Supplementary Material for
“AniGS: Animatable Gaussian Avatar from a Single Image with Inconsistent Gaussian Reconstruction”

Lingteng Qiu[1*] Shenhao Zhu[1,3*] Qi Zuo[1*] Xiaodong Gu[1*]
Yuan Dong[1] Junfei Zhang[1] Chao Xu[1] Zhe Li[1,4] Weihao Yuan[1] Liefeng Bo[1]
Guanying Chen[2†] Zilong Dong [1†]

[1]Tongyi Lab, Alibaba Group [2]Sun Yat-sen University
[3]Nanjing University [4]Huazhong University of Science and Technology

Contents

1. More Details for the Method

1.1. Implementation Details

Multi-view Generation For training the multi-view canonical image generation model, we first pre-train our RGB-Normal DiT model on in-the-wild video clips. To supervise the normal map output, we utilize Sapiens [3], an off-the-shelf normal estimation prior, to generate pseudo ground-truth normals from in-the-wild data. The model is trained using the Adam optimizer [4] with a learning rate of 2×10^{-4} and a batch size of 1. We employ 16 Nvidia A100 80G GPUs for training, with the pre-training process comprising 100,000 optimization iterations. Subsequently, the model is fine-tuned on a synthetic dataset using the same hyperparameters, performing an additional 50,000 iterations of optimization. To preserve the model’s generalizability, we adopt a data-mixing strategy during fine-tuning, assigning a 10% probability to sampling in-the-wild data and a 90% probability to synthetic data.

3D Reconstruction from Inconsistent Images. In the multi-view reconstruction phase, after obtaining the deformed coarse mesh from the original SMPL-X as the initialization for 4DGS, we first performed 3,000 iterations of optimization the 3DGS parameters. Sequentially, we continue to conduct 4,000 iterations of optimization in the temporal dimension to address multi-view inconsistency. In the multi-view reconstruction phase, we initialize with a deformed coarse mesh derived from the original SMPL-X model for the 4DGS process. The first step is optimizing the 3DGS parameters over 3,000 iterations. Subsequently, we perform 4,000 iterations of optimization considering the temporal dimension to address multi-view inconsistency.

1.2. RGB-Normal Diffusion Transformer

Figure S3 illustrates the architecture of our multi-view diffusion transformer model for canonical image and normal map generation. For simplicity, we omit SPML-X conditioning in the figure. Both ‘I-DiT-E’ and ‘N-DiT-E’ denote two independent DiT encoder blocks conditioned on image and normal input, respectively, while ‘I-DiT-D’ and ‘N-DiT-D’ refer to two independent decoders responsible for generating multi-view canonical images and normal maps. Additionally, ‘I-N’ within the intermediate DiT blocks represents a multi-modal attention module that effectively encodes joint image and normal features.

[1]Equal contribution.
[2]Corresponding author.

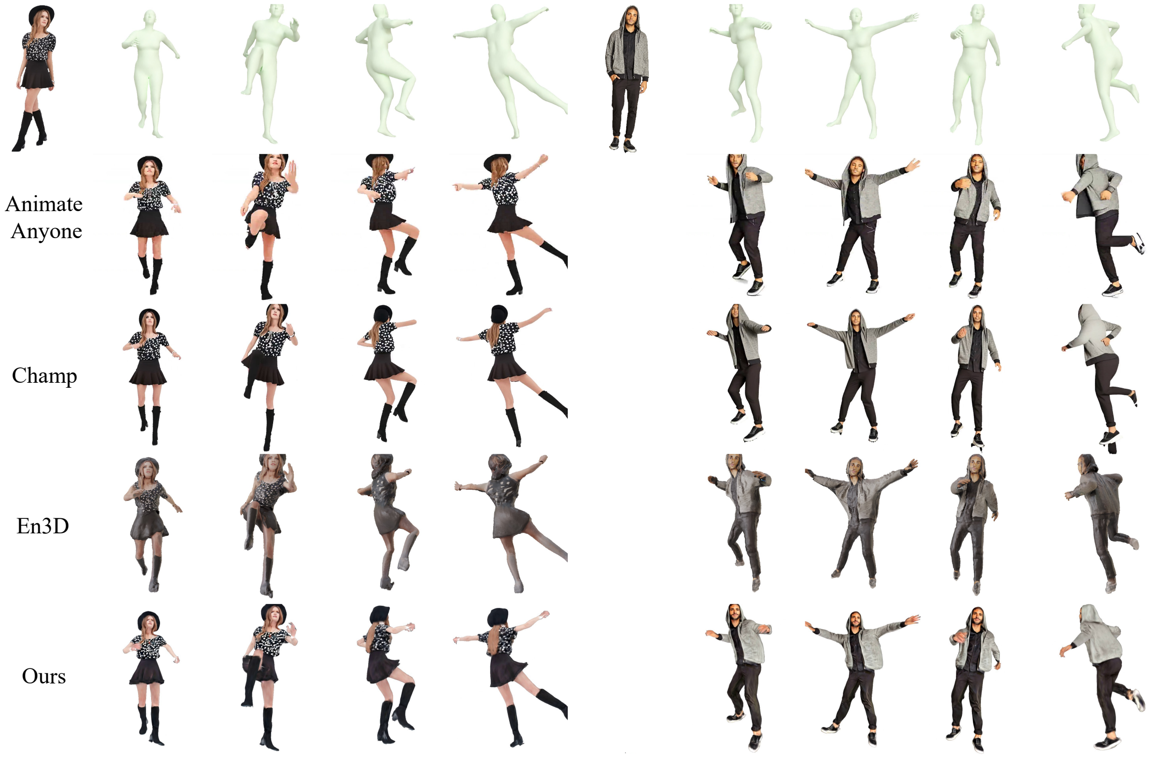


Figure S1. Visual comparison of animation results for the reconstructed 3D avatars. Best viewed with zoom-in.

1.3. Coarse Shape Initialization

We optimize the following objective function to obtain the initial coarse mesh $\mathcal{M}'$ for 3DGS initialization:

$$\begin{aligned}\mathcal{L}_{init} = \lambda_{mask} \cdot \mathcal{L}_{mask} + \lambda_n \cdot \mathcal{L}_{normal} \\ + \lambda_{lap} \cdot \mathcal{L}_{lap}(\mathcal{M}') + \lambda_{edge} \cdot \mathcal{L}_{edge}(\mathcal{M}').\end{aligned} \tag{1}$$

where $\lambda_{mask} = 1.0$, $\lambda_n = 0.5$, $\lambda_{lap} = 0.1$, and $\lambda_{edge} = 0.05$.

Figure S2 demonstrates the coarse mesh results reconstructed from the generated images. As illustrated in the figure, the coarse mesh provides only a rough geometric surface, with several noticeable artifacts remaining on its surface.

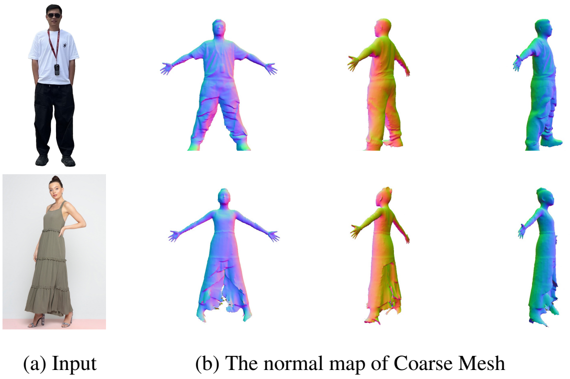


Figure S2. Sample results for the about coarse mesh reconstruction from multi-view images.

1.4. Skinning-based Animation

We model large body motions using linear blend skinning (LBS) transformations based on the SMPL-X [6] model. Specifically, given an SMPL body with shape parameter β and pose parameter θ_i in the i-th frame, a point p on the body surface in canonical space with skinning weights $w(p)$ can be warped to camera view space via the skinning transformation W.

Notably, the skinning weights $w(p)$ are only defined for points on the SMPL-X surface. To handle shapes with large deformations (*e.g.*, skirts) and to better facilitate the warping of arbitrary points in canonical space to the camera view, we employ the diffused skinning strategy [5] to propagate the skinning weights of the SMPL-X body vertices to the entire canonical space. These weights are stored in a voxel grid of size 256 × 256 × 256. Skinning weights for arbitrary points are then obtained through trilinear interpolation.

1.5. More Details for the Synthetic Dataset

We leverage a combination of public synthetic 3D datasets to render multi-view images for fine-tuning the multi-view

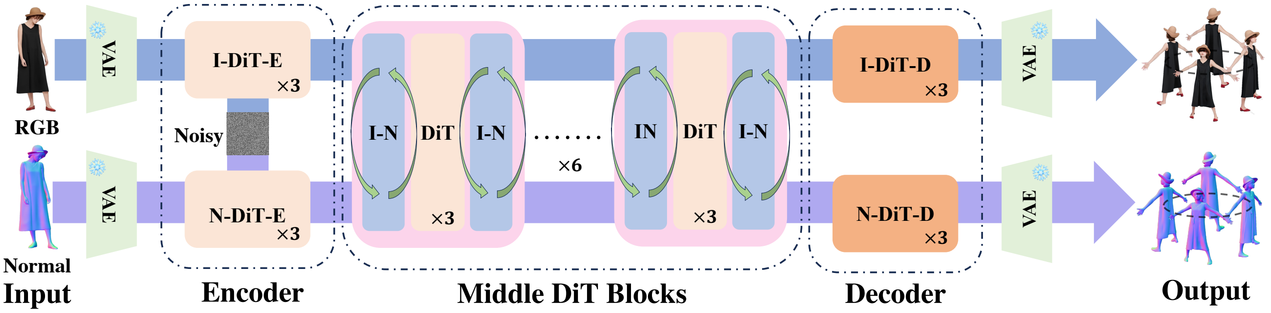
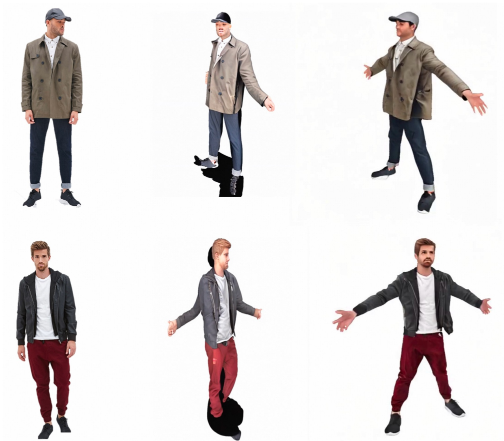


Figure S3. The architecture of the joint RGB-Normal Diffusion Transformer designed for generating multi-view canonical images and normal maps. For simplicity, SPML-X conditioning is omitted from the depiction.



Reference (b) w/o pretraining (c) w/ pretraining

Figure S4. Effectiveness of pre-training on in-the-wild videos.

canonical image and normal generation model. These datasets include 2K2K [1], Thuman2.0, Thuman2.1 [7], and CustomHumans [2], along with commercial datasets such as Thwindom and RenderPeople. In total, we utilize 6,124 synthetic human scans.

For the synthetic data, we render each object from 30 different viewpoints by rotating the object. To improve the quality of multi-view reconstruction, images are rendered at varying elevations, which helps to regularize the optimization of the 3D Gaussian Splatting (3DGS) method. Specifically, the elevation range oscillates between $-20°$ and $20°$, following a sine function over a cycle of 30 views.

2. More Details for the Ablation Study

Body Identity The bodily proportions of the human figure depend on the human body model predicted by the off-the-shelf method. Figure S6 demonstrates that the reconstructed 3D human can adapt to different human body conditions.

Multi-view Reconstruction from Varying Inputs We conducted an ablation study on multi-view reconstruction with varying inputs, which revealed that simply reducing the number of views does not resolve the associated issues due to inconsistencies among frames; in fact, it may even degrade the results, as shown in Table S1 and Figure S5. This experiment empirically demonstrates that continuously increasing the number of input images for 4DGS reconstruction, does not yield infinite improvements in the quality of the reconstruction, with an optimal empirical range identified to be around 24 to 30 images.

Table S1. Results of 3DGS and 4DGS with different input views.

	10 Views		20 Views		30 Views	
Metrics	3DGS	4DGS	3DGS	4DGS	3DGS	4DGS
FID ↓	120.52	**79.196**	112.841	**76.086**	114.645	**77.879**
PSNR ↑	20.868	**22.513**	21.424	**22.519**	21.910	**22.954**
SSIM ↑	0.867	**0.882**	0.873	**0.887**	0.877	**0.893**
LIPIS ↓	0.131	**0.113**	0.119	**0.108**	0.120	**0.109**

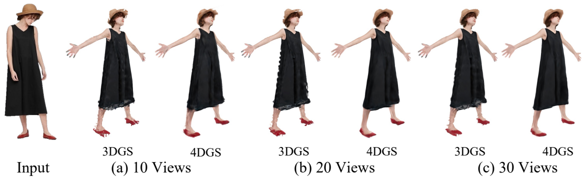


Figure S5. Novel-view result of 3DGS/4DGS with varying inputs.

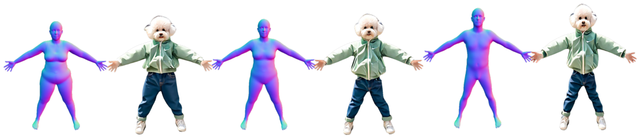


Figure S6. Controlling body size conditioned on different shapes.

3. More Results

3.1. Comparison of animation results

As is visualized in Fig. S1, our method produces accurate and photorealistic animation results than the baseline methods.

3.2. Pre-training on In-the-wild Data

Figure S4 underscores the critical role of pre-training on in-the-wild data. Models pre-trained on diverse and realworld datasets demonstrate substantially enhanced generalization capabilities compared to models trained without pre-training, verifying the training strategy of our method.

3.3. Animation Results

Figure S7–Figure S8 showcase the animation results of input human images with diverse appearances and a wide range of poses. Our method demonstrates the ability to generate animations that are both robust and photorealistic, preserving fine details of the human appearance while ensuring smooth and natural motion transitions. These results highlight the generalizability and effectiveness of our approach in handling varying levels of complexity in human avatars.

3.4. Reconstruction and Animation from Any Input

Figure S11 and Fig. S12 illustrate reconstructions and animation results from a diverse set of images collected from the internet. Notably, the reference image is a non-human image input, demonstrating the model's still maintain original diffusion model's generalizability.

3.5. Canonical Shape Reconstruction

To further validate the effectiveness of the proposed method, we provide additional results for canonical shape reconstruction from single images. Figure S13–Fig. S15 present reconstruction results on the DeepFashion dataset, showcasing accurate recovery of canonical shapes from fashion images. Meanwhile, Figure S16–Fig. S18 illustrate reconstructions from a diverse set of images collected from the internet, demonstrating the model's adaptability to various image sources and styles.

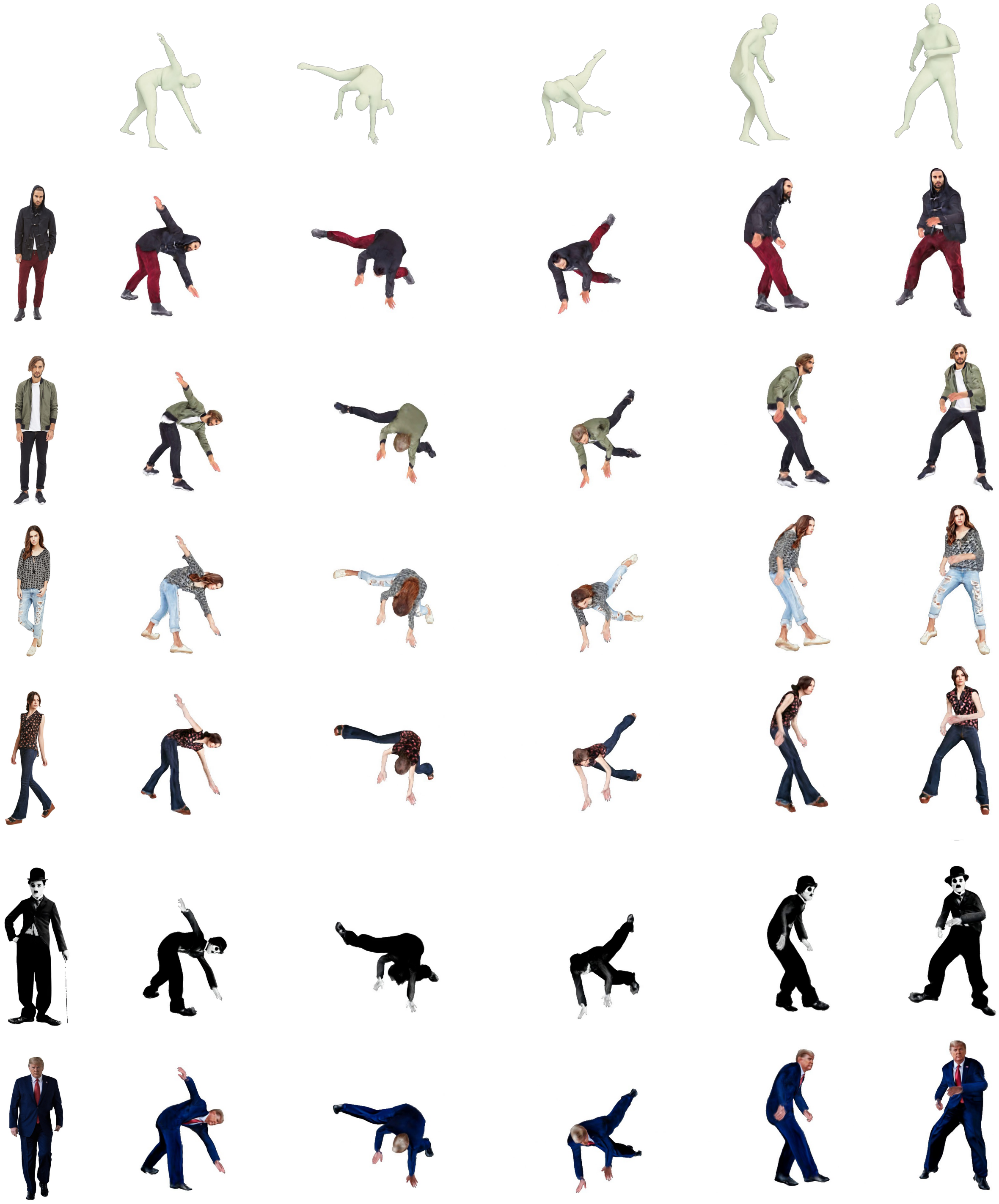


Figure S7. Visual results of human animation results (Part I) from any input. Best viewed with zoom-in.

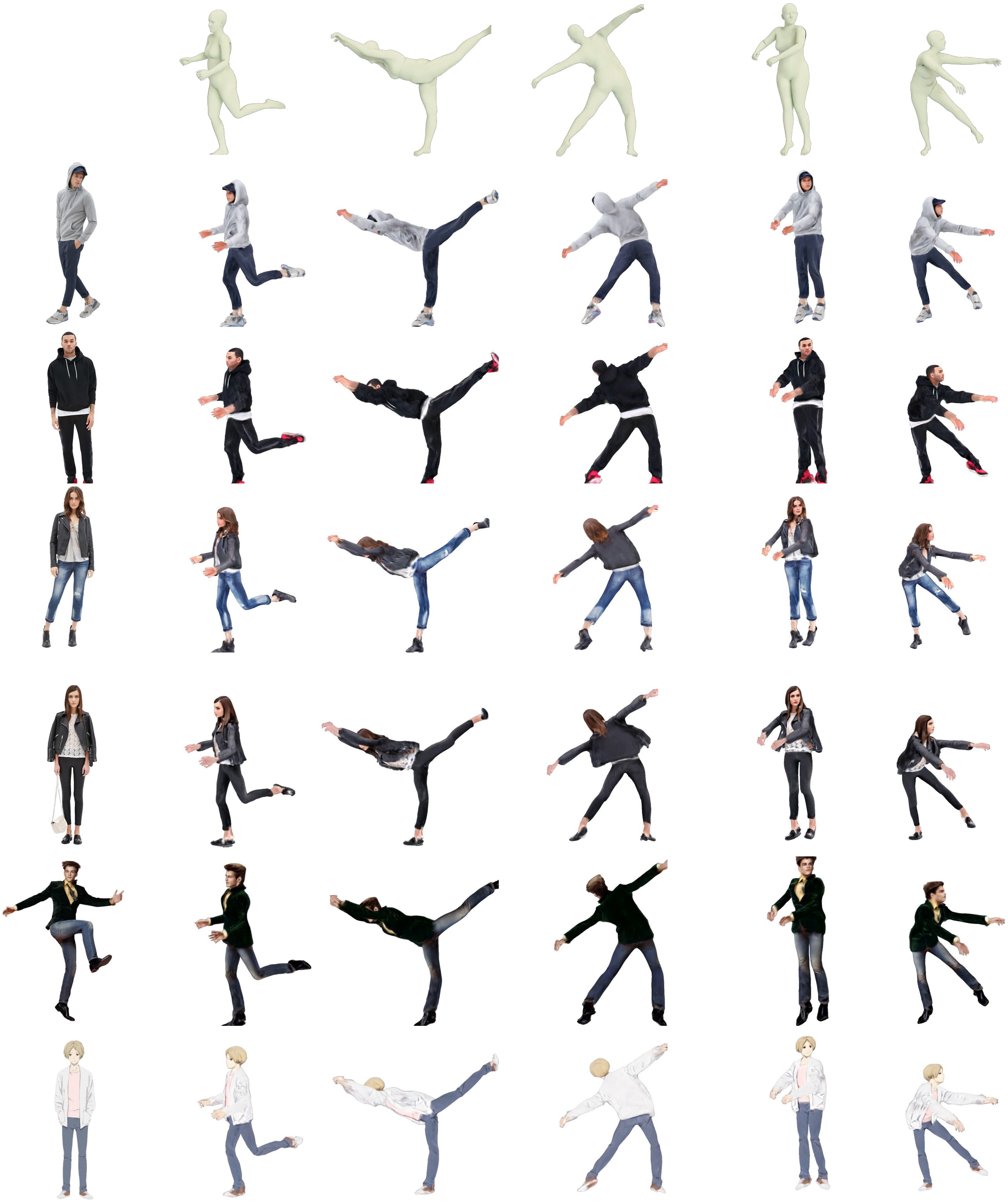


Figure S8. Visual results of human animation results (Part II) from any input. Best viewed with zoom-in.

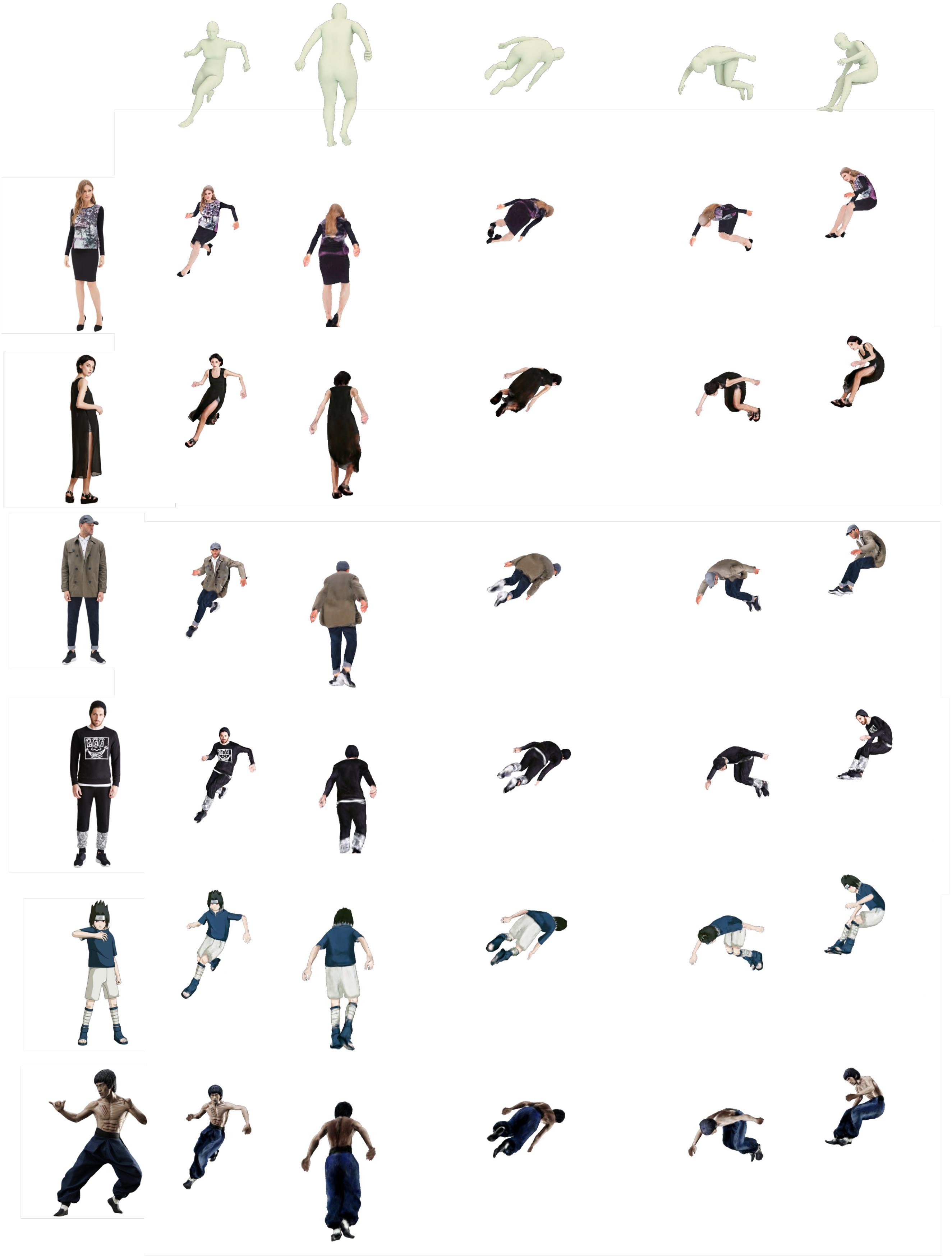


Figure S9. Visual results of human animation results (Part III) from any input. Best viewed with zoom-in.

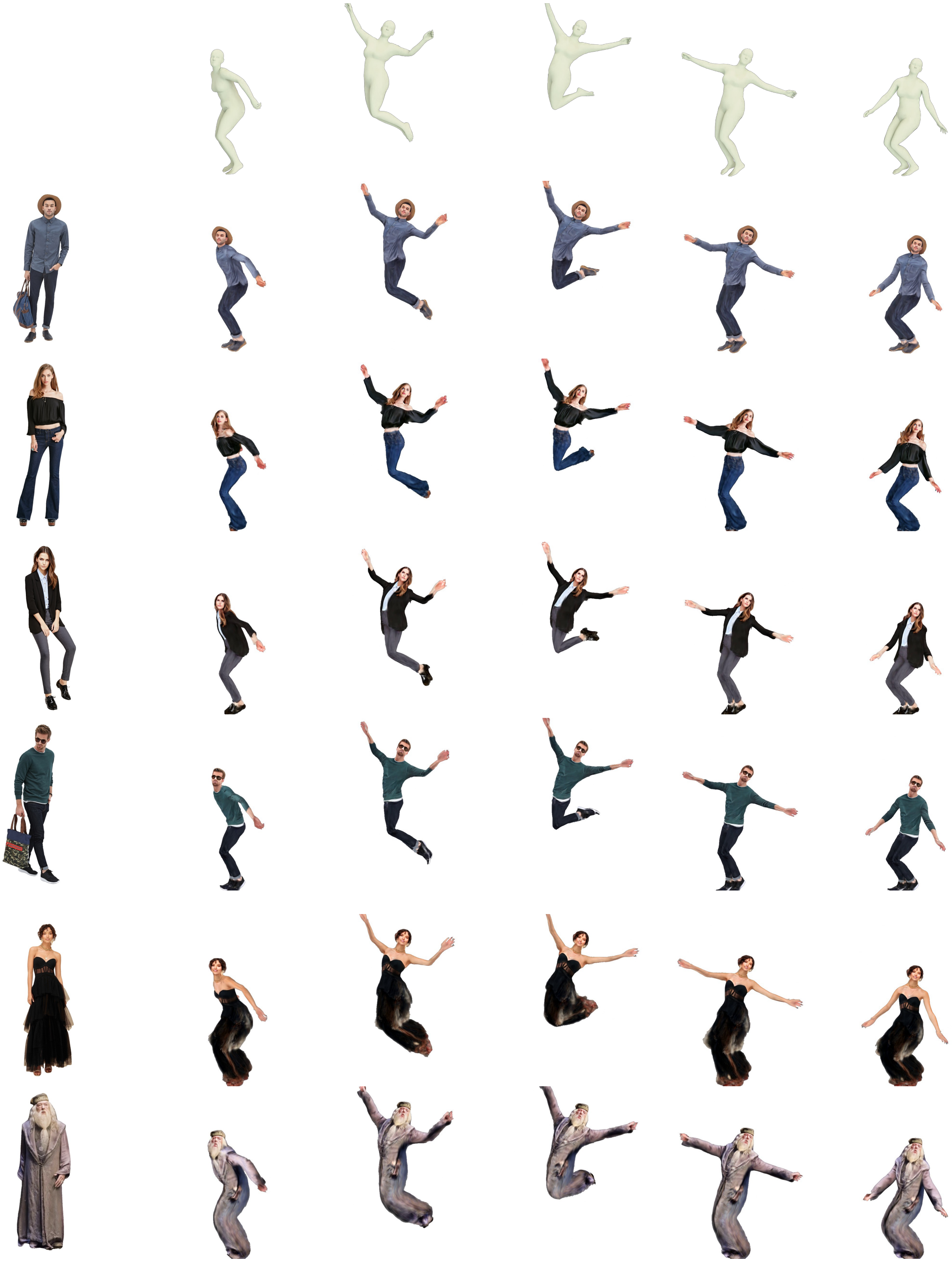


Figure S10. Visual results of human animation results (Part IV) from any input. Best viewed with zoom-in.

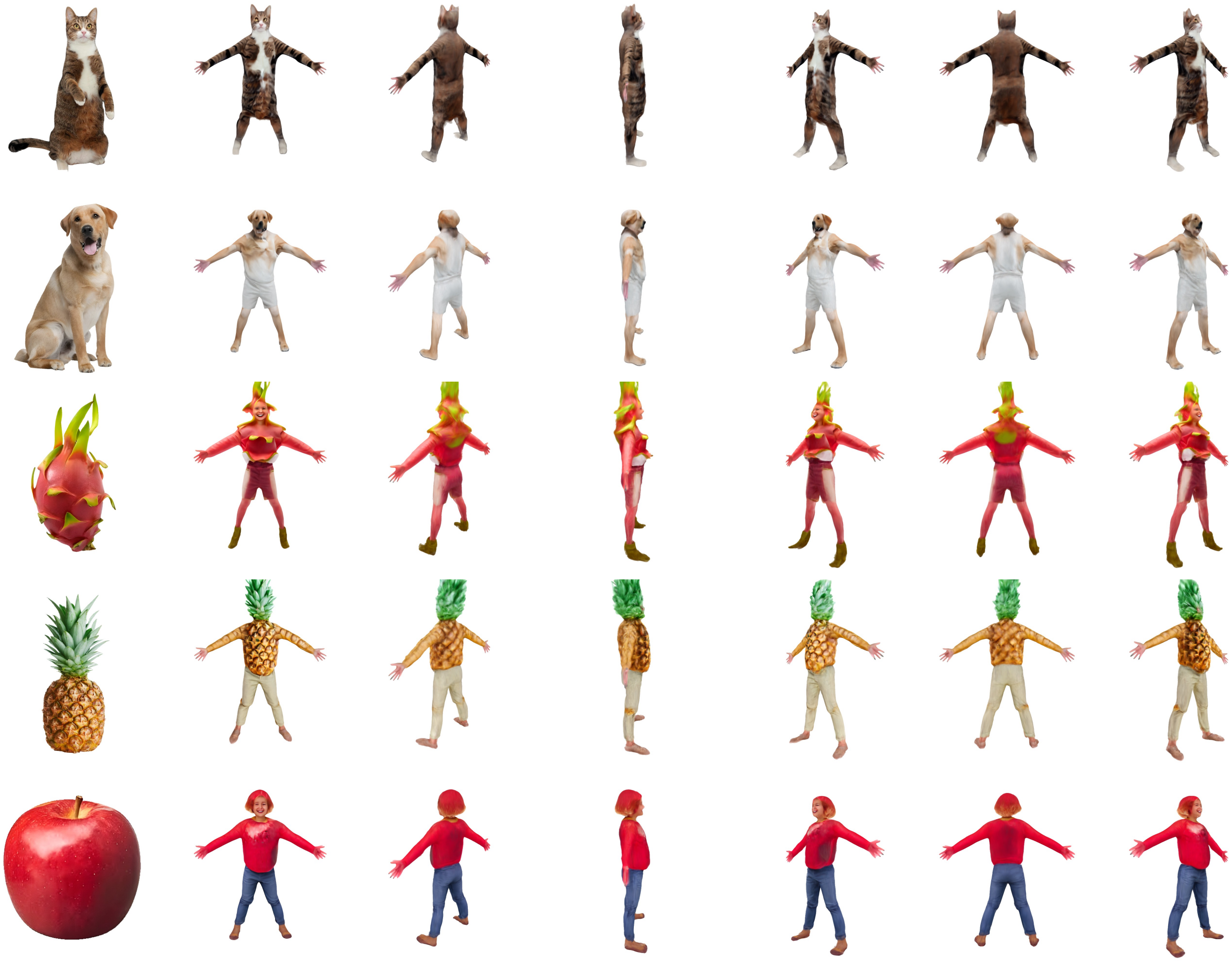


Figure S11. Visual results of canonical shape reconstruction from “Any Input”. Best viewed with zoom-in.

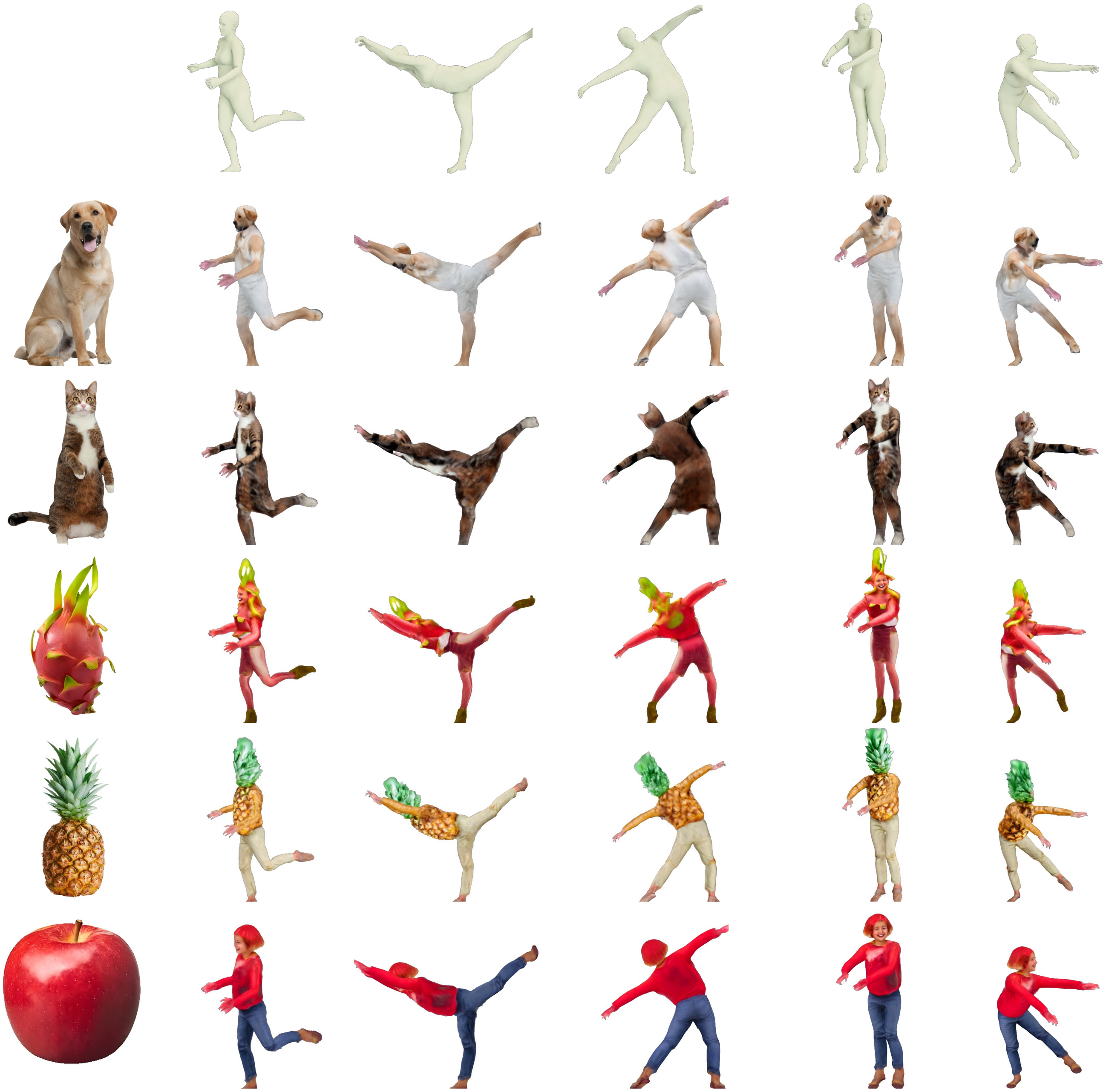


Figure S12. Visual results of canonical shape reconstruction from “Any Input”. Best viewed with zoom-in.

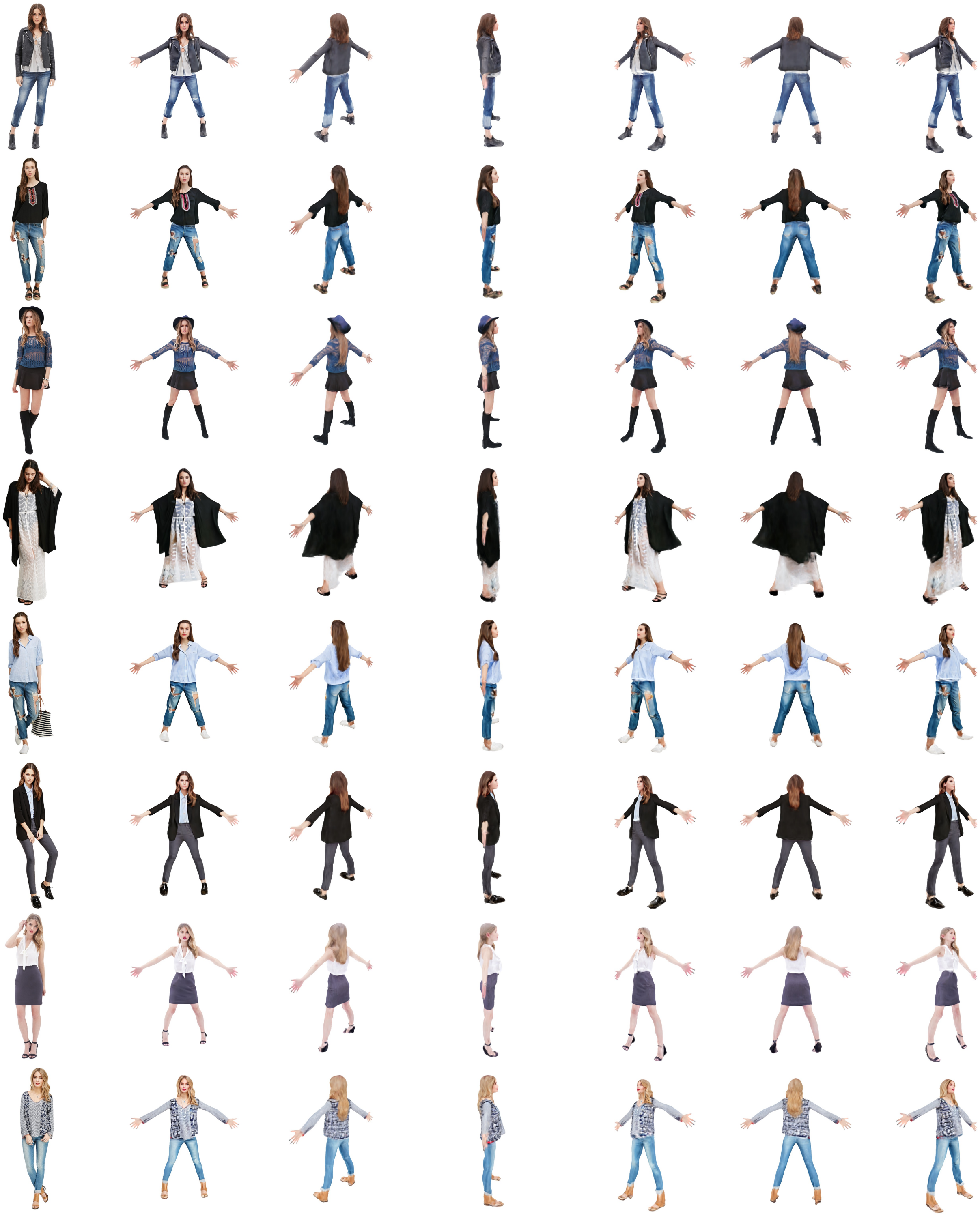


Figure S13. Visual results of canonical shape reconstruction (Part I). Best viewed with zoom-in.

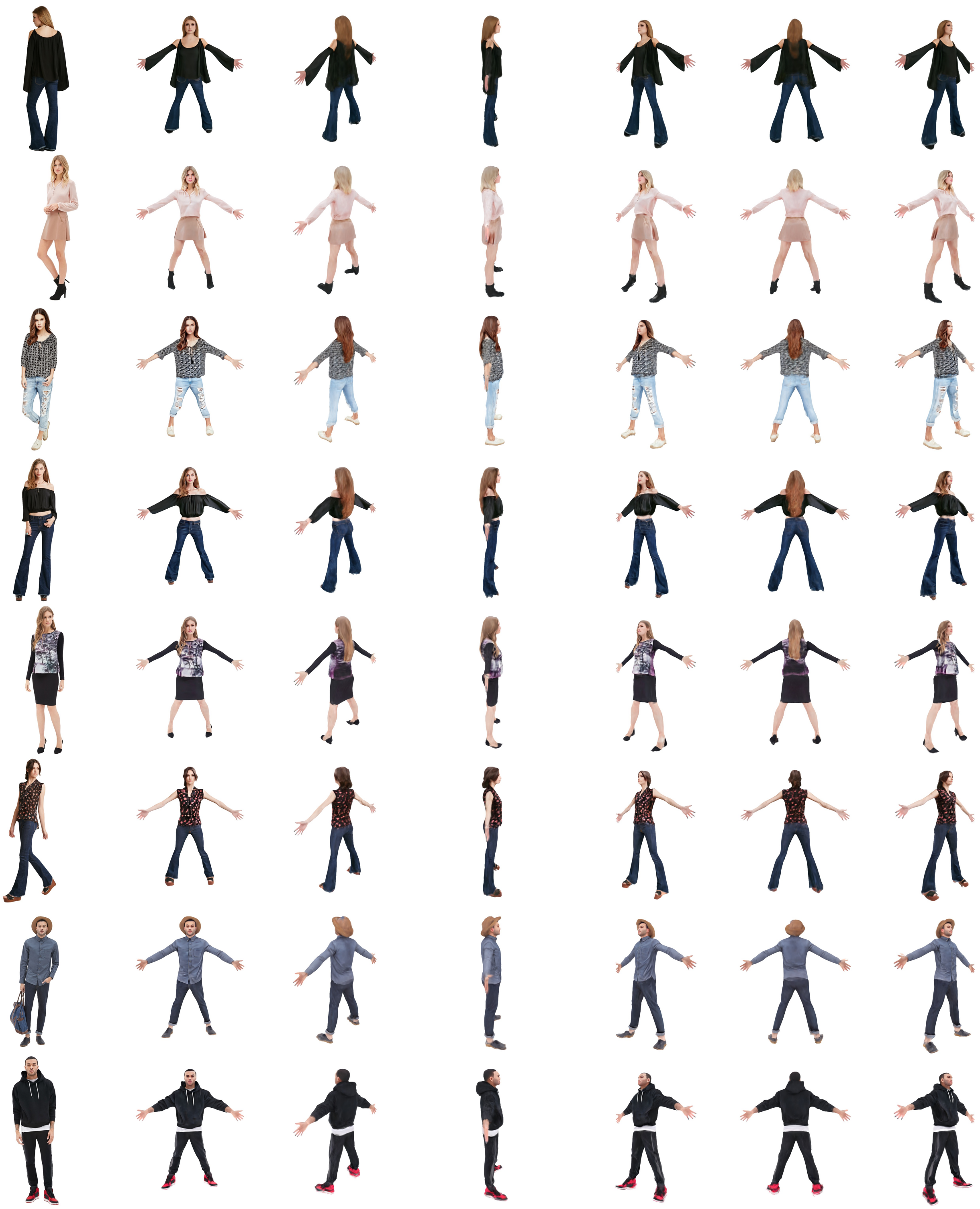


Figure S14. Visual results of canonical shape reconstruction (Part II). Best viewed with zoom-in.

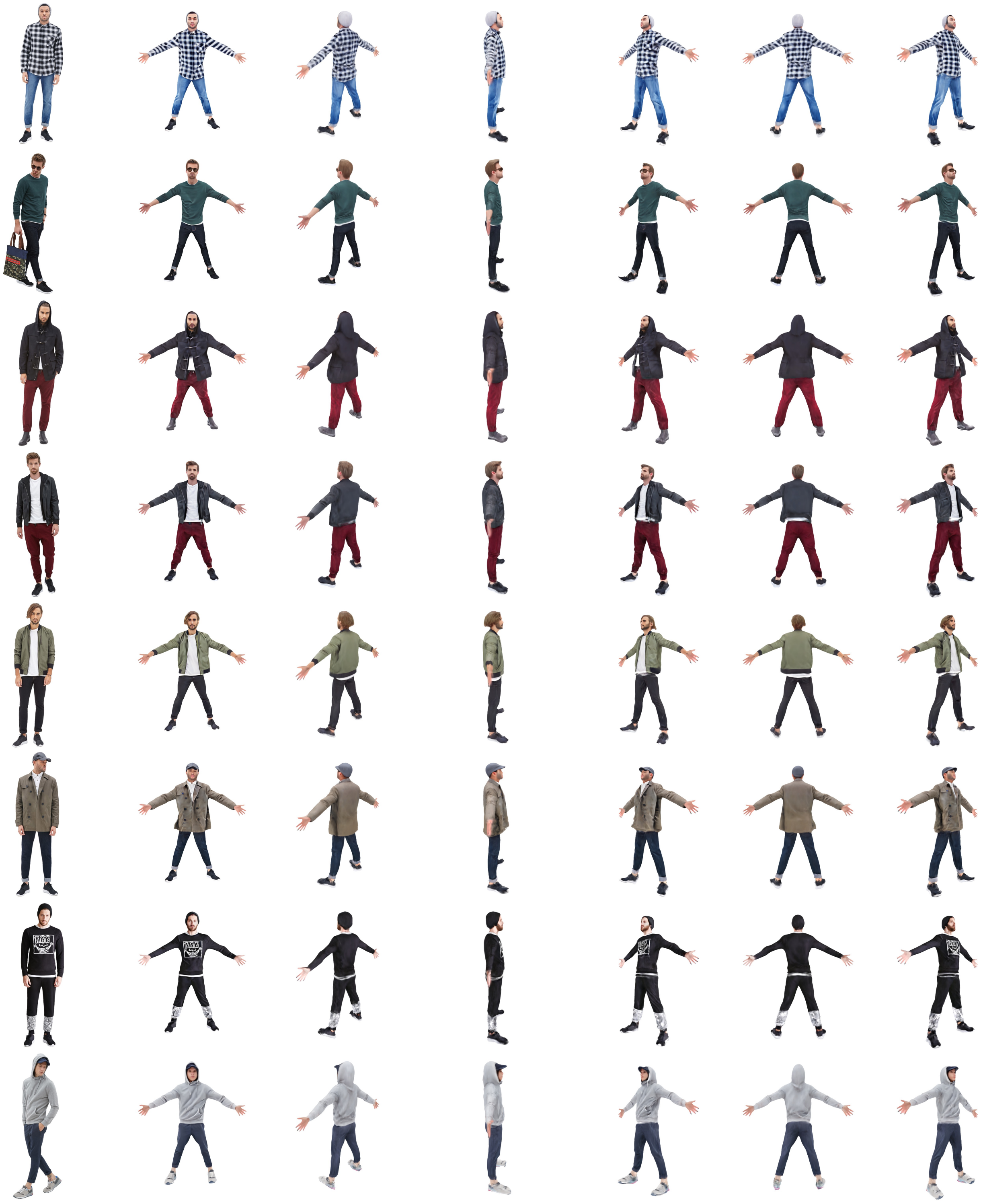


Figure S15. Visual results of canonical shape reconstruction (Part III). Best viewed with zoom-in.

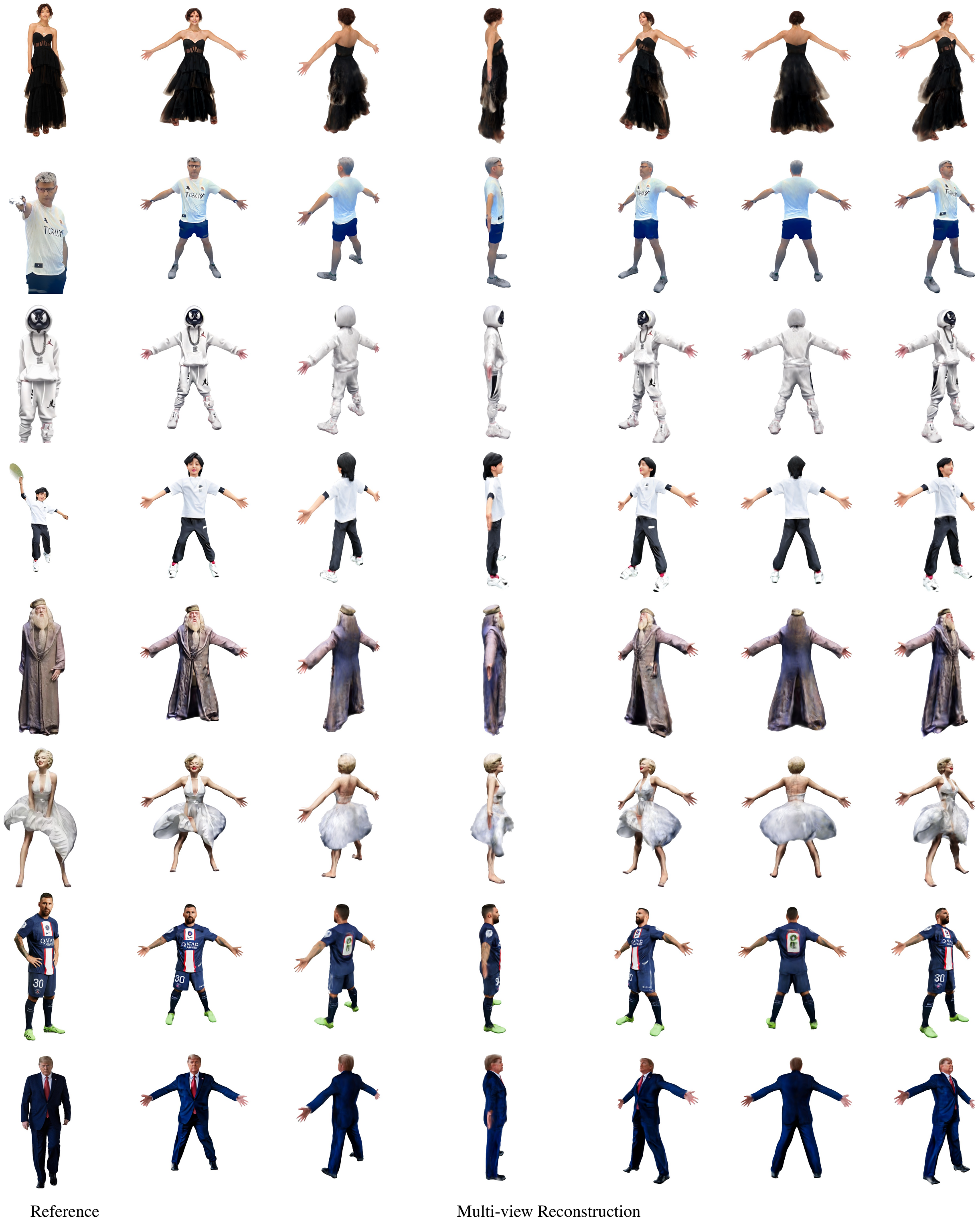


Figure S16. Visual results of canonical shape reconstruction (Part IV). Best viewed with zoom-in.

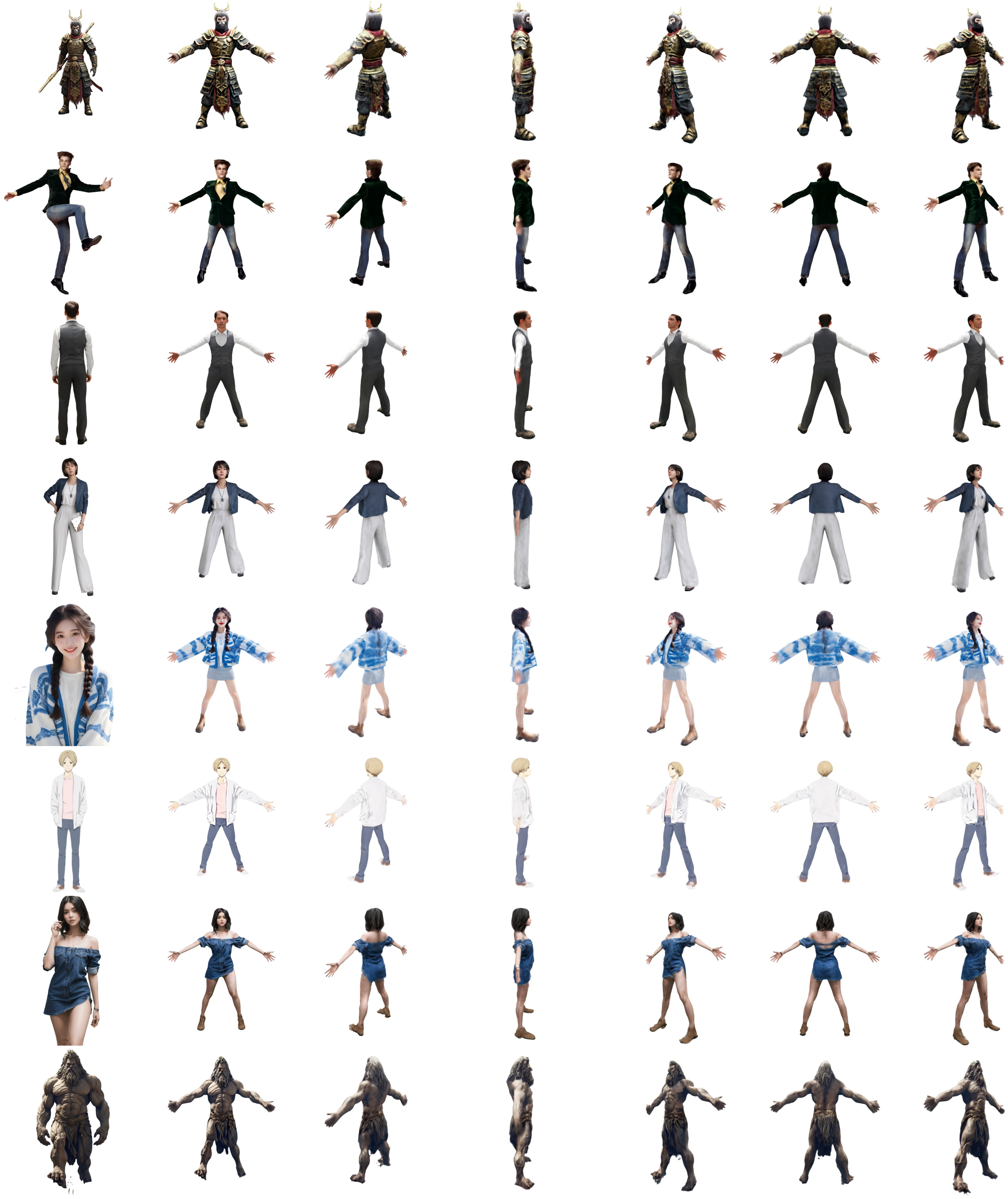


Figure S17. Visual results of canonical shape reconstruction (Part V). Best viewed with zoom-in.

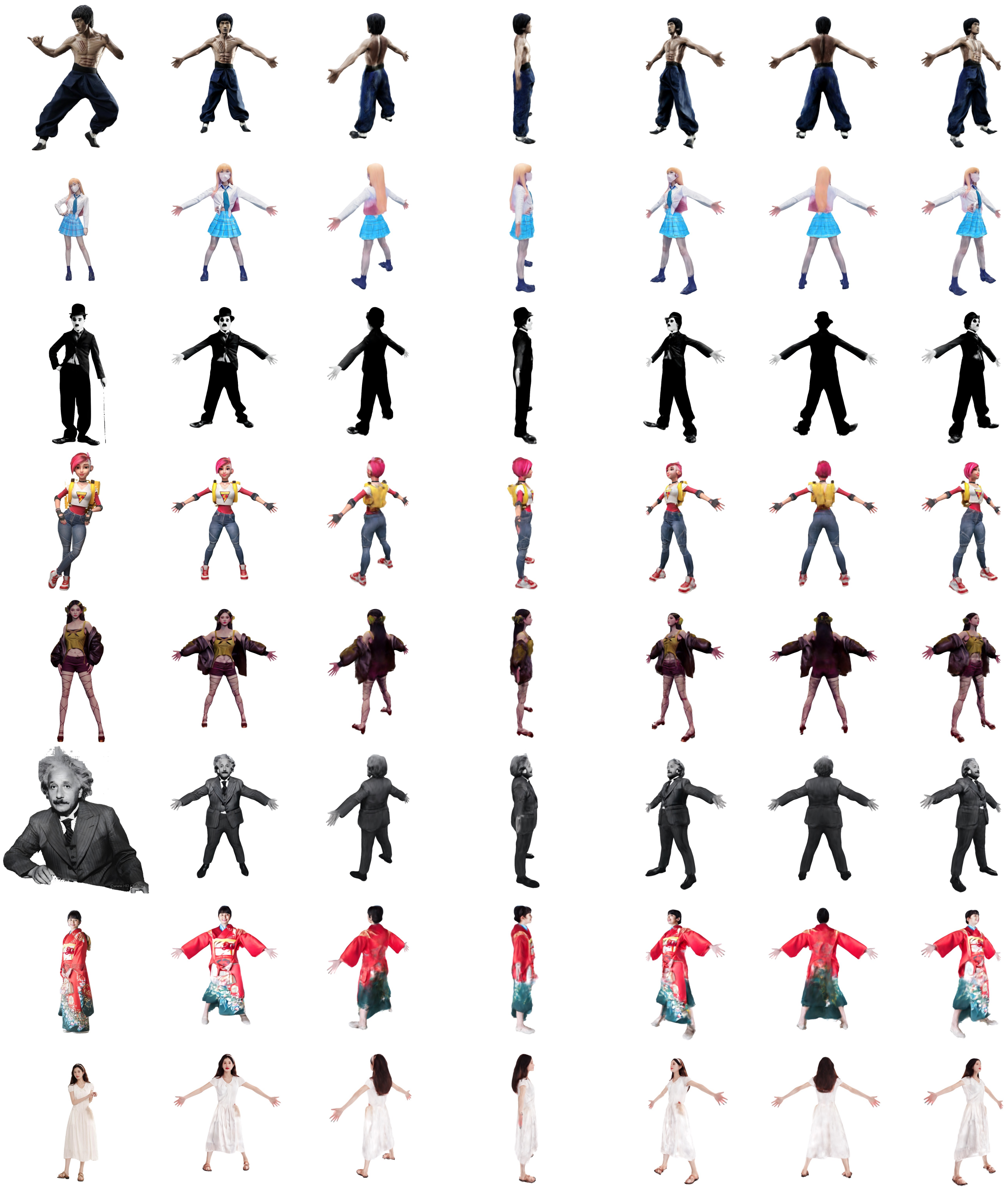


Figure S18. Visual results of canonical shape reconstruction (Part VI). Best viewed with zoom-in.

References

[1] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *CVPR*, 2023. 3

[2] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In *CVPR*, 2023. 3

[3] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *ECCV*, 2024. 1

[4] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, 2014. 1

[5] Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu. Learning implicit templates for point-based clothed human modeling. In *ECCV*, 2022. 2

[6] 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 *CVPR*, 2019. 2

[7] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *CVPR*, 2021. 3