Diffuman4D: 4D Consistent Human View Synthesis from Sparse-View Videos with Spatio-Temporal Diffusion Models

Supplementary Material

A. Model Details

Training. We initialize our model from Stable Diffusion 2.1 [12] developed by Diffusers [15]. We apply full-parameter fine-tuning to the diffusion model for 200k iterations with a batch size of 32 and a learning rate of 10^{-5} using 32 NVIDIA H20 GPUs. To accommodate input images with the conditions, we expand the input channels of the model's first convolutional layer from 4 to 15, consisting of 4 channels for image latents, 4 for skeleton latents, 6 for Plücker embeddings, and 1 for a conditional mask. Following the previous work [4], the conditional mask is a binary indicator specifying whether an image serves as a conditioning input or a target.

Sampling. Following Stable Diffusion 2.1 [12, 15], we use DPM-Solver++ [11] with 24 sampling steps and a classifier-free guidance scale of 3.0. Our sliding iterative denoising strategy takes approximately 2 minutes to generate a sample sequence of length 48 when executed on a single A100 GPU. To improve efficiency, we parallelize the denoising process across 8 A100 GPUs.

4D reconstrcution. We employ LongVolcap [17] to reconstruct the 4D human performances from the generated multi-view videos. LongVolcap is an enhanced version of 4DGS [18] with the ability of effectively reconstructing long volumetric videos with a temporal Gaussian hierarchy representation. We initialize the 4D Gaussian primitives with the coarse geometry obtained using the predicted foreground masks and the space carving algorithm [9, 16]. We then follow the same training and evaluation settings as in the original paper [17] to reconstruct the 4D human performances. Specifically, we optimize the model with the Adam optimizer [8] with a learning rate of $1.6e^{-4}$, each model is trained for 100k iterations for a sequence of 7200 frames, which takes around 1 hour on a single NVIDIA RTX 4090 GPU.

B. Datasets Details

We conduct extensive processing on the original DNA-Rendering [1] dataset to generate high-quality multi-view videos along with additional masks and skeletons for training and evaluation. The processing pipeline includes camera re-calibration, color correction matrices (CCMs) optimization, foreground mask prediction, and human skeleton estimation. We provide detailed descriptions of each step below.

Camera calibration. We empirically found that the camera parameters provided in the DNA-Rendering dataset are not accurate enough for reconstruction verified with 3D Gaussian Splatting (3DGS) [6]. In order to achieve pixel-level accuracy, we first re-calibrated the camera parameters using Colmap [13, 14]. We then optimized the color correction matrix for each camera to ensure consistent color across different views.

Foreground mask prediction. There are only a few (around 1/6) sequences in the DNA-Rendering dataset that provide ground truth foreground masks. To obtain accurate foreground masks, we leverage three state-of-the-art background removal methods, namely RMBG-2.0 [19], BiRefNet-Portrait [19], and BackgroundMattingV2 [10], and combine their predictions using a voting mechanism to fully leverage the strengths of each approach. Specifically, we found that RMBG-2.0 may incorrectly recognize background objects as foreground, BiRefNet-Portrait may segment small objects as background, and BackgroundMattingV2 may produce inaccurate results for certain human poses. We demonstrate the effectiveness of the voting strategy in Fig. 1.

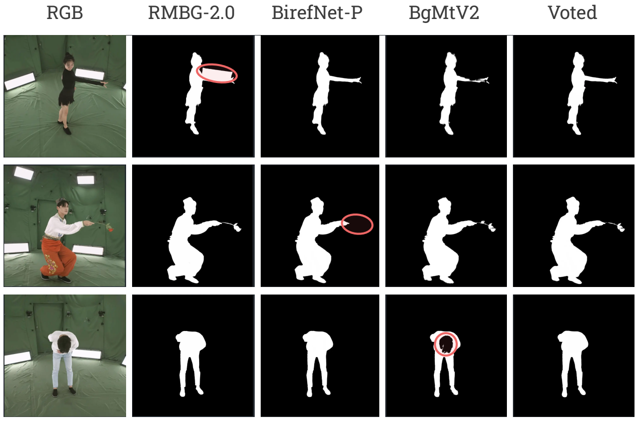

Figure 1. Our voting strategy effectively leverages the strengths of different background removal methods to produce robust foreground masks.

Human skeleton estimation. Similar to the foreground mask, only a few sequences have ground truth human skeletons. We thus adopt the state-of-the-art human skeleton estimation model, Sapiens [7], to predict the 2D human skeleton for each frame. We additionally adjust the transparency of the skeleton colors based on the confidence scores of the

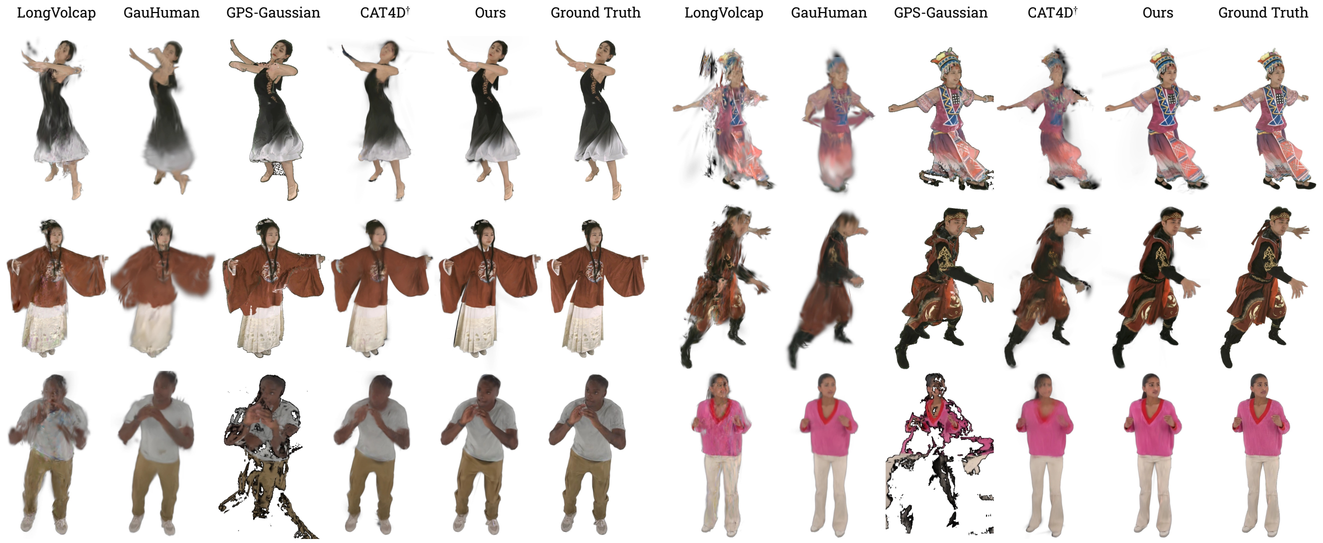

Figure 2. **More qualitative comparisons on DNA-Rendering [1] and ActorsHQ [5].** GPS-Gaussian uses 8 input views while all others use 4 input views, and CAT4D† is our reproduced version. Our Diffuman4D consistently outperforms baselines with higher visual quality and better spatio-temporal consistency.

skeleton, which helps to visualize and encode the uncertainty of the skeleton estimation. After obtaining the 2D human skeletons, we then triangulate them to obtain the 3D human skeleton sequence, which can be further used for projection and evaluation.

We demonstrate the processed data samples in Fig. 3. We plan to release the additionally processed data under the DNA-Rendering open-source license to facilitate future research within the community.

Dataset filtering. DNA-Rendering [1] contains many scenes involving human-object interactions, such as writing on a desk, playing guitar, or organizing items. Since the diversity of objects is significantly greater than that of humans, training generative models typically requires extensive object datasets (e.g., Objaverse [2, 3]). To address the relatively limited scale of the DNA-Rendering dataset, we employed the Llama Vision 3.2 model to classify all scenes and filtered out those containing large objects to avoid potential model collapse during training.

Nevertheless, we observe that even though the training dataset does not include objects, our model successfully generalizes to scenes featuring simple objects, such as the basketball player shown in Fig. 4.

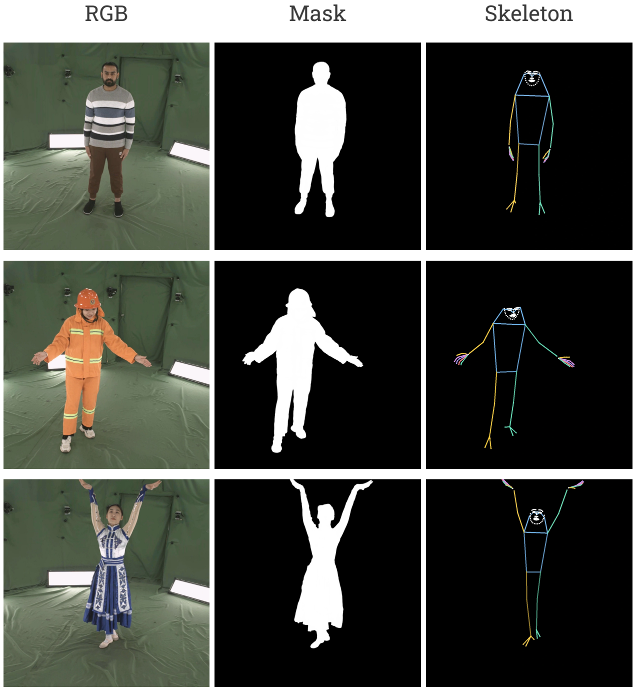

Figure 3. High-quality foreground masks and human skeletons predicted using state-of-the-art methods.

C. Additional Comparisons

More qualitative results. We provide additional comparisons with baselines in Fig. 2. Results show that our method consistently outperforms the baselines in terms of visual quality and fine details.

Diffusion generation vs. 4DGS rendering. Although our model already supports novel-view synthesis, we choose to optimize a 4DGS model using LongVolcap [17] to enable real-time rendering. As shown in Fig. 4, our model can generate high-fidelity human videos, but they still inevitably exhibit spatio-temporal inconsistencies. Recon-

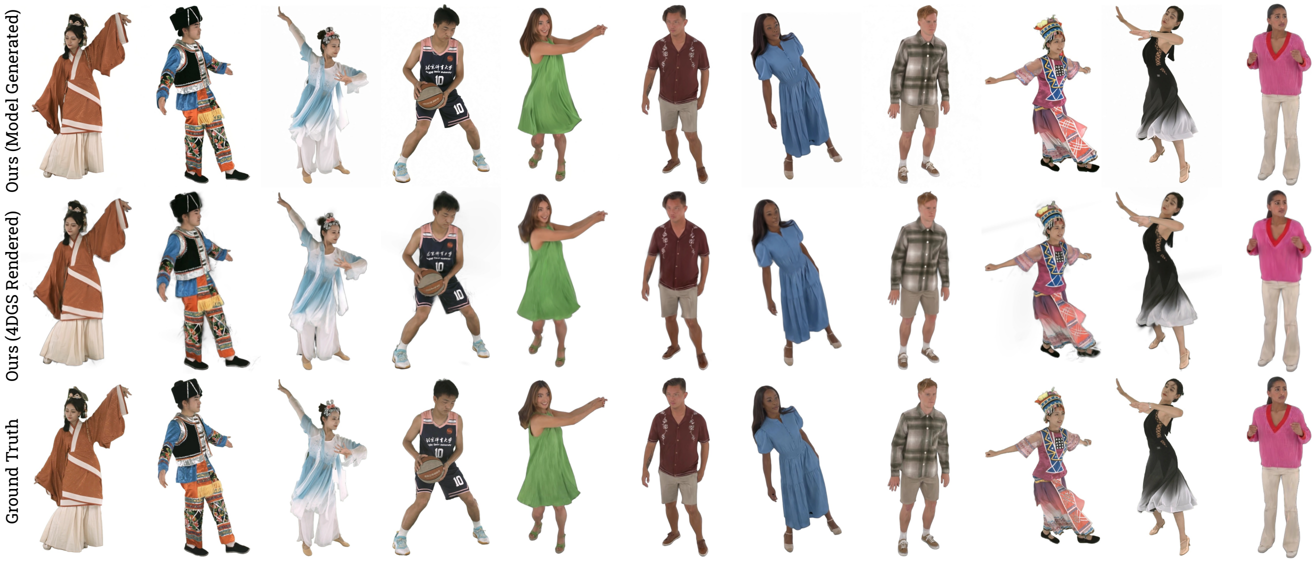

Figure 4. Qualitative comparisons between novel views generated by our model and those rendered from the 4DGS model reconstructed using LongVolcap [17].

structing a 4DGS model further alleviates these inconsistencies, though at the cost of reduced sharpness compared to the originally generated images.

References

[1] Wei Cheng, Ruixiang Chen, Wanqi Yin, Siming Fan, Keyu Chen, Honglin He, Huiwen Luo, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, Daxuan Ren, Lei Yang, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, Bo Dai, and Kwan-Yee Lin. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. *arXiv preprint*, arXiv:2307.10173, 2023. 1, 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. *arXiv preprint arXiv:2212.08051*, 2022. 2

[3] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 2

[4] 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*, 2024. 1

[5] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 2

[6] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1

[7] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569*, 2024. 1

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

[9] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *International journal of computer vision*, 38:199–218, 2000. 1

[10] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8762–8771, 2021. 1

[11] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 1

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1

[13] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[14] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 1

[15] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022. 1

[16] Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 4k4d: Real-time 4d view synthesis at 4k resolution. In *CVPR*, 2024. 1

[17] Zhen Xu, Yinghao Xu, Zhiyuan Yu, Sida Peng, Jiaming Sun, Hujun Bao, and Xiaowei Zhou. Representing long volumetric video with temporal gaussian hierarchy. *ACM Transactions on Graphics*, 43(6), 2024. 1, 2, 3

[18] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Realtime photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In *International Conference on Learning Representations (ICLR)*, 2024. 1

[19] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, 3:9150038, 2024. 1