Supplementary Materials of HiReFF: High-Resolution Feedforward Human Reconstruction from Uncalibrated Sparse-View Video

Yiming Jiang[1⋆], Hanzhang Tu[2], Wenfeng Song[3], Siyou Lin[2], Liang An[2], Shuai Li[1,4], Aimin Hao[1(✉)], and Yebin Liu[2]

[1] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
{jiangyimingjym, lishuai, ham}@buaa.edu.cn
[2] Tsinghua University, Beijing, China
{thz22, linsy21}@mails.tsinghua.edu.cn
{anliang, liuyebin}@mail.tsinghua.edu.cn
[3] College of Computer Science, Beijing Information Science and Technology University, Beijing, China
songwenfenga@gmail.com
[4] Zhongguancun Laboratory, Beijing, China

1 Experiment

1.1 Experimental Settings

Datasets. We conduct training mainly on the DNA-Rendering [2] dataset. To improve generalization, we also utilize the ZJU-MoCap [3, 8] and MVHumanNet [5, 9] datasets. Specifically, we use 7 videos from the ZJU-MoCap dataset and 439 videos from the MVHumanNet dataset. Each video in MVHumanNet contains 16 views with an image resolution of 1024×1024, while each video in ZJU-MoCap contains 23 views with the same 1024×1024 resolution. During training, we set the data sampling ratio among DNA-Rendering, MVHumanNet, and ZJU-MoCap to 0.45:0.45:0.1. For MVHumanNet and ZJU-MoCap, we pre-interpolate their images to 2072×2072 as network inputs during training.

Implementation Details. Since our task involves foreground human segmentation and the comparison methods lack built-in foreground masks, we employ an open-source video portrait segmentation algorithm [7] to generate masks. GPS-Gaussian receives segmented foreground images as input, whereas other methods (including ours) take full images with background intact. To ensure fair comparison, we multiply all methods' predicted results by a consistent mask before computing evaluation metrics.

We select four views as input and employ eight views for testing. Our method and AnySplat [4] use identical input views. For NoPoSplat [11], we provide two input views at a time and render from test viewpoints positioned between them.

⋆ Y.Jiang—Work done during an internship at Tsinghua University.

For 4DGT [10], to maintain a streaming-reconstruction setting, we feed one view with one frame at a time and duplicate it into a 16-frame clip, and then render images from near test viewpoints. Since GPS-Gaussian [13] cannot handle the 90° four-view reconstruction task, we supplement it with four additional views, resulting in eight input views at 45° intervals. Critically, we ensure that all methods are evaluated on the same eight test viewpoints with consistent camera parameters, none of which are used as input views.

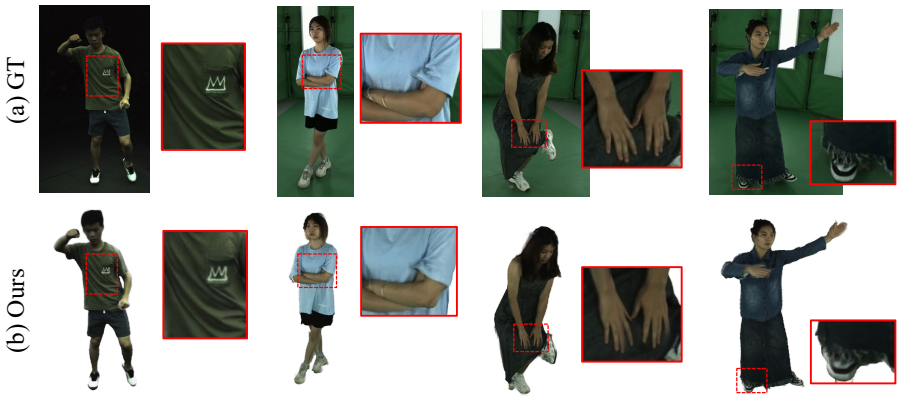

Fig. 1: Performance on more datasets (§1.2). Our method can faithfully recover fine details of the images on ZJU-MoCap and MVHumanNet datasets. Please 🔍 **zoom in** to see details.

Fig. 2: Performance on HuMMan datasets (§1.2). Our method performs well on the out-of-domain datasets. L: one of the inputs. R: novel view. Please 🔍 **zoom in** to see details.

1.2 Performance on More Datasets

As shown in Fig. 1, the leftmost column corresponds to the ZJU-MoCap dataset, and the right three columns correspond to the MVHumanNet dataset. Fig. 2

shows results on the HuMMan [1] dataset. None of these data were used during training. Our method can faithfully recover fine details of the images on those datasets.

1.3 Number of Gaussians

Our method takes four 2K images as input, predicts the foreground mask of the human subject, and outputs 3D Gaussians. The number of Gaussians generated by our method typically ranges from 350,000 to 500,000. This is because the human foreground region occupies only a small portion of the full image. Specifically, the output of the feature dimension of the AA Transformer is $B \times V \times H \times W \times C$, where $H = W = 518$. In the High-resolution Side-tuning module, we fuse the 2K image features with the attention-fused features from the AA Transformer, and upsample them to $B \times V \times H' \times W' \times C'$, where $H' = W' = 1036$. The human region accounts for approximately 10% of the total image area, meaning the valid mask ratio μ is 10%. Accordingly, the number of 3D Gaussians is roughly $V \times H' \times W' \times \mu$. With $V = 4$, this results in about 430,000 Gaussians. With 400,000 Gaussians, our approach achieves over 120 FPS at 2K resolution rendering [12] on an RTX 4090 GPU, enabling real-time rendering.

Table 1: Quantitative results of novel-view synthesis of reconstruction. (§1.4). Our method outperforms Depth Anything 3 across all metrics.

Method	Cam. Pose	Views Num.	Aligned	PSNR↑	SSIM↑	LPIPS↓
DepthAnything3	×	4	×	21.4966	0.8805	0.2006
Ours	×	4	×	**26.5138**	**0.9164**	**0.1277**

1.4 Comparison Results with Depth Anything 3.

We compared our method with Depth Anything 3 [6]. Depth Anything 3 is a foundational model for uncalibrated camera depth estimation and geometry reconstruction using sparse viewpoints, while also supporting Gaussian reconstruction from sparse viewpoints for uncalibrated cameras. Specifically, we select the best-performing DA3NESTED-GIANT-LARGE model, which has 1.40 billion parameters, significantly exceeding the 0.91 billion parameters of VGGT. Although DA3 performs well in natural scenes with high view overlap, our method excels at novel human view rendering under ring-shaped camera settings with 90-degree intervals.

Our method adopts the same input settings as Depth Anything 3. Quantitative results are presented in Table 1, where our method outperforms all competing approaches across all evaluation metrics. Novel-view visualization results

Fig. 3: Qualitative results of novel-view synthesis of reconstruction (§1.4). Our method surpasses the Depth Anything 3 in global shape, garment details, and facial fidelity. Please **zoom in** to see details.

are shown in Fig. 3. Since Depth Anything 3 does not have a Gaussian mask head, we retain the background in its output. It can be observed that Depth Anything 3 only supports 504×504 image output with low clarity, whereas our method achieves superior performance in terms of facial detail sharpness and overall human body integrity.

Table 2: Camera Pose Estimation on our test set (§1.5).

Method	AUC@5↑	AUC@10↑	AUC@20↑	AUC@30↑
AnySplat [4]	22.3	60.5	80.0	86.6
DepthAnything3 [6]	88.6	93.9	96.7	97.8
Ours	76.3	88.2	94.1	96.1

Table 3: Camera Pose Estimation(§1.5).

Method	AUC@5↑	AUC@10↑	AUC@20↑	AUC@30↑
VGGT w/o bg	20.1	43.0	67.9	78.3
VGGT w/ bg	80.0	89.9	94.9	96.6

1.5 Camera Pose Estimation.

We compare the camera pose estimation results of AnySplat [4] and Depth Anything 3 [6] on the test set. As shown in Table 3, our method achieves significantly higher camera pose estimation accuracy than AnySplat, but is slightly inferior to Depth Anything 3. This is because we compare against the DA3NESTED-GIANT-LARGE variant of Depth Anything 3, which has 1.4 billion parameters, whereas our method uses the VGGT backbone with only 0.91 billion parameters. Although Depth Anything 3 surpasses ours in camera pose estimation accuracy, our method exhibits clear advantages in sparse-view novel view rendering.

We also conduct experiment on VGGT of camera estimation on foreground humans to explain why we use whole image with background as input. Tab. 3 shows the camera estimation of VGGT with 4 input views. Given the large 90° view interval, camera poses predicted from the foreground alone lack sufficient accuracy and cannot serve as valid model initialization.

1.6 Temporal Stability.

We evaluate temporal stability via MediaPipe-extracted COCO-17 keypoints, and compute RMSE between original trajectories and 5-frame centered moving-average smoothed ones. GT's RMSE captures inherent human motion jitter, while predictions reveal natural fluctuation and temporal drift. Our RMSE is 0.0093, close to GT (0.0108) and outperforms AnySplat (0.0137). Although our method lacks explicit temporal modules, its temporal stability implicitly stems from camera optimization. We will add discussion and refrain from overclaiming temporal modeling.

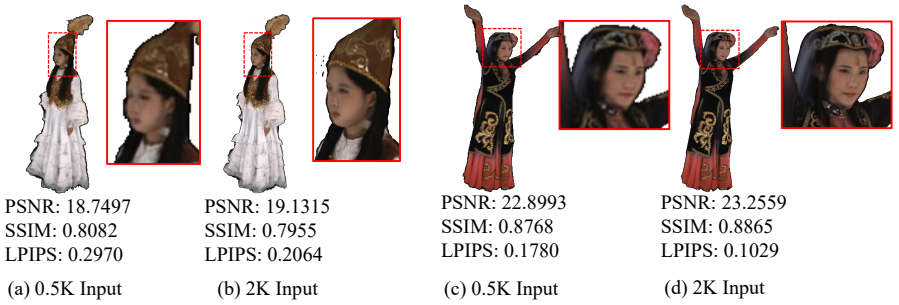

Fig. 4: Ablation on High-resolution Side-tuning (§1.7). By supporting 2K-resolution input, High-resolution Side-tuning effectively enhances rendering quality. Please **zoom in** to see details.

1.7 Ablation Study.

Ablation on High-resolution Side-tuning. We trained our model using 0.5K and 2K resolution images as input, respectively. As shown in Fig. 4, using High-

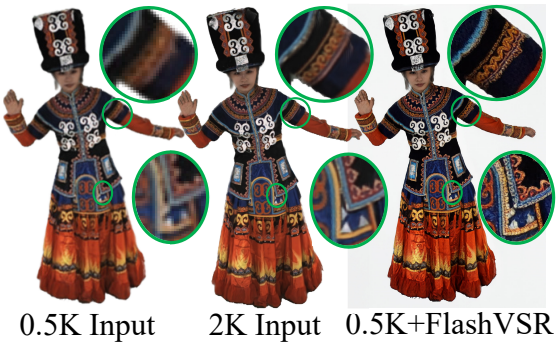

Fig. 5: Comparison with 2D super resolution method FlashVSR [14]. (§1.7). Ours avoids hallucination, faithfully preserving original texture topology.

resolution Side-tuning with 2K images as input yields significant quality improvements over using 0.5K images, demonstrating the effectiveness of the High-resolution Side-tuning module. The 2K inputs produce richer visual details and achieve higher quantitative metrics.

Our High-resolution Side-tuning extracts real 2K image features to complement the geometric features from AA Trans. As shown in Fig. 5, the sharpness and details of 0.5K+FlashVSR [14] stem from hallucinations, which distort intrinsic texture structures and cause unnatural artifacts. In comparison, ours avoids hallucination, faithfully preserving original texture topology.

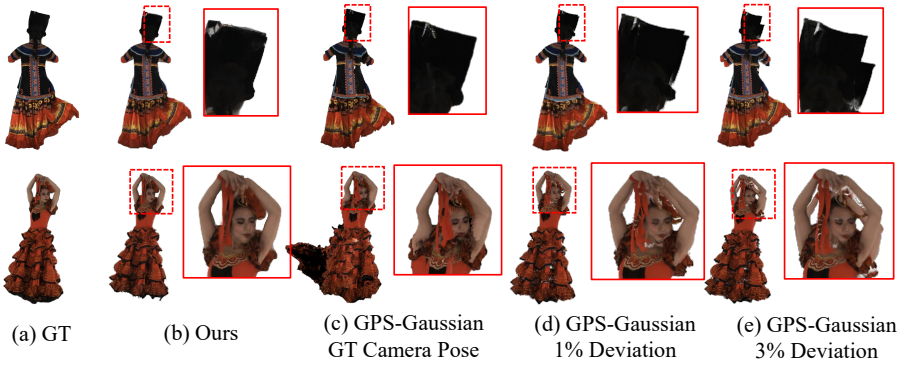

Fig. 6: Advantages of Uncalibrated Methods over Calibrated Methods. (§2). Methods requiring accurate camera pose estimation are highly sensitive to calibration noise: a 1% deviation from ground-truth already causes noticeable reconstruction errors, while 3% deviation results in severe degradation and complete failure. Please **zoom in** to see details.

2 Discussion

Advantages of Uncalibrated Methods over Calibrated Methods. For sparse-view human reconstruction tasks, uncalibrated methods not only save the time and labor costs of camera calibration but also avoid reliance on accurate camera parameters, which are difficult to obtain in real-world applications. We conducted experiments using GPS-Gaussian [13] as a representative calibrated method. For camera extrinsics, we injected random deviation of 1% and 3% into the ground-truth parameters, while keeping the intrinsic parameters unchanged. As shown in Fig. 6, only 1% deviation in camera parameters leads to obvious artifacts and degradation in reconstruction quality, and 3% deviation further causes severe ghosting and duplicated human structures. In contrast, our method requires no camera parameters as input and is thus completely robust to calibration errors. For calibrated methods, the calibration error must be controlled within 1% to maintain satisfactory reconstruction quality in practical scenarios.

References

1. Cai, Z., Ren, D., Zeng, A., Lin, Z., Yu, T., Wang, W., Fan, X., Gao, Y., Yu, Y., Pan, L., Hong, F., Zhang, M., Loy, C.C., Yang, L., Liu, Z.: HuMMan: Multi-modal 4d human dataset for versatile sensing and modeling. In: 17th European Conference on Computer Vision, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII. pp. 557–577. Springer (2022)
2. Cheng, W., Chen, R., Fan, S., Yin, W., Chen, K., Cai, Z., Wang, J., Gao, Y., Yu, Z., Lin, Z., Ren, D., Yang, L., Liu, Z., Loy, C.C., Qian, C., Wu, W., Lin, D., Dai, B., Lin, K.: Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In: IEEE/CVF International Conference on Computer Vision, ICCV. pp. 19925–19936 (2023)
3. Fang, Q., Shuai, Q., Dong, J., Bao, H., Zhou, X.: Reconstructing 3d human pose by watching humans in the mirror. In: CVPR (2021)
4. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., Lin, D., Dai, B.: Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. CoRR **abs/2505.23716** (2025)
5. Li, C., Liao, H., Zhi, Y., Yang, X., Sun, Z., Chang, J., Cui, S., Han, X.: Mvhumannet++: A large-scale dataset of multi-view daily dressing human captures with richer annotations for 3d human digitization. arXiv preprint arXiv:2505.01838 (2025)
6. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
7. Lin, S., Yang, L., Saleemi, I., Sengupta, S.: Robust high-resolution video matting with temporal guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 238–247 (2022)
8. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: CVPR (2021)

9. Xiong, Z., Li, C., Liu, K., Liao, H., Hu, J., Zhu, J., Ning, S., Qiu, L., Wang, C., Wang, S., 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. pp. 19801–19811 (2024)
10. Xu, Z., Li, Z., Dong, Z., Zhou, X., Newcombe, R.A., Lv, Z.: 4dgt: Learning a 4d gaussian transformer using real-world monocular videos. CoRR **abs/2506.08015** (2025)
11. Ye, B., Liu, S., Xu, H., Li, X., Pollefeys, M., Yang, M., Peng, S.: No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In: The Thirteenth International Conference on Learning Representations, ICLR (2025)
12. Ye, V., Li, R., Kerr, J., Turkulainen, M., Yi, B., Pan, Z., Seiskari, O., Ye, J., Hu, J., Tancik, M., Kanazawa, A.: gsplat: An open-source library for gaussian splatting. Journal of Machine Learning Research **26**(34), 1–17 (2025)
13. Zheng, S., Zhou, B., Shao, R., Liu, B., Zhang, S., Nie, L., Liu, Y.: Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19680–19690 (2024)
14. Zhuang, J., Guo, S., Cai, X., Li, X., Liu, Y., Yuan, C., Xue, T.: Flashvsr: Towards real-time diffusion-based streaming video super-resolution. arXiv preprint arXiv:2510.12747 (2025)