AdaHuman: Animatable Detailed 3D Human Generation with Compositional Multiview Diffusion

Yangyi Huang Ye Yuan Xueting Li Jan Kautz Umar Iqbal
NVIDIA
https://nvlabs.github.io/AdaHuman

Figure 1. Given a single input image, AdaHuman reconstructs pixel-aligned a 3DGS avatar with detailed appearance. It can also generate the same avatar in novel poses, or in a standard animation-friendly A-pose to build an animatable avatar.

Abstract

Existing methods for image-to-3D avatar generation struggle to produce highly detailed, animation-ready avatars suitable for real-world applications. We introduce AdaHuman, a novel framework that generates high-fidelity animatable 3D avatars from a single in-the-wild image. AdaHuman incorporates two key innovations: (1) A pose-conditioned 3D joint diffusion model that synthesizes consistent multi-view images in arbitrary poses alongside corresponding 3D Gaussian Splats (3DGS) reconstruction at each diffusion step; (2) A compositional 3DGS refinement module that enhances the details of local body parts through image-to-image refinement and seamlessly integrates them using a novel crop-aware camera ray map, producing a cohesive detailed 3D avatar. These components allow AdaHuman to generate highly realistic standardized A-pose avatars with minimal self-occlusion, enabling rigging and animation with any input motion. Extensive evaluation on public benchmarks and in-the-wild images demonstrates that AdaHuman significantly outperforms state-of-the-art methods in both avatar reconstruction and reposing. Code and models will be publicly available for research purposes.

1. Introduction

Generating high-quality animatable 3D human avatars is crucial for numerous applications in gaming, animation, and virtual reality. Recent advances in diffusion-based image generation models have significantly accelerated research in this domain. Early approaches tackled this challenge using score distillation sampling (SDS) [33], where a 3D model is distilled from a diffusion-based image generation model [16, 18]. While SDS-based methods offer flexibility and compatibility with various 3D representations, they suffer from oversaturation artifacts and slow generation speed, making them impractical for large-scale avatar creation. More recent methods have shifted towards multi-view generation and reconstruction pipelines [55], where diffusion models first synthesize multi-view images from text or image inputs, followed by a reconstruction phase that converts these images into a 3D avatar. This feed-forward approach improves both realism and generation speed. However, significant challenges remain. First, the avatars are typically generated in the same pose as the in-

put image, leading to self-occlusion issues that complicate rigging and animation. Second, the resulting avatars often lack fine details and appear blurry, limiting their utility in real-world applications.

Motivated by these challenges, we introduce AdaHuman, a new framework for generating animatable high-fidelity 3D human avatars from a single input image. At its core, AdaHuman employs a pose-conditioned joint 3D diffusion model that seamlessly integrates multi-view image synthesis with 3D Gaussian Splats (3DGS)-based reconstruction during the diffusion process. By performing 3D reconstruction at each diffusion step, our approach ensures strong multi-view consistency across generated images, resulting in high-quality 3DGS avatars. A key advantage of our multi-view diffusion model is its ability to generate images in any arbitrary pose by simply conditioning on the desired pose. To enable animation, we leverage this capability to generate the 3DGS avatar in a standard A-pose, which minimizes self-occlusion, inpaints missing details, and naturally facilitates rigging and animation. Notably, our method achieves this without requiring training images in such standard poses.

To enhance the fidelity and detail of the generated avatars, AdaHuman further introduces a compositional 3DGS refinement module. This module first renders zoomed-in views of local body parts (e.g., head, upper body, lower body) from the initial 3DGS avatar. These local views then undergo an image-to-image refinement process using our multi-view diffusion model to improve detail and resolution. Using these refined local views, we propose a novel approach that seamlessly integrates the local views and global full-body views to produce a highly detailed holistic 3D avatar. This is enabled through two innovations: (1) a crop-aware camera ray map that establishes precise correspondences between 3D locations in local and global views, and (2) a visibility-aware composition scheme that intelligently merges partial 3DGS reconstructions based on view coverage and visibility salience. Our approach effectively prevents floating artifacts while preserving fine details and coherency, resulting in high-quality 3D avatars with enhanced local and global consistency.

In summary, our main contributions are as follows: (1) We introduce a new image-to-avatar framework leveraging pose-conditioned 3D joint diffusion, enabling both avatar reconstruction and reposing for seamless rigging and animation. (2) We develop an innovative compositional 3DGS refinement approach that produces highly detailed and globally consistent avatars using a crop-aware camera ray map and a visibility-aware composition scheme. (3) Through comprehensive evaluation on public benchmarks and challenging in-the-wild images, we demonstrate that our method substantially outperforms state-of-the-art approaches in both avatar reconstruction and reposing tasks.

2. Related Work

3D Avatar Reconstruction. Early methods for monocular RGB-based 3D avatar reconstruction typically rely on the SMPL [7, 30] model, predicting per-vertex offsets to capture clothing and hair details, but are limited by SMPL's fixed topology. Consequently, recent approaches adopt implicit representations allowing arbitrary topologies [3, 13, 18, 19, 38, 39, 51, 52, 56, 63, 64], however, they depend heavily on extensive 3D training data. Moreover, they struggle with occlusion handling in complex poses making it difficult to animate the reconstructed avatars. Methods enabling animation through pose canonicalization usually require ground-truth standard-pose meshes or rigged avatars [11, 19, 32]. In contrast, our method generalizes reposing from diverse multiview video data, directly generating avatars in arbitrary poses without relying on standard-pose training data.

3D Avatars Generation via 2D Foundation Model. Advances in 2D diffusion models [35, 37] have driven significant progress in 3D avatar generation [4, 6, 16, 17, 20, 22, 24, 25, 33, 36, 42, 44, 49, 60–62]. These methods adopt the Score Distillation Sampling (SDS) technique to extract 3D knowledge from these models. SDS-based methods, however, suffer from unrealistic outputs and slow iterative optimization, resulting in lower quality avatars and prohibitively long run time for wider adoption.

Joint Diffusion and Reconstruction. Recent methods combine diffusion models with reconstruction networks to improve efficiency and quality for 3D avatar generation [26, 27, 29, 41, 53, 54, 66]. Zero123 [28] and its variants [26, 27, 29, 41] generate consistent multi-view images that facilitate accurate 3D avatar reconstruction. More recent works [14, 45, 47, 57] predict implicit 3D representations directly from multi-view images, enabled by advancements in implicit representations such as triplanes [5] and Gaussian Splats [21]. Similar strategies have been applied to human avatars [2, 18, 23, 40], though they remain limited by the quality of generated views. Recently, Xue et al. [55] proposed a method that jointly trains diffusion and reconstruction models in an end-to-end manner, allowing for mutual enhancement.

Our method follows this direction but introduces key innovations: (1) a pose-conditioned multi-view joint diffusion model that synthesizes avatars in arbitrary poses to handle occlusions and facilitate animation; (2) a compositional 3DGS refinement strategy integrating global and local views via a crop-aware camera ray embedding, significantly enhancing avatar detail and coherence.

Concurrent works. Some of the latest research, IDOL [65] and LHM [34] also trying reconstruct high-resolution 3DGS avatars with large scale training data. While they develops feed-forward models for efficiency, we build our

model based on diffusion models to utilize the strong generative priors.

3. Approach

Problem Specification. As illustrated in Fig. 2, given a full-body input image $\mathbf{x}_I$ depicting a person, AdaHuman aims to build a 3D avatar that supports two key functionalities: (1) **Avatar Reconstruction**: Without requiring any additional inputs, AdaHuman reconstructs an avatar $\mathcal{G}_R$ represented by 3D Gaussian Splats (3DGS) [21] that precisely matches the pose of the input image, enabling high-fidelity novel view synthesis; (2) **Avatar Synthesis**: Using an estimated input pose P_s and an arbitrary target 3D pose P_t, AdaHuman generates a reposed 3DGS avatar $\mathcal{G}_{P_t}$ in the target pose while faithfully preserving the person's appearance and identity. This capability enables pose canonicalization, where we generate a standardized A-posed avatar $\mathcal{G}_A$ that minimizes self-occlusion. The canonicalized avatar can then be rigged automatically for animation and used to render temporally coherent 4D videos with high visual quality.

Our method consists of two key modules that enable the generation of detailed and animatable avatars: (1) Pose-Conditioned 3D Joint Diffusion (Sec. 3.1), which generates multiview images and the corresponding 3DGS avatar of the person in arbitrary poses by interleaving image synthesis and 3D reconstruction inside the diffusion process; (2) Compositional 3DGS Refinement (Sec. 3.2), which enhances the visual quality by first refining local body part renderings at high resolution and then seamlessly composing them into a holistic detailed avatar.

3.1. Pose-Conditioned 3D Joint Diffusion

As shown in Fig. 2, given a full-body input image $\mathbf{x}_I$, we first generate local view images of different body parts (e.g., head, upper body, and lower body). These local views, along with the input, form our input views $\mathcal{I}_{i=1}^V$, which are fed to the 3D joint diffusion module as in Fig. 2 (right). The module then synthesizes images of the target views $\mathcal{T}_{j=1}^K$ which look at the full-body and local body parts of the person from different viewpoints than the input. Combining both full-body and local perspectives enables our method to achieve detailed and globally consistent generation of multi-view images and their corresponding 3DGS avatar.

Each input view is represented by a tuple $\mathcal{I}_i = \{\mathbf{x}_i, \mathbf{p}_i, \mathbf{c}_i\}$, consisting of an RGB image $\mathbf{x}_i$, an *optional* pose condition $\mathbf{p}_i$, and camera parameters $\mathbf{c}_i$. The pose condition $\mathbf{p}_i$ takes the form of a 2D semantic pose map derived from the 3D input pose θ, created by rendering the semantic segmentation of the SMPL model [30] from the camera's perspective. The camera parameters $\mathbf{c}_i$ are encoded into a camera ray map using sinusoidal embeddings of the camera rays' origins and directions. Similarly, each target view is defined by $\mathcal{T}_j = \{\mathbf{x}_j^t, \mathbf{p}_j, \mathbf{c}_j\}$, where $\mathbf{x}_j^t$ represents the noisy target RGB image at diffusion step t, $\mathbf{p}_j$ is the optional pose condition, and $\mathbf{c}_j$ encodes the target view's camera parameters. The primary objective of our pose-conditioned 3D joint diffusion is to model the conditional denoising distribution of the target RGB images $\{\mathbf{x}_j^{t-1}\}_{j=1}^K$:

$$p(\{\mathbf{x}_j^{t-1}\}_{j=1}^K | \{\mathbf{p}_j, \mathbf{c}_j\}_{j=1}^K, \{\mathbf{x}_i, \mathbf{p}_i, \mathbf{c}_i\}_{i=1}^V, t)\,, \quad (1)$$

where we assume V input views and K target views. Inspired by recent work [10], we employ a multi-view image latent diffusion model (LDM) to model the denoising distribution. Specifically, we modify the U-Net architecture of a single-image LDM by replacing the 2D self-attention layers with 3D attention layers. The 2D pose semantic map $\mathbf{p}_i$ and camera ray map $\mathbf{c}_i$ are concatenated with the RGB images as additional conditions before being fed to the U-Net.

To enhance the 3D consistency of the generated multi-view images and produce the underlying 3DGS avatar, we incorporate a 3DGS generator $\mathbf{G}$ [45] into the denoising diffusion process. Following [55], at each denoising step t, we generate a 3DGS avatar $\mathcal{G}^t$ from the image predictions:

$$\mathcal{G}^t = \mathbf{G}(\{\mathbf{x}_j^{t\to 0}, \mathbf{x}_j^t, \mathbf{p}_j, \mathbf{c}_j\}_{j=1}^K, \{\mathbf{x}_i, \mathbf{p}_i, \mathbf{c}_i\}_{i=1}^V, t)\,, \quad (2)$$

where $\mathbf{x}_j^{t\to 0}$ represents the "clean" target image obtained through one-step denoising by the LDM at diffusion step t. Once $\mathcal{G}^t$ is obtained, we render it under the target views to generate new *3D-consistent* clean target images $\hat{\mathbf{x}}_j^{t\to 0}$. Using these 3D-consistent images, we then sample the noisy images $\mathbf{x}_j^{t-1}$ for the next diffusion step according to: $\mathbf{x}_j^{t-1} \sim q(\mathbf{x}_j^{t-1} | \mathbf{x}_j^t, \hat{\mathbf{x}}_j^{t\to 0})$, where q denotes the reverse diffusion process [43]. The final output of our pose-conditioned 3D joint diffusion model is the 3DGS avatar $\mathcal{G}^0$ produced at the end of the diffusion process.

Unlike previous works [10, 55], our approach incorporates pose conditioning to enable pose-conditioned multi-view image synthesis. This key enhancement empowers our model to not only reconstruct pixel-aligned 3DGS avatars but also generate reposed avatars that are well-suited for animation and other applications. This capability is particularly valuable since subjects in input images often exhibit severe self-occlusion, which makes rigging in the original body pose challenging and suboptimal. Through pose-conditioned multi-view image synthesis, our method can transition the avatar into a rigging-friendly pose while simultaneously recovering geometry and appearance details that were previously occluded.

View Selection and Model Training. During training, we first randomly select either the full body or a local body parts from upper body, lower body, or head. For the selected body part, we choose an input view from a training video frame. The key distinction between reconstruction and reposing lies in the selection of target views: for re-

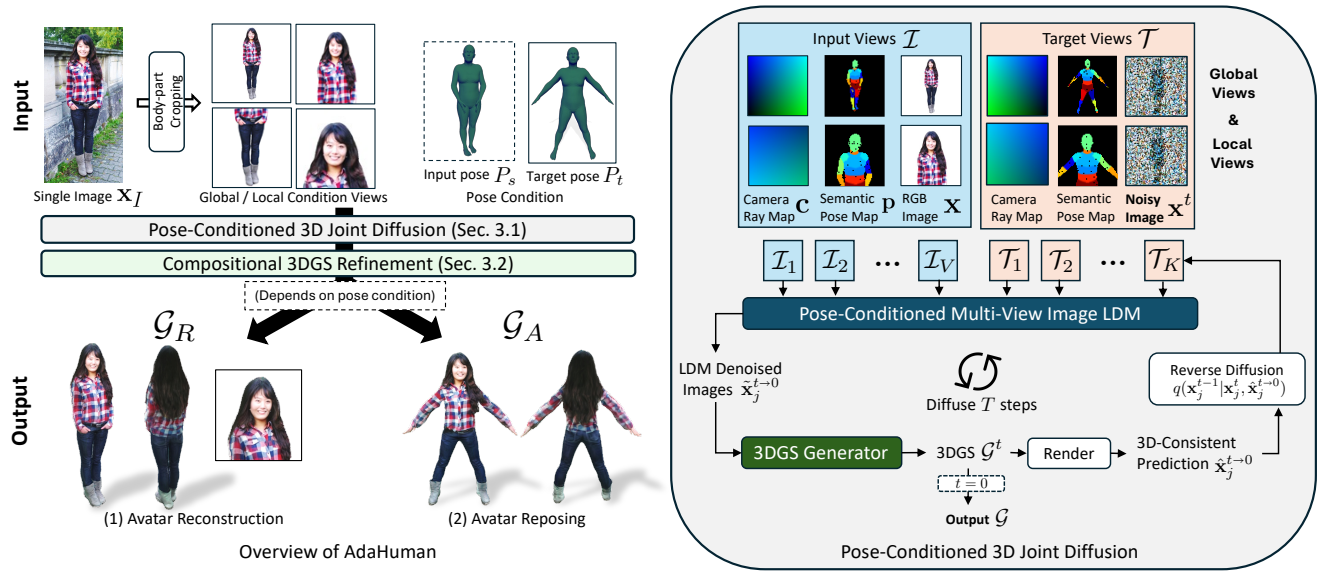

Figure 2. **Method Overview.** Left: Given an RGB image of an unseen person as input, AdaHuman could (1) reconstruct a high-fidelity pixel-aligned 3D Gaussian Splat (3DGS) avatar, as well as (2) generate an reposed 3DGS avatar with a target pose condition, enable building animatable avatar in a standard A-pose. Right: A pose-conditioned joint 3D diffusion process is utilized to generate global or local 3DGS reconstruction or reposing results. This process ensures 3D consistency of the reconstruction by utilizing generated 3DGS results in each reverse diffusion process of multi-view avatar images.

construction, we select three canonical target views (separated by 90° azimuth angles) of the body part from the *same* frame as the input view; for reposing, we select four canonical target views from a *different* frame showing the subject in a different pose, where the additional target view coincides with the input view to account for the pose difference.

We jointly train the multi-view image LDM and the 3DGS generator $\mathbf{G}$ using multi-view video data from MVHumanNet [50] and image renderings from CustomHuman [12]. To leverage powerful generative priors learned from large-scale datasets, both models are initialized from official pretrained weights [37, 45]. We first train the model for avatar reconstruction for 30k steps and then fine-tune the model for reposing for 10k steps. Camera ray embeddings are computed relative to the input view. The LDM is supervised using MSE loss between predicted and ground truth image latents, while the 3DGS generator $\mathbf{G}$ is supervised following [55] using MSE, LPIPS rendering losses, and surface regularization loss. In addition to the target views, we sample 12 additional views to provide dense supervision to the 3DGS generator. Additional implementation details are provided in the appedix.

3.2. Compositional 3DGS Refinement

Recent feed-forward 3D reconstruction models [14, 45] have demonstrated promising results in generating 3D models of general objects from sparse-view images. However, these approaches are constrained by their networks' fixed output resolution (e.g., 256×256 3D Gaussians in LGM [45]), limiting their ability to capture the fine-grained details essential for realistic human avatar reconstructions. To address this limitation, we introduce a new compositional 3DGS refinement module, as illustrated in Fig. 3. The module leverages an image-to-image local body refinement scheme as well as a novel crop-aware camera ray map to enable detailed and coherent reconstructions of individual local body parts. During inference, it takes the coarse 3DGS avatar $\mathcal{G}_{\text{coarse}}$ from the 3D joint diffusion module as input and refines it to produce a detailed 3DGS avatar $\mathcal{G}_{\text{refined}}$.

Local body part refinement. To achieve enhanced details for local body parts, we begin by rendering N_v=4 90-degree separated canonical views (front, left, back, and right) for each of N_b=3 local body parts (head, upper body, and lower body) of the coarse avatar $\mathcal{G}_{\text{coarse}}$. Each local view is produced using a crop-view camera that zooms into the local body region inside the original global view (by manipulating the camera intrinsics). This zoom-in region is computed using the 2D body joints and segmentation masks. We then employ our multi-view LDM introduced in Section 3.1 to refine the local renderings via an image-to-image editing process similar to SDEdit [31], significantly enhancing their detail. This approach enables the high-fidelity generation of local body parts. To properly handle the modified camera perspective for these local views, we provide the LDM with a specialized cropped version of the camera ray map, which we detail in the following section.

Crop-aware local ray map. A key challenge in the refinement process is effectively combining the $N_v \times N_b$ refined local view images and N_v global full-body view images

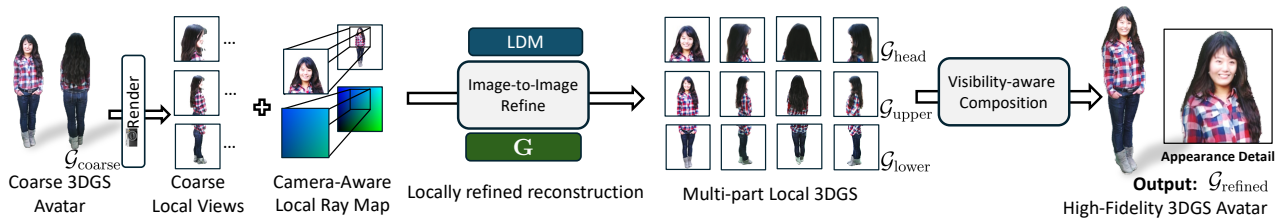

Figure 3. **Compositional 3DGS Refinement.** Given the coarse 3DGS reconstruction $\mathcal{G}_{\text{coarse}}$ as input, we render initial coarse views, and refine them with image-to-image editing for enhancing local 3DGS $\mathcal{G}_{\text{upper}}, \mathcal{G}_{\text{lower}}, \mathcal{G}_{\text{head}}$. Finally, a refined holistic 3DGS avatar $\mathcal{G}_{\text{refined}}$ is generated from these results by our proposed visibility-aware 3DGS Composition.

into a holistic 3DGS avatar. The 3DGS generator in [45] uses four fixed canonical camera views as inputs to generate a global 3DGS in unit space, but this fixed camera setup does not naturally accommodate additional local views.

To address this challenge, we propose a simple yet effective solution: a crop-aware local ray map that establishes correspondences between the 3D coordinates of local and global views. This approach extends [45] by incorporating additional local views as inputs, enabling high-resolution reconstruction of local body parts with fine details. Specifically, for a pixel at coordinates (u, v) in a local view image of size (H, W), where the local view is obtained by cropping a box region $(x_{tl}, y_{tl}, x_{br}, y_{br})$ from the global view, we map its coordinates back to the global view using:

$$(i, j) = \left(x_{tl} + \frac{(x_{br} - x_{tl}) \cdot u}{W}, y_{tl} + \frac{(y_{br} - y_{tl}) \cdot v}{H} \right). \tag{3}$$

Using these mapped coordinates, we compute the camera ray embedding for the local view pixel using the 3DGS generator's global camera ray map equation:

$$\mathcal{R}(i, j) = (\mathbf{o}(i, j), \mathbf{o}(i, j) \times \mathbf{d}(i, j)) \tag{4}$$

where $\mathbf{o}$ and $\mathbf{d}$ represent the origin and direction of the camera rays based on the camera extrinsics. The crop-aware local ray map is utilized during both training and inference to help the 3DGS generator establish correspondences between the 3D locations in local and global views. Using the crop-aware local ray map, we can directly use the 3DGS generator $\mathbf{G}$ to map refined local views to 3DGS in the global avatar space. In the following, we will describe a strategy to combine the 3DGS produced by the local and global views into a holistic 3DGS avatar $\mathcal{G}_{\text{refined}}$.

Visibility-aware 3DGS Composition. As we will show in Fig. 10, naively combining these partial 3DGS leads to floating artifacts and degraded appearance details. To address this challenge, we introduce a visibility-aware 3DGS composition scheme that intelligently merges the parts into a coherent, high-quality avatar. Our approach employs two key criteria to determine which 3D Gaussians to preserve during composition: (1) *View Coverage* quantifies how many input views capture each 3D Gaussian point within their field of view, and (2) *Visibility Salience* measures the gradient magnitude of the alpha channel across all rendered input views. Intuitively, Gaussians with low view coverage lack multi-view consensus and are likely unreliable, while those with low visibility salience contribute minimally to the final appearance and likely represent noise. Specifically, given globally or locally reconstructed body part 3DGS $\mathcal{G}_{\text{p}}$ and the canonical views for each body part $\mathcal{I}_{\text{p}}^{j}$, where $p \in \{\text{full}, \text{upper}, \text{lower}, \text{head}\}$ and $j = 0 \ldots 3$, we evaluate each splat $\mathcal{G}_{\text{p}}^{i}$ as follows:

First, we calculate the number of covered input views of the splat in different local parts $n_c(\mathcal{G}_{\text{p}_1}^{i}, \mathcal{I}_{\text{p}_2})$. A splat is considered reliable if it is covered by more than 2 input views in its own body part (or 3 views if it is generatd by the head part. If the splat is also well-covered by input views of another more detailed body part (e.g., head is more detailed than upper-body), it is deemed redundant and removed.

Second, we assess visibility salience using rendering gradients. If a splat has higher visibility in the input views of another body parts with similar level of detail (e.g., between upper and lower body), it is likely redundant and should be dropped to avoid conflicts or redundancy.

This approach ensures efficient composition while maintaining visual fidelity, focusing on the most reliable and visually significant splats.

4. Experiments

In order to comprehensively evaluate the performance of AdaHuman, we conduct experiments on avatar reconstruction and avatar reposing tasks, comparing our method with state-of-the-art (SOTA) approaches both quantitatively and qualitatively. Additionally, we perform a user study to assess the perceptual quality of the generated avatars.

Datasets. Unlike most existing 3D avatar reconstruction methods that rely on 3D human mesh data for training, AdaHuman leverages multi-camera video data from MVHumanNet [50], which captures 3D appearances of humans in real-world settings and diverse poses. We sample 6,209 unique subjects for training and 50 unseen subjects for evaluating the novel pose synthesis task. Additionally, we mixed the training data with multiview images rendered from 589 human meshes in the CustomHumans [12] dataset

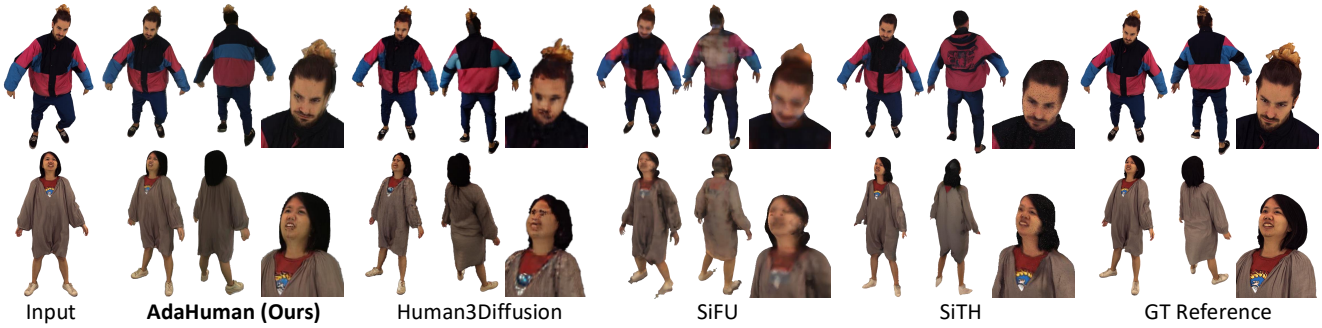

Figure 4. **Qualitative comparison on vatar reconstruction task on CustomHumans[12] dataset.**

Figure 5. **Comparison on in-the-wild images.** AdaHuman generalizes well to images with diverse appearances, body shapes, and clothing styles, while SIFU[63] and SiTH[13] fail on loose and complex clothing, and Human3Diffusion [55] fail to preserve appearance details. Coarse 3DGS is an ablation variant of AdaHuman without compositional 3DGS refinement, which fails to capture fine avatar details.

with a more diverse camera distribution to improve generalizability. 50 testing subjects from the CustomHumans [12] dataset and 97 subjects from Sizer [46] dataset are used to quantitatively compare our method against SOTA approaches. To further assess visual quality, we use 53 in-the-wild human images from the SHHQ [9] dataset to conduct a user study on perceptual quality.

Runtime. Our whole pipeline takes around 70s for inference on a NVIDIA A100 GPU.

4.1. Avatar Reconstruction

For novel view synthesis, we compare AdaHuman with SOTA mesh reconstruction methods (SiTH [13] and SIFU [63]) and 3DGS-based methods (LGM [45] and Human3Diffusion [55]) on the CustomHumans dataset [12]. For each test subject, we use a frontal camera view as the input image and render 20 novel views (1024×1024) by rotating around the body. We follow [55] to extract mesh from 3DGS results, and evaluate 3D reconstruction quality using Chamfer Distance (CD), Normal Consistency (NC) and F1 score. We evaluate rendering quality using PSNR, SSIM, and LPIPS scores for all novel views and the frontal view. FID scores are assessed to measure the perceptual quality of the avatars. We provide qualitative and quantitative comparisons of AdaHuman against SOTA methods in Fig. 4 and Tab. 1. Our method generates significantly higher-quality avatars, with a better performance on all of image quality metrics, while keeping a comparable performance in the 3D reconstruction metrics.

4.2. Perceptual Study

To fully evaluate the perceptual quality and generalizability of our method, we conducted a user study on 53 in-the-wild images from the SHHQ [9] dataset. We compared AdaHuman with SiTH [13], SIFU [63], Human3Diffusion [55], and an ablation of AdaHuman using the coarse 3DGS avatar without compositional 3DGS refinement. Each survey con-

Model	PSNR↑		SSIM↑		LPIPS↓		FID ↓		CD(cm)↓		F-score↑		Normal↑	
	CH	Sizer	CH	Sizer	CH	Sizer	CH	Sizer	CH	Sizer	CH	Sizer	CH	Sizer
LGM [45]	18.99	17.58	0.8445	0.8909	0.1664	0.1188	122.3	124.20	2.175	1.832	0.3941	0.4897	0.6431	0.6451
SiTH [13]	20.77	20.67	0.8727	0.9219	0.1277	0.0883	42.9	37.11	1.389	1.229	0.4701	0.5688	**0.7978**	**0.7915**
SIFU [63] †	20.59	20.56	0.8853	0.9196	0.1359	0.0987	92.6	101.79	2.009	1.560	0.3438	0.4787	0.7539	0.7768
Human3Diffusion [55]	21.08	19.50	0.8728	0.9211	0.1364	0.0953	35.3	20.69	1.230	1.174	0.5324	**0.6336**	0.7338	0.7389
AdaHuman (Ours)	**21.46**	**21.42**	**0.8925**	**0.9258**	**0.1087**	**0.0856**	**27.3**	**19.15**	**0.962**	**1.135**	**0.6083**	0.6075	0.7597	0.7477

Table 1. **Quantitative comparison on avatar reconstruction task.** On CustomHumans (CH) [12] and Sizer [46] datasets, AdaHuman surpasses all baselines on rendering quality metrics (PSNR, SSIM, LPIPS and FID), and also achieves best the shape reconstruction metrics (CD, F-score), except getting slightly lower F-score with Human3Diffusion [55] on the Sizer dataset. However, since we borrow the same normal estimation method from [55], AdaHuman got similar performance on Normal Consistency. The **best** scores are highlighted. †: not using SIFU's text-guided texture refinement since prompts are unavailable.

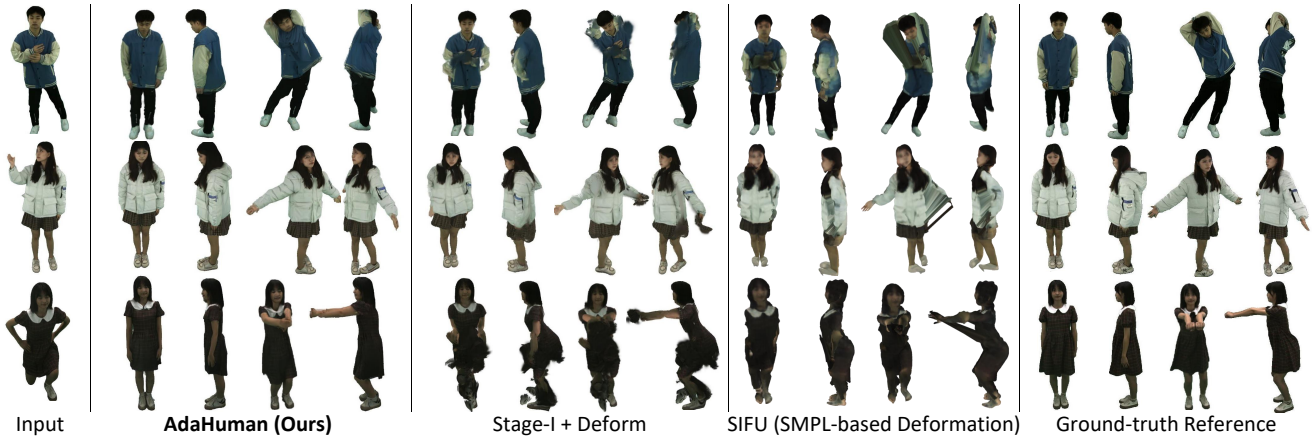

Figure 6. **Qualitative comparison on novel pose synthesis task.**

Baseline methods	SiTH[13]	SIFU[63]	H3D[55]	Coarse 3DGS
Preference of AdaHuman (%)	88.3	99.2	79.7	93.8

Table 2. **User preference of AdaHuman.** Our method achieves substantially higher preference against all baseline methods.

sisted of 40 pairs of generated avatars, and 28 participants were asked to select the avatar with better overall quality.

As shown in Tab. 2, AdaHuman was preferred by a significant margin over other methods. Fig. 5 demonstrates that SIFU [63] and SiTH [13] often produce lower texture quality for side views and struggle to recover accurate geometry, likely due to the limitations of template-based mesh reconstruction. Our method generates avatars with substantially higher quality and generalizes well across diverse appearances, clothing styles, and body poses. Compared to Human3Diffusion [55], which fails to capture fine appearance details, our method recovers significantly better details thanks to our local refinement approach. More results on in-the-wild images are provided on the website.

4.3. Avatar Reposing and Animation

Avatar Reposing. For avatar reposing evaluation, we sample one input pose $\mathbf{p}_{\text{in}}$ and six target novel poses $\mathbf{p}_{\text{target}}$ from the video sequence for each unseen subject in the MVHumanNet dataset. Our method takes a single input image $\mathbf{x}_{\text{in}}$ and pose conditions $\mathbf{p}_{\text{in}}, \mathbf{p}_{\text{target}}$ as inputs, directly synthesizes the avatar in the target poses using the Pose-Conditioned Joint 3D Diffusion. We compare our approach with SOTA mesh-based methods SiTH [13] and SIFU [63] using the same inputs, which repose characters into target poses using linear blend skinning and the SMPL-X body model. As an additional baseline, we also evaluate results from directly deforming the input pose reconstructed 3DGS avatar into target poses by SMPL blending weights. In particular, other 3DGS-based methods, such as LGM [45] and Human3Diffusion [55] are excluded from this evaluation because they do not have aligned body models to support reposing of their reconstructed avatars. As shown in Tab. 3, AdaHuman significantly outperforms competing methods across all metrics. Fig. 6 illustrates that our pose-conditioned model generalizes effectively to challenging input and target poses, benefiting from the diverse motions present in multi-view video datasets. Notably, AdaHuman excels at synthesizing realistic cloth deformations in target poses, while other methods struggle due to limitations of

Method	PSNR↑	SSIM↑	LPIPS ↓
SiTH	21.21	0.8742	0.1261
SIFU	21.27	0.8722	0.1244
AdaHuman $\mathcal{G}_R$+ deform	23.01	0.8825	0.1100
AdaHuman $\mathcal{G}_{P_t}$ (Ours)	**24.64**	**0.9046**	**0.0863**

Table 3. **Comparison on novel pose synthesis task.** Our model achieves the best rendering similarity (PSNR, SSIM, LPIPS), showcasing the ability of our pose-conditioned model to generalize to diverse input and target poses.

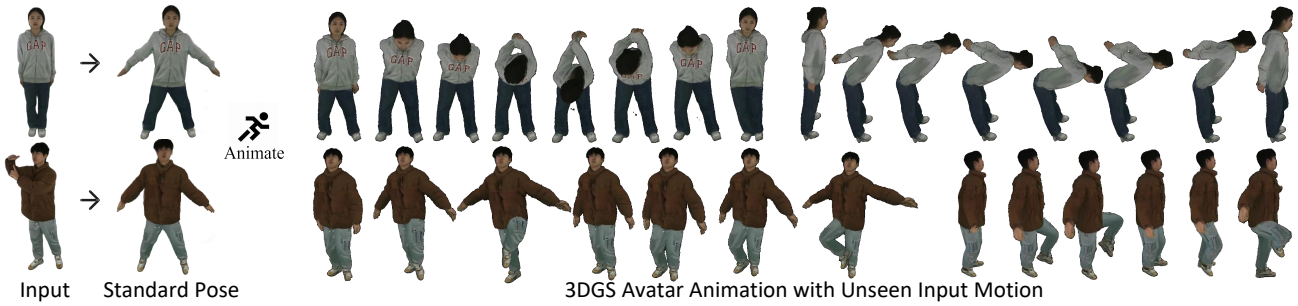

Figure 7. AdaHuman generates animation-ready avatar in a standard pose, which can be animated with unseen input motion.

Exp.	JointDiff	$\mathcal{G}_{\text{refined}}$	Body parts	$\mathbf{p}_{\text{gt}}$	PSNR ↑	SSIM ↑	LPIPS ↓	FID↓
Coarse 3DGS $\mathcal{G}_{\text{coarse}}$	✓	✗	F	✗	20.84	0.8789	0.1296	31.9
Direct Composition	✓	✓	U,L,H	✗	20.41	0.8700	0.1350	36.2
Learnable Composition	✓	✓	F,U,L,H	✗	20.87	0.8788	0.1270	28.0
No Joint Diffusion	✗	✓	F,U,L,H	✗	20.79	0.8762	0.1283	27.6
Additional body part	✓	✓	F,U,M,L,H	✗	21.43	0.8922	0.1104	27.6
Ours	✓	✓	F,U,L,H	✗	21.46	0.8925	0.1087	27.3
Ours + GT Pose Condition	✓	✓	F,U,L,H	✓	**23.00**	**0.9028**	**0.1086**	**27.0**

Table 4. **Ablation study.** Without ground-truth pose, our full method achieves the best scores compared to the ablation baselines, showcasing the effectiveness of joint diffusion (JointDiff), compositional 3DGS with local refinement ($\mathcal{G}_{\text{refined}}$), and the selection of body parts (F: fullbody, U: upper, L: lower, H: head, M: middle). Using ground-truth pose ($\mathbf{p}_{\text{gt}}$) with our pose-conditioned model can further improve the alignment and provide better results.

SMPL-based deformation and the fixed topology of mesh-based reconstruction methods.

Additionally, in Fig. 8, we show results of reposing SHHQ[9] characters with complex loose clothing to standard poses. Our reposing model successfully generalize to these OOD garments with realistic deformation effects.

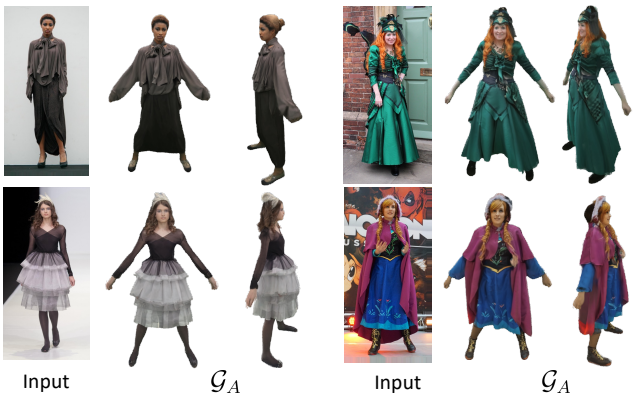

Figure 8. **Reposing avatars with challenging garments.**

Avatar Animation. Fig. 7 showcases the animation results of AdaHuman using the animatable avatar from Avatar Reposing with a standard pose condition. Although the model is not directly trained with standard pose data, it learns to generalize to the standard poses with the help of the diverse distribution of poses in MVHumanNet[50].

Avatar Reposing vs. LBS-based Animation As AdaHuman supporting two modes to synthesize novel pose avatars, Fig. 9 compares the performance of these two modes. Here we analysis by comparing their pros and cons.

Mode 1: Direct Avatar Reposing - This mode directly generates reposed Gaussians for a target pose. Pros: (1) Captures pose-dependent effects for non-rigid clothing, (2) More realistic deformation of loose clothing, (3) No need for rigging. Cons: More computationally expensive and less temporal coherent.

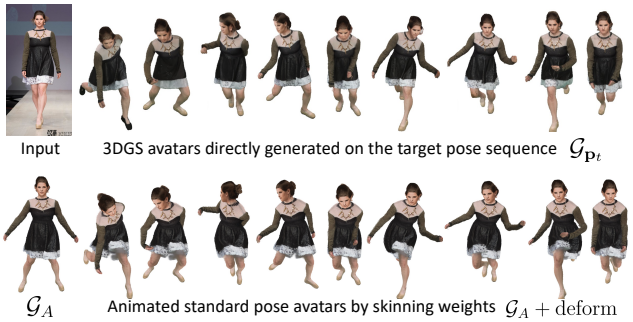

Figure 9. **Comparison of direct avatar reposing and standard posed avatar with skinning weight animation.**

Mode 2: SMPL-based LBS Animation - This mode first reconstructs a standard pose avatar, then applies SMPL-based skinning weights for motion deformation. Pros: (1) Enables real-time rendering, (2) Better temporal consistency. Cons: Limited loose clothing deformation.

4.4. Ablation Study

To evaluate the effectiveness of our design choices, we conduct various ablation studies on avatar reconstruction using the CustomHumans dataset. Tab. 4 and Fig. 10 compare variants of our method, focusing on rendering quality.

Coarse 3DGS $\mathcal{G}_{\text{coarse}}$ uses only the generated coarse avatar without refinement, failing to capture fine details, particularly in facial regions. Our full method achieves bet-

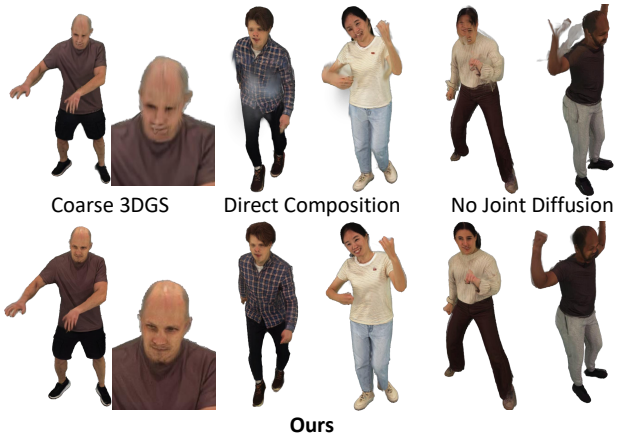

Figure 10. **Comparison of our method and ablation variants.**

ter FID while maintaining comparable PSNR, SSIM, and LPIPS scores, demonstrating that compositional refinement improves details without sacrificing accuracy.

Composition Strategy. We compare our visibility-aware approach with: (1) **Direct Composition**, which ensembles all local 3DGS without filtering unreliable splats, yet this variant results in significant artifacts; (2) **Learnable Composition**, which uses a network with self-attention between parts to predict the holistic avatar. Despite showing slight improvement, this variant still encounters artifacts and requires more computation. This demonstrates the importance and effectiveness of our visibility-aware 3DGS composition.

Body Part Selection. To evaluate the design of body part selection, we compare with variants that use an additional body part in the middle of the body for local refinement and 3D composition. This comparison demonstrates that using 4 parts (fullbody, upper, lower and head) is a good balance between performance and efficiency.

No Joint Diffusion is a variant that applies the 3DGS generator only to multiview images from the last diffusion step. Results show that it leads to view inconsistencies and performance drops, confirming the importance of 3D joint diffusion for consistent avatar generation.

GT Pose Condition shows that using ground-truth SMPL annotations significantly improves reconstruction quality through better pose alignment, indicating potential for further improvement.

5. Discussion and Limitations

In this paper, we introduced AdaHuman, a novel framework for generating highly-detailed and animatable 3DGS avatars from a single input image. Our approach integrates 3DGS reconstruction within the multi-view diffusion process, ensuring 3D-consistent generation of multiview images as well as 3DGS avatars in both input and novel poses. Furthermore, our visibility-aware compositional 3DGS refinement module significantly enhances the appearance details of the avatars and seemlessly integrates local and global body parts into a coherent 3DGS avatar. Extensive experiments on public benchmarks and in-the-wild images showed that AdaHuman substantially outperforms state-of-the-art methods in both novel view synthesis and novel pose synthesis tasks.

Despite these advancements, some limitations of our method warrant further exploration. The local refinement strategy may encounter difficulties with occluded or poorly covered regions, particularly around hands and arms, leading to artifacts and limiting fine-grained animation in these areas. Additionally, while our model can generate avatars in an animation-friendly standard pose, the animation capability still relies on the alignment of the SMPL body models and their skinning weights, which poses challenges in detailed animation such as facial expressions, hand gestures, and garment deformation. Future work could explore better integration of body models and simulation-based methods, as well as the use of video diffusion model to enhance the animation quality.

References

[1] Easymocap - make human motion capture easier. Github, 2021. 12

[2] 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. 2

[3] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In CVPR, 2022. 2

[4] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. DreamAvatar: Text-and-Shape Guided 3D Human Avatar Generation via Diffusion Models. In CVPR, 2024. 2

[5] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In CVPR, 2022. 2

[6] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In ICCV, 2023. 2

[7] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In ECCV, pages 20–40, 2020. 2

[8] 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. In IEEE Conf. Comput. Vis. Pattern Recog., pages 13142–13153, 2023. 13

[9] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen-Change Loy, Wayne Wu, and Ziwei Liu.

StyleGAN-Human: A data-centric odyssey of human generation. In ECCV, 2022. 6, 8

[10] 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, 12

[11] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In CVPR, 2021. 2

[12] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In CVPR, 2023. 4, 5, 6, 7, 13

[13] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In CVPR, 2024. 2, 6, 7, 13

[14] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3d. In ICLR, 2024. 2, 4

[15] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In SIGGRAPH. Association for Computing Machinery, 2024. 12

[16] Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, and Ying Feng. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. In CVPR, 2024. 1, 2

[17] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. Dreamwaltz: Make a scene with complex 3d animatable avatars. In NeurIPS, 2023. 2

[18] 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. 1, 2

[19] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable reconstruction of clothed humans. In CVPR, 2020. 2

[20] Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control. In ICCV, 2023. 2

[21] 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. 2, 3

[22] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. In NeurIPS, 2023. 2

[23] Nikos Kolotouros, Thiemo Alldieck, Enric Corona, Eduard Gabriel Bazavan, and Cristian Sminchisescu. Instant 3d human avatar generation using image diffusion models. In ECCV, 2024. 2

[24] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J Black. TADA! text to animatable digital avatars. In 3DV, 2024. 2

[25] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3D: High-Resolution Text-to-3D Content Creation. In CVPR, 2023. 2

[26] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In NeurIPS, 2023. 2

[27] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In CVPR, 2024. 2

[28] 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 CVPR, 2023. 2

[29] 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. 2

[30] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. SIGGRAPH Asia, 2015. 2, 3

[31] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In ICLR, 2022. 4, 13

[32] Hao-Yang Peng, Jia-Peng Zhang, Meng-Hao Guo, Yan-Pei Cao, and Shi-Min Hu. Charactergen: Efficient 3d character generation from single images with multi-view pose canonicalization. ACRM Trans. Graph., 43(4):1–13, 2024. 2

[33] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3d using 2d diffusion. In ICLR, 2022. 1, 2

[34] Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, and Liefeng Bo. Lhm: Large animatable human reconstruction model from a single image in seconds. In arXiv preprint arXiv:2503.10625, 2025. 2

[35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022. 2

[36] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. SIGGRAPH, 2023. 2

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, 2022. 2, 4, 12

[38] 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. 2

[39] Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In CVPR, 2020. 2

[40] Akash Sengupta, Thiemo Alldieck, Nikos Kolotouros, Enric Corona, Andrei Zanfir, and Cristian Sminchisescu. DiffHuman: Probabilistic Photorealistic 3D Reconstruction of Humans. In CVPR, 2024. 2

[41] 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. 2

[42] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3d generation. In ICLR, 2023. 2

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In ICLR, 2021. 3

[44] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. DreamGaussian: Generative gaussian splatting for efficient 3d content creation. In ICLR, 2024. 2

[45] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. LGM: Large multi-view gaussian model for high-resolution 3d content creation. In ECCV, 2025. 2, 3, 4, 5, 6, 7, 12, 13

[46] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In ECCV, pages 1–18. Springer, 2020. 6, 7

[47] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, , Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. arXiv preprint arXiv:2403.02151, 2024. 2

[48] A Vaswani. Attention is all you need. Adv. Neural Inform. Process. Syst., 2017. 12

[49] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. ProlificDreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In NeurIPS, 2023. 2

[50] Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, et al. MVHumanNet: A large-scale dataset of multi-view daily dressing human captures. In CVPR, 2024. 4, 5, 8, 12

[51] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit clothed humans obtained from normals. In CVPR, 2022. 2

[52] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit clothed humans optimized via normal integration. In CVPR, 2023. 2

[53] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. arXiv preprint arXiv:2404.07191, 2024. 2

[54] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. DMV3D: Denoising multi-view diffusion using 3d large reconstruction model. In ICLR, 2024. 2

[55] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard Pons-Moll. Human 3diffusion: Realistic avatar creation via explicit 3d consistent diffusion models. In NeurIPS, 2024. 1, 2, 3, 4, 6, 7, 12, 13

[56] Ze Yang, Shenlong Wang, Sivabalan Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, and Raquel Urtasun. S3: Neural shape, skeleton, and skinning fields for 3d human modeling. In CVPR, 2021. 2

[57] Xu Yinghao, Shi Zifan, Yifan Wang, Chen Hansheng, Yang Ceyuan, Peng Sida, Shen Yujun, and Wetzstein Gordon. GRM: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In ECCV, 2024. 2

[58] 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 IEEE Conf. Comput. Vis. Pattern Recog., 2021. 13

[59] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. ACRM Trans. Graph., 2024. 12

[60] Ye Yuan, Xueting Li, Yangyi Huang, Shalini De Mello, Koki Nagano, Jan Kautz, and Umar Iqbal. GAvatar: Animatable 3d gaussian avatars with implicit mesh learning. In CVPR, 2024. 2

[61] Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Kang Du, and Min Zheng. Avatarverse: High-quality & stable 3d avatar creation from text and pose. In AAAI, 2024.

[62] Xuanmeng Zhang, Jianfeng Zhang, Chacko Rohan, Hongyi Xu, Guoxian Song, Yi Yang, and Jiashi Feng. Getavatar: Generative textured meshes for animatable human avatars. In ICCV, 2023. 2

[63] Zechuan Zhang, Zongxin Yang, and Yi Yang. SIFU: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In CVPR, 2024. 2, 6, 7, 13

[64] Tiancheng Zhi, Christoph Lassner, Tony Tung, Carsten Stoll, Srinivasa G Narasimhan, and Minh Vo. Texmesh: Reconstructing detailed human texture and geometry from rgb-d video. In ECCV, 2020. 2

[65] Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3d human creation from a single image, 2024. 2

[66] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In CVPR, 2024. 2

A. Implementation Details

Network Structure. In Fig. 11, we illustrate the architecture of our Pose-Conditioned Multi-View Image LDM model, along with the 3DGS generators $\mathbf{G}$ and $\mathbf{G}_{\text{comp}}$. For the LDM model, following [10], we enable 3D cross-view attention only in layers with a feature map resolution of $\leq 32 \times 32$. We also add extra input channels to the latent maps for camera ray maps, condition masks, and semantic pose maps. For $\mathbf{G}$, we adopt the architecture of the pre-trained LGM-big model [45] and include additional input channels for noisy images $\mathbf{x}_t$.

Additional, as an ablation mentioned at Tab. 4, we hvae tried training a compositional 3DGS generator $\mathbf{G}_{\text{comp}}$ for Learnable Composition. Based on the LGM network, we insert an additional cross-part self-attention layer after each original cross-view self-attention layer in the LGM network. Note that the output image resolution of our LDM model is 512×512, which is then downsampled to 256×256, the input resolution for the 3DGS generator $\mathbf{G}$.

Ray Map Embedding. We use different methods to embed ray map information for the image LDM model and the 3DGS generators $\mathbf{G}$ and $\mathbf{G}_{\text{comp}}$. For the 3DGS generators, to effectively utilize the pretrained weights of LGM, we scale the entire scene to ensure a camera distance of $r = 1.5$ meters and use Plücker ray embeddings as described in Eq. 4 of the main text.

For the LDM model, we employ sinusoidal positional embeddings [48] to encode ray origins and directions, providing rich information about 3D locations across different cropping scales:

$$\mathcal{R}_{\text{LDM}}(i,j) = \text{PE}(\mathbf{o}(i,j), \mathbf{d}(i,j)) \tag{5}$$

where PE is the sinusoidal positional encoding function, with the number of octaves N_{octaves} set to 8.

View Sampling. Since our training data consists of multi-camera video captures in a 3D scene, the avatar is not always positioned at a standard location. We use 2D joint locations and foreground mask areas to crop global and local training views, resizing them to a resolution of 512×512. In Tab. 5, we list the OpenPose joints used to determine the cropping centers and relative size ratios of the local crops. During inference, after obtaining coarse reconstruction results with global views, we render $N_v = 20$ views to estimate 3D joints using EasyMocap [1], which helps sample local views for our compositional 3DGS refinement.

Parts	Full body	Upper Body	Lower Body	Head
Joints	Pelvis	Neck	Left Ankle, Right Ankle	Left Ear, Right Ear
Scale	1.0	0.5	0.5	0.25

Table 5. Body part sampling details.

Training Schedule. We initialize our LDM model with the official weights of `stable-diffusion-v1-5`[1] [37] and our 3DGS generator $\mathbf{G}$ with LGM-big[2] [45].

For training the LDM model weights θ, the model first learns to predict $K = 3$ canonical views from one input view ($V = 1$) without pose conditioning. We fine-tune the model on predicting global full-body views for $20,000$ iterations, followed by fine-tuning on all $N_p + 1 = 4$ global and local view for another $30,000$ iterations to obtain $\theta_{\text{no\_pose}}$. Finally, we fine-tune the pose-conditioned model weights $\theta_{\text{novel\_pose}}$ from $\theta_{\text{no\_pose}}$. This model learns to predict $K = 4$ canonical views of a novel pose avatar from $V = 1$ input views sampled from different frames in the same video sequence. The novel pose synthesis model is fine-tuned for $1,0000$ iterations using all $N_p + 1 = 4$ global and local views.

For training the 3DGS generator model $\mathbf{G}$, we first fine-tune it from pre-trained weights using clean full-body images in MVHumanNet [50] for $2,000$ iterations to adapt it for human reconstruction. Then, we randomly sample diffusion timesteps to train with both noisy inputs $\mathbf{x}_t$ and clean inputs $\mathbf{x}_0$ for $20,000$ iterations. The 3DGS model $\mathbf{G}$ is also fine-tuned on local views for an additional $20,000$ iterations. We use $N_{\text{ref}} = 12$ reference views of each part to supervise the predicted 3DGS.

All training processes are conducted on 16 NVIDIA A100 80GB GPUs, with a total batch size of $n_{\text{batch}} = 128$ and a learning rate of $\eta = 5 \times 10^{-5}$.

Training Losses. The training losses for the pose-conditioned LDM and the 3DGS generator are as follows:

$$\mathcal{L}_{\text{LDM}} = \mathcal{L}_{\text{MSE}}(\epsilon, \epsilon_\theta) \tag{6}$$

$$\mathcal{L}_{\mathbf{G}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} \tag{7}$$

$$\begin{aligned}\mathcal{L}_{\text{recon}} = {} & \lambda_{\text{MSE}}\mathcal{L}_{\text{MSE}}(\hat{\mathbf{x}}^{t\to 0}_{\text{novel}}, \mathbf{x}_{\text{novel}}) \\ & + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}}(\hat{\mathbf{x}}^{t\to 0}_{\text{novel}}, \mathbf{x}_{\text{novel}})\end{aligned} \tag{8}$$

where the training loss of LDM, denoted as $\mathcal{L}_{\text{LDM}}$, is the MSE loss of the predicted latent noise. The training loss of $\mathbf{G}$ consists of rendering reconstruction loss computed using MSE and LPIPS. Following [55], we also incorporate the 3DGS regularization loss from [15, 59] to enhance surface quality.

Inference. This section details the inference pipeline of avatar reconstruction and avatar reposing our method. In both settings, we perform 3D joint diffusion on global views only when $t \in (500, 900]$ to maintain the stability of the diffusion process. The earlier steps focus on pure

[1] https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5

[2] https://huggingface.co/ashawkey/LGM/resolve/main/model_fp16_fixrot.safetensors

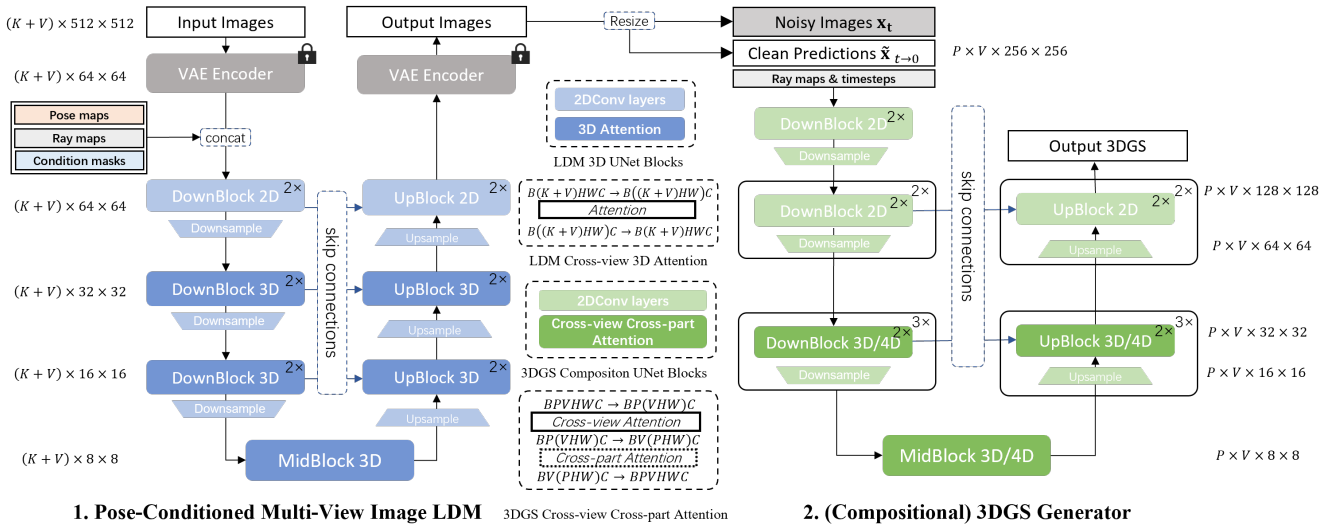

Figure 11. **Network Architectures** of (1) Pose-Conditioned Multi-View LDM Model and (2) Compositional 3DGS Generator.

2D diffusion to generate more detailed appearances. During image-to-image local refinement, we utilize SDEdit [31] with a strength of $s = 0.5$, meaning that denoising begins at $t = 500$ and 3D joint diffusion is performed when $t \in (350, 500]$.

B. Evaluation Settings

Baseline Models. Our baseline methods, including Human3Diffusion [55], LGM [45], SiTH [13], and SIFU [63], have been trained on various 3D mesh datasets [8, 12, 58]. In this work, our aim is to demonstrate the advantages of training models on both mesh datasets and video datasets for better pose generalization and the synthesis of novel pose characters. We utilize their official weights for comparison. We also note that some models (e.g. [55]) rely on private data or synthesized meshes for training.

Avatar Reconstruction. We selected front views of the mesh avatar as input views, rendered by horizontal perspective cameras for a fair and realistic comparison. The results of the quantitative evaluation are rendered at a resolution of 1024×1024 using 20 perspective cameras.

Avatar Reposing. For SiTH [13] and SIFU [63], we deform their avatars to the target pose and align the avatar meshes with the ground-truth SMPL meshes to render images for evaluation.