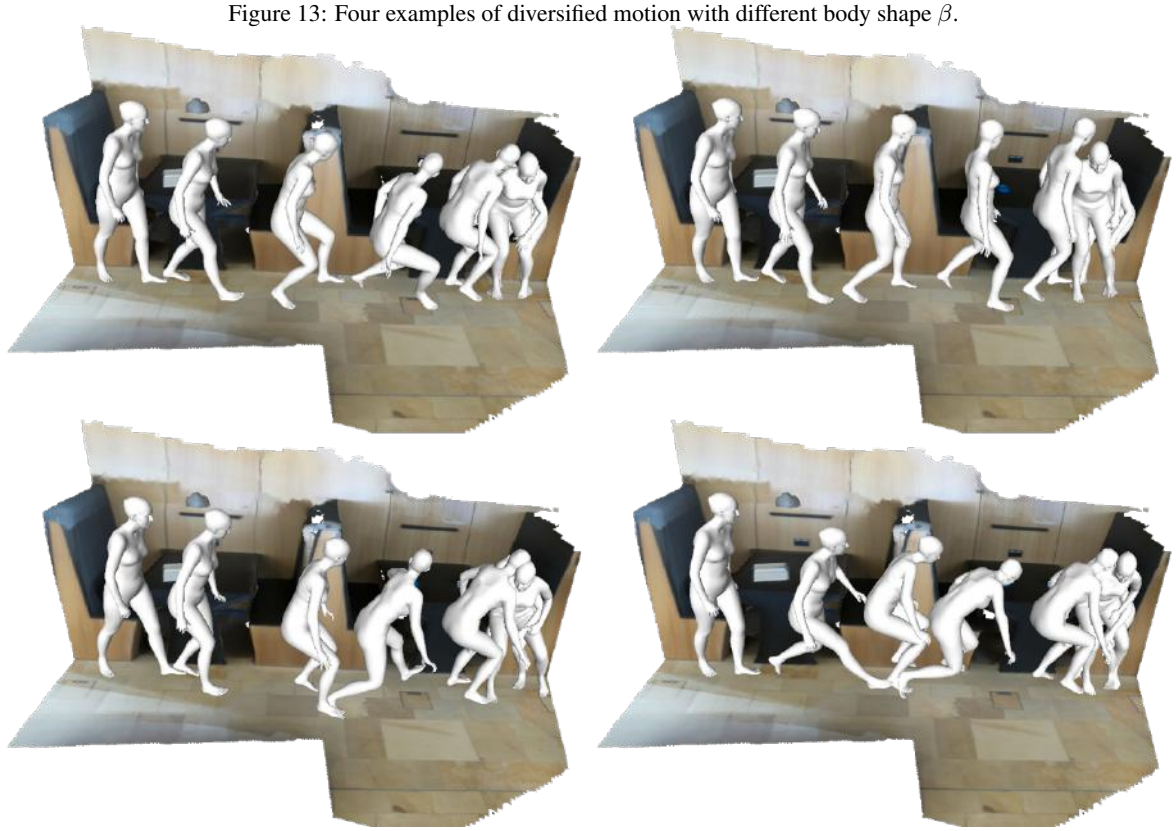


Figure 14: Four examples of diversified motion with different latent z for sub-goal body.