

# MEAT: Multiview Diffusion Model for Human Generation on Megapixels with Mesh Attention

## Supplementary Material

Yuhan Wang<sup>1</sup> Fangzhou Hong<sup>1</sup> Shuai Yang<sup>2</sup> Liming Jiang<sup>1</sup> Wayne Wu<sup>3</sup> Chen Change Loy<sup>1</sup>  
<sup>1</sup>S-Lab, Nanyang Technological University <sup>2</sup>WICT, Peking University <sup>3</sup>UCLA

### Abstract

*In the supplementary material, we discuss further details and provide more results that are not included in the main paper. In Appendix A, we provide more details of our model setting and structure. In Appendix B, we discuss further details and provide visualization for our dataset processing pipeline. In Appendix C, we present more results on qualitative comparison with monocular reconstruction methods and illustration of our cross-view consistency preservation ability.*

### A. Implementation Details

In this section, we specify the details regarding the model implementation and the experiment settings.

#### A.1. Model Implementation

**Keypoints Conditioning.** We use a small 3-layer convolutional network to process the keypoints condition, down-sampling the keypoints visualization image by 8x and aligning it with the channel of the denoiser U-Net after the `conv_in` block. Each downsampling is achieved with two convolutional layers. The final output is processed with a `conv_out` convolutional layer, which is zero-initialized to allow this condition to be smoothly integrated into the U-Net. We found that an additional branch like ControlNet [17] is unnecessary. Directly adding the processed condition to the U-Net features yields satisfactory training results.

**VAE Feature Encoder.** The VAE feature encoder is very similar to the diffusion U-Net down-sampling blocks without *Attention* layers. At each resolution scale, there are 2 layers of `ResnetDownsampleBlock2D`, whose number of channels is matched with that in the U-Net. We use the last features before down-sampling in each residual block to be fused into the U-Net through VAE attention.

**Implementation Details.** Our model is initialized with Stable Zero123 [2] pretrained weights, and optimized using

$\epsilon$ -prediction. Notably, since the SDXL-VAE [13] can produce NaN under `fp16` precision, we utilize the `fp16-fix` version [1] to support mixed-precision training. Our model supports sparse-view training. We randomly sample seven views, including the reference, in each training batch. The batch size on each GPU is 1, and we use 8 NVIDIA-A100-80GB GPUs to train 150,000 iterations without gradient accumulation, which takes about 7 days. Our model can generate 16 views simultaneously during inference. It employs a *Trailing* sample steps selection method to minimize the signal-to-noise ratio (SNR) at the beginning of the denoising process. We use DDIM sampler with 50 steps and a CFG scale of 3.0.

#### A.2. Experiment Setting

**Baselines.** For quantitative experiments, we compare our method with Stable Zero123 [2], SyncDremer [11], Wonder3D [12], and SV3D [15]. For Wonder3D with pretrained weights, as it generates six views at a time, we split the 15 non-reference test views into three batches, each combined with the reference view for the generation. We re-train Stable Zero123 and Wonder3D on DNA-Rendering at the resolution of  $256 \times 256$ . Wonder3D is only trained in the color domain since ground-truth normal maps are not available. We only compare the results of MagicMan [7] qualitatively as its preset views cannot align with the test setting.

**Metrics.** Since most of the previous multi-view diffusion models only generate at a resolution of 256, we also resize our results to calculate metrics at this resolution for fair comparison. Moreover, to show the advantage of high-resolution generation, we also compute metrics at a resolution of 1024. For both resolutions, we include PSNR, SSIM [16], and LPIPS [18] metrics to compare the generated results with the ground-truth images. For the 1024 category, we use Patch-FID (P-FID) [3, 6, 10] instead of FID [8] as a metric for generation quality. FID resizes images to  $299 \times 299$  before calculation, which does not reflect MEAT's advantage at high resolutions. Instead, we split

each image into a  $4 \times 4$  grid of  $256 \times 256$  patches and select the middle two columns, yielding eight patches per image. The calculation is based on the patch set. In the 256 categories, we also use the PPLC metric proposed by Free3D [20] to evaluate cross-view consistency in multiview generation. We exclude it in the 1024 category because upsized blurry results gain an unfair advantage in this metric.

## B. DNA-Rendering for Multiview Generation

In this section, we present the full details of the novel ideas proposed to harness multiview human video dataset DNA-Rendering [4] for multiview diffusion training. We construct our training data using the multiview human dataset DNA-Rendering [4], which provides 15 FPS multiview videos of human motion. By sampling one set of frames every five frames, we generate over 20,000 sets of multiview images. The first partition, containing 2,000 samples, is reserved for testing, while the second partition is used for training. While this larger dataset offers a significant advantage, the multiview setting brings additional challenges. We address three primary issues: (1) selecting the front view for monocular reconstruction, (2) adapting the monocular reconstructed mesh to the calibrated coordinate system, and (3) cropping the images with corresponding adjustments to the camera calibration parameters.

### B.1. Frontal Camera Selection

For each frame of multiview images in the DNA-Rendering [4] dataset, we need to first determine which view is the ‘‘frontal’’ one. This config is utilized in monocular reconstruction, training views sampling, and inference. Since the dataset provides the SMPL-X coefficients and camera calibration parameters  $R_v$  and  $T_v$  for each view, we can derive the global orientation  $\mathbf{d}$  of the human body, the 3D coordinates  $\mathbf{G}$  of the pelvis, and the camera coordinates  $\mathbf{C}_v$ , where

$$\mathbf{C}_v = -R_v^{-1}T_v.$$

We define the frontal view as the viewpoint where the angle between the line connecting the camera’s optical center to the pelvis and the global orientation is minimized, *i.e.*

$$\text{front view} \leftarrow \arg \max_v \frac{\mathbf{d} \cdot \mathbf{GC}_v}{\|\mathbf{d}\| \|\mathbf{GC}_v\|} \quad (1)$$

### B.2. Mesh Adaptation

To ensure consistent mesh quality during both training and inference and to prevent the model from overly relying on the accuracy of the centric geometric representation, we use monocular reconstruction from the selected frontal image above to extract the centric mesh for training. We use PIFuHD[14] for its balance of speed and quality. However, monocular reconstruction typically assumes a specific position and orthographic projection for the frontal camera,

which differs from our dataset where the frontal camera is perspective and can be positioned variably. Consequently, we need to determine a transformation TF to align the mesh with the world coordinate system of the dataset.

Our adaptation approach is based on the following rule:  $P_p$  of each pixel  $p$  in the reference view, after transformation TF and reprojection, should return to its original position in its own view and reach the feature-matching point in adjacent views. These two relationships establish an optimization objective for TF with a unique optimal solution. We use RoMa [5] to detect all feature-matching pairs and apply gradient descent to solve TF.

Specifically, we assume that the transformation TF for each vertex  $P$  consists of a scaling  $S$ , rotation  $R$ , and translation  $t$ :

$$S = \text{diag}(\mathbf{s}), \mathbf{s} = [s_x, s_y, s_z], \quad (2)$$

$$R = \text{rot6d}(\mathbf{c}_1, \mathbf{c}_2), \quad (3)$$

$$p' = \text{TF}(P) = R(SP) + t. \quad (4)$$

We use `rot6d` rotation representation [21] for more stable optimization. We can then define the re-projection process  $\tilde{\Pi}_v$  of a frontal-view pixel  $p$  into the view  $v$ .

$$\tilde{\Pi}_v(p) = \Pi_v(\text{TF}(p \rightarrow P)). \quad (5)$$

Here  $p \rightarrow P$  indicates the inverse orthographic rasterization process and  $\Pi_v$  is the projection to view  $v$  as is described in Eq.(6) in the main paper. Let  $v = 1$  be the frontal view. We use two types of alignment to build the optimization target:

1.  $\tilde{\Pi}_1(p)$  - Pixels return to their original positions.
2.  $\tilde{\Pi}_v(p)$  - Pixel  $p$  on the frontal view is matched with pixel  $q_v$  on view  $v$ .

We use RoMa [5] to detect such  $(p, q_v)$  pairs. All the pixels  $p$  that do not intersect with the mesh are filtered out. The pixel values are normalized to  $[0, 1]$  based on the resolution of the raw image. Finally, we can solve the transformation TF through:

$$\arg \min_{\mathbf{s}, \mathbf{c}_1, \mathbf{c}_2, t} \sum_p \|\mathbf{p} - \tilde{\Pi}_1(p)\|_2^2 + \sum_{p, q_v} \|\mathbf{q}_v - \tilde{\Pi}_v(p)\|_2^2. \quad (6)$$

We initialize these parameters with the assumption of zero translation, identical scaling, and an aligned coordinate system. It yields  $\mathbf{s}_0 = [1, 1, 1]$ ,  $\mathbf{t}_0 = \mathbf{0}$ , and

$$R_0 = \left( \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \cdot R_{v=1} \right)^{-1} \quad (7)$$

Here  $R_{v=1}$  is the calibrated extrinsic rotation matrix of the frontal camera in the DNA-Rendering [4] dataset. DNA-Rendering adopts the `opencv` camera coordinate system convention, which has an opposite direction of  $y$ -axis and  $z$ -axis. We show visualization results in Fig. 1.

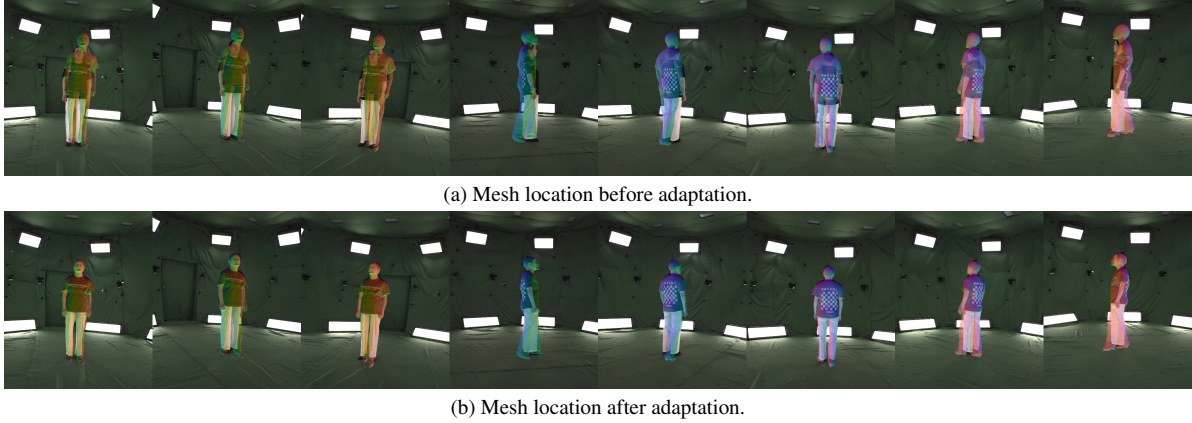


Figure 1. **Mesh Adaptation.** Although the monocular reconstructed human mesh inevitably exhibits certain deviations from the ground truth, our mesh adaptation method can robustly align it to the dataset’s coordinate system. Our MEAT model, trained using this data, effectively mitigates the interference of geometric noise in human meshes during multi-view image generation.

### B.3. Image Cropping

Existing multiview diffusion models place the object at the origin of the world coordinate system when rendering datasets, and position the camera on a fixed-radius sphere centered at this origin. This approach simplifies the view-point representation to just azimuth and elevation, reducing training complexity.

During training, we use the 1-meter-high circular camera array of DNA-Rendering to simulate the zero-elevation rendered data. These cameras are all oriented toward the calibrated center of the world coordinate system. However, this center often does not align precisely with the person’s position, resulting in variable positioning within the images. This variability introduces ambiguity when using the camera representation of existing multiview diffusion models.

To address this issue, we propose cropping the images based on the pelvis position. We align the pelvis joint from SMPL-X in each frame to the center of the pixel grid. To maintain consistency with the spherical camera arrangement, we assume the subject has the same height in each pixel plane since all cameras have the same height. We set the cropping radius to  $1.3 \times$  the maximum height difference between any keypoint and the pelvis in each pixel plane:

$$R_v = 1.3 \cdot \max_{\mathbf{P}} |\Pi_v(\mathbf{P})_y - \Pi_v(\mathbf{P}_{\text{pelvis}})_y|. \quad (8)$$

The cropped images from each view are then resized to the same resolution. Since only cropping and resizing are involved, we only need to adjust the principal point coordinates in the camera intrinsics and normalize the camera to the NDC (Normalized Device Coordinate) system.

## C. More Results

### C.1. Cross-view Consistency Preservation

We show the generated results of models with and without mesh attention modules in Fig. 2. In the multiview diffusion model, the generation of front-facing regions leverages information from reference viewpoints, resulting in reduced randomness. Conversely, the generation of the backside relies more heavily on the model’s generative capabilities, thereby exhibiting greater randomness inherent to diffusion models. As is shown in Fig. 2, one-view-at-a-time models lacking mesh attention frequently make random selections among different modes in local structures, resulting in inconsistencies across viewpoints. The mesh attention module effectively mitigates this issue, achieving better cross-view consistency preservation.

### C.2. Monocular Reconstruction Methods

In this section, we compare the novel view generation results of our MEAT diffusion model with monocular reconstruction methods like SiTH [9] and SIFU [19]. The qualitative comparison results are shown in Fig. 3. For monocular reconstruction methods, novel view images are rendered from textured human meshes, thereby inherently ensuring perfect cross-view consistency.

However, due to the challenges associated with accurate geometric estimation, monocular reconstructed human meshes often exhibit reduced realism when dealing with relatively loose clothing, thus the results after texture mapping are unsatisfactory. Our MEAT model utilizes such coarse human meshes solely as a medium for cross-view feature fusion; the generated images themselves are not rendered from any explicit geometric representations, resulting in a noticeable enhancement in realism.



Figure 2. **Cross-view Consistency Preservation.** Models without mesh attention adhere to a one-view-at-a-time approach. Due to the stochastic nature of diffusion models, generating the backside often fails to maintain local structural consistency across different viewpoints. The mesh attention module significantly enhances the cross-view consistency preservation.

## References

- [1] SDXL-VAE-FP16-Fix. <https://huggingface.co/madebyollin/sdxl-vae-fp16-fix>, 2023. 1
- [2] Stable Zero123. <https://stability.ai/news/stable-zero123-3d-generation>, 2024. 1
- [3] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis. In *ECCV*, 2022. 1
- [4] 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 *ICCV*, 2023. 2
- [5] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust dense feature matching. In *CVPR*, 2024. 2
- [6] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Wayne Wu, and Ziwei Liu. UnitedHuman: Harnessing multi-source data for high-resolution human generation. In *ICCV*, 2023. 1
- [7] Xu He, Xiaoyu Li, Di Kang, Jiangnan Ye, Chaopeng Zhang, Liyang Chen, Xiangjun Gao, Han Zhang, Zhiyong Wu, and Haolin Zhuang. MagicMan: Generative novel view synthesis of humans with 3D-aware diffusion and iterative refinement. *arXiv preprint*, arXiv:2408.14211, 2024. 1
- [8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, 2017. 1
- [9] I Ho, Jie Song, Otmar Hilliges, et al. SiTH: Single-view textured human reconstruction with image-conditioned diffusion. In *CVPR*, 2024. 3
- [10] Shikai Li, Jianglin Fu, Kaiyuan Liu, Wentao Wang, Kwan-Yee Lin, and Wayne Wu. CosmicMan: A text-to-image foundation model for humans. In *CVPR*, 2024. 1
- [11] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. SyncDreamer: Generating multiview-consistent images from a single-view image. In *ICLR*, 2024. 1
- [12] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3D: Single image to 3D using cross-domain diffusion. In *CVPR*, 2024. 1
- [13] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 1

- [14] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *CVPR*, 2020. [2](#)
- [15] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. SV3D: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion. In *ECCV*, 2024. [1](#)
- [16] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13:600–612, 2004. [1](#)
- [17] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. [1](#)
- [18] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. [1](#)
- [19] Zechuan Zhang, Zongxin Yang, and Yi Yang. SIFU: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *CVPR*, 2024. [3](#)
- [20] Chuanxia Zheng and Andrea Vedaldi. Free3D: Consistent novel view synthesis without 3D representation. In *CVPR*, 2024. [2](#)
- [21] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. [2](#)



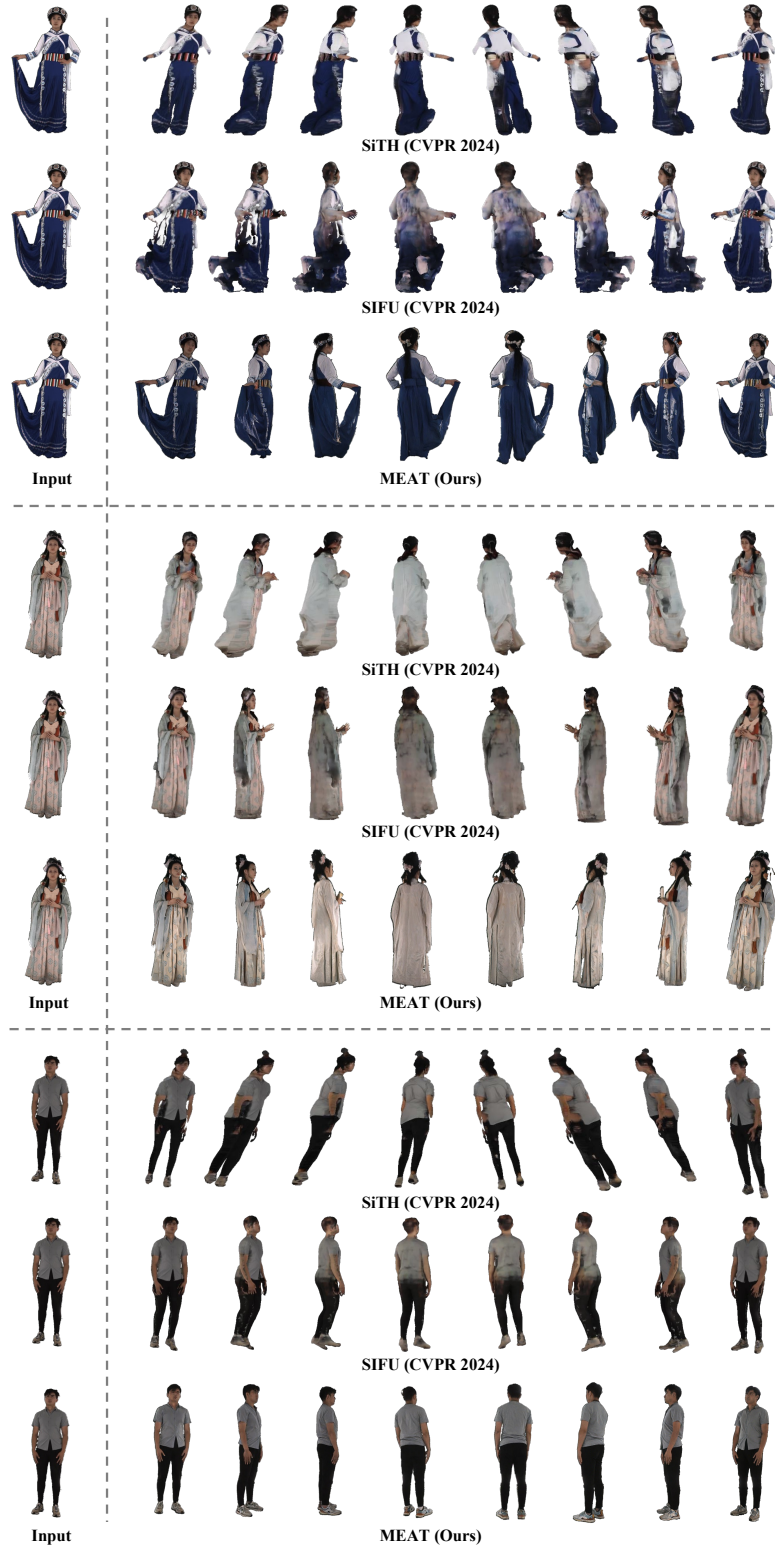


Figure 3. **Comparison with Monocular Reconstruction Methods.** In the novel view generation results for human bodies, compared to monocular reconstructed meshes, the multiview images generated by our MEAT diffusion model exhibit significant advantages in geometric plausibility, geometric details, texture details, and clarity. Please **zoom in** for details.