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

Yuhan Wang[1] Fangzhou Hong[1] Shuai Yang[2] Liming Jiang[1] Wayne Wu[3] Chen Change Loy[1]
[1]S-Lab, Nanyang Technological University [2]Peking University [3]UCLA

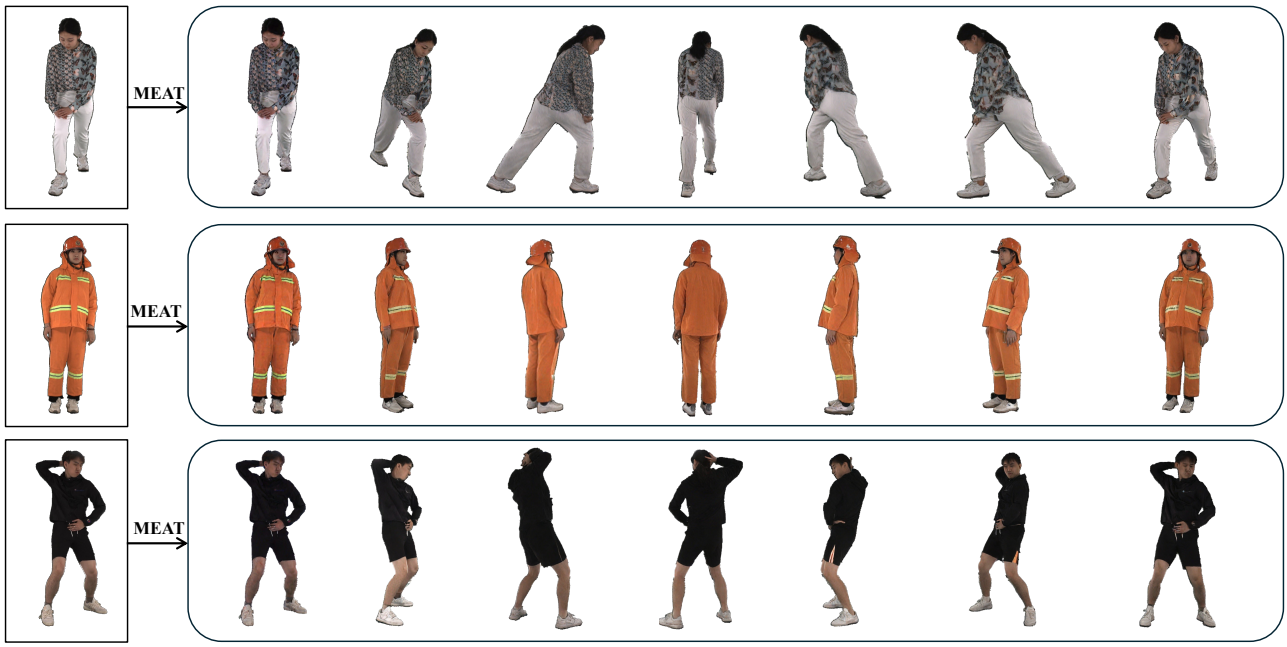


Figure 1. Given a frontal human image, **MEAT** can generate dense, view-consistent multiview images at a resolution of 1024^2.

Abstract

*Multiview diffusion models have shown considerable success in image-to-3D generation for general objects. However, when applied to human data, existing methods have yet to deliver promising results, largely due to the challenges of scaling multiview attention to higher resolutions. In this paper, we explore human multiview diffusion models at the megapixel level and introduce a solution called **mesh attention** to enable training at 1024^2 resolution. Using a clothed human mesh as a central coarse geometric representation, the proposed mesh attention leverages rasterization and projection to establish direct cross-view coordinate correspondences. This approach significantly reduces the complexity of multiview attention while maintaining cross-view consistency. Building on this foundation, we devise a mesh attention block and combine it with keypoint conditioning to create our human-specific multiview diffusion model, **MEAT**. In addition, we present valuable insights into applying multiview human motion videos for diffusion training, addressing the longstanding issue of data scarcity. Extensive experiments show that MEAT effectively generates dense, consistent multiview human images at the megapixel level, outperforming existing multiview diffusion methods. Code and model will be publicly available.*

1. Introduction

In this paper, we address the problem of multiview human generation, which aims to generate realistic, consistent multi-angle renderings of a human figure. We assume a single frontal image is provided. Recent advancements in diffusion models offer a promising new approach to this task, as they excel at generating high-quality images conditioned on various inputs. However, achieving realistic human renderings remains highly challenging due to the importance of resolution for capturing fine details. Specifically, existing multiview diffusion models [14, 17, 19, 26] for general objects are typically trained at a resolution of 256^2, with a few recent methods increasing this to 512^2 [15] or 578^2 [30]. However, this remains insufficient for human data. As shown in Fig. 2, under the latent diffusion setting, a reso-

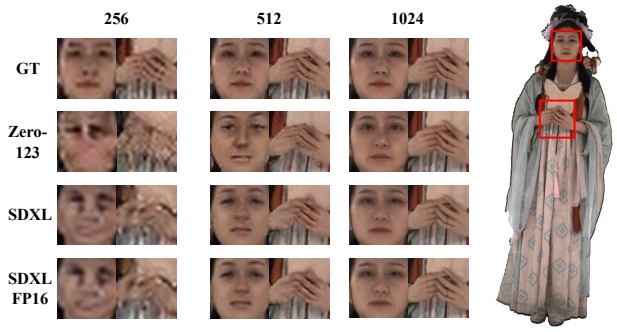


Figure 2. **VAE and Resolution.** Each row represents the same version of VAE, while each column corresponds to the same resolution of the full-body image after VAE reconstruction. Although the full-body image rendered at 512×512 shows good visual quality, it falls short when used in diffusion models with VAE. We find that a resolution of 1024×1024 is necessary for optimal results.

Table 1. **Multiview Attention Comparison.** (1) Dense multiview attention requires each pixel to integrate all other pixels in different views, consuming $N\times$ more memory than self-attention. (2) Row-wise attention is based on the orthographic assumption, making it unsuitable for videos shot with an arbitrary perspective. (3) Epipolar attention is related to our approach. It requires sampling 3D point candidates for each pixel, with the density K balancing multiview accuracy and complexity. (4) Our mesh attention eliminates this sampling with a centric mesh. We assume the feature map dimensions are $H = W = S$, with each view interacting with all N views. d represents the grid sampling constant.

Attn. Type	Q	K,V	Attn. Map	Persp.
Self-Attn	NCS^2	NCS^2	NS^4	-
Dense MV	NCS^2	$NC(NS^2)$	N^2S^4	✓
Row-wise	$(NH)CW$	$(NH)C(NW)$	N^2S^3	×
Epipolar	$(NS^2)C \cdot 1$	$(NS^2)C(NKd)$	N^2S^2Kd	✓
Mesh Attn	$(NS^2)C \cdot 1$	$(NS^2)C(Nd)$	$\boldsymbol{N^2S^2d}$	✓

lution of 1024^2 is necessary to achieve satisfactory results, as the result is highly sensitive to details in areas such as the face, hands, and clothing. Any lack of detail, unnatural appearance, or inconsistency in these regions significantly diminishes the realism. Since these areas each occupy only a small portion of the overall pixel space, variational autoencoder (VAE) reconstructions at resolutions below 1024×1024 are suboptimal, making it challenging to train an effective multiview diffusion model.

Directly increasing the working resolution of existing multiview diffusion models to 1024×1024 is impractical either. To maintain multiview consistency, current methods generate all views simultaneously and add cross-view attention within the denoising U-Net to integrate features from different views. Table 1 summarizes the attention map complexity of existing multiview attention methods. Dense multiview attention [19, 27, 31] has extremely high memory requirements, making it difficult to apply directly at megapixel resolutions. Meanwhile, row-wise attention [15] relies on an orthographic projection assumption, which significantly restricts the applicable training data.

To address these challenges, we propose MEAT, a multiview diffusion model designed for human novel view generation on megapixels, conditioned on a frontal image. In particular, we wish to address the high computational complexity of multiview attention in existing diffusion models. Our key idea is to leverage a rough central 3D representation that enables our method to directly establish correspondences between pixels across different viewpoints using rasterization and projection. We refer to this pixel correspondence-based feature fusion as **mesh attention**. This optimization allows us to sample sufficiently dense viewpoints on each GPU and train the model using 1024×1024 images. As shown in Table 1, our method achieves the lowest complexity and offers graceful complexity growth as resolution increases. Building on the design principles of Zero-1-to-3 [17], we generate all target views in parallel and introduce mesh attention blocks to maintain cross-view consistency. In addition, we enhance texture and geometric consistency by incorporating multi-scale VAE latent features and keypoints conditioning.

Apart from introducing the MEAT approach, we also present a new training source. The typical data source for multiview diffusion models is textured mesh data. However, high-quality human scan data at 1024×1024 resolution is extremely scarce and mostly limited to static poses. Even the largest dataset, THUman2.1 [35], includes only around 2,500 multiview subjects, making multiview diffusion model training highly susceptible to overfitting. To address this, we propose a data processing pipeline that leverages DNA-Rendering [5], a multiview human motion video dataset, as a training source. The data greatly increases the diversity of poses available during training. We will discuss a series of techniques for adapting this dataset to train our mesh-attention-based multiview diffusion model.

To summarize, our main contributions are as follows:

- We propose mesh attention, which establishes correspondences between pixels using rasterization and projection of a centric mesh, making it the most efficient cross-view attention method to date.
- Based on mesh attention, we introduce a human-specific multiview diffusion model, MEAT, capable of generating consistent 16-view images at megapixel resolution.
- We present techniques for adapting a large-scale multiview human motion video dataset as a training source for multiview diffusion.

2. Related Work

Multiview Diffusion. Research of multiview diffusion models began with Zero-1-to-3 [17], which first proposed using camera viewpoints as control conditions for image diffusion models to achieve novel view synthesis. As a one-view-at-a-time approach, it often produces inconsistencies in the generated views due to the stochastic nature of diffusion models. Subsequent approaches shifted to all-view-at-once generation to mitigate the inconsistency issue.

The first category of methods [14, 15, 19, 27, 29, 31] treats the generation of each view as a separate branch of image generation, using multiview attention across branches to achieve feature fusion and consistency constraints. MVDream [27] introduces dense multiview attention for single-object text-to-multiview generation. ImageDream [31] expands this approach to image-conditioned generation. Wonder3D [19] incorporates normal data and cross-domain attention to enhance geometric consistency. Recent methods have started optimizing the complexity of multiview attention. EpiDiff [14] uses epipolar attention for efficient pixel-matching candidate retrieval. Era3D [15] proposes row-wise attention based on the orthographic projection assumption. Other methods treat multiview images in alternative forms, such as a tiled big image [26] or a video [8, 30], leading to different approaches. Our work, MEAT, further extends parallel multiview generation by enabling direct cross-view feature integration through rasterization and projection using a central 3D mesh representation.

Monocular Human Reconstruction. Monocular human reconstruction methods can be categorized into two groups based on whether they rely on optimizing a 3D representation. Optimization-based approaches, like ICON [33] and ECON [34], achieve purely geometric clothed human reconstruction with aligned SMPL-X [21] parameters, while TeCH [13] and SIFU [38] additionally support faithful texture generation. The other category of methods [24, 25, 40] use feed-forward networks to estimate the 3D occupancy field and extract the human mesh using the Marching Cubes algorithm [20], then attach textures through shape-guided inpainting [3]. A concurrent work, MagicMan, like our approach, combines a 512-resolution multiview diffusion model with monocular human reconstruction. MagicMan and our MEAT can generate dense multiview results that can be directly applied to 2DGS [12] reconstruction.

3. Methodology

3.1. Preliminaries

Multiview Diffusion Models. Following the structure of the Latent Diffusion Model (LDM) [23], existing multiview diffusion models typically consist of a VAE encoder $\mathcal{E}$, a denoiser U-Net ϵ_θ, and a VAE decoder $\mathcal{D}$. The encoder maps the image x_0 into a low-resolution latent space as $z_0 = \mathcal{E}(x_0)$. The decoder $\mathcal{D}$ then reconstructs the image from the latent feature. Multiview diffusion models can be categorized into two main types: one-view-at-a-time approaches [17, 26] and all-view-at-once methods [14, 15, 19, 27, 31, 39].

The first category, represented by Zero-1-to-3 [17], trains the denoiser ϵ_θ to process one target view at a time. It predicts the noise ϵ from the noisy latent z_t of the target view image x_0, conditioned on the reference view image y and the associated relative camera rotation R and translation T. The training objective is

$$\min_\theta \mathbb{E}_{\mathcal{E}(x_0),\epsilon\sim\mathcal{N}(0,I),t}\Big[\|\epsilon - \epsilon_\theta(z_t, t, y, R, T)\|_2^2\Big]. \quad (1)$$

Such models can generate multiple target views sequentially but lack explicit consistency constraints across views.

The second category processes all target views simultaneously and integrates features across views using attention modules, improving cross-view consistency. The training objective is extended to include N target views:

$$\min_\theta \mathbb{E}_{\mathcal{E}(x_0^{1:N}),\epsilon\sim\mathcal{N}(0,I),t}\Big[\|\epsilon - \epsilon_\theta(z_t^{1:N}, t, y, [R,T]^{1:N})\|_2^2\Big]. \quad (2)$$

While this approach achieves better cross-view visual consistency, it comes at the cost of significant memory and computational overhead during the cross-view attention. Our model also follows the all-view-at-once setup but efficiently produces the non-trivial dense, 1024×1024 high-resolution multiview generation through a novel mesh attention mechanism, which we will detail in Sec. 3.2.

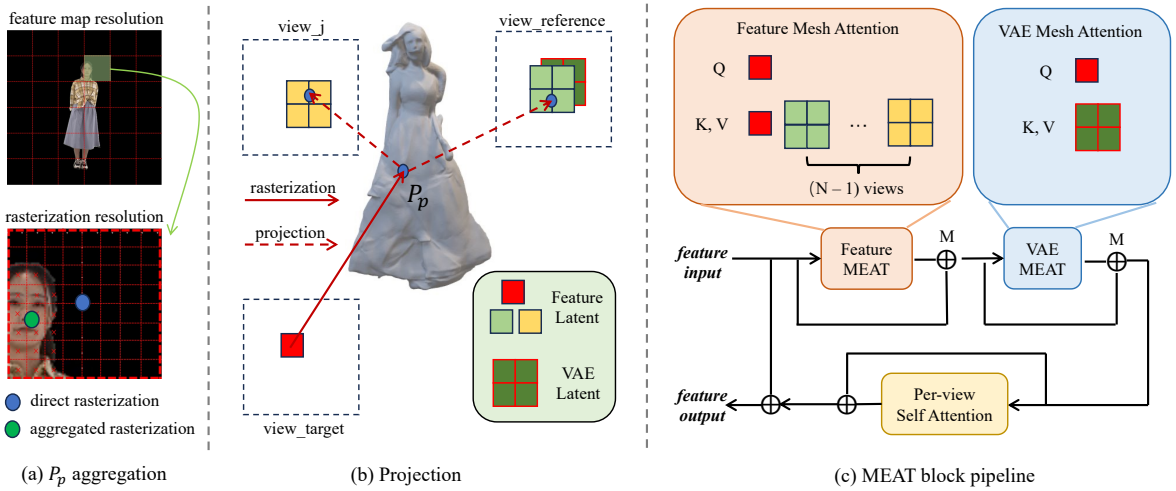
Rasterization. In mesh-based rasterization, each pixel on the 2D image plane is associated with a ray cast from the camera into 3D space, intersecting with the mesh surface. For each pixel $\boldsymbol{p}$, the rasterization output includes the intersection mask $M_{\boldsymbol{p}}$, the intersected triangle face index ϕ, and the barycentric coordinates $\boldsymbol{\lambda_p} = (\lambda_{p1}, \lambda_{p2}, \lambda_{p3})$. With the barycentric coordinates $\boldsymbol{\lambda_p}$ and the triangle face vertex coordinates $\boldsymbol{P}_\phi$, we can derive the 3D coordinates of the intersected point on mesh

$$\boldsymbol{P_p} = \text{interp}(\boldsymbol{\lambda_p}, \boldsymbol{P}_\phi). \quad (3)$$

Our mesh attention takes advantage of the aggregation and projection of $\boldsymbol{P_p}$.

3.2. Mesh Attention

We introduce mesh attention, MEAT, to overcome the inefficiencies of traditional cross-view attention, where each pixel must access and integrate information from all other pixels in different views, resulting in substantial redundant computation. In practice, pixels across views correspond to each other according to the 3D structure of the object. Given an approximate clothed mesh as the centric coarse geometric representation of the human object, our approach leverages the 3D coordinate transformations to directly identify



(a) P_p aggregation (b) Projection (c) MEAT block pipeline

Figure 3. **Mesh Attention Block.** (a) P_p aggregation. When the resolution of the feature map is very low, the ray cast from the center of a pixel may not intersect with the mesh, although the pixel area itself overlaps with it. (b) Projection. Each projected point is rounded to four integer pixels, corresponding to $d = 4$ in Table 1. The projected points on the reference view are also used to retrieve the encoded VAE features. (c) MEAT block pipeline. We use mesh attention to fuse U-Net features from all N views, and VAE features from the reference. An additional per-view self-attention block is applied to process the captured multiview features. M stands for masked skip connection.

corresponding 2D pixel locations across different views. This allows us to aggregate information from these matched pixels, reducing redundancy and improving cross-view consistency. Details of MEAT are explained below.

Aggregated Rasterization. We can obtain the 3D coordinates of the intersection on the mesh for each pixel $\boldsymbol{p}$ through rasterization and Eq. (3). However, due to the potentially low resolution of the diffusion features (*e.g.*, 16×16 mid-block feature maps for 1024^2 images), pixels near the object edges, which may contain useful information, can be misclassified as having no intersection with the mesh when using direct rasterization, as shown in Fig. 3(a). To address this, we aggregate higher-resolution rasterization results to obtain the intersection point $\boldsymbol{P_p}$ and the mask $M_{\boldsymbol{p}}$ at the resolution of the feature map.

Consider a pixel $\boldsymbol{p}$ on the feature map that corresponds to a pixel region S in the higher-resolution rasterization. We treat $\boldsymbol{P_p}$ as the average of all valid $\boldsymbol{P_s}$ within the region S:

$$\boldsymbol{P_p} = \frac{\sum_{\boldsymbol{s}\in S} M_{\boldsymbol{s}} \boldsymbol{P_s}}{\sum_{\boldsymbol{s}\in S} M_{\boldsymbol{s}}}, \tag{4}$$

$$M_{\boldsymbol{p}} = \vee_{\boldsymbol{s}\in S} M_{\boldsymbol{s}}, \tag{5}$$

where $\vee$ is the "logical or" operation. The higher-resolution rasterization only needs to be performed once and can be reused for aggregation at different target resolutions.

Projection and Grid Sampling. After obtaining $\boldsymbol{P_p}$ for a target view pixel $\boldsymbol{p}$, we can use the calibration matrices K_v, R_v, T_v of each view v to locate the corresponding pixel of $\boldsymbol{P_p}$ in other views:

$$\boldsymbol{p}_v = [K_v(R_v \boldsymbol{P_p} + T_v)]_{xy}. \tag{6}$$

The corresponding features can then be retrieved using grid sampling. Instead of interpolating the features of neighboring pixels based on $\boldsymbol{p}_v = (x, y)$, we round x, y up and down to extract the corresponding four features $\boldsymbol{f}_v$ from the feature map $\boldsymbol{F}_v$ of view v:

$$\boldsymbol{f}_v = \text{grid\_sample}(\boldsymbol{F}_v, \{\lfloor x \rfloor, \lceil x \rceil\} \times \{\lfloor y \rfloor, \lceil y \rceil\}). \tag{7}$$

Cross-view Attention. For pixel $\boldsymbol{p}$ on the target view with U-Net feature $\boldsymbol{f}$, we use cross attention to fuse the features from other views. To provide location priors, we concatenate the harmonic-embedded view camera pose $\boldsymbol{c}_v$ to the raw U-Net features $\boldsymbol{f}_v$. The masked skip connections are applied to omit pixels that do not intersect with the mesh from participating in mesh attention.

$$Q = W_Q(\boldsymbol{f} \oplus \boldsymbol{c}_{tgt}) \quad K, V = W_{K,V}(\boldsymbol{f}_{1:N} \oplus \boldsymbol{c}_{1:N}), \tag{8}$$

$$\text{MEAT}_{feat}(\boldsymbol{f}, \boldsymbol{p}) = M_{\boldsymbol{p}} \cdot \text{Attention}(Q, K, T) + \boldsymbol{f}, \tag{9}$$

where $\oplus$ denotes channel-wise concatenation.

In addition to the fusion of U-Net features across views, we use a fully convolutional residual encoder to process VAE latent z_0 of the reference view into multi-scale feature tensors $\boldsymbol{F}_\gamma$ and inject them through mesh attention. Specifically, for pixel $\boldsymbol{p}$ on the target view, we use the projection of $\boldsymbol{P_p}$ on the reference view as the pixel location $\boldsymbol{p}_{ref}$ and employ grid sampling as defined in Eq. (7) to extract $\boldsymbol{f}_\gamma$ from the VAE features. Mesh attention is applied exclusively to the reference view in this step.

$$Q_\gamma = W_{Q_\gamma}(\boldsymbol{f} \oplus \boldsymbol{c}_{tgt}) \quad K_\gamma, V_\gamma = W_{K_\gamma, V_\gamma}(\boldsymbol{f}_\gamma \oplus \boldsymbol{c}_{ref}), \tag{10}$$

$$\text{MEAT}_{vae}(\boldsymbol{f}, p) = M_p \cdot \text{Attention}(Q_\gamma, K_\gamma, V_\gamma) + \boldsymbol{f}. \tag{11}$$

Here, ref and tgt indicate the reference and target view. The above operations are applied to each pixel of each view.

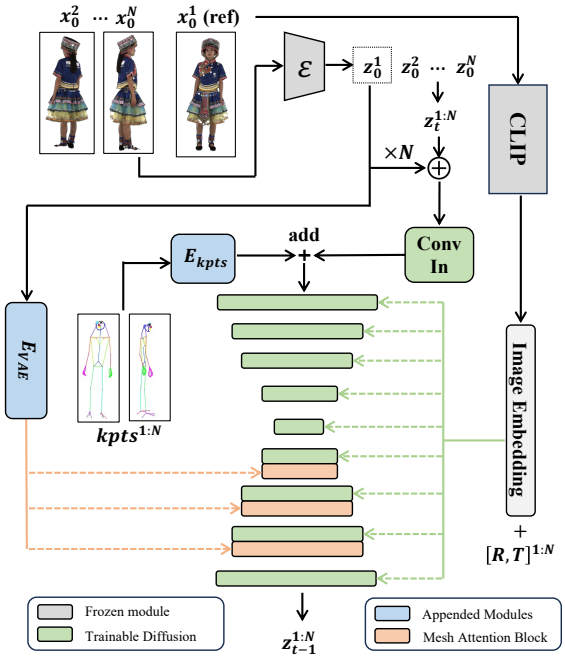


Figure 4. **Pipeline of MEAT.** We insert mesh attention blocks into up-sampling blocks of the U-Net to fuse multiview features.

After the two per-pixel attention operations, we apply a self-attention mechanism for each view to process the fused features. The complete pipeline is shown in Fig. 3(c). In the classifier-free guidance training scheme, we always retain the mesh attention module and set 15% of the data's camera embeddings and concatenated x_0 to null, encouraging the model to fully leverage the mesh attention.

3.3. Multiview Diffusion Model with MEAT

Figure 4 shows the proposed framework for multi-view human image generation. The framework incorporates some design principles from Zero-1-to-3 [17]. Specifically, we employ a view-conditioned diffusion model to synthesize novel views of an object by learning controls over camera viewpoints. Unlike Zero-1-to-3, which processes one view at a time and may encounter view consistency issues, our framework processes all target views simultaneously and integrates features across views using the proposed mesh attention mechanism, detailed in Sec. 3.2. In addition, our framework incorporates the following designs to improve performance: 1) Keypoint conditioning, 2) Resolution upscaling and choice of VAE, and 3) Linear noise schedule.

Keypoint Conditioning. To develop a model with strong generalization capabilities, we use DNA-Rendering [5] as our training dataset. This dataset comprises real human videos captured from multiple views, offering a diverse range of poses. However, this also adds complexity to model learning. To handle these complex poses, we propose incorporating detected skeleton keypoints of the target views into the model. Specifically, we add the keypoint features (after adjusting their spatial resolutions and channel numbers) to the U-Net features as a condition. By explicitly providing such keypoint conditioning, our model no longer needs to rely solely on camera parameters to estimate human poses in new views and can instead focus on ensuring cross-view consistency and generating detailed outputs.

Resolution Upscaling and Choice of VAE. As analyzed in Sec. 1 and Table 1, most multiview diffusion models are limited to a low-resolution of 256×256, with only a few recent studies reaching 512×512. As shown in Fig. 2, higher resolutions and improved VAE models are crucial for capturing highly detailed human data. To minimize cross-view inconsistencies and quality degradation caused by VAE reconstruction, we train our model using 1024×1024 images and use SDXL VAE [22] in our framework.

Noise Schedule. Following the recommendation from Zero123++ [26], we use a linear schedule for the denoising process instead of a scaled-linear schedule to achieve better global consistency across multiple views.

3.4. Inference

For in-the-wild image inputs, we crop the image according to the dataset setting, which we detail in Sec. 4. We then apply ECON [34] to produce the clothed human mesh and the corresponding SMPL-X [21] parameters.

Orthographic to Perspective. Since ECON operates under an orthographic camera assumption, we first obtain a frontal perspective camera by optimization. Based on the "Look-At" transformation, we assume a fixed field of view (FoV) for all cameras, directed at the pelvis. We optimize the frontal camera position to align the rendered SMPL-X keypoints with those in the image. After that, we sample camera parameters that cover a 360-degree view of the human body, maintaining a fixed elevation and distance.

Generation and Reconstruction. With these cameras, we render a keypoints visualization image for each view and perform rasterization and aggregation for mesh attention. We use DDIM [28] scheduler to generate multiview images and apply 2DGS [12] for direct reconstruction.

4. Adapting DNA-Rendering for Training

We construct our training data using the multiview human dataset DNA-Rendering [5], 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 two primary issues: (1) adapting the monocular reconstructed mesh to the calibrated coordinate system, and (2) cropping the images with corresponding adjustments to the camera calibration parameters. For further

details, please refer to the *Supp. Mat*.

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 a pre-selected frontal image to extract the centric mesh for training. We use PIFuHD[25] 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. $\boldsymbol{P_p}$ of each pixel $\boldsymbol{p}$ 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 [6] to detect all feature-matching pairs and apply gradient descent to solve TF.

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 viewpoint 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:

$$R_v = 1.3 \cdot \max_{\boldsymbol{P}} |\Pi_v(\boldsymbol{P})_y - \Pi_v(\boldsymbol{P}_{pelvis})_y|. \tag{12}$$

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.

5. Experiments

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

Baselines. For quantitative experiments, we compare our method with Stable Zero123 [2], SyncDreamer [18], Wonder3D [19], and SV3D [30]. 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 256×256. Wonder3D is only trained in the color domain since ground-truth normal maps are not available. We only compare the results of MagicMan [9] qualitatively as its preset views cannot align with the test setting.

Metrics. Since most of previous multi-view diffusion models only generate at 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 resolution of 1024. For both resolutions, we include PSNR, SSIM [32], and LPIPS [37] metrics to compare the generated results with the ground-truth images. For the 1024 category, we use Patch-FID (P-FID) [4, 7, 16] instead of FID [10] as a metric for generation quality. FID resizes images to 299×299 before calculation, which does not reflect MEAT's advantage at high resolutions. Instead, we split each image into a 4×4 grid of 256×256 patches and select the middle two columns, yielding eight patches per image. The calculation is based on the patch set. In the 256 category, we also use the PPLC metric proposed by Free3D [39] 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.

5.1. Main Results

Quantitative. Table 2 presents the quantitative comparison with the baselines. For each method, we generate 16 pre-set viewpoints and compare them with the ground-truth images. Our method achieves the best results across both resolutions in reconstruction metrics and leads in generation quality. Notably, MEAT significantly outperforms existing methods on the Patch-FID metric, highlighting the value of megapixel-resolution training. For cross-view consistency metric (PPLC), Wonder3D, without retraining, achieves the best performance, with our method closely following. The results of Wonder3D highlight the significant improve-

Figure 5. **Qualitative Results.** MEAT (Ours) demonstrates significant advantages in resolution, detail, and cross-view consistency in novel view synthesis tasks. * Methods are re-trained on the DNA-Rendering dataset for fair comparison. Please **zoom in** for details.

ment in cross-view consistency made possible by combining cross-domain attention. However, Wonder3D is highly memory-intensive and difficult to scale to megapixel resolutions. In contrast, our method is much more efficient.

Qualitative. We show the qualitative comparison with other baselines in Fig. 5. All methods operating at 256^2 resolution fail to produce any facial details, and their texture clarity is noticeably inferior to that of MEAT. The pre-trained Wonder3D frequently generates highly consistent back views with limited perspective variation, potentially giving it an unfair advantage in the PPLC metric. SV3D shows a clear improvement in resolution but falls short of our method in geometric consistency, lacking perceptual awareness of human structure. MagicMan, as a concurrent work, stands out among the baselines but still struggles with visible artifacts and incomplete limbs when generating side views (*e.g.*, in the third example). Our method achieves high-resolution, detail-rich, and view-consistent human novel view synthesis. More examples are available in the *Supp. Mat.*

5.2. Ablations and Discussions

The qualitative and quantitative ablation results are shown in Fig. 6 and Table 3, respectively.

Resolution Upscaling. Directly increasing the training resolution to 1024 causes Stable Zero123 to generate numer-

Table 2. **Main Quantitative Results.** We highlight the best value in blue, and the second-best value in green.

Method	Type	Res.	1024				256				
			PSNR ↑	SSIM ↑	LPIPS ↓	P-FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PPLC ↓
Stable Zero123 [2]	Infer.	256	9.039	0.7839	0.3299	74.24	9.056	0.7033	0.3966	55.16	0.4549
SyncDreamer [18]	Infer.	256	12.12	0.8653	0.2331	102.8	12.13	0.7998	0.3231	71.42	0.2017
Wonder3D [19]	Infer.	256	16.58	0.9084	0.1456	59.79	16.68	0.8649	0.1359	39.32	0.0897
SV3D [30]	Infer.	578	13.32	0.8843	0.1830	24.99	13.43	0.8175	0.2372	20.14	0.1333
Stable Zero123 [2]	Train	256	17.52	0.9139	0.1345	62.71	17.62	0.8768	0.1173	34.53	0.1010
Wonder3D [19]	Train	256	16.73	0.9081	0.1449	67.11	16.82	0.8684	0.1356	47.59	0.1042
MEAT (Ours)	1-stage	1024	18.91	0.9271	0.0751	10.60	19.41	0.9043	0.0791	16.56	0.0991

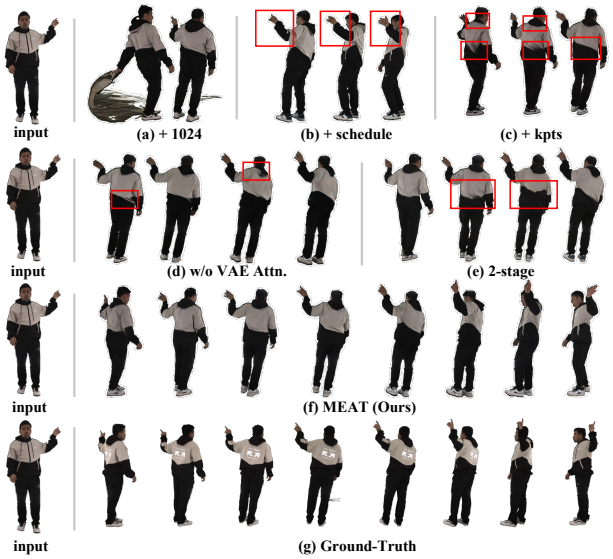


Figure 6. **Qualitative Ablation.** MEAT achieves the best cross-view consistency.

Table 3. **Quantitative Ablation.** Best value in blue, second-best in green. *Here PPLC is calculated on 1024 × 1024 resolution.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	P-FID ↓	PPLC* ↓
SZ123 - 256	17.51	0.9139	0.1344	23.71	62.71	–
+ res. 1024	14.41	0.8873	0.1480	21.41	14.56	0.1805
+ schedule	16.56	0.9114	0.1023	16.81	11.21	0.1170
+ keypoints	18.78	0.9238	0.0776	16.19	10.79	0.0995
× VAE Attn	18.50	0.9233	0.0788	16.91	10.76	0.0981
2-stage	19.11	0.9266	0.0755	15.37	9.983	0.0973
Ours	18.91	0.9271	0.0751	17.08	10.60	0.0928

ous artifacts, as shown in Fig. 6-(a). Adjusting the noise scheduler to reduce the SNR at the beginning of the denoising process is key to mitigating this issue (see Fig. 6-(b)).

Keypoint Conditioning. Without the keypoint condition, in Fig. 6-(b), the generated results show noticeable misalignment in the left arm, when compared against the reference view and ground truth. The keypoint conditioning reduces the model's difficulty in understanding the human geometric structure.

Mesh Attention. Adding only keypoint conditioning does not ensure cross-view consistent texture generation, as each view is still generated independently (see Fig. 6-(c)). Mesh attention is the key to address the consistency issue. We compared three variants with mesh attention. Models without VAE attention tend to produce local consistency anomalies, as is shown in Fig. 6-(d). We examine a 2-stage training strategy for MEAT, where we first train the U-Net without mesh attention for 100k iterations, then freeze these parameters and train the mesh attention block for another 50k iterations. We find that this model shows slightly better generation quality in terms of FID, but exhibits noticeable issues with cross-view consistency. It usually shows inconsistencies in texture patterns, such as color blocks. See Fig. 6-(e). As is reflected by Fig. 6-(f) and the PPLC metric in Table 3, 1-stage-trained MEAT shows the best cross-view consistency. See *Supp. Mat.* for more comparison.

6. Conclusion

In this paper, we propose **MEAT**, a human-specific multiview diffusion model that generates dense novel views of humans on megapixels conditioned on a frontal image. Our proposed **mesh attention** uses the monocular-reconstructed human mesh as a coarse central geometric representation, establishing cross-view coordinate correspondences through rasterization and projection. It enables highly memory-efficient cross-view attention, which overcomes the high complexity that hinders increasing the resolution to 1024^2 for existing multiview attention methods. Through a series of techniques, we have, for the first time, enabled training a multiview diffusion model using multiview human motion videos, effectively enhancing the pose diversity of the training dataset. Extensive experiments demonstrate that our generated multi-view human images exhibit significant advantages in cross-view consistency, clarity, and detail quality.

Acknowledgement. This work is supported by the National Research Foundation, Singapore under its AI Singapore Programme

(AISG Award No: AISG2-PhD-2022-01-030), the RIE2020 Industry Alignment Fund Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] SDXL-VAE-FP16-Fix. https://huggingface.co/madebyollin/sdxl-vae-fp16-fix, 2023. 11

[2] Stable Zero123. https://stability.ai/news/stable-zero123-3d-generation, 2024. 6, 8

[3] Badour AlBahar, Shunsuke Saito, Hung-Yu Tseng, Changil Kim, Johannes Kopf, and Jia-Bin Huang. Single-image 3D human digitization with shape-guided Diffusion. In *SIGGRAPH Asia*, 2023. 3

[4] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis. In *ECCV*, 2022. 6

[5] 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, 11

[6] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust dense feature matching. In *CVPR*, 2024. 6, 11

[7] 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. 6

[8] 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. 3

[9] 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. 6

[10] 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. 6

[11] I Ho, Jie Song, Otmar Hilliges, et al. SiTH: Single-view textured human reconstruction with image-conditioned diffusion. In *CVPR*, 2024. 12

[12] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2D gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH*, 2024. 3, 5

[13] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-guided reconstruction of lifelike clothed humans. In *3DV*, 2024. 3

[14] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. EpiDiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *CVPR*, 2024. 1, 3

[15] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. Era3D: High-resolution multiview diffusion using efficient row-wise attention. In *NeurIPS*, 2024. 1, 2, 3

[16] 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. 6

[17] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In *ICCV*, 2023. 1, 2, 3, 5

[18] 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. 6, 8

[19] 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, 2, 3, 6, 8

[20] William E Lorensen and Harvey E Cline. Marching Cubes: A high resolution 3D surface construction algorithm. *ACM SIGGRAPH Computer Graphics*, 21:163–169, 1987. 3

[21] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 3, 5

[22] 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. 5, 11

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[24] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019. 3

[25] 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. 3, 6, 11

[26] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: A single image to consistent multi-view diffusion base model. *arXiv preprint*, arXiv:2310.15110, 2023. 1, 3, 5

[27] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3D generation. In *ICLR*, 2024. 2, 3

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 5

[29] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. MVDiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *NeurIPS*, 2023. 3

[30] 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, 3, 6, 8

[31] Peng Wang and Yichun Shi. ImageDream: Image-prompt multi-view diffusion for 3D generation. *arXiv preprint*, arXiv:2312.02201, 2023. 2, 3

[32] 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. 6

[33] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. ICON: Implicit clothed humans obtained from normals. In *CVPR*, 2022. 3

[34] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. ECON: Explicit clothed humans optimized via normal integration. In *CVPR*, 2023. 3, 5

[35] 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 *CVPR*, 2021. 2

[36] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 11

[37] 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. 6

[38] Zechuan Zhang, Zongxin Yang, and Yi Yang. SIFU: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *CVPR*, 2024. 3, 12

[39] Chuanxia Zheng and Andrea Vedaldi. Free3D: Consistent novel view synthesis without 3D representation. In *CVPR*, 2024. 3, 6

[40] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction. *TPAMI*, 44: 3170–3184, 2021. 3

[41] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 11

Appendix

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 further specify the model implementation details.

VAE Version. Notably, since the SDXL-VAE [22] can produce `NaN` under `fp16` precision, we utilize the `fp16-fix` version [1] to support mixed-precision training.

Keypoints Conditioning. We use a small 3-layer convolutional network to process the keypoints condition, downsampling 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-[36] 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.

B. Dataset

In this section, we discuss further details of the novel ideas proposed to harness multiview human video dataset DNA-Rendering [5] for multiview diffusion training.

Frontal Camera Selection. For each frame of multiview images in the DNA-Rendering [5] 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 $\boldsymbol{d}$ of the human body, the 3D coordinates $\boldsymbol{G}$ of the pelvis, and the camera coordinates $\boldsymbol{C}_v$, where

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

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.*

$$\text{front view} \leftarrow \arg\max_v \frac{\boldsymbol{d} \cdot \boldsymbol{G}\boldsymbol{C}_v}{\|\boldsymbol{d}\|\|\boldsymbol{G}\boldsymbol{C}_v\|} \tag{13}$$

Mesh Adaptation. With the selected frontal image, we use PIFuHD [25] to predict a clothed human mesh. To adapt this mesh into the DNA-Rendering camera system, we need to determine the transformation TF to align the mesh with the world coordinate system. We assume that the transformation TF for each vertex $\boldsymbol{P}$ consists of a scaling S, rotation R, and translation $\boldsymbol{t}$:

$$S = \text{diag}(\boldsymbol{s}), \boldsymbol{s} = [s_x, s_y, s_z], \tag{14}$$

$$R = \texttt{rot6d}(\boldsymbol{c}_1, \boldsymbol{c}_2), \tag{15}$$

$$p' = \text{TF}(\boldsymbol{P}) = R(S\boldsymbol{P}) + \boldsymbol{t}. \tag{16}$$

We use `rot6d` rotation representation [41] for more stable optimization. We can then define the re-projection process $\tilde{\Pi}_v$ of a frontal-view pixel $\boldsymbol{p}$ into the view v.

$$\tilde{\Pi}_v(\boldsymbol{p}) = \Pi_v(\text{TF}(\boldsymbol{p} \rightarrow \boldsymbol{P})). \tag{17}$$

Here $\boldsymbol{p} \rightarrow \boldsymbol{P}$ indicates the inverse orthographic rasterization process and Π_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. $\tilde{\Pi}_1(\boldsymbol{p})$ - Pixels return to their original positions.
2. $\tilde{\Pi}_v(\boldsymbol{p})$ - Pixel $\boldsymbol{p}$ on the frontal view is matched with pixel $\boldsymbol{q}_v$ on view v.

We use RoMa [6] to detect such $(\boldsymbol{p}, \boldsymbol{q}_v)$ pairs. All the pixels $\boldsymbol{p}$ 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:

$$\arg\min_{\boldsymbol{s},\boldsymbol{c}_1,\boldsymbol{c}_2,\boldsymbol{t}} \sum_{\boldsymbol{p}} \|\boldsymbol{p} - \tilde{\Pi}_1(\boldsymbol{p})\|_2^2 + \sum_{\boldsymbol{p},\boldsymbol{q}_v} \|\boldsymbol{q}_v - \tilde{\Pi}_v(\boldsymbol{p})\|_2^2. \tag{18}$$

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

$$R_0 = \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \cdot R_{v=1} \right)^{-1} \tag{19}$$

Here $R_{v=1}$ is the calibrated extrinsic rotation matrix of the frontal camera in the DNA-Rendering [5] 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. 7.

C. More Results

C.1. Cross-view Consistency Preservation

We show the generated results of models with and without mesh attention modules in Fig. 8. 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. 8, 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.

(a) Mesh location before adaptation.

(b) Mesh location after adaptation.

Figure 7. **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.

Figure 8. **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.

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 [11] and SIFU [38]. The qualitative comparison results are shown in Fig. 9. 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.

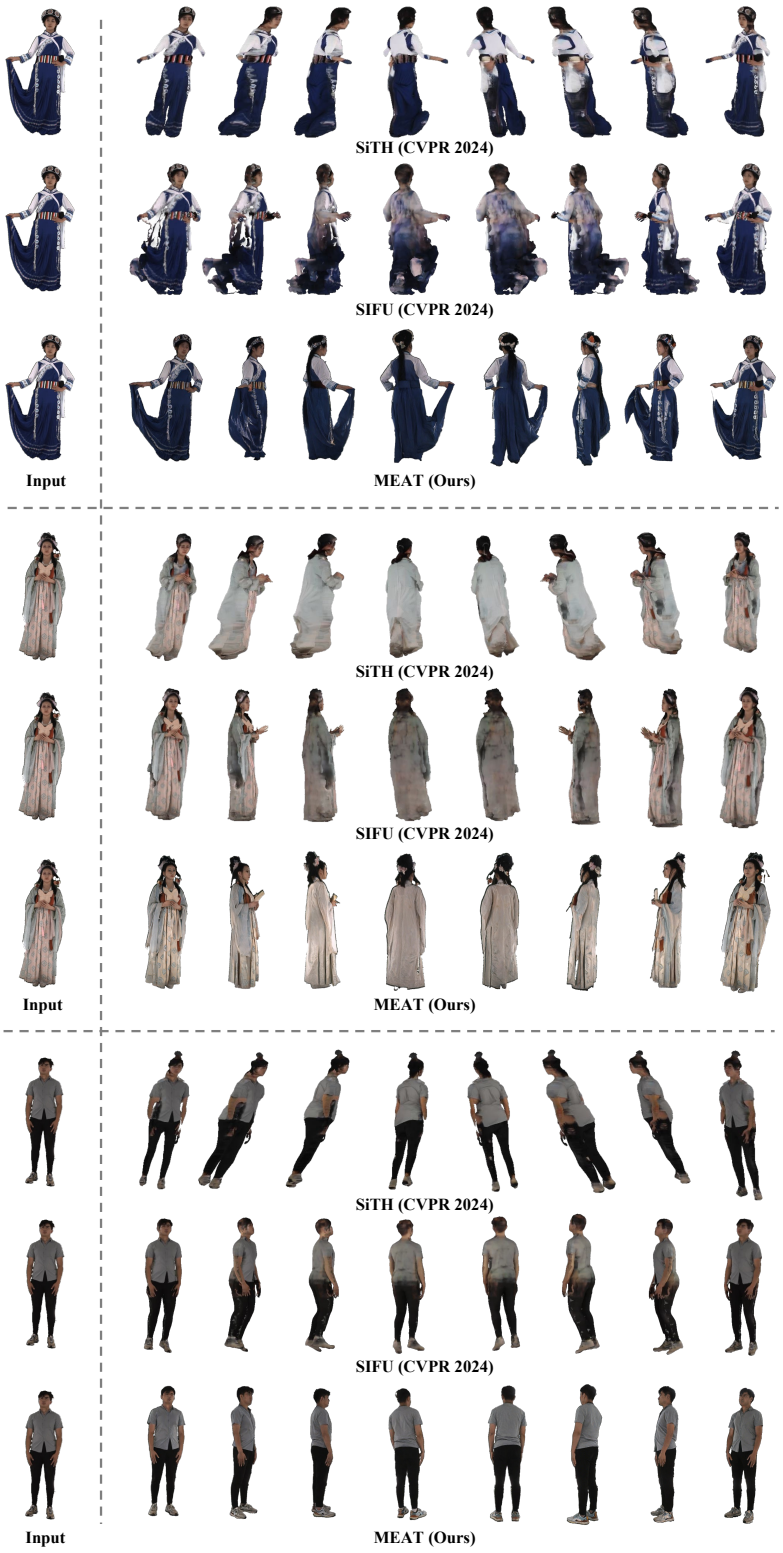


Figure 9. **Comparison with Monocular Reconstruction Methods.** In the novel view generation results for human bodies, compared to monocular reconstructed meshes, the multiview images generated by our MEAT diffusion model exhibit significant advantages in geometric plausibility, geometric details, texture details, and clarity. Please **zoom in** for details.