Supplementary Material of AdaHuman: Animatable Detailed 3D Human Generation with Compositional Multiview Diffusion

Yangyi Huang[1,2] Ye Yuan[1] Xueting Li[1] Jan Kautz[2] Umar Iqbal[1]
[1]NVIDIA [2]The Chinese University of Hong Kong

A. Additional Results

Qualitative Comparison on Avatar Reposing. As shown in Fig. A1, direct deformation of characters from the reconstruction results in input poses could fail on body parts and clothing due to self-contact and self-occlusion. Directly generating the reposed 3DGS avatar $\mathcal{G}_{P_t}$ by AdaHuman could better generalize to a novel target pose P_t, synthesizing realistic details of the deformation of the clothe.

Animating Standard Pose Avatars. Fig. A2 showcases the animation results of using the animatable avatar from Avatar Reposing with a standard pose condition. Although the model is not directly trained with standard pose data, it learns to generalize to the standard poses with the help of the diverse distribution of poses in MVHumanNet[11].

B. Limitations and Future Directions

Despite the advancements, some limitations of our method warrant further exploration. The local refinement strategy may encounter difficulties with occluded or poorly covered regions, particularly around hands and arms, leading to artifacts and limiting fine-grained animation in these areas. Additionally, while our model can generate avatars in an animation-friendly standard pose, the animation capability still relies on the alignment of the SMPL body models and their skinning weights, which poses challenges in detailed animation such as facial expressions, hand gestures, and garment deformation. Future work could explore better integration of body models and simulation-based methods, as well as the use of video diffusion model to enhance the animation quality.

C. Implementation Details

Network Structure. In Fig. A3, we illustrate the architecture of our Pose-Conditioned Multi-View Image LDM model, along with the 3DGS generators $\mathbf{G}$ and $\mathbf{G}_{\text{comp}}$. For the LDM model, following [3], we enable 3D cross-view attention only in layers with a feature map resolution of $\leq 32 \times 32$. We also add extra input channels to the latent maps for camera ray maps, condition masks, and semantic pose maps. For $\mathbf{G}$, we adopt the architecture of the pre-trained LGM-big model [9] and include additional input channels for noisy images $\mathbf{x}_t$.

Additional, as an ablation mentioned at **??**, we hvae tried training a compositional 3DGS generator $\mathbf{G}_{\text{comp}}$ for Learnable Composition. Based on the LGM network, we insert an additional cross-part self-attention layer after each original cross-view self-attention layer in the LGM network. Note that the output image resolution of our LDM model is 512×512, which is then downsampled to 256×256, the input resolution for the 3DGS generator $\mathbf{G}$.

Ray Map Embedding. We use different methods to embed ray map information for the image LDM model and the 3DGS generators $\mathbf{G}$ and $\mathbf{G}_{\text{comp}}$. For the 3DGS generators, to effectively utilize the pretrained weights of LGM, we scale the entire scene to ensure a camera distance of $r = 1.5$ meters and use Plücker ray embeddings as described in Eq. 4 of the main text.

For the LDM model, we employ sinusoidal positional embeddings [10] to encode ray origins and directions, providing rich information about 3D locations across different cropping scales:

$$\mathcal{R}_{\text{LDM}}(i,j) = \text{PE}(\mathbf{o}(i,j), \mathbf{d}(i,j)) \tag{1}$$

where PE is the sinusoidal positional encoding function, with the number of octaves N_{octaves} set to 8.

View Sampling. Since our training data consists of multi-camera video captures in a 3D scene, the avatar is not always positioned at a standard location. We use 2D joint locations and foreground mask areas to crop global and local training views, resizing them to a resolution of 512×512. In Tab. 1, we list the OpenPose joints used to determine the cropping centers and relative size ratios of the local crops.

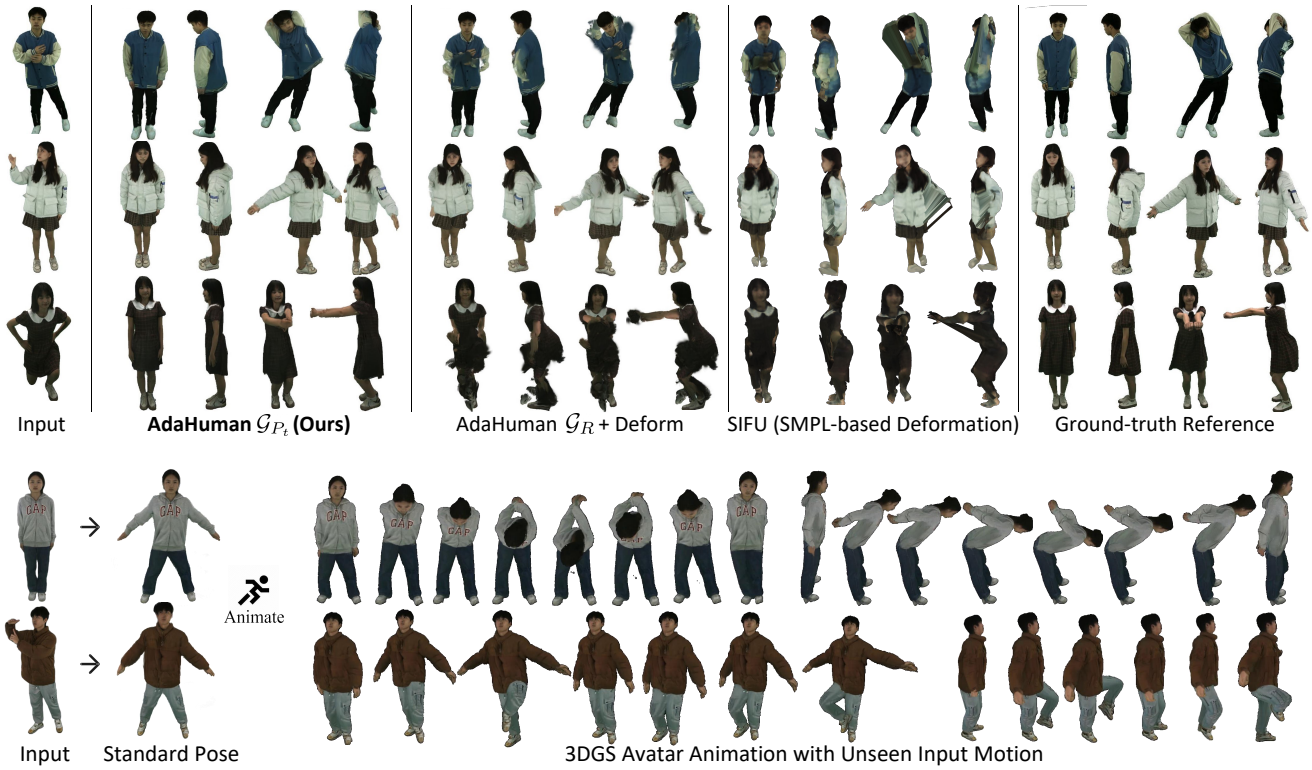

Figure A2. AdaHuman generates animation-ready avatar in a standard pose, which can be animated with unseen input motion.

During inference, after obtaining coarse reconstruction results with global views, we render $N_v = 20$ views to estimate 3D joints using EasyMocap [1], which helps sample local views for our compositional 3DGS refinement.

Parts	Full body	Upper Body	Lower Body	Head
Joints	Pelvis	Neck	Left Ankle, Right Ankle	Left Ear, Right Ear
Scale	1.0	0.5	0.5	0.25

Table 1. Body part sampling details.

Training Schedule. We initialize our LDM model with the official weights of `stable-diffusion-v1-5`[1] [8] and our 3DGS generator $\mathbf{G}$ with LGM-big[2] [9].

For training the LDM model weights θ, the model first learns to predict $K = 3$ canonical views from one input view ($V = 1$) without pose conditioning. We fine-tune the model on predicting global full-body views for $20,000$ iterations, followed by fine-tuning on all $N_p + 1 = 4$ global and local view for another $30,000$ iterations to obtain $\theta_{\text{no\_pose}}$. Finally, we fine-tune the pose-conditioned model weights $\theta_{\text{novel\_pose}}$ from $\theta_{\text{no\_pose}}$. This model learns to predict $K = 4$ canonical views of a novel pose avatar from $V = 1$ input views sampled from different frames in the same video sequence. The novel pose synthesis model is fine-tuned for $1,0000$ iterations using all $N_p + 1 = 4$ global and local views.

For training the 3DGS generator model $\mathbf{G}$, we first fine-tune it from pre-trained weights using clean full-body images in MVHumanNet [11] for $2,000$ iterations to adapt it for human reconstruction. Then, we randomly sample diffusion timesteps to train with both noisy inputs $\mathbf{x}_t$ and clean inputs $\mathbf{x}_0$ for $20,000$ iterations. The 3DGS model $\mathbf{G}$ is also fine-tuned on local views for an additional $20,000$ iterations. We use $N_{\text{ref}} = 12$ reference views of each part to supervise the predicted 3DGS.

All training processes are conducted on 16 NVIDIA A100 80GB GPUs, with a total batch size of $n_{\text{batch}} = 128$ and a learning rate of $\eta = 5 \times 10^{-5}$.

Training Losses. The training losses for the pose-conditioned LDM and the 3DGS generator are as follows:

$$\mathcal{L}_{\text{LDM}} = \mathcal{L}_{\text{MSE}}(\epsilon, \epsilon_\theta) \tag{2}$$

$$\mathcal{L}_{\mathbf{G}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} \tag{3}$$

$$\begin{aligned}\mathcal{L}_{\text{recon}} = {} & \lambda_{\text{MSE}}\mathcal{L}_{\text{MSE}}(\hat{\mathbf{x}}^{t\to 0}_{\text{novel}}, \mathbf{x}_{\text{novel}}) \\ & + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}}(\hat{\mathbf{x}}^{t\to 0}_{\text{novel}}, \mathbf{x}_{\text{novel}})\end{aligned} \tag{4}$$

where the training loss of LDM, denoted as $\mathcal{L}_{\text{LDM}}$, is the MSE loss of the predicted latent noise. The training loss of $\mathbf{G}$ consists of rendering reconstruction loss computed using

[1] https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5

[2] https://huggingface.co/ashawkey/LGM/resolve/main/model_fp16_fixrot.safetensors

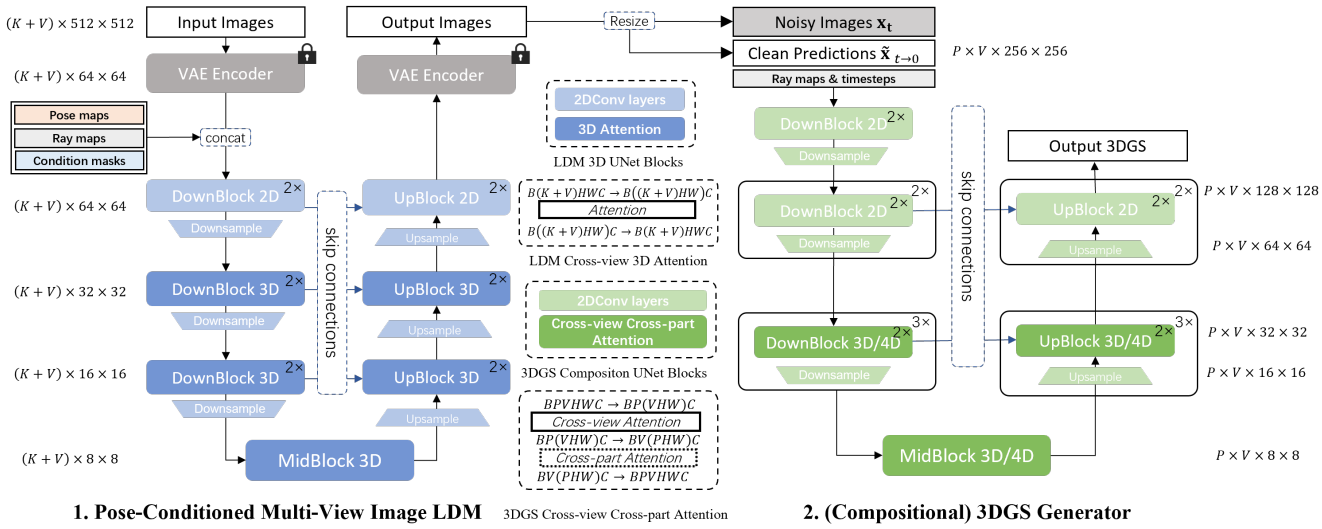

Figure A3. **Network Architectures** of (1) Pose-Conditioned Multi-View LDM Model and (2) Compositional 3DGS Generator.

MSE and LPIPS. Following [12], we also incorporate the 3DGS regularization loss from [6, 14] to enhance surface quality.

Inference. This section details the inference pipeline of avatar reconstruction and avatar reposing our method. In both settings, we perform 3D joint diffusion on global views only when $t \in (500, 900]$ to maintain the stability of the diffusion process. The earlier steps focus on pure 2D diffusion to generate more detailed appearances. During image-to-image local refinement, we utilize SDEdit [7] with a strength of $s = 0.5$, meaning that denoising begins at $t = 500$ and 3D joint diffusion is performed when $t \in (350, 500]$.

D. Evaluation Settings

Baseline Models. Our baseline methods, including Human3Diffusion [12], LGM [9], SiTH [5], and SIFU [15], have been trained on various 3D mesh datasets [2, 4, 13]. In this work, our aim is to demonstrate the advantages of training models on both mesh datasets and video datasets for better pose generalization and the synthesis of novel pose characters. We utilize their official weights for comparison. We also note that some models (e.g. [12]) rely on private data or synthesized meshes for training.

Avatar Reconstruction. We selected front views of the mesh avatar as input views, rendered by horizontal perspective cameras for a fair and realistic comparison. The results of the quantitative evaluation are rendered at a resolution of 1024×1024 using 20 perspective cameras.

Avatar Reposing. For SiTH [5] and SIFU [15], we deform their avatars to the target pose and align the avatar meshes with the ground-truth SMPL meshes to render images for evaluation.

References

[1] Easymocap - make human motion capture easier. Github, 2021. 2

[2] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13142–13153, 2023. 3

[3] Ruiqi Gao*, Aleksander Holynski*, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole*. CAT3D: Create anything in 3d with multi-view diffusion models. In *NeurIPS*, 2024. 1

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

[5] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *CVPR*, 2024. 3

[6] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH*. Association for Computing Machinery, 2024. 3

[7] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 3

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[9] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. LGM: Large multi-view gaussian model for high-resolution 3d content creation. In *ECCV*, 2025. 1, 2, 3

[10] A Vaswani. Attention is all you need. *Adv. Neural Inform. Process. Syst.*, 2017. 1

[11] Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, et al. MVHumanNet: A large-scale dataset of multi-view daily dressing human captures. In *CVPR*, 2024. 1, 2

[12] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard Pons-Moll. Human 3diffusion: Realistic avatar creation via explicit 3d consistent diffusion models. In *NeurIPS*, 2024. 3

[13] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 3

[14] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. *ACRM Trans. Graph.*, 2024. 3

[15] Zechuan Zhang, Zongxin Yang, and Yi Yang. SIFU: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *CVPR*, 2024. 3