Supplementary Document for Joint2Human: High-quality 3D Human Generation via Compact Spherical Embedding of 3D Joints

Muxin Zhang[1,†], Qiao Feng[1,†], Zhuo Su[2], Chao Wen [2], Zhou Xue[3], Kun Li[1,*]
[1]College of Intelligence and Computing, Tianjin University
[2]PICO IDL, ByteDance [3]Li Auto

In the supplemental material, we provide more details about the paper, including:

- Implementation Details.
- Running Time.
- More Visual Results.
- More Comparison Results.
- User Study.
- Failure Cases and Analysis.

We also provide a demo video to show the main idea of our method and more 3D results, which can be found on our project page at http://cic.tju.edu.cn/faculty/likun/projects/Joint2Human.

1. Implementation Details

1.1. Training

We use an EMA rate of 0.9999 for all experiments. For the training stage of the FOF autoencoder, we trained it for three days on 8 NVIDIA A100 GPUs, with a batch size of 32. To represent and generate a high-quality 3D human geometry, the channel number N of the FOF feature maps needs to be set to at least 32. However, it is hard to model high-dimensional and multi-channel data for diffusion models. Following the latent diffusion model [9] to address this issue, we utilize an auto-encoder to compress the raw high-dimensional data space into a lower-dimensional latent space, which encodes each shape into a normal distribution. In detail, we adopted a VAE-like auto-encoder, which contains the encoder $\mathcal{E}$ and decoder $\mathcal{D}$. Given a FOF feature $x \in \mathbb{R}^{512\times512\times32}$, the encoder $\mathcal{E}$ encodes FOF into latent vectors $z = \mathcal{E}(x)$, and the decoder $\mathcal{D}$ decodes the latent vectors back to the original FOF space $\tilde{x} = \mathcal{D}(\mathcal{E}(x))$, where $\tilde{x} \in \mathbb{R}^{128\times128\times8}$. We pretrained the auto-encoder with the reconstruction loss and KL-regularization loss. The loss for the auto-encoder training process can be expressed as

$$\mathcal{L}_{vae} = \left\|\tilde{x} - x\right\|_1 + \lambda\left(D_{KL}\left(q_\phi\left(z|x\right)||p\left(z\right)\right)\right), \quad (1)$$

the first item is reconstruction loss, calculating the l1 loss between ground-truth FOF and results from the decoder $\mathcal{D}$. The second term calculates KL-divergence loss between the target $p(z)$ distribution and the latent space. $q_\phi(z|x)$ is the approximation to the true posterior.

Next, we use the vectors in the latent space as training data for the diffusion model. For the diffusion model training stage, we trained it for eight days on 8 NVIDIA A100 GPUs, with a batch size of 64. Following the DDPM [2], we also use the U-net [10] as our backbone. We adopt the standard Adam optimizer, maintaining a learning rate 1e-4 and a linear learning rate warm-up schedule spanning 10,000 iterations. The total number of timesteps is set as $T = 1000, T' = 200$ for the main diffusion in our pipeline. Our pipeline involves training multiple models; both the human pose embedding and image embedding branches can be switched on or off and applied simultaneously. We implement this by training multiple models.

1.2. U-Net Related Settings

Our diffusion model adopts the U-Net architecture. The hyperparameters of U-Net are set with 128 channels and three residual convolution blocks. Apart from that, we impose a self-attention mechanism at the fusion of information interactions between feature layers. During upsampling, the number of feature channels at each level changes from 128 to 1024. To make the model more deeply aware of the conditional semantic information, we learn an additional condition encoder to map the conditional input into the latent vector with more centralized information.

2. Running Time

We have tested the running time of each stage of our method shown in Tab. 1. In addition, our approach is more efficient in computation. Different from the previous approaches [3–5, 7, 11], we don't need to use differentiable rendering for additional optimizations. Our method can directly generate detailed 3D human geometry using 2D dif-

Table 1. The running time of each stage of our method.

	Main diffusion	High-frequeney enhancer	Recarving strategy
Time(s)	50.01	0.04	8.92

fusion models. In fairness, we compare methods that specifically concentrate on generating 3D human geometry. For consistency, we assess the time required to generate a human body in a single sample using the same hardware device. Chupa [5] requires an average of 2min39s for generating a human body, whereas our method achieves the same task in just 59 seconds on average.

3. More Visual Results

Here, we provide additional generation results of our model. Please note that our approach can simultaneously guarantee global structure and local details with low computational cost. Figure. Fig. 1 shows more results guided by the same pose with the help of our compact spherical embedding of 3D joints, Fig. 2 shows the generation results by the text-guided control. Fig. 3 shows the generation results for the loose clothes.

4. More Comparison Results

Modeling of Loose-fitting Clothes. Our method outperforms previous methods in modeling loose clothing like dresses and loose coats. Because we do not have strong dependencies on human parameterization models such as SMPL[6]/SMPL-X[8]. Fig. 3 shows the results of the generation of loose clothing human body, and compares it to other methods. Our generated results are more natural and realistic; the other two methods model clothes more inclined to be close to the skin.

Diversity of Human Generation. With the pose input (SMPLx or our compact spherical embedding of 3D joints) as our condition, our model can generate results aligned to input with better diversity compared to Chupa [5]. The code of AG3D [1] is not friendly to the support of pose control, so we only compare our model with Chupa. Fig. 4 shows the generation results produced by different algorithms under the given human pose guidance.

5. User Study

To better evaluate our model, we conduct a perceptual study to ask the users about their preferences for the generated body geometry. Users are asked to choose among our approach and two other current state-of-the-art methods. The results of the survey are shown in Tab. 2 .

In the study, we present the human generation results of AG3D (Method A) [1], Chupa (Method B) [5], and our method (Method C) in video form. The study is divided into two sections, with a total of 7 cases. Cases 1 through 5 focused on examining individual generation performance, with each case comprising three consistent questions. These questions pertained to generating a 3D human shape with respect to 1. Global Structure, 2. Local Detail, and 3. Overall Impression. Each participant was asked to rank the results shown in the video from best to worst based on these three metrics. Tab. 2 demonstrates that our approach emerged as the most popular choice.

In the second section, our focus shifts to examining the diversity of human generation. Since AG3D does not incorporate guidance from the human body pose, our analysis in this section is limited to methods B and C. Unlike section 1, where video presentations were used, we presented users with multiple image results in two cases, prompting them to compare the diversity between the two methods. Each case included six results for both methods. During the user study, we collected a total of 123 responses, comprising 54 females and 69 males across different age groups (6 users under 18, 113 users between 18 and 40, 2 users between 40 and 60, and 2 users above 60). In section 2, 70.73% of the users considered the diversity of our methods to be superior to Chupa (Method B).

Table 2. Proportion of popularity of different methods in different metrics.

Method	AG3D (A)	Chupa (B)	Ours (C)
Global Structure	18.70%	32.20%	**49.10%**
Local Detail	18.05%	29.92%	**52.03%**
Overall Impression	19.02%	29.27%	**51.71%**
Diversity of Generation	-	29.27%	**70.73%**

6. Failure Cases and Analysis.

Fig. 5 illustrates examples of unrealistic artifacts appearing on the generated results due to the extreme poses, which fall outside the distribution of the training data. Additionally, there are a small number of cases involving broken legs and arms, which are limited by the fact that the FOF data was generated by sampling from a single viewpoint.

References

[1] Zijian Dong, Xu Chen, Jinlong Yang, Michael J Black, Otmar Hilliges, and Andreas Geiger. AG3D: Learning to Generate 3D Avatars from 2D Image Collections. In *Int. Conf. Comput. Vis.*, 2023. 2

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Adv. Neural Inform. Process. Syst.*, 2020. 1

[3] Fangzhou Hong, Zhaoxi Chen, Yushi LAN, Liang Pan, and Ziwei Liu. EVA3d: Compositional 3d human generation from 2d image collections. In *Int. Conf. Learn. Represent.*, 2023. 1

[4] Suyi Jiang, Haoran Jiang, Ziyu Wang, Haimin Luo, Wenzheng Chen, and Lan Xu. Humangen: Generating human radiance fields with explicit priors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

[5] Byungjun Kim, Patrick Kwon, Kwangho Lee, Myunggi Lee, Sookwan Han, Daesik Kim, and Hanbyul Joo. Chupa: Carving 3d clothed humans from skinned shape priors using 2d diffusion probabilistic models. *arXiv preprint arXiv:2305.11870*, 2023. 1, 2, 6

[6] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.*, 2015. 2

[7] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 1

[8] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 1

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Springer. Medical Image Computing and Computer-Assisted Intervention*, 2015. 1

[11] Zhangyang Xiong, Di Kang, Derong Jin, Weikai Chen, Linchao Bao, Shuguang Cui, and Xiaoguang Han. Get3dhuman: Lifting stylegan-human into a 3d generative model using pixel-aligned reconstruction priors. In *Int. Conf. Comput. Vis.*, 2023. 1

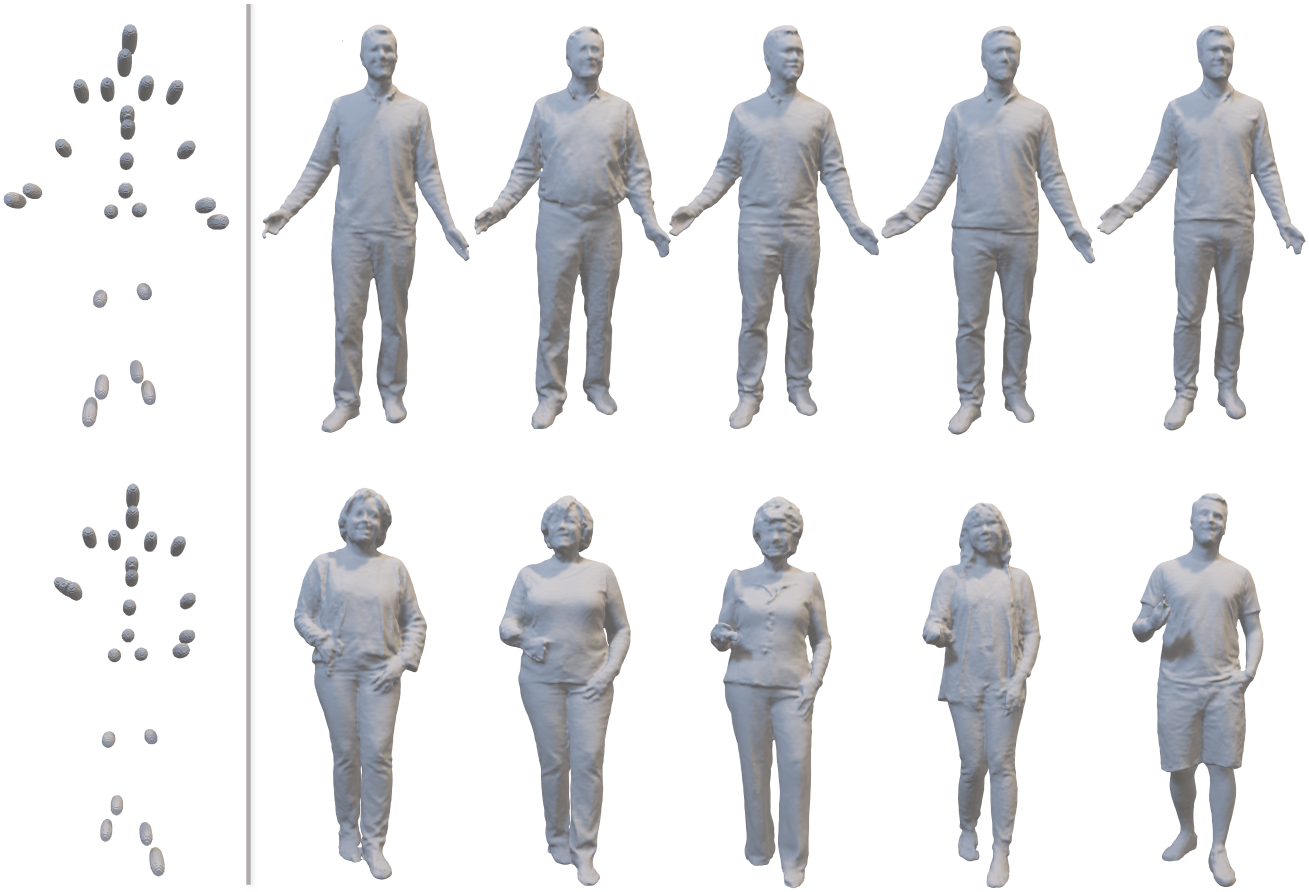
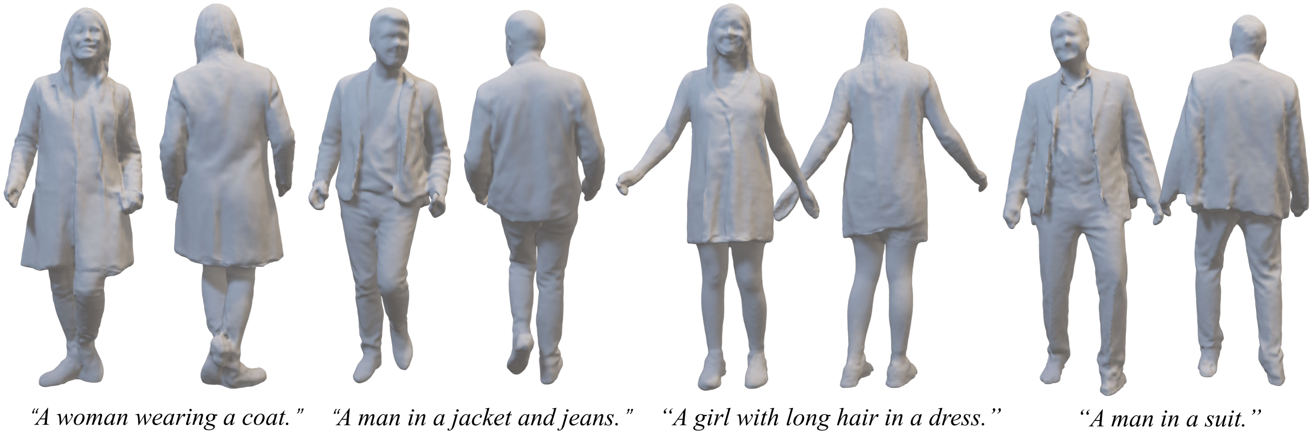


Figure 1. More generated results guided by specific poses with the help of our compact spherical embedding of 3D joint (left).



"A woman wearing a coat."
"A man in a jacket and jeans."
"A girl with long hair in a dress."
"A man in a suit."

Figure 2. Results of text-guided human generation.

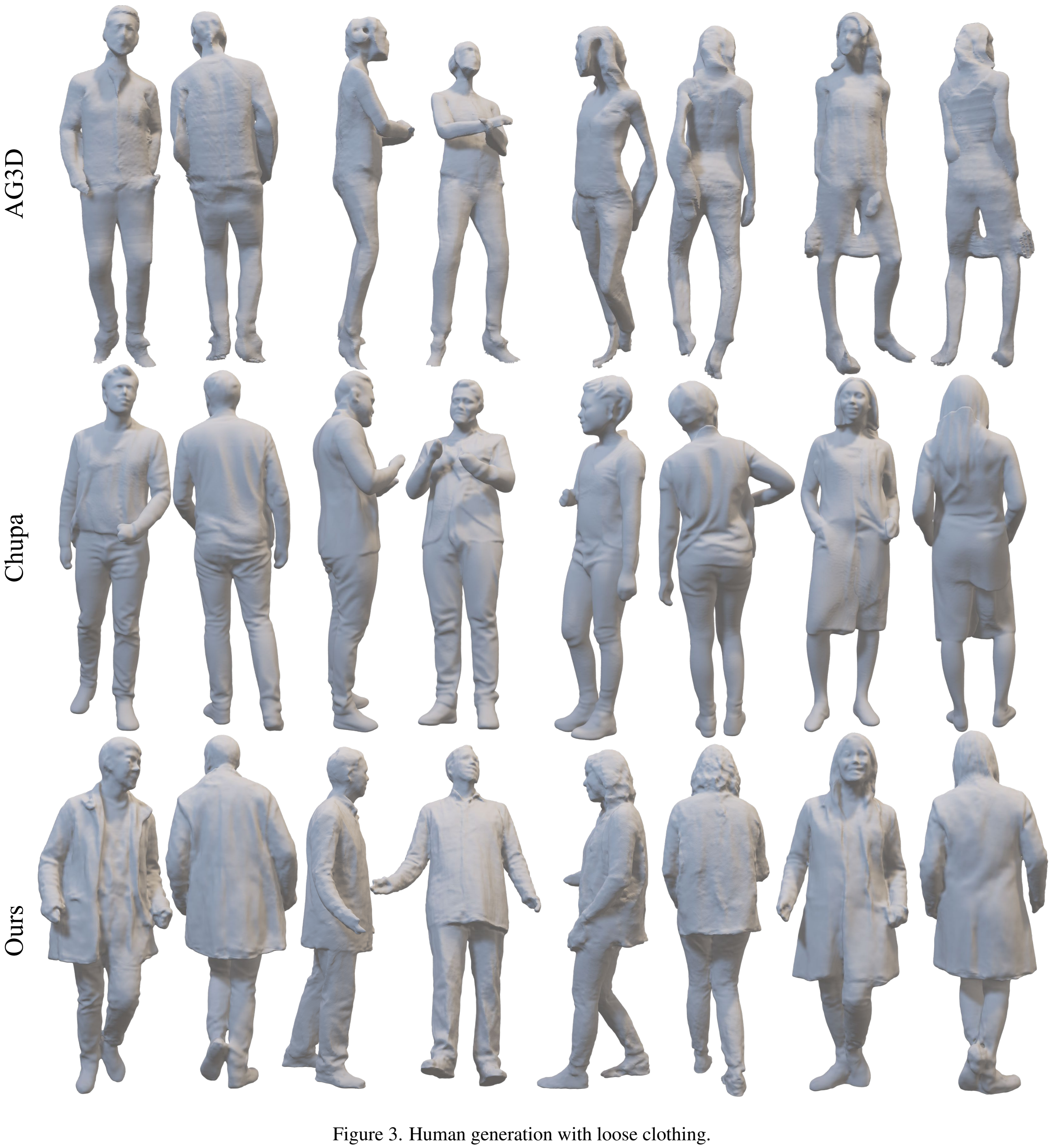


Figure 3. Human generation with loose clothing.

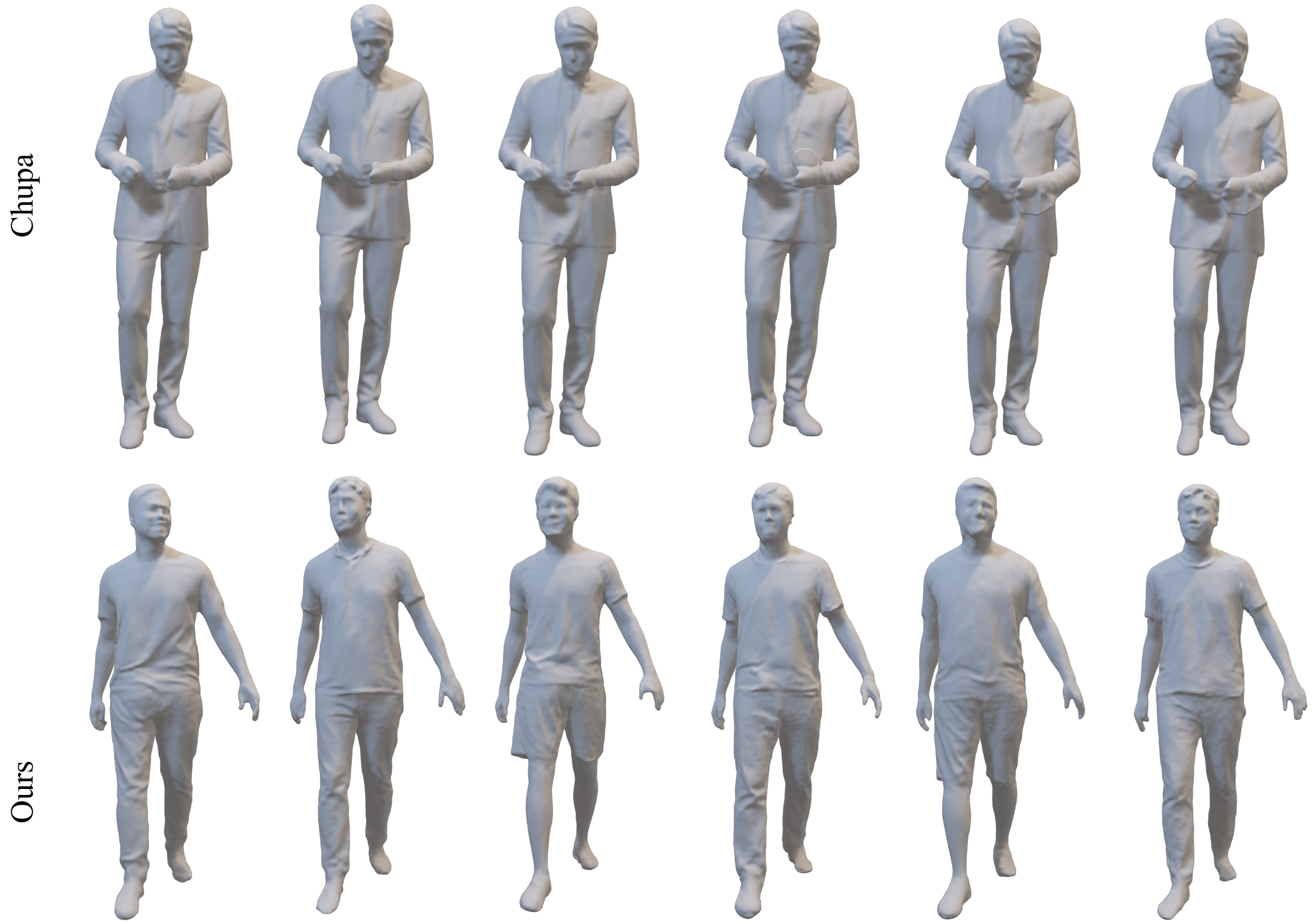


Figure 4. We randomly select fixed poses as the respective conditional information inputs for each model. Without fixing the random number seed, the model randomly samples to generate human bodies, our model has better diversity compared to Chupa [5].

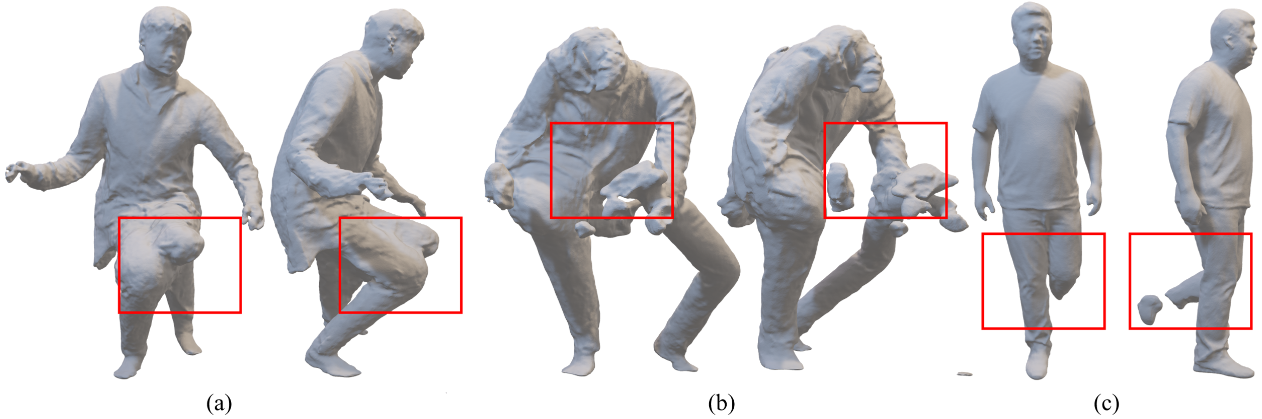


Figure 5. **Failure cases.** (a) and (b) depict the results with the human pose outside the data distribution, and (c) portray instances of leg breaks.