Supplementary Materials of SIGMAN: Scaling 3D Human Gaussian Generation with Millions of Assets

Yuhang Yang[1,2,*], Fengqi Liu[2,3*], Yixing Lu[4], Qin Zhao[2], Pingyu Wu[1], Wei Zhai[1,†], Ran Yi[3]
Yang Cao[1], Lizhuang Ma[3], Zheng-Jun Zha[1], Junting Dong[2,†]
[1] USTC [2] Shanghai AI Lab [3] SJTU [4] CMU
∗Equal Contribution †Corresponding Author
https://yyvhang.github.io/SIGMAN_3D/

1. Appendix

1.1. Details of HGS-1M Dataset

To construct the HGS-1M dataset, we first aggregate publicly available multi-view human datasets [2, 3, 11] and process them with the AnimatableGaussians framework [7], which optimizes dynamic human sequences into static 3D Gaussian representations. For each human sequence, we perform per-subject optimization for approximately 20 hours on an NVIDIA 4090 GPU, leveraging differentiable rendering and skeletal priors from the SMPL-X model to ensure geometric consistency across poses. The optimization process aligns each Gaussian sequence to a canonical space by extracting SMPL-X root rotation and translation parameters, then rigidly transforming all Gaussians to the origin. This step standardizes positional coordinates across diverse datasets, eliminating inconsistencies in global orientation and scale. For datasets lacking SMPL-X annotations (e.g., synthetic assets), we fit SMPL-X parameters using EasyMoCap [1] before alignment.

After alignment, we render each human Gaussian from 90 viewpoints to maximize supervision coverage. The camera setup includes 30 horizontal views (azimuth angles spaced at 12° intervals), 30 upward views (elevation +15° to +90°, azimuth spaced at 30° intervals), and 30 downward views (elevation -15° to -90°), ensuring dense angular sampling for full 360° reconstruction. For small-scale datasets like Thuman2.1 [12] and 2K2K [5], we apply identical rendering protocols to maintain consistency. To further enhance diversity, we integrate 100k synthetic human assets generated via parametric body models (e.g., SMPL-X) and procedurally augmented with varied textures, clothing meshes, and lighting conditions. These synthetic assets are converted into Gaussians using the same optimization pipeline, with their root transformations reset to the origin. The final dataset combines optimized real-world sequences, rendered multi-view data, and synthetic samples, totaling 1 million human 3D Gaussians.

We report the statistical content of the following aspects in the dataset to better display the content of the dataset, including the gender, race, and age composition, shown in the table 1.

Man	Woman	Asian	White	Black	Hisp.	$< 20s$	20-50s	$> 50s$
51.6%	48.4%	69.3%	13.6%	9.0%	8.1%	34.0%	59.2%	6.8%

Table 1. Distribution of certain aspects of the HGS-1M dataset.

1.2. Method Details

VAE. The encoder of VAE that accepts multi-view input consists of 4 3D-convolutional blocks, which downsamples the H and W of the original image by 8 times. The view dimension is not downsampled, and the input image size is 512×512. For learnable tokens, the initial width and height are the same as the size after the VAE 3D convolution encoder downsamples. In addition, for the initialized UV map, we selected 16 viewpoints to project back the RGB value to the mesh and extract the output UV map value. After that, this UV map is encoded through a 1×1 2D convolutional block into the feature dimension of the learnable token and concatenated with it as the final initialized token. After cross attention, we use 6 Conv-Attn dual branch blocks to model the latent, and the final latent size is $64 \times 64 \times 16$. The VAE decoder includes 4 2D convolution blocks, and finally upsamples the UV map with Gaussian attributes to 512×512 for sampling. After this, multiple decode heads corresponding to various Gaussian attributes are employed to obtain the final Gaussian.

MM-DiT We use 2D rotation position encoding (RoPE) and RMS-norm in the DiT architecture. According to our observation, it is necessary to add RMS-Norm to the training from scratch. After that, the model is more robust to the learning rate and can converge normally under multi-

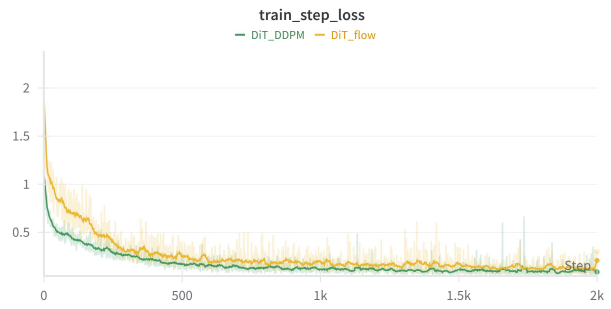

Figure 1. Changes in DDPM v-prediction and Rectified flow training loss curves in the early stages of training

ple learning rates. For the final 2B model, a total of 30 MM-DiT blocks are included, with 32 attention heads in each block, for a total of 64 heads. For the DIT training objective, in our experiments, the difference between using DDPM and flow matching is not obvious. For flow matching, we use the same noise addition method and sampling shift strategy as SD3 [4], and finally do not observe a significant performance improvement over DDPM. We observe that DDPM converges faster in the early stage when training from scratch, as shown in the Fig. 1.

Baselines. **1) GHG** [6]: The official checkpoint released by GHG [6] is trained on the THuman2.0 [12] dataset, which contains approximately 500 3D human subjects. To ensure a fair comparison, we fine-tune GHG on part of the HGS-1M dataset. For each subject, we render nine views evenly distributed across azimuth and altitude. For the inpainting network, we directly use the official checkpoint. During training, following GHG, we use three fixed horizontal views as input and apply multi-view supervision on three randomly sampled views with evenly distributed azimuth. During inference, we use the same three input views to generate novel viewpoints. **2) LGM** [10], if we remove the step of generating multi-view images from a single image in front of LGM, LGM is a process of outputting Gaussian images from multiple perspectives. Given this, we train its second stage and directly give 4 GT perspectives during inference, removing the multi-view generation step. In theory, LGM can output the best results. **3)** SIFU [13] is a per-subject optimization method for single-image 3D human reconstruction. For each subject, we use the front view as input to generate a textured 3D human model. **DiffSplat** [8], We use Diffsplat to compare the text-to-3D Gaussian. We use its original VAE and fine-tune it with our text and 3D data pairs. Finally, we calculated the CLIP score [9] indicator on 100 test samples. The result of our model is 25.89, while the result of DiffSplat is 24.62. We also provide some visual results of the text-to-3D, generated by our methods, please refer to Fig. 2.

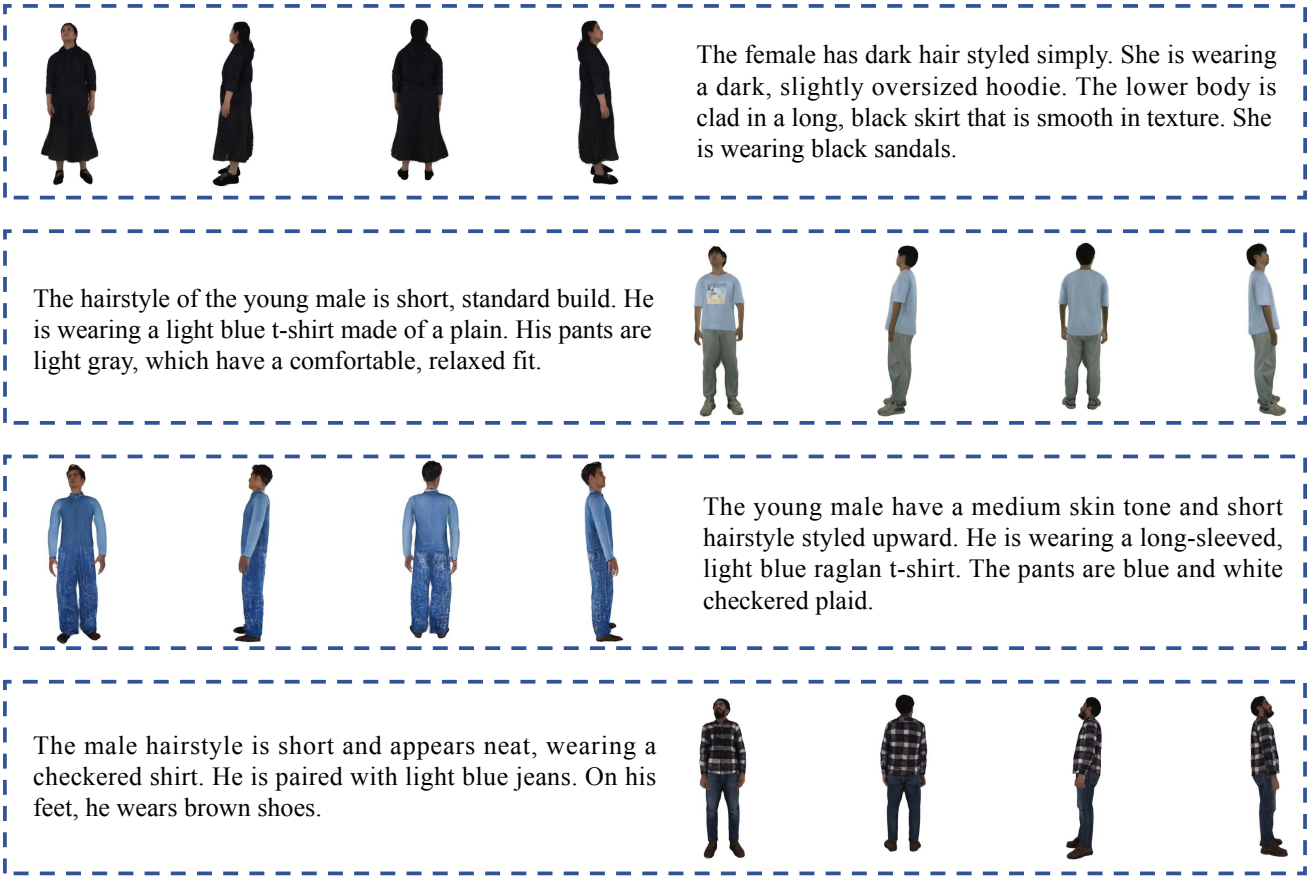

Figure 2. Results of Text-to-3D human Gaussians.

Figure 3. Visualization of sampled cases from our HGS-1M Dataset.

Figure 4. Results of single image-to-3D human Gaussians.

References

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

[2] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision*, pages 557–577. Springer, 2022. 1

[3] Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, et al. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19982–19993, 2023. 1

[4] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12869–12879, 2023. 1

[6] Youngjoong Kwon, Baole Fang, Yixing Lu, Haoye Dong, Cheng Zhang, Francisco Vicente Carrasco, Albert Mosella-Montoro, Jianjin Xu, Shingo Takagi, Daeil Kim, et al. Generalizable human gaussians for sparse view synthesis. In *European Conference on Computer Vision*, pages 451–468. Springer, 2024. 2

[7] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[8] Chenguo Lin, Panwang Pan, Bangbang Yang, Zeming Li, and Yadong Mu. Diffsplat: Repurposing image diffusion models for scalable gaussian splat generation. *arXiv preprint arXiv:2501.16764*, 2025. 2

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[10] 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 *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 2

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19801–19811, 2024. 1

[12] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5746–5756, 2021. 1, 2

[13] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9936–9947, 2024. 2