Avatar++: Fast and Pose-Controllable 3D Human Avatar Generation from a Single Image

Seonghee Han[1,#] Minchang Chung[1,#] Gyeongsu Cho[1] Kyungdon Joo[1,*] Taehwan Kim[1,*]
[1]Ulsan National Institute of Science and Technology (UNIST), Republic of Korea
{seonghee, minchang, threedv, kyungdon, taehwankim}@unist.ac.kr [*†]

Abstract

We introduce Avatar++ as an optimization-free pipeline that converts a single frontal photograph into a 3D representation in a single forward pass, taking less than 15 seconds. Generating a human avatar from a single image is challenging due to the complex structure of the human body and the intricacies of facial features, and most existing models employ Score Distillation Sampling [9, 42] or iterative refinement [6] methods to progressively enhance the generated textures. However, they have limitations of relying on computationally expensive and time-consuming optimization steps. To address these challenges, we propose a novel approach, named Avatar++*, that generates a human avatar through a fast and efficient single forward pass. Our model uses two different types of embeddings, one is facial identity and the other one is visual embedding. By combining two embeddings, our multi-view Diffusion Transformer (DiT) generates viewpoint-aligned images that preserve the subject's facial identity. Additionally, we introduce an attention mechanism that propagates information from the input image during sampling to enhance visual quality. We additionally give guidance on the pose. This pose guidance allows the model to generate either a canonical pose (e.g., T-pose or A-pose) or replicate the pose from the input image using OpenPose [2]. In addition to offering control over the pose in the generated multi-view images, this mechanism also enables the creation of animatable human avatars by generating canonical poses compatible with Gaussian Articulated Template Models [14]. Canonical poses are especially advantageous for the animating process, as they typically provide less occluded views of the body, thereby improving reconstruction quality. These contributions position Avatar++ as a unified and efficient framework for generating identity-consistent and pose-controllable 3D human avatars from a single image. The proposed model achieves state-of-the-art performance on Thuman2.0 and RenderPeople benchmarks across all evaluation metrics, while delivering a 5× faster inference time than the fastest existing method.*

1. Introduction

The rise of digital interaction has intensified the demand for automated 3D avatar generation, especially in applications such as mixed reality and virtual communication. Manually creating a realistic human avatar needs multi-view image capture and manual modeling, making it labor-intensive and costly. To address this, generative methods that synthesize 3D avatars from limited input, such as a single image, have gained significant attention [9, 29, 30, 38, 42].

Prior approaches to generating a human avatar from a single image can be broadly categorized into explicit and implicit methods. Explicit [12] methods rely on mesh-based parametric models to estimate 3D human shape and pose, but they often struggle to represent complex clothing or non-standard body poses. In contrast, implicit [19] function-based models represent human geometry as a continuous field, providing greater flexibility in capturing loose garments and intricate poses.

To overcome the limited information provided by a single input image, many recent methods utilize pre-trained 2D diffusion models (e.g., Stable Diffusion) to guide texture generation. Techniques such as Score Distillation Sampling (SDS) [24] iteratively optimize a 3D representation using the gradients from these diffusion models. In some cases, DreamBooth [28] is applied beforehand to personalize the diffusion model, which is widely used for human avatar generation.

A wide range of existing avatar generation approaches—whether based on implicit representation or explicit representation (e.g., Gaussian splatting)—depend heavily on optimization-based pipelines such as SDS or iterative refinement to synthesize high-fidelity 3D results. In addition to being time-consuming, these optimization-based methods often produce results that are difficult to predict, as their outcomes are highly dependent on initialization and the behavior of the optimization process.

[*]# Equal contribution.
[†]* Corresponding authors.

Figure 1. comparison of in-the-wild avatar generation. Each pair shows our result (left) and Human3Diffusion [38] (right).

In this work, we propose *Avatar*++, a fast, unified, and optimization-free pipeline that generates a pose-controllable, animation-ready 3D avatar from a single image in under 15 seconds. By leveraging identity-aware embeddings for diffusion transformer and pose-guided multi-view generation, Avatar++ bridges the gap between personalization and efficient 3D avatar synthesis. As illustrated in Figure 1, our method successfully generates human avatars with better facial details in less than 15 seconds with a consumer-grade GPU while achieving state-of-the-art performance on rendered novel view images.

Firstly, our model leverages two distinct types of embeddings: facial identity and visual features. For the Face ID embedding, we extract facial features using ArcFace [3] and map them into a semantic space via a pre-trained text encoder [21], which is designed to interpret Face ID features as text-like embeddings. This allows our model to generate a consistent face based on the input identity. In addition, we use a CLIP vision encoder to capture clothing, body pose, and overall appearance from the input image, enabling the avatar to reflect both the subject's identity and visual context.

Secondly, we propose a modified cross-attention mechanism in which temporary guidance is injected into the query. While less conventional than conditioning keys and values, this formulation allows the model to focus attention in a controlled manner without permanently modifying the latent representations. Since both the diffusion latent and the reference latent aim to represent the same subject, we concatenate them to form the attention query. This enables the model to attend to visual features with enhanced identity awareness. After the attention operation, the added dimension of reference latent is removed, allowing the model to maintain a generative trajectory while still being guided by the reference input.

Lastly, we employ a pose-guided multi-view generation mechanism based on ControlNet [40] to provide explicit human body anatomy. This additional pose-based guidance is critical in 3D generation, as it directly impacts the structural anatomy of the human body and ensures geometric consistency across synthesized views. This additional control is important in single-image settings, where the model lacks prior knowledge of the human body. By conditioning on 2D pose annotations [2], our method compensates for this limitation by injecting precise structural cues. Furthermore, it enables generating canonical poses (e.g., T-pose or A-pose) that are vital for downstream animation tasks such as rigging, and replicating the pose of the input image to preserve fidelity when canonicalization is not desired.

By integrating these components, Avatar++ enables efficient and personalized 3D avatar generation from a single front-view image. Combined with our multi-view synthesis model and a fine-tuned Large Gaussian Model (LGM) [34], the two-stage pipeline produces high-quality, full-body avatars while significantly reducing computational demands.

2. Related Work

2.1. Diffusion Models

The field of text-to-image generation has seen significant progress, driven by several key advances. Stable Diffusion [27], a powerful diffusion model, excels in efficiency and high-quality image generation. To control image generation, ControlNet [40] introduces additional encoding layers with auxiliary inputs, such as pose, masks, edges, and depth, allowing for more precise control of visual content.

Recently, Stable Diffusion 3 [4] presents a text-to-image multi-modal diffusion transformer architecture that integrates a rectified flow model. This architecture employs sepa-

rate weight parameters for image and text representations, facilitating bidirectional information exchange. Consequently, it significantly enhances text comprehension and accuracy, enabling better handling of complex prompts. Moreover, the implementation of the rectified flow model supports an efficient and stable training and inference process, which gives an advantage in high-resolution image generation.

2.2. Multi-view Diffusion Models

A foundational contribution to the multi-view diffusion field is zero-1-to-3, which first proposed using camera viewpoints as control conditions for image diffusion models to achieve novel view synthesis. However, its outputs often produce inconsistencies across views due to the stochastic nature of diffusion models. Subsequent approaches, such as Zero123++ [31], shifted to an all-view-at-once generation to mitigate the inconsistency. MVDream [32] introduces dense multi-view attention for single-object text to multi-view generation. ImageDream [35] expands this approach to image-conditioned generation. Wonder3D [17] incorporates normal data and cross-domain attention to enhance geometric consistency. However, these methods are trained on non-human objects, leading to challenges in rendering human faces. Recent work has started to optimize the complexity of multi-view attention. EpiDiff [10] employs epipolar attention to restrict matching candidates along epipolar lines, while Era3D [16] proposes row-wise attention under orthographic projection assumptions. These strategies greatly reduce attention costs, but EpiDiff can struggle with complex scenes or wide fields of view where epipolar geometry is ambiguous, and Era3D's performance degrades once the orthographic assumption no longer holds. Unlike prior methods that suffered from view inconsistency and relied on SMPL [18]/text prompts, Avatar++ leverages an Identity-Aware Diffusion Transformer, a Pose-ControlNet for canonical poses to generate high-quality and cross-view consistent multi-views from a single frontal photo.

2.3. Single image human reconstruction

Reconstructing a human from a single image generally involves two primary approaches: using explicit parametric models and relying on implicit representations.

2.3.1. Explicit parametric models

Explicit approaches rely on mesh-based parametric models, e.g., SMPL [18] or SMPL-X [23], which aim to estimate a minimally clothed human body mesh from the input image. These methods employ neural networks to predict the SMPL shape and pose parameters, enabling the generation of a base body mesh. To incorporate clothing details, subsequent techniques have proposed applying 3D offsets to the body surface or using pre-defined garment templates. Although these approaches achieve reasonable reconstructions, their mesh topology is limited by the underlying parametric model, posing challenges in reconstructing loose or complex clothing.

2.3.2. Implicit function based models

Implicit representation methods, on the other hand, capture human geometry as continuous fields, such as occupancy [29] functions, signed distance fields [22], or neural radiation fields (NeRF) [19], offering greater flexibility in topology. PiFU [29] utilizes pixel-aligned image features to predict 3D occupancy values and colors at sampled points within a predefined grid. The PIFuHD [30] enhances geometric and textural detail by incorporating front- and back-facing normal maps as additional inputs. GTA [41] employs Transformers with fixed learnable embeddings to transform single-image features into 3D tri-plane representation. Similarly, SIFU [42] refines 3D features by conditioning them on side-view information. TeCH [9] leverages diffusion models to synthesize invisible regions, though they demand extensive optimization and a precise SMPL-X alignment. HumanSGD [1] also adopts diffusion models for texture inpainting but remains dependent on the mesh estimation model, which introduces inaccuracies. In parallel, NeRF-based approaches such as SHERF [7] and ELICIT [8] represent the human body as radiance fields, enabling photorealistic rendering from single images. SHERF addresses incomplete data by filling gaps using 2D cues, while ELICIT employs a pre-trained CLIP [25] model to guide the reconstruction process via semantic understanding. Although these methods produce highly detailed renderings, they often require significant optimization time and can struggle with explicit mesh extraction.

Despite the advancements in implicit methods, achieving an optimal balance between rendering quality and computational efficiency remains a challenge. Recent innovations in 3D Gaussian splatting (3DGS) have demonstrated promising progress in this regard. HumanSplat [20] exemplifies this trend, introducing a generalizable framework that predicts 3D Gaussian properties from a single image without per-instance optimization. It integrates a 2D multi-view diffusion model to hallucinate unseen regions and a latent reconstruction Transformer with SMPL-based structure priors to enhance geometric and appearance consistency. However, the reliance on SMPL introduces difficulties in generalizing across diverse body types and clothing styles and a dependency on specific data for accurate pose and shape estimation. In contrast, our approach utilizes a diffusion Transformer to generate highly detailed novel multi-view images without depending on SMPL priors. By directly learning from image data, our model overcomes the generalization and data dependency issues of SMPL-based methods, resulting in more accurate and versatile human reconstruction across a broader range of scenarios.

3. Method

Our proposed method, as illustrated in Figure 2, introduces an effective framework for reconstructing a 3D human representation from a single front-view input image. Our approach consists of a two-stage process that integrates multi-view synthesis and 3D avatar generation, achieving high-quality reconstructions. The first stage generates multi-view images of the human subject from a single input by leveraging the 2D diffusion transformer, while the second stage leverages these synthesized views to synthesize a 3D avatar using Gaussian splatting.

3.1. Preliminary

3D Gaussian Splatting [13] is an efficient representation for novel view synthesis. 3D Gaussian splatting employs a collection of 3D Gaussians. Each Gaussian is parameterized by a center position $\mathbf{x}_i \in \mathbb{R}^3$, a scale vector $\mathbf{s}_i \in \mathbb{R}^3$, a rotation quaternion $\mathbf{r}_i \in \mathbb{R}^4$, an opacity $\alpha_i \in \mathbb{R}$, and a color feature $\mathbf{c}_i \in \mathbb{R}^C$. The complete set of parameters for the i-th Gaussian is denoted as $\mathcal{G}_i = \{\mathbf{x}_i, \mathbf{c}_i, \mathbf{s}_i, \mathbf{r}_i, \alpha_i\}$, encompassing its spatial configuration, appearance, and transparency. The 3D Gaussians are rendered by projecting them onto the image plane as 2D Gaussians and performing alpha composition on each pixel in front-to-back depth order, resulting in the final color and alpha value.

3.2. Image generation diffusion model for multi-view generation

We leveraged a powerful text-to-image diffusion model, Stable Diffusion 3.5 medium [33]. Stable Diffusion 3.5 is a strong image generation model that can effectively reflect text conditions.

Our multi-view generation transformer integrates both image and face information within a single framework. As our model does not require text conditioning, we employ a separate vision encoder (OpenCLIP bigG [11]) to produce global image embeddings from the input. We employed a CLIP [25] vision encoder to get the pooled output of the vision encoders and then project it to the text encoder space. Additionally, the Arc2Face [21] encoder takes ArcFace-based face embeddings to capture fine-grained facial information. To further make a robust representation, we combine the face representation and the global image representations, and concatenate multiple representations as illustrated in Figure 2.

The model takes pooled embeddings and prompt embeddings. The pooled prompt embeddings are concatenation of pooled output of Arc2Face encoder and pooled output of vision embedding. To align with the prompt embeddings, the image embeddings are first projected and then concatenated with Arc2Face prompt embeddings. This results in a fused conditioning embedding that integrates multi-modal information from both facial features and the input image. This approach contributes to generating more robust multi-view outputs.

3.3. Image conditioning mechanism

Our model is designed to generate consistent multi-view images from a given front-view image. To cooperate with this, our model has two mechanisms to preserve the face and outfit of the conditioning front-view image. First, our model leverages a multiple-embedding strategy as described above. Second, it directly manipulates the attention mechanism to inject information from the conditioning image. This mechanism incorporates a reference image into the existing attention-based architecture, enabling the generation process to leverage additional contextual cues. Consequently, our model benefits from enhanced contextual information, leading to improved output quality and relevance.

Our transformer-based Diffusion Transformer block uses Cross-Attention and Self-Attention. Our attention injection process is involved when processing Cross-Attention. During the conditional forward pass, our attention processor captures the encoded conditional information and stores it in a reference memory. Then, during the actual generation or reference injection process, the processor retrieves the saved reference image's hidden states from the memory and concatenates them with the current encoded conditional information. After computing Cross-Attention, it removes the added dimension where the retrieved reference image's hidden state is concatenated. By dynamically managing reference states and augmenting the attention mechanism, our approach supports reference-based generation while maintaining compatibility with existing architectures. With two conditional conditioning mechanisms, our model generates multi-view images while maintaining the identity and appearance of the input face and clothing.

3.4. Pose Guidance with ControlNet for better Avatar Generation

When reconstructing or synthesizing human avatars from images, not all views provide full visibility of all body parts. To address this, it is beneficial to generate multi-view images under specific poses where key body regions are clearly visible. Controllable pose when generating multi-view with canonical pose provides maximum visibility for animation and rigging, and direct replication of the input pose to preserve appearance fidelity. This capability allows the model to synthesize multi-view images in a structurally consistent and pose-aware manner, effectively covering occluded or ambiguous regions in the original input.

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z}_0,t,\boldsymbol{c}_i,\boldsymbol{c}_\mathrm{f},\epsilon\sim\mathcal{N}(0,1)}\left[w_t\lambda_t^{'}\|\epsilon-\epsilon_\theta\left(z_t,t,\boldsymbol{c}_i,\boldsymbol{c}_\mathrm{T}\right)\right\|_2^2] \tag{1}$$

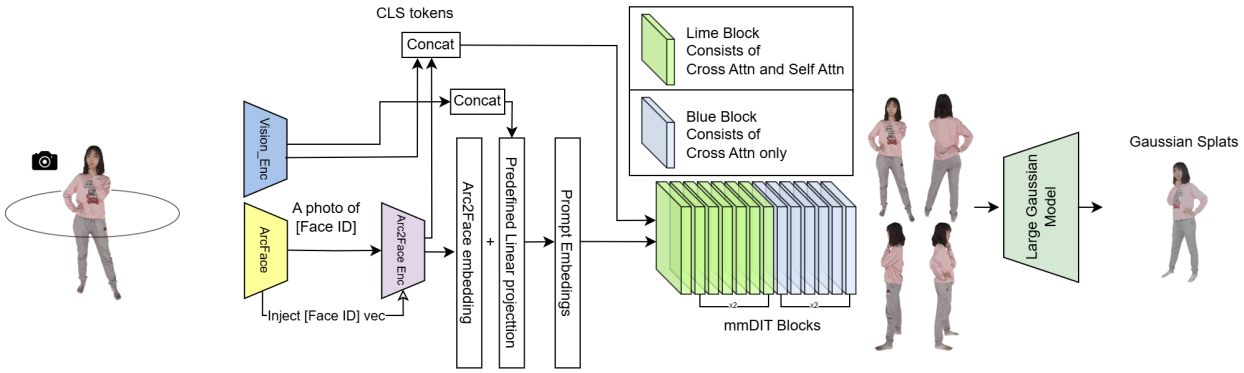

Figure 2. The overall architecture of our proposed model. The symbol + denotes concatenation. The predefined linear projection is a fixed mapping that projects the pooled output of the vision encoder into the prompt embedding space.

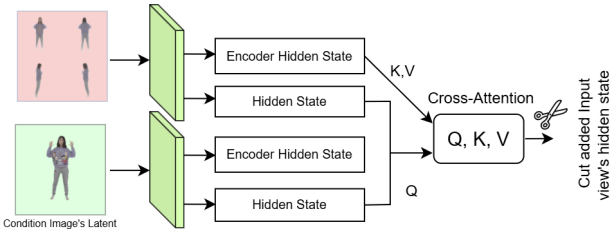

Figure 3. Hidden state injection during Cross-Attention for injecting conditioning image's information.

To train Pose ControlNet, we use a combination of the MVHumanNet [36] dataset and the Renderpeople [26] dataset. Our Renderpeople subset contains rigged human models originally in canonical poses, which we retarget to cover a wide range of diverse body poses. MVHumanNet provides multi-frame sequences of the same subject with varying poses. We train Pose ControlNet by using the front-view image at a given timestep as input and conditioning on a target pose from another timestep, enabling the model to generate multi-view images aligned with the desired pose.

3.5. Large Multi-View Gaussian Model for Avatar Generation

To further enhance our avatar reconstruction quality, we integrate the Large Multi-View Gaussian Model (LGM) [34] in our second-stage pipeline. LGM efficiently generates high-resolution 3D Gaussian representations from synthesized multi-view images provided by our diffusion transformer. Specifically, we adopt its asymmetric U-Net architecture, which predicts a set of compact and expressive Gaussian features from viewpoint-aligned images. These Gaussian splats are then fused through differentiable rendering to form a detailed and animatable 3D representation, significantly improving the fidelity and structural accuracy of our generated avatars.

4. Experiment

	Inference Time (on RTX 3090)
TeCH	10428s
SIFU	80.81s
SIFU Texture Refinement	349s + 80.81s = 429.81s
MagicMan	497.6 s
Human3Diffusion GS	68.96s
Ours(multi-view)	11.28 s
Ours(GS)	11.28 s + 2.52 s = 13.79 s

Table 1. Inference time measured on RTX 3090

4.1. Implementation details

In multi-view generation stage, we trained the model on four Nvidia A100 GPUs (80GB) for approximately 560000 iterations. We used the AdamW optimizer with a learning rate of $\mathbf{3 \times 10^{-5}}$ and batch size of 1, together with a CosineAnnealingWarmRestarts scheduler. We used mixed precision(bfloat16 for Transformer and CLIP; float32 for the VAE). For the Pose controlnet stage, we simply augmented the trained multi-view generation model with a pose-conditioned ControlNet module and continued training under the same hyperparameters for approximately 280000 iterations.

Finally, LGM[34] was finetuned using a single Nvidia A6000 GPU over two weeks on the combination of the 2K2K, Renderpeople, and Thuman 2.1 dataset.

4.2. Training dataset

We train our model using a comprehensive collection of 3D scans and reconstructions drawn from multiple datasets.

For training, we leverage a diverse set of 3D human data, combining both high-quality scans and real-world multi-view captures. Specifically, we use 2,050 scans from the 2K2K dataset [5], 500 scans from THuman2.0 [39], an additional 1,464 samples from the extended THuman2.1, and 478 commercial scans from Renderpeople [26], resulting in 4,497 high-quality 3D human models with relatively simple clothing. In addition, we incorporate 3,172 identities and 10,135 multi-view captures from the MVHumanNet [36] dataset, which provides more diverse clothing and pose variations from real subjects captured using calibrated multi-camera systems. To obtain 3D representations from MVHumanNet, we apply the SplatFacto algorithm to reconstruct Gaussian splats from the captured multi-view images. In total, our model is trained on 14,632 3D human samples.

For consistency, all 3D assets from 2K2K, THuman, THuman2.1, Renderpeople, and MVHumanNet reconstructions are rendered into four standardized views using a uniform camera setup, with cameras placed 90 degrees apart around the object center. This consistent projection protocol ensures that all datasets contribute comparable training samples, enabling effective learning of our 3D reconstruction framework.

We render all views at a resolution of 4096 x 4096 and resize to 512×512, and for every dataset, we extract face embeddings using ArcFace to capture detailed facial attributes.

4.3. Evaluation

We evaluated our proposed model against existing baselines to measure overall performance and further conduct ablation studies to analyze the contribution of each component. The baselines include TeCH [9], SIFU [42], and MagicMan [6], which are all optimization-based methods relying on iterative refinement. In contrast, Human-3Diffusion [38], similar to our scenario, employs a single feed-forward pass, offering faster inference. Our model requires 2D pose annotations as input, which we obtain using OpenPose [2]. During inference, we extract 2D poses for the remaining four views by projecting the reconstructed ECON [37] mesh into those views and applying OpenPose on the projected images.

4.3.1. Quantitative Evaluation

We quantitatively evaluate our approach on the 21 samples on THuman2.0 [39] and 10 samples on the RenderPeople datasets. For every test subject, we render two view sets: (i) four canonical views—frontal, right, rear, and left—obtained by rotating the virtual camera in 90° steps, and (ii) twenty dense views generated every 18°. Each cell in Tables 2, 3, 4, and 5 contains a pair of values in the format: 4-view / 20-view. We average PSNR, SSIM, and LPIPS over the corresponding views to measure image fidelity and perceptual quality.

On THuman2.0, our method reaches 21.086 / 20.634 PSNR, 0.896 / 0.891 SSIM, and 0.0970 / 0.101 LPIPS, markedly outperforming TeCH, SIFU, MagicMan, and the recent Gaussian-Splatting baseline Human-3Diffusion.

The advantage persists on RenderPeople: our pipeline records 24.637 / 24.200 PSNR, 0.904 / 0.900 SSIM, and 0.088 / 0.092 LPIPS, again surpassing Human-3Diffusion as well as all other baselines.

These consistent improvements demonstrate the effectiveness of our single forward-pass pipeline, which fuses identity-aware ArcFace embeddings, pose conditioning, and multi-view diffusion to achieve superior fidelity without expensive iterative optimization. The model's ability to output pose-controllable, animation-ready canonical avatars further enhances reconstruction quality, especially when driven by 2D pose inputs.

Table 1 shows inference times measured on an RTX 3090. Optimization based methods like TeCH take over 10,000 seconds, while SIFU requires around 80 seconds plus an additional 349 seconds for texture refinement. MagicMan and Human3Diffusion GS take 498 and 69 seconds, respectively. Our model achieves a significant speed-up, running inference in about 14 seconds, which is about 5 to 900 times faster than previous methods.

4.3.2. Qualitative Evaluation

Figure 4 shows the qualitative comparison against three other baselines on the Thuman2.0 dataset. Human3Diffusion struggles particularly around facial regions, producing distorted or blurred features, whereas our results preserve facial details. This limitation arises because its multi-view diffusion backbone is capped at 256×256 resolution, which smooths out details of facial features such as eyes, nose, and mouth. Also, there is no mechanism to deliver facial information.

SIFU deterministically regresses a 3D avatar directly from an input image, generating texture via a separate optimization process, which leads to misalignment between textual annotations and the generated mesh. As a result, it generates frontal regions seen in the input image, but fails to produce coherent texture on unseen regions.

Finally, TeCH, which employs BLIP [15] to annotate input images before optimization which often fails to generate accurate annotations on the input image, leading to textures that do not match the input or failure to generate coherent textures. For example, the input image is asian, but the caption generated by BLIP is caucasian.

4.4. Ablation studies

As part of our ablation studies, we evaluate structurally modified versions of our model by removing face embeddings and attention injection to assess their impact.

	Thuman2.0		
	PSNR	SSIM	LPIPS
TeCH^{+}	17.004 / 16.496	0.852 / 0.843	0.137 / 0.146
SIFU (W/O Texture Refinement)	18.33 /18.01	0.878 / 0.875	0.115 / 0.115
SIFU (Texture Refinement)$^{+}$	19.699 / 19.482	0.872 / 0.870	0.118 / 0.123
MagicMan$^{\circ}$	19.186 / 18.07	0.863 / 0.853	0.122 / 0.13
Human-3Diffusion (GS)	20.146 / 19.729	0.884 / 0.880	0.108 / 0.106
Ours (GS)	**21.086* / 20.634***	**0.896* / 0.891***	**0.097*/ 0.101***

Table 2. Quantitative results on the THuman2.0 dataset. '$^{+}$' indicates optimization-based models and '$^{\circ}$' indicates iterative methods.

	RenderPeople		
	PSNR	SSIM	LPIPS
TeCH^{+}	18.683 / 18.212	0.870 / 0.862	0.119 / 0.127
SIFU (W/O Texture Refinement)	20.389 /19.975	0.894 /0.888	0.096 /0.102
SIFU (Texture Refinement)$^{+}$	23.701 / 23.078	0.894 / 0.888	0.094 / 0.101
MagicMan$^{\circ}$	24.224 / 23.535	0.890 / 0.889	0.090 / 0.098
Human-3Diffusion (GS)	24.262 / 23.744	0.900 / 0.894	0.091 / 0.096
Ours (GS)	**24.637* / 24.200***	**0.904* / 0.900***	**0.088* / 0.092***

Table 3. Quantitative results on the RenderPeople dataset. '$^{+}$' indicates optimization-based models and '$^{\circ}$' indicates iterative methods.

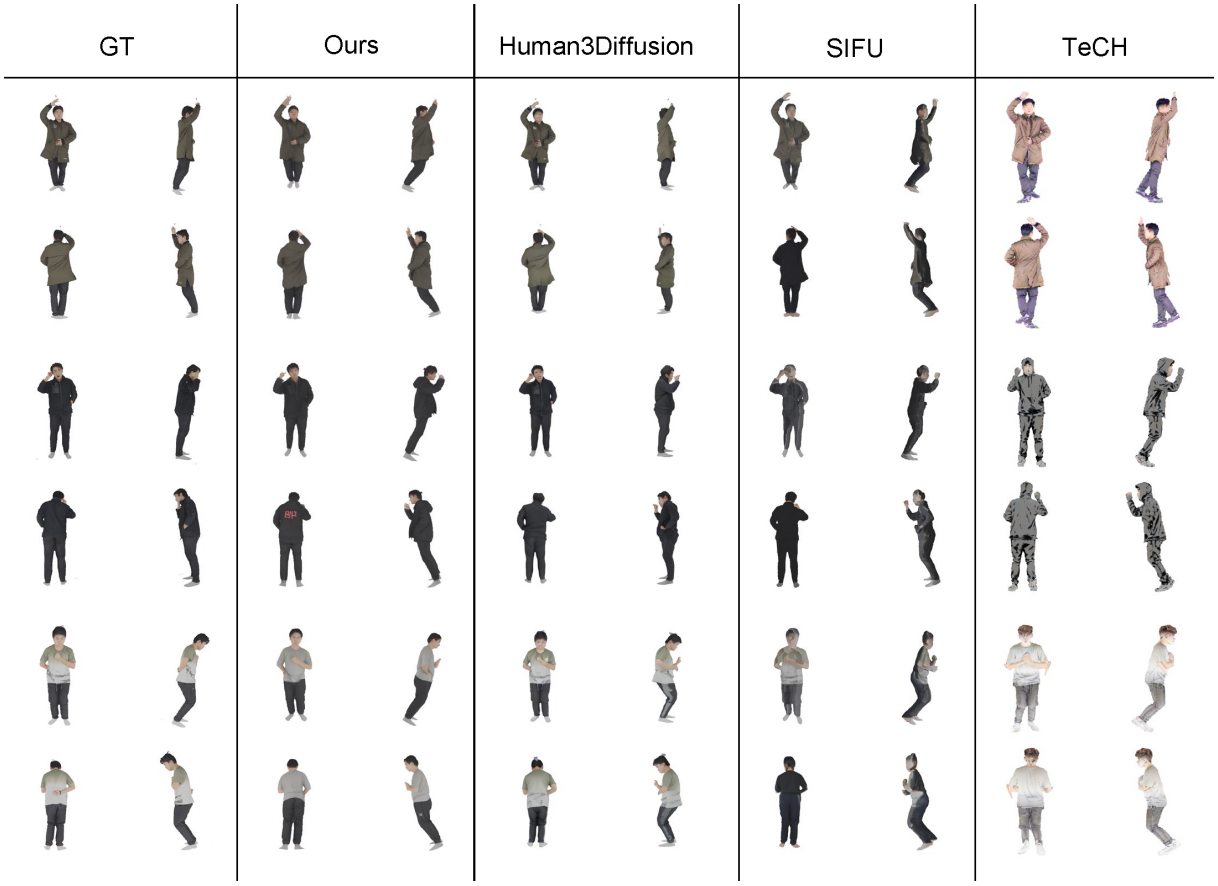

Figure 4. Qualitative results on THuman2.0 dataset. Our method captures appearance faithfully. Please zoom in for a detailed inspection.

As shown in Table 4 and Table 5, both components contribute significantly to the final performance. On the THuman2.0 dataset, removing attention injection causes PSNR to drop from 21.086 / 20.634 to 18.894 / 18.697, SSIM from

	Thuman2.0		
	PSNR	SSIM	LPIPS
W/O attn injection	18.894 / 18.697	0.850 / 0.862	0.133 / 0.126
W/O Face embeddings	18.789 / 18.542	0.866 / 0.862	0.124 / 0.128
Ours (GS)	**21.086* / 20.634***	**0.896* / 0.891***	**0.097*/ 0.101***

Table 4. Ablation studies on THuman2.0 dataset

	RenderPeople		
	PSNR	SSIM	LPIPS
W/O attn injection	20.976 / 20.824	0.884 / 0.882	0.109 / 0.112
W/O Face embeddings	22.000 / 21.770	0.887 / 0.884	0.106 / 0.109
Ours (GS)	**24.637 / 24.200**	**0.904 / 0.900**	**0.088 / 0.092**

Table 5. Ablation studies on the RenderPeople dataset

0.896 / 0.891 to 0.850 / 0.862, and LPIPS increases from 0.097 / 0.101 to 0.133 / 0.126. Similarly, removing face embeddings also degrades performance across all metrics, though to a slightly lesser extent.

On the RenderPeople dataset, a similar trend is observed: without attention injection, PSNR drops by around 3.7 points, SSIM by 0.02, and LPIPS increases by 0.021 compared to the full model. Removing face embeddings also yields lower PSNR (from 24.637 / 24.200 to 22.000 / 21.770), lower SSIM (from 0.904 / 0.900 to 0.887 / 0.884), and higher LPIPS (from 0.088 / 0.092 to 0.106 / 0.109).

These results clearly demonstrate that both face embeddings and attention injection are essential for achieving high-fidelity, identity-consistent multi-view image synthesis, contributing to sharper textures, better structural similarity, and perceptual quality.

4.5. Application

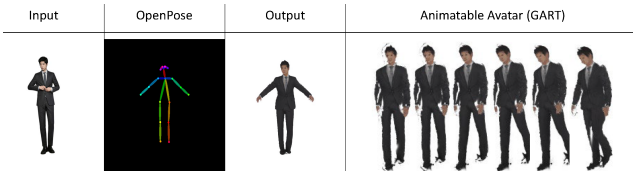

Figure 5. Animatable avatar generation using GART

To animate the static 3D avatars generated by Avatar++, we adopt the GART (Gaussian Articulated Template) [14] framework as a post-processing stage, as illustrated in Figure5. While Avatar++ generates high-quality Gaussian representations from a single image, these outputs are inherently static and lack temporal articulation. GART offers an efficient and explicit method to bring motion into these avatars by leveraging forward skinning and latent bone modeling, enabling realistic animation from sparse monocular cues.

Specifically, we initialize the GART model using the Gaussian parameters produced by Avatar++. These Gaussians are aligned with a canonical pose, and we associate them with the SMPL template skeleton to define an articulated structure. We then optimize GART to learn deformation fields and skinning weights that adaptively model both rigid body motion and non-rigid deformations (e.g., clothing). This process allows our avatars to be animated across diverse poses while preserving the original geometry and appearance fidelity generated by Avatar++.

Through this integration, we enable a seamless transition from static 3D avatar creation to fully animatable avatars capable of dynamic pose rendering at high frame rates. Fast inference, efficient rendering via 3D Gaussian Splatting, and expressiveness in capturing complex motions significantly extend the usability of Avatar++ in downstream applications such as gaming, virtual try-on, and real-time communication in VR/AR environments.

5. Conclusion

In conclusion, Avatar++ presents an efficient and robust solution for generating high-quality, animation-ready 3D human avatars from a single image. By employing dual embeddings—facial identity through ArcFace and global visual features via CLIP—combined with an attention injection mechanism and pose-guided multi-view synthesis, our method significantly surpasses existing models in both speed and quality. Avatar++ achieves state-of-the-art performance on benchmark datasets such as THuman2.0 and RenderPeople, demonstrating improved image fidelity, structural coherence, and perceptual quality, all while maintaining inference speeds that are 4 times faster than the fastest alternatives. This advancement not only simplifies the process of avatar generation but also expands potential applications in virtual reality, animation, gaming, and real-time communication, marking a substantial step forward in accessible and high-fidelity digital human representation.

6. Limitation and future works

Avatar++ has limitations on Pose estimation. To synthesize multi-view images with the same pose as the input, our method relies on the external ECON [37] framework for 3D pose estimation. However, this dependency can introduce inaccuracies during inference, especially under challenging articulations or occlusions. Also, its performance is constrained by limited and biased training data. THuman2.1 and MVHumanNet datasets are heavily biased toward East Asian subjects, resulting in possible racial bias. Future work will therefore focus on developing more robust pose predictors to address ECON-related error. Additionally, we aim to curate more diverse and representative datasets to mitigate racial and demographic biases, ensuring equitable performance across varied user groups. We expect Avatar++ to evolve into a more faithful solution for real-time avatar creation across all users by overcoming these limitations.

References

[1] Badour AlBahar, Shunsuke Saito, Hung-Yu Tseng, Changil Kim, Johannes Kopf, and Jia-Bin Huang. 2023. Single-image 3d human digitization with shape-guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers*. 1–11.

[2] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[3] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *CVPR*.

[4] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. arXiv:2403.03206 [cs.CV] https://arxiv.org/abs/2403.03206

[5] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. 2023. High-fidelity 3d human digitization from single 2k resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12869–12879.

[6] Xu He, Zhiyong Wu, Xiaoyu Li, Di Kang, Chaopeng Zhang, Jiangnan Ye, Liyang Chen, Xiangjun Gao, Han Zhang, and Haolin Zhuang. 2025. MagicMan: Generative Novel View Synthesis of Humans with 3D-Aware Diffusion and Iterative Refinement. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 3 (Apr. 2025), 3437–3445. doi:10.1609/aaai.v39i3.32356

[7] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. 2023. Sherf: Generalizable human nerf from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9352–9364.

[8] Yangyi Huang, Hongwei Yi, Weiyang Liu, Haofan Wang, Boxi Wu, Wenxiao Wang, Binbin Lin, Debing Zhang, and Deng Cai. 2023. One-shot implicit animatable avatars with model-based priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8974–8985.

[9] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. 2024. Tech: Text-guided reconstruction of lifelike clothed humans. In *2024 International Conference on 3D Vision (3DV)*. IEEE, 1531–1542.

[10] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, and Lu Sheng. 2024. EpiDiff: Enhancing Multi-View Synthesis via Localized Epipolar-Constrained Diffusion . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 9784–9794. doi:10.1109/CVPR52733.2024.00934

[11] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. *OpenCLIP*. doi:10.5281/zenodo.5143773 If you use this software, please cite it as below..

[12] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4 (July 2023). https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

[13] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics* 42, 4 (2023), 1–14.

[14] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. 2024. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 19876–19887.

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[16] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. 2024. Era3D: High-Resolution Multiview Diffusion using Efficient Row-wise Attention. *arXiv preprint arXiv:2405.11616* (2024).

[17] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. 2023. Wonder3D: Single Image to 3D using Cross-Domain Diffusion. *arXiv preprint arXiv:2310.15008* (2023).

[18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

[19] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

[20] Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. 2024. Humansplat: Generalizable single-image human gaussian splatting with structure priors. *Advances in Neural Information Processing Systems* 37 (2024), 74383–74410.

[21] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. 2024. Arc2Face: A Foundation Model for ID-Consistent Human Faces. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[22] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 165–174.

[23] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[24] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *ICLR*. https://openreview.net/forum?id=FjNys5c7VyY

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

[26] Renderpeople. 2024. Renderpeople 3D Human Model Dataset. https://renderpeople.com/ Accessed: 2025-05-20.

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[28] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.

[29] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2304–2314.

[30] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 2020. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 84–93.

[31] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. 2023. Zero123++: a Single Image to Consistent Multi-view Diffusion Base Model. arXiv:2310.15110 [cs.CV]

[32] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. 2024. MVDream: Multi-view Diffusion for 3D Generation. In *International Conference on Learning Representations (ICLR)*.

[33] Stability AI. 2024. *Introducing Stable Diffusion 3.5*. https://stability.ai/news/introducing-stable-diffusion-3-5 Accessed: 2025-05-22.

[34] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. 2024. LGM: Large Multi-view Gaussian Model for High-Resolution 3D Content Creation. In *ECCV (4)*. 1–18. https://doi.org/10.1007/978-3-031-73235-5_1

[35] Peng Wang and Yichun Shi. 2023. ImageDream: Image-Prompt Multi-view Diffusion for 3D Generation. arXiv:2312.02201 [cs.CV] https://arxiv.org/abs/2312.02201

[36] Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, et al. 2024. Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19801–19811.

[37] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. 2023. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[38] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard. Pons-Moll. 2024. Human 3Diffusion: Realistic Avatar Creation via Explicit 3D Consistent Diffusion Models, In NIPS. *NeurIPS 2024*.

[39] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. 2021. Function4D: Real-time Human Volumetric Capture from

Very Sparse Consumer RGBD Sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*.

[40] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543 [cs.CV] https://arxiv.org/abs/2302.05543

[41] Zechuan Zhang, Li Sun, Zongxin Yang, Ling Chen, and Yi Yang. 2023. Global-correlated 3D-decoupling Transformer for Clothed Avatar Reconstruction. arXiv:2309.13524 [cs.CV] https://arxiv.org/abs/2309.13524

[42] Zechuan Zhang, Zongxin Yang, and Yi Yang. 2024. SIFU: Side-view Conditioned Implicit Function for Real-world Usable Clothed Human Reconstruction. arXiv:2312.06704 [cs.CV] https://arxiv.org/abs/2312.06704