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[3] as training scenes and we use 'MPH16', 'MPH1Library', 'N0SittingBooth', 'N3OpenArea' in PROX[3] and the family room, living room and bedroom of '17DRP5sb8fy' in MP3D[2] 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 [5] 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. 1, CVAE interpolation can not generate a complete human motion.

Another related work is [1], 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 [4] error in PROX [3] 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.

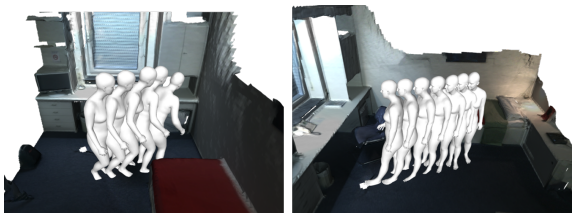

Figure 1: Two examples of CVAE interpolation results. Left example shows that if the start and end bodies we give are without legs walking motion, the result of interpolation is more like standing but being pushed forward. The right example shows that if there is legs changing motion but with a certain 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 [7].

Human evaluation details. Different from [6] giving two examples once and asking user to compare which is better and [7] 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. 2 and Fig. 3. 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. 4 and also examples of randomly sampled latent variables for sub-goal bodies in Fig. 5. It can be seen that our method can synthesize diversified motion with different body shape and different motion style.

Figure 2: **Our results.** We show the generated sub-goal bodies and motion between in sub-goal bodies.

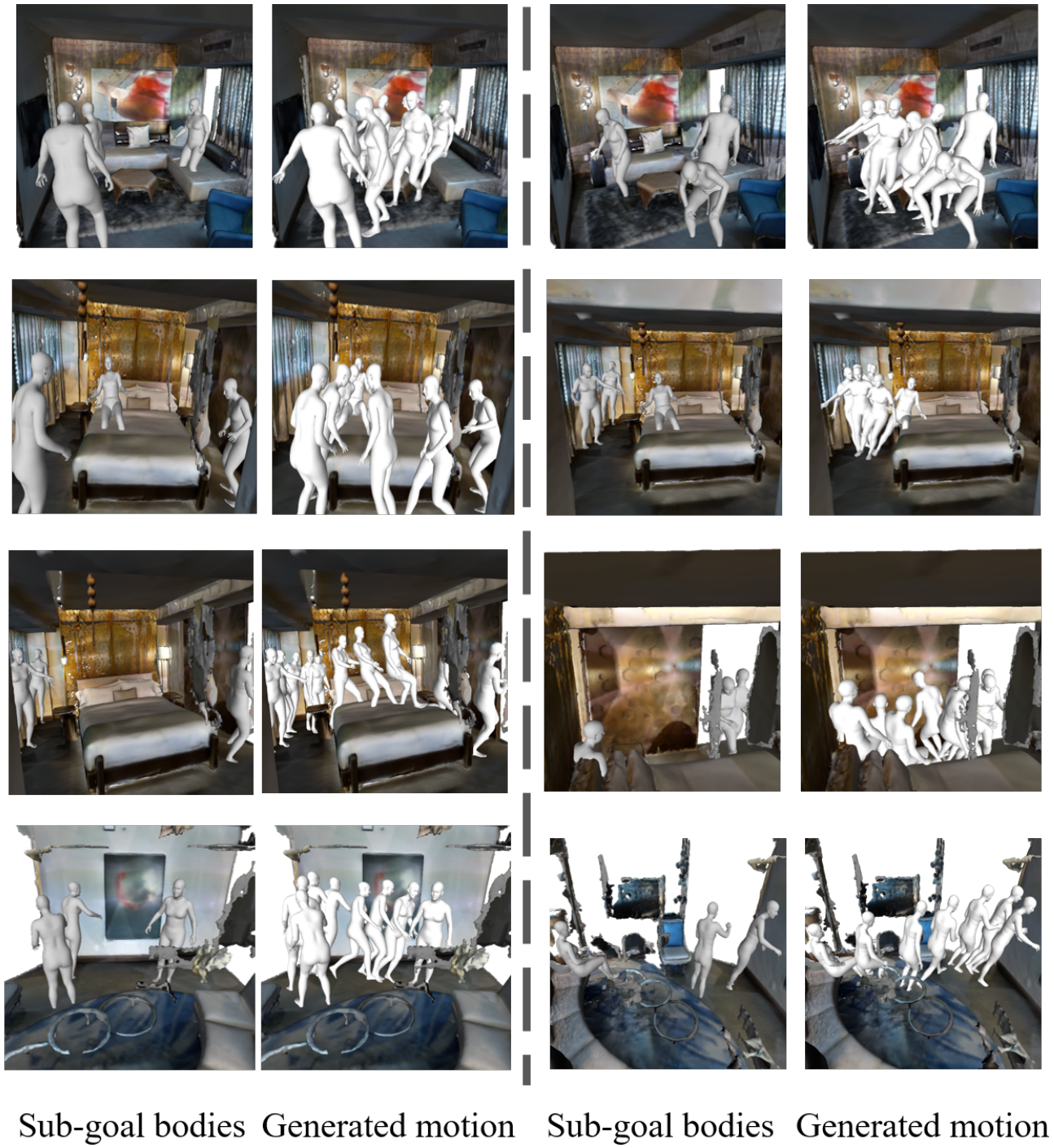

Figure 3: **Our results.** We show the generated sub-goal bodies and motion between in sub-goal bodies.

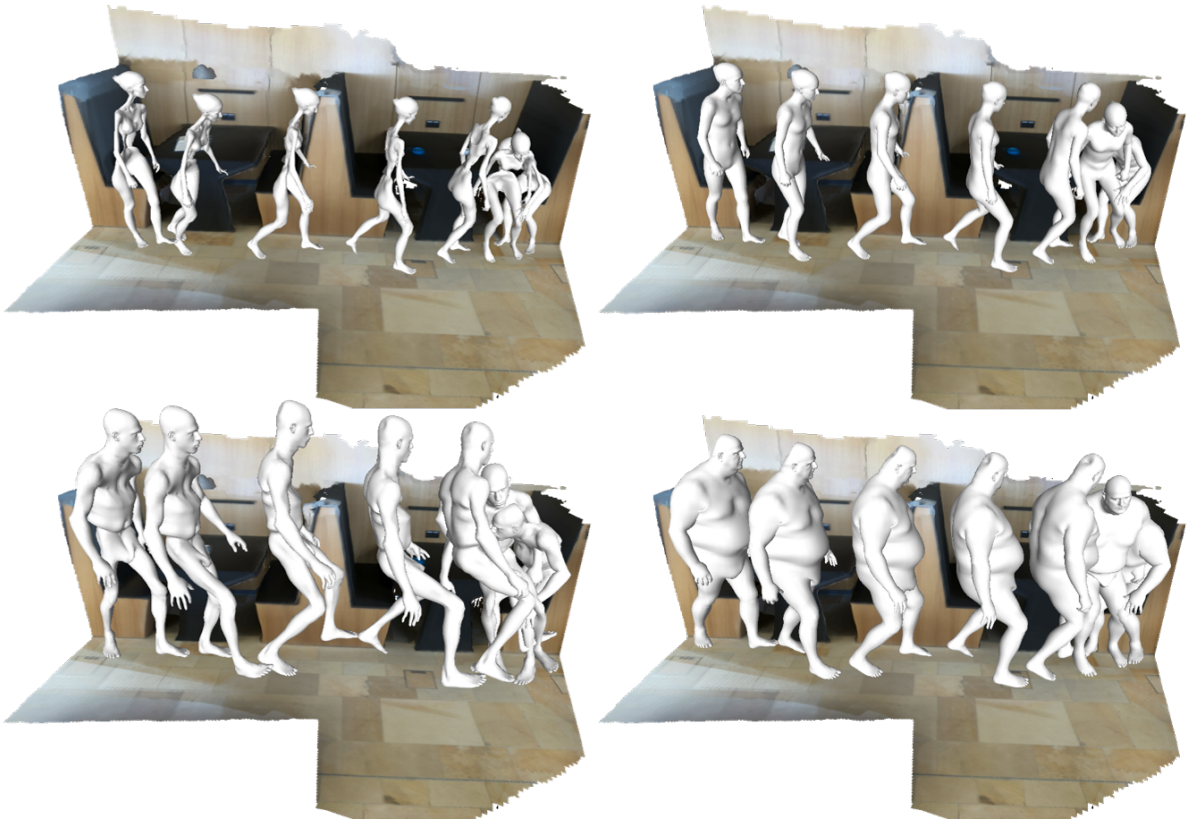

Figure 4: Four examples of diversified motion with different body shape β.

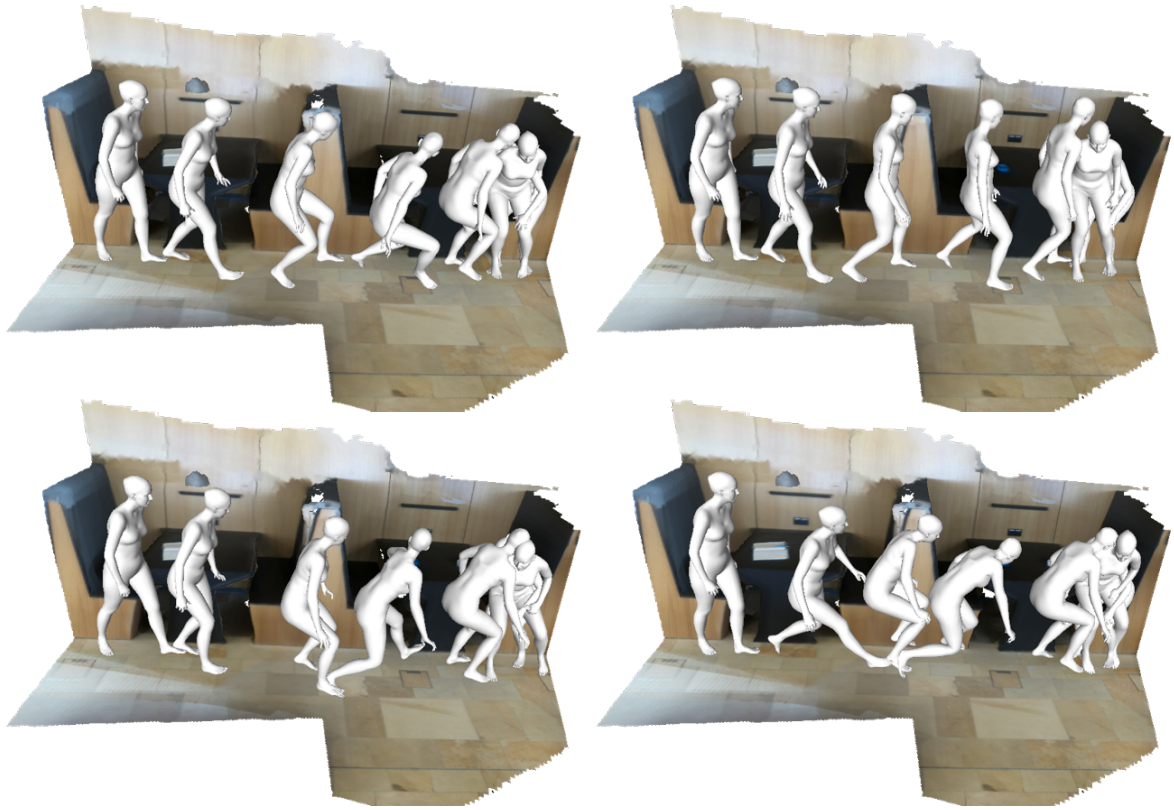

Figure 5: Four examples of diversified motion with different latent z for sub-goal body.

References

[1] 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. 1

[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. *arXiv preprint arXiv:1709.06158*, 2017. 1

[3] 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. 1

[4] 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. 1

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[6] 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

[7] 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