SplatCtrlA: Generalizable Single Image to Fully Controllable 3D Avatar

Jun Xiang[1], Yudong Guo[1*], Boyang Guo[1], Yancheng Yuan[2], and Juyong Zhang[1]

[1] University of Science and Technology of China
[2] The Hong Kong Polytechnic University

Abstract. Expressive human avatar creation from a single image is highly challenging due to the inherently ill-posed nature of the problem, as well as the complexities of appearance preservation and dynamic human modeling. To address these challenges, this work presents a comprehensive pipeline for training a large feed-forward model that efficiently generates fully controllable 3D avatars from single-image. Due to the scarcity of consistent training data, we construct a large-scale 3D Gaussian avatar dataset to support model training. To better recover appearance details, we propose an input-aware decoding scheme that fully leverages information from the input image. Furthermore, to achieve comprehensive full-body control, we introduce a Gaussian blending–based facial enhancement module and apply Gaussian geometric constraints to stabilize expressive avatar generation. Extensive experiments demonstrate that our method enables one-shot reconstruction of photorealistic avatars with whole-body control, outperforming existing works in terms of human dynamic and appearance details.

Keywords: Fine-grained Avatar · Feed-forward Generation · One-Shot Input

1 Introduction

Creating expressive human avatars from a single image is a key capability for immersive virtual reality and creative content generation. Despite its importance, the task remains challenging due to the complexities of dynamic human modeling, appearance preservation, and the fundamentally ill-posed nature of the problem.

Recently, works [7, 22, 26, 33, 40, 57] based on 2D diffusion models are capable of producing realistic human animation videos. Nevertheless, despite extensive progress in driving mechanisms [5, 15, 45, 72], network architectures [13, 46, 51, 76], and temporal consistency techniques [3, 4, 17, 76], these approaches remain limited by inefficient multi-step denoising and suffer from inconsistent temporal or identity coherence during long sequences or under large pose variations. Methods

*Corresponding author: Yudong Guo.

Fig. 1: Our method reconstructs a fully controllable 3D human avatar from a single image within seconds through a single feed-forward pass, supporting natural control over body pose, gestures, and facial expressions.

based on 3D representations [29,42], such as 3D Gaussian Splatting (3DGS) [29], inherently offer stronger consistency across views and time. While one-shot human reconstruction approaches [34, 43, 81] can achieve high-quality novel view synthesis, they are mostly restricted to static scenarios. To compensate for the limited pose and deformation information available in the input image and to enable dynamic motion, several works [12, 50, 54, 64] combine diffusion-prior-based personalized reconstruction with human parametric models [35, 45] for controllable avatar animation; however, such pipelines require heavy optimization and long training times. More recently, researchers have explored bringing the success of large reconstruction models (LRMs) [21,56] in the 3D reconstruction field into the drivable human avatar field, yielding initial progress through efficient feed-forward generation. However, these methods either struggle with general full-body inputs due to dataset scale and bias [74, 78], or focus primarily on coarse body pose while neglecting detailed appearance capture and fine-grained control of facial and hand regions [49, 85].

In this work, we propose a comprehensive pipeline for training a scalable feed-forward model that produces expressive 3D human avatars from a single image in seconds. We employ a supplementary UV Gaussian Avatar based on the SMPL-X [45] model as our fundamental 3D representation, taking into account both the real-time rendering capability of 3DGS and the strong compatibility between UV maps and transformer-based 2D regression networks. Our model takes a single image as input and directly predicts the UV attribute maps corresponding to the human Gaussian field. Given the scarcity of high-quality, 3D-consistent training data, we construct a large-scale 3D Gaussian avatar dataset. Inspired by [54,64,85], we first employ high-quality text-to-image diffusion models [32] to generate image collections with diversity. For each image, we leverage 2D human video diffusion priors [57,79] along with SMPL-X–based geometric constraints to reconstruct the corresponding Gaussian avatar, which enables the rendering of 3D-consistent multi-view and multi-pose images for model training. In addition, we incorporate existing public human datasets [73, 85] and a large collection of

real human head videos into a unified hybrid dataset, enhancing the model's generalization to novel viewpoints and fine-grained human details.

For the model architecture, we adopt a pretrained feature extractor [30] and a transformer-based encoder to encode image features into the UV space. Considering the inevitable information loss during feature extraction, we propose an Input-aware Decoder that fully exploits the information contained in the input image, leading to improved training efficiency and texture completeness. We explicitly inject color information from the input image into the UV space as decoding conditions, using the rasterization correspondence between the mesh and the UV space, as well as the 3D-to-image projection mapping. To mitigate misalignment between the tracked mesh and the input, we decouple the decoding of Gaussian geometry and color attributes by applying a UV Mesh Inverse module to predict the Gaussian deformation field, followed by the calibrated UV Gaussian Inverse to assist in color decoding. Moreover, we introduce semantic guidance during the Gaussian Inverse stage to ensure semantic correctness under tracking and deformation prediction errors. To further enhance the expressiveness of the human avatar, we introduce an additional facial refinement module and apply stronger regularization to hand modeling, ensuring comprehensive and fine-grained control over the entire avatar.

In summary, our contributions include the following aspects:

- We develop a comprehensive and efficient feed-forward framework capable of producing photo-realistic whole-body avatars with precise texture preservation and full control over body pose, hand gestures, and facial expressions.
- We present a data generation pipeline that constructs a large-scale 3D Gaussian avatar dataset emphasizing 3D consistency, which is then integrated with existing datasets to create a hybrid training set for large model training.
- Our proposed input-aware decoding scheme fully leverages the color, geometric, and semantic information from the input image, and when paired with a transformer-based encoder, enables more accurate texture capture and completion.

2 Related Work

Single-Image Human Reconstruction. Some works [11,27,39,65] train large 3D human generative models and apply 3D GAN inversion for one-shot human reconstruction, but they struggle with generalization and preserving full personalized details. PIFu [52] and subsequent works [20,53,83] introduce pixel-aligned features for image-based human reconstruction. Subsequent loop optimization techniques [9,66,67,80] integrate neural fields with hybrid human priors to improve robustness when reconstructing complex clothing. An alternative direction leverages diffusion priors to fill in missing information, and works do this in different ways, including training a human-centered diffusion model [2,19,43], using enhanced novel-view diffusion results as extra supervision [34,36,81], and lifting 2D priors to 3D by employing the Score Distillation Sampling (SDS) [48]

technique [25, 68, 71, 75]. Meanwhile, some works [6, 43, 61] integrate large reconstruction models [21, 56] with diffusion models for inference acceleration or quality refinement. Human-Splat [43] applies a latent reconstruction transformer to latent features generated by a multi-view diffusion model to predict human Gaussians. HGM [6] employs a generate-then-refine pipeline guided by diffusion priors. In contrast to these methods, which focus on static scenes and overlook dynamic human motion, our approach captures both human appearance and dynamic, enabling expressive whole-body animations.

One-Shot Animatable Avatar. Many works [8, 10, 16, 18, 58, 70, 82] achieve realistic one-shot animation but focus exclusively on head modeling. For human avatar animation, 2D approaches have emerged as a compelling direction due to the powerful synthesis ability of diffusion models, but they still face challenges such as image distortion, identity changes and pose misalignment, resulting from loose 3D constraints. Some approaches [14,22,33,40,69,84] incorporate temporal layers from AnimateDiff [17] to ensure smooth transitions, while others [47, 55, 57, 79] use video diffusion models [3, 4] for dynamic sequences. More recent works [7,38,59] adopt the Diffusion Transformer (DiT) [13,46] as the core network backbone, while others [26, 41] emphasize multi-modal controllability. Early 3D approaches [23, 24] were constrained by limited reconstruction quality. Recent methods [12,50,54,62,64] have achieved higher fidelity by incorporating diffusion priors and personalized optimization, though at the cost of heavy computation. Meanwhile, another line of research [37, 49, 74, 78, 85] investigates more efficient feed-forward generation frameworks. Although IDOL [85] achieves impressive reconstruction results, it fails to fully capture the identity and texture details from the input image, mainly due to feature extraction loss and the inconsistency of 2D synthetic datasets. LHM [49], trained on large-scale real human video datasets, produces relatively better results but still struggles with fine-grained facial expressions and hand modeling. GUAVA [74], which applies inverse texture mapping, captures finer texture details but is limited to the frontal upper-body regions. In contrast, our method leverages a hybrid training dataset — including our Gaussian avatar dataset — to improve generalization and incorporates an input-aware decoder and a facial enhancement module to achieve more detailed and expressive modeling.

3 Large Avatar Reconstruction Model

Given a single input image I of a human body, we aim to reconstruct a whole-body Gaussian avatar in an efficient feed-forward manner. Sec. 3.1 describes the geometric representation of our Gaussian avatar. In Sec. 3.2 and Sec. 3.3, we present the details of our large avatar reconstruction model. Sec. 3.4 shows how we further enhance the quality of the facial region. Sec. 3.5 details the construction of our training dataset, while Sec. 3.6 introduces the carefully designed constraints and loss terms used to train the large-scale model. The entire pipeline is illustrated in Fig. 2.

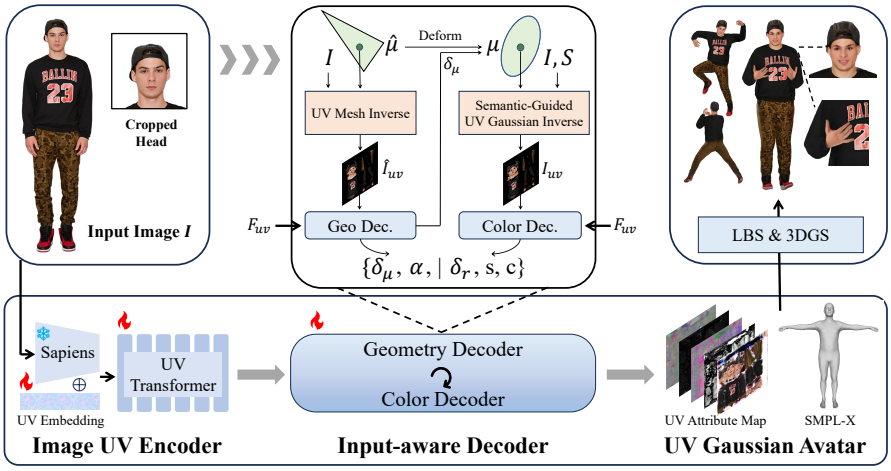

Fig. 2: Overview. The input image is processed in two parallel streams: together with the cropped head image, it is fed into a transformer-based image encoder to extract UV feature maps; meanwhile, via the proposed input-aware scheme, it is explicitly projected into the SMPL-X UV space and combined with semantic information to guide the separate decoding of Gaussian geometry and color attributes. The Input-aware Decoder directly predicts the Gaussian UV attribute maps, serving as a 2D representation of the 3D human avatar in the canonical space, which can then be deformed to the target pose via Linear Blend Skinning (LBS) for free animation.

3.1 UV Gaussian Avatar

For 3D human representation, we couple 3D Gaussian Splatting with the whole-body parametric mesh model SMPL-X, which provides a strong initialization and facilitates human animation. Inspired by [44, 63], we maintain the 3D Gaussian field in the SMPL-X UV space. This UV parameterization not only enables flexible control of Gaussian density but also connects seamlessly to a transformer-based 2D regression network. Specifically, the Gaussian field is parameterized as $G = \{\boldsymbol{\mu}, \boldsymbol{\alpha}, \mathbf{r}, \mathbf{s}, \mathbf{c}\}$, where $\boldsymbol{\mu}$ is the Gaussian center, $\boldsymbol{\alpha}$ is the opacity, $\mathbf{r}$ is the rotation, $\mathbf{s}$ is the scale, and $\mathbf{c}$ is the RGB color. Rather than predicting all UV attribute maps directly, we predict spatial offsets for Gaussian positions and rotations to better exploit the geometric priors provided by the SMPL-X model, i.e., $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}} + \delta_{\boldsymbol{\mu}}$ and $\mathbf{r} = \hat{\mathbf{r}} \cdot \delta_{\mathbf{r}}$, where $\hat{\boldsymbol{\mu}}, \hat{\mathbf{r}}$ are initialized from the canonical SMPL-X mesh. Given a driving pose, the canonical Gaussian avatar is transformed to the target space using Linear Blend Skinning (LBS).

3.2 Image UV Encoder

Image Feature Encoding. We use the backbone of Sapiens [30], pretrained on millions of human images, to extract human features. Considering the importance of the head region, we obtain a head image I_{head} from I through cropping and super-resolution [60]. We obtain transformer-compatible global tokens

by concatenating the body and head image tokens extracted using the frozen Sapiens-1B encoder $\mathcal{E}_{\text{sapiens}}$, formulated as:

$$\mathbf{F} = \mathcal{E}_{\text{sapiens}}(\text{I}) \oplus \mathcal{E}_{\text{sapiens}}(\text{I}_{\text{head}}), \tag{1}$$

where $\oplus$ denotes token concatenation.

UV Transformer. Similar to [85], we implement a scaling transformer to map the global human token from screen space to UV space. We first concatenate the image tokens $\mathbf{F}$ with a learnable positional UV embedding $\mathbf{F}_{\text{pos}}$ along the patch channel. The concatenated tokens are then aligned and refined through D transformer blocks. The self-attention mechanism used in the transformer helps aggregate related information and fill in invisible areas. Finally, the output is passed through a CNN-based upsampling network to generate the final high-resolution UV feature maps, $\mathbf{F}_{\text{uv}}$.

3.3 Input-aware Decoder

While Sapiens effectively captures the majority of features from the input, some inevitable information loss remains. To further enhance reconstruction quality, it is essential to exploit the rich information contained in the input image I, such as color, semantic context, and tracking-derived geometric priors.

Geometry Decoding. Regarding color and geometric priors, we incorporate the inverse UV RGB map as an additional input signal to the decoder, similar to [74]. Specifically, for each valid UV pixel localized by a Gaussian g_k, we determine its corresponding position, $\hat{\mu}_k$, on the tracked SMPL-X mesh via UV rasterization. This position is then projected onto the input image using the camera parameters to obtain the RGB value of the UV pixel. After applying the visible-region mask, $\hat{M}_{\text{uv}}$, we obtain the desired inverse UV RGB, $\hat{I}_{\text{uv}}$, which is used for geometry decoding. This process can be formulated as:

$$\begin{aligned} \hat{I}_{\text{uv}} &= \text{MeshInverse}(I, \hat{\boldsymbol{\mu}}), \\ \delta_{\boldsymbol{\mu}},\, \boldsymbol{\alpha} &= \mathcal{D}_{\text{geo}}\big(F_{\text{uv}} \oplus \hat{I}_{\text{uv}}\big). \end{aligned} \tag{2}$$

Semantic-Guided Color Decoding. Due to the spatial mismatch, δ_μ, between the Gaussian and mesh fields, we use the inverse UV RGB, I_{uv}, corresponding to the Gaussian field as the conditioning signal for color attribute decoding (with the same visibility mask, $\hat{M}_{\text{uv}}$, for simplicity), which helps capture color details outside the mesh surface. To further reduce projection errors caused by imperfect tracking and deformation fields, δ_μ, especially near object boundaries, we refine the projected positions using 2D semantic guidance. Each Gaussian is associated with a semantic label, and mismatched projections are corrected to the nearest semantically consistent location. This process can be formulated as:

$$\begin{aligned} I_{\text{uv}} &= \text{GauInverse}\big(I, \hat{\boldsymbol{\mu}} + \delta_{\boldsymbol{\mu}}, S\big), \\ \delta_{\mathbf{r}}, \mathbf{s},\, \mathbf{c} &= \mathcal{D}_{\text{color}}\big(F_{\text{uv}} \oplus I_{\text{uv}}\big), \end{aligned} \tag{3}$$

where S denotes the semantic image of the input I.

While geometry and color are decoded separately, the predicted geometric attributes serve as conditions for color decoding, promoting consistency and accuracy between the two decoders. Furthermore, by explicitly incorporating input information, the Input-aware design mitigates over-reliance on the training distribution and improves generalization across diverse human appearances.

3.4 Facial Enhancement

Head Avatar Model. Due to the limited quality and diversity of facial data in our synthetic human dataset, we train an additional head reconstruction model on a collected large-scale, high-resolution real head video dataset to enhance facial reconstruction. The model adopts the same architecture as the whole-body network, but the UV Transformer processes only head tokens extracted from cropped head images, with the Transformer depth reduced to $D/2$.

Gaussian Blending Network. Since the head model is initialized from the pretrained body model, the two exhibit high consistency around facial boundaries. Nevertheless, we further train a Gaussian blending network to seamlessly fuse the two models. Specifically, the Gaussian blending network takes as input the concatenated UV attribute maps generated by the two models (body and head) according to the facial-region UV mask. The network outputs a smoothed, seamless final Gaussian UV attribute map. The Gaussian blending network consists of a StyleUNet followed by a zero convolution layer. The introduction of the zero convolution layer ensures that the network learns only a small residual correction for fine adjustment, rather than relearning the entire UV Gaussian representation from scratch.

3.5 Human Dataset Construction

For human avatar reconstruction, there is still a lack of publicly available large-scale datasets containing diverse real-world human videos, which poses a major obstacle to data-driven generalization. To address this limitation, we develop a pipeline to obtain large-scale Gaussian avatars from single images in batches and construct a Gaussian avatar dataset with more than 25k identities that emphasizes 3D consistency and diversity to support model training. Sec. 3.1 details the Gaussian Avatar representation we use. Next, we will describe how we construct a human avatar from a single image in an optimization-based manner.

Gaussian Avatar Dataset. To enable Gaussian avatar reconstruction from a single image, we follow a diffusion-prior-based framework similar to [54, 64]. Given an input image from our image collection, we generate pose-rich frames as pseudo supervision using the diffusion-based human animation method, MimicMotion [79]. The generated frames provide additional pose variations and complete the invisible parts missing from the input image. We use them, along with the input image, to form the training set for our UV Gaussian Avatar

reconstruction. Once the Gaussian avatar is reconstructed, we can render 3D-consistent multi-view and multi-pose human images to train our generalizable model. Since reconstructing a Gaussian avatar fully aligned with the input image takes several hours, while training the generalization model only requires diverse 3D-consistent human images, we reduce both the number of generated frames and the training iterations to accelerate the reconstruction process, enabling large-scale dataset construction.

We create the initial image collections using the text-to-image model Flux [32] by designing descriptive prompts. While IDOL [85] generates multi-view images using a 2D video diffusion model, inconsistencies between frames are inevitable, and complex regions, such as fingers, remain problematic (as shown in Fig. 3). These issues hinder the model's ability to learn 3D-consistent features from inputs and to capture fine-grained details. In contrast, our Gaussian avatar dataset naturally provides strong 3D consistency and geometrically stable supervision, significantly enhancing the upper bound of models trained on synthetic data.

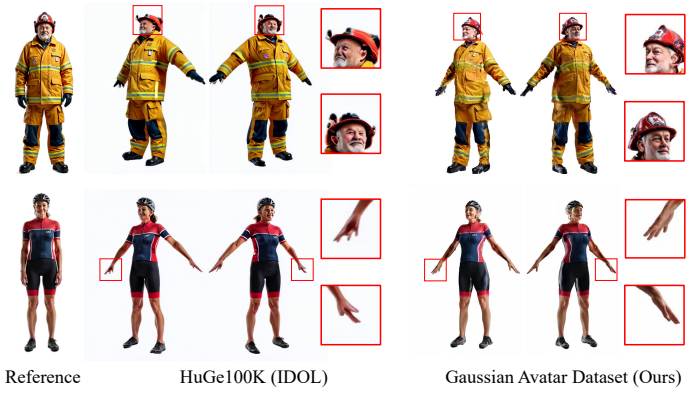

Fig. 3: Qualitative comparisons between two types of datasets. Our Gaussian avatar dataset exhibits strong 3D consistency and rich details, which are crucial for training LRMs. Note that neither dataset needs to be fully aligned with the reference image.

3.6 Objective Functions

Our framework directly generates a UV Gaussian avatar in a feed-forward manner, which is then rendered to the target view and pose to support end-to-end training. The objective function consists of two parts: photometric supervision to ensure rendering fidelity, and Gaussian regularization terms to improve training robustness.

Photometric supervision. In the view space corresponding to the target pose, we adopt the conventional L1 loss and perceptual loss [77] between the ground-

truth images, I_{gt}, and the predicted splatted images, $\mathrm{I}_{\mathrm{pred}}$:

$$\mathcal{L}_{\mathrm{photometric}} = \lambda_{\mathrm{rgb}}\mathcal{L}_1(\mathrm{I}_{\mathrm{gt}}, \mathrm{I}_{\mathrm{pred}}) + \lambda_{\mathrm{per}}\mathcal{L}_{\mathrm{lpips}}(\mathrm{I}_{\mathrm{gt}}, \mathrm{I}_{\mathrm{pred}}). \tag{4}$$

Gaussian Regularization. Given the ill-posed nature of single-image reconstruction and the large pose-induced deformations, geometric constraints are crucial for maintaining stability in the Gaussian field. We adopt three types of Gaussian regularization terms. The first term constrains the positional deformation field, encouraging each Gaussian to remain close to its initial position on the SMPL-X mesh:

$$\mathcal{L}_{\mathrm{pos}} = \frac{1}{N_{\mathrm{gs}}}\sum_{i}^{N_{\mathrm{gs}}} \max(\|\delta_{\mu_i}\|_2 - \epsilon_{\mathrm{pos}}, 0). \tag{5}$$

The second term regularizes the Gaussian scale to avoid needle-shaped ellipsoids:

$$\mathcal{L}_{\mathrm{sca}} = \frac{1}{N_{\mathrm{gs}}}\sum_{i}^{N_{\mathrm{gs}}} \max(\frac{s_i^{\max}}{s_i^{\min}} - \epsilon_{\mathrm{sca}}, 0). \tag{6}$$

Both ϵ_{pos} and ϵ_{sca} represent empirically determined thresholds. Finally, we compute the Laplacian smoothing loss, $\mathcal{L}_{\mathrm{lap}}$, for δ_μ along with the scaling and RGBs of the 3D Gaussians, based on their initial connectivity on the canonical mesh surface, to further enhance the spatial consistency of Gaussians. In particular, we assign a larger weight ($\times 100$) to the Laplacian smoothing term for hand regions. The combined regularization operates as:

$$\mathcal{L}_{\mathrm{reg}} = \lambda_{\mathrm{pos}}\mathcal{L}_{\mathrm{pos}} + \lambda_{\mathrm{sca}}\mathcal{L}_{\mathrm{sca}} + \lambda_{\mathrm{lap}}\mathcal{L}_{\mathrm{lap}}. \tag{7}$$

The total loss function is the sum of the photometric supervision term and the Gaussian regularization term:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{photometric}} + \mathcal{L}_{\mathrm{reg}}. \tag{8}$$

4 Experiments

4.1 Implementation Details

Training Dataset. We construct a large-scale 3D Gaussian avatar dataset containing 28.5k identities, each rendered with 15 frames under different views and poses for training. In addition, we incorporate several public datasets, including the 3D scan dataset THuman2.1 [73] and a subset of the synthetic dataset HuGe100K [85], to improve generalization across identities, viewpoints, and real-world appearances. Furthermore, we collect 5,253 real human head videos to train a dedicated head model for enhanced facial detail reconstruction. All datasets are unified into a common motion space, forming a comprehensive hybrid dataset that is essential for achieving strong generalization and fine-grained control in the final model.

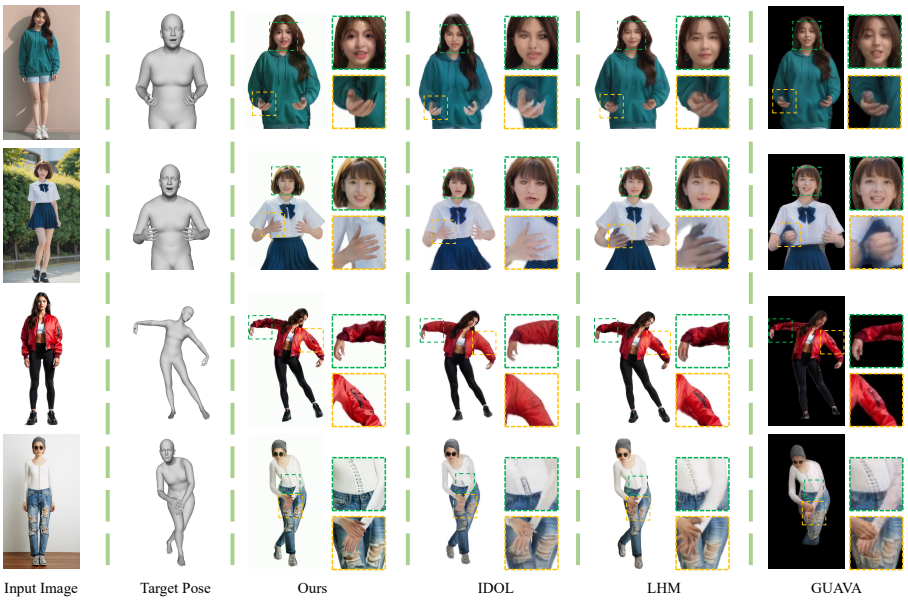

Fig. 4: Qualitative comparisons with representative methods [49, 74, 85] in the novel pose animation task. Our method accurately models facial and hand regions and produces more precise, photo-realistic animation results, with almost all fine details preserved, even compared with methods trained on real video datasets.

Model Configuration. We set the UV map resolution to 768×768 and use approximately 320,000 Gaussians in total. We train our models using the Adam optimizer [31] with an initial learning rate of $4e-4$. A warm-up schedule of 6,000 steps is employed to stabilize training in the initial stages. The entire training process takes about 5 days on 8 NVIDIA A100 GPUs with a batch size of 8. For Gaussian regularization, we set the thresholds to $\epsilon_{\text{pos}} = 5e-2$ and $\epsilon_{\text{sca}} = 5$. For the loss weights, we set λ_{rgb}, λ_{per}, λ_{pos}, λ_{sca}, λ_{lap} to $3e^1$, $1e^1$, $1e^1$, $5e^1$ and $1e^4$, respectively. And for head avatar model training, λ_{rgb} and λ_{per} are reduced to 10 and 5. The head avatar model and the Gaussian blending network are trained sequentially for 15,000 steps on the head and full-body datasets, respectively. The learning rate for the Gaussian blending network is $4e-6$.

4.2 Comparison with Representative Methods

We compare our method with several representative works, including IDOL [85], LHM [49], and GUAVA [74]. All of these methods follow a feed-forward paradigm to generate Gaussian-based 3D avatars from a single image. Since they share the same problem setting as our approach, they serve as appropriate baselines for comparison.

Qualitative Comparisons. Fig. 4 and Fig. 5 present a qualitative comparison

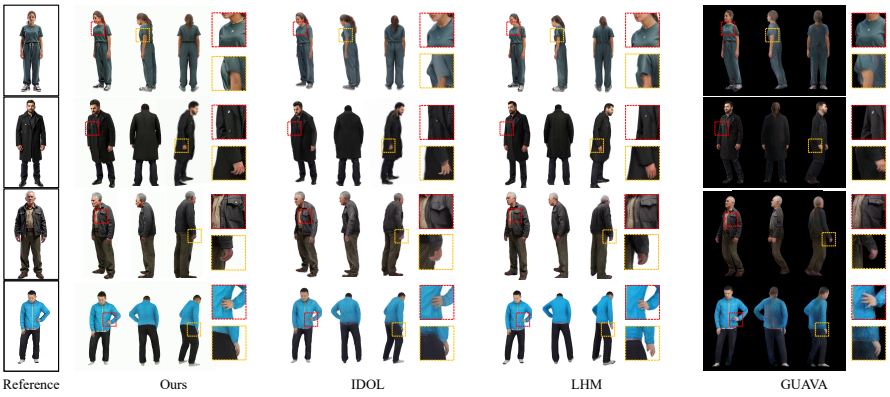

Fig. 5: Qualitative comparisons with representative methods [49, 74, 85] on the novel view synthesis task show that our method achieves superior texture sharpness and appearance fidelity, particularly in clothing textures and hand regions.

between our method and other representative approaches. Fig. 4 illustrates the single-view human animation task, in which the reconstructed avatar is animated under a different driving pose, whereas Fig. 5 shows the results of static novel-view synthesis.

As observed, IDOL relies on a 2D synthetic dataset with incomplete consistency and fails to fully utilize the input information, resulting in coarse reconstructions that struggle to preserve identity and texture. In addition, its optimization stage lacks geometric constraints on the Gaussian field, making it difficult to recover thin structures, such as fingers. LHM similarly suffers from texture degradation and, due to the lack of dynamic expression and gesture modeling, produces severe artifacts on the hands, making it incapable of handling highly expressive talking scenes. GUAVA focuses primarily on frontal talking scenarios and, therefore, cannot achieve full 360-degree body modeling. Its strong reliance on tracking alignment makes it fragile in challenging cases, such as occluded hands. It is further limited by its fixed low resolution and struggles to synthesize detailed textures, even in upper-body driving settings.

In contrast, our method generates highly expressive 3D avatars from a single input image. The resulting avatars support consistent 360° reconstruction and diverse pose-driven animations while maintaining high-fidelity textures and enabling precise modeling of facial expressions, teeth, and hand gestures.

Quantitative Comparison. Tab. 1 presents a quantitative comparison between our model and other methods on THuman2.0 [73] (100 identities in test-set), NeuMan [28] (all 6 identities) and Seamless Interaction [1] (randomly selected 100 videos) datasets. We use four common metrics for comparison: Mean Squared Error, PSNR, SSIM, and LPIPS. Our method outperforms all others across these metrics, demonstrating superior realism and naturalness in the results. However, we believe that these metrics do not fully reflect the quality or capabilities of the methods, especially for the one-shot image animation task. We

conduct a user study for subjective evaluation with 32 participants, showing the best percentage for each method in Tab. 2. We also encourage readers to refer to the visualization results for a more comprehensive and objective evaluation.

Table 1: Quantitative comparisons with representative methods [49, 74, 85] on various datasets [1, 28, 73].

Dataset	Method	MSE($\times 10^{-3}$) ↓	PSNR ↑	SSIM ↑	LPIPS ↓
THuman2.0	IDOL	10.841	19.995	0.912	0.0907
	GUAVA	9.727	20.239	0.886	0.0765
	LHM	9.951	20.276	0.917	0.0858
	Ours	**6.440**	**22.542**	**0.931**	**0.0653**
NeuMan	IDOL	7.262	22.072	0.961	0.0424
	GUAVA	7.365	21.875	0.904	0.0666
	LHM	6.469	23.139	**0.964**	0.0332
	Ours	**5.904**	**23.489**	**0.964**	**0.0329**
Seamless	IDOL	4.551	23.685	0.929	0.0650
	GUAVA	4.483	24.125	0.929	0.0645
	LHM	3.776	24.383	0.935	0.0582
	Ours	**3.595**	**24.858**	**0.936**	**0.0542**

Table 2: User study: Percentage of methods rated the best.

Method	Face	Hand	Clothed Body	Overall Quality
IDOL	4.8	3.9	2.9	1.9
GUAVA	13.5	20.4	11.5	9.6
LHM	17.3	7.8	10.6	10.6
Ours	**64.4**	**67.9**	**75.0**	**77.9**

4.3 Ablation Studies

Facial Enhancement. Models trained on large-scale full-body data can reconstruct basic facial structure and texture. However, our facial enhancement module further improves the modeling of facial dynamics, enabling clearer reconstruction of fine details, such as eyes and teeth, as shown in Fig. 6. The resulting fully controllable 3D avatars, combined with naturally modeled hands, can support expressive talking scenarios that require high realism. It is worth noting that in other ablation studies, we did not apply the facial enhancement module, and for quantitative comparisons, we masked out the facial regions to ensure fairness.

Gaussian Avatar Dataset. As shown in Fig. 6, when the model is primarily

trained on the inconsistent 2D dataset [85], it tends to produce blurry results with noticeable texture-bleeding artifacts. Moreover, since hand regions are often problematic in such datasets, the model struggles to reconstruct natural and coherent hand appearances.

Fig. 6: Visualization of ablation studies on the importance of Gaussian avatar dataset and the facial enhancement module.

Input-aware Decoder. We conduct a detailed ablation study on the design of the Input-aware Decoder, as illustrated in Fig. 7. It can be observed that, without leveraging the rich information from the input image during decoding, the model produces reasonable yet coarse results. When using only the direct mesh-based inverse process, tracking errors lead to noticeable texture artifacts and prevent the recovery of regions distant from the mesh surface, such as hair. Introducing the Gaussian-based inverse process for color decoding alleviates this issue, but without semantic correction, residual texture misalignments still remain. Our full model achieves natural reconstruction, with accurate texture capture across both the body and fine-grained regions.

Besides the visual results shown in Figs. 6 and 7, quantitative evaluation is also conducted, as presented in Tab. 3.

Table 3: Quantitative results of ablation studies on [1].

Metrics	MSE ($\times 10^{-3}$) ↓	PSNR ↑	SSIM ↑	LPIPS ↓
w/o-GA-data	5.283	23.037	0.9326	0.0676
w/o-input-aware	5.281	23.114	0.9328	0.0684
w/o-GauInverse	5.846	22.733	0.9320	0.0669
w/o-semantic	3.953	24.477	0.9328	0.0583
Full-model	**3.527**	**24.949**	**0.9365**	**0.0527**

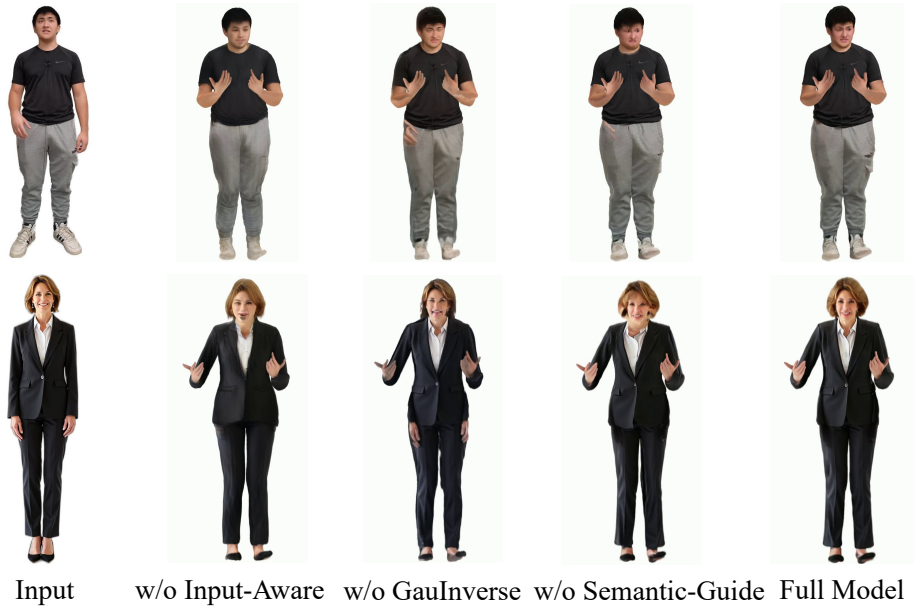

Fig. 7: Visualization of ablation studies on the input-awareness of the decoder and the critical components of the Input-aware Decoder.

5 Conclusion

Summary. In this paper, we introduced a comprehensive pipeline for creating fully controllable avatars from a single image. This was achieved through innovations in dataset construction, model architecture, and algorithmic design, including a large-scale, 3D-consistent Gaussian avatar dataset, an input-aware decoding scheme that establishes a tight coupling between Gaussian geometry and color and fully exploits input information to recover precise appearance, and a Gaussian-blending-based facial enhancement module that further improves the expressiveness of reconstructed human avatars. Experimental results demonstrate that our method outperforms existing techniques in fine-grained human animation and appearance detail preservation. With its high-quality reconstruction and ability to generate expressive animations, our approach holds strong potential for practical 3D avatar applications across a wide range of scenarios.

Limitations. Since our method is primarily trained on synthetic data, it may struggle to robustly handle inputs with complex lighting conditions. In addition, the current generalization model cannot faithfully reconstruct fine protruding structures, such as individual hair strands. Future work could focus on enhancing the dataset construction pipeline to incorporate richer illumination variations and more diverse cases, as well as exploring more efficient strategies for personalized adaptation to better capture subject-specific details.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (No.62272433, No.62402468, No.U25A20390), Anhui Provincial Natural Science Foundation (No.2508085ZD011), Key Science & Technology Project of Anhui Province (No.202523o09050004), Industry-University-Research Fund Project (IA20260624014), the Fundamental Research Funds for the Central Universities and the advanced computing resources provided by the Supercomputing Center of the USTC.

References

1. Agrawal, V., Akinyemi, A., Alvero, K., Behrooz, M., Buffalini, J., Carlucci, F.M., Chen, J., Chen, J., Chen, Z., Cheng, S., et al.: Seamless interaction: Dyadic audio-visual motion modeling and large-scale dataset. arXiv preprint arXiv:2506.22554 (2025) 11, 12, 13
2. AlBahar, B., Saito, S., Tseng, H.Y., Kim, C., Kopf, J., Huang, J.B.: Single-image 3d human digitization with shape-guided diffusion. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–11 (2023) 3
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023) 1, 4
4. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 1, 4
5. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 1
6. Chen, J., Li, C., Zhang, J., Zhu, L., Huang, B., Chen, H., Lee, G.H.: Generalizable human gaussians from single-view image. arXiv preprint arXiv:2406.06050 (2024) 4
7. Cheng, G., Gao, X., Hu, L., Hu, S., Huang, M., Ji, C., Li, J., Meng, D., Qi, J., Qiao, P., et al.: Wan-animate: Unified character animation and replacement with holistic replication. arXiv preprint arXiv:2509.14055 (2025) 1, 4
8. Chu, X., Harada, T.: Generalizable and animatable gaussian head avatar. Advances in Neural Information Processing Systems **37**, 57642–57670 (2024) 4
9. Corona, E., Zanfir, M., Alldieck, T., Bazavan, E.G., Zanfir, A., Sminchisescu, C.: Structured 3d features for reconstructing controllable avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16954–16964 (2023) 3
10. Deng, Y., Wang, D., Wang, B.: Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. In: European Conference on Computer Vision. pp. 316–333. Springer (2024) 4
11. Dong, Z., Chen, X., Yang, J., Black, M.J., Hilliges, O., Geiger, A.: Ag3d: Learning to generate 3d avatars from 2d image collections. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14916–14927 (2023) 3

12. Dong, Z., Duan, L., Song, J., Black, M.J., Geiger, A.: Moga: 3d generative avatar prior for monocular gaussian avatar reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13304–13314 (2025) 2, 4
13. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024) 1, 4
14. Feng, M., Liu, J., Yu, K., Yao, Y., Hui, Z., Guo, X., Lin, X., Xue, H., Shi, C., Li, X., et al.: Dreamoving: A human video generation framework based on diffusion models. arXiv e-prints pp. arXiv–2312 (2023) 4
15. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7297–7306 (2018) 1
16. Guo, J., Zhang, D., Liu, X., Zhong, Z., Zhang, Y., Wan, P., Zhang, D.: Liveportrait: Efficient portrait animation with stitching and retargeting control. arXiv preprint arXiv:2407.03168 (2024) 4
17. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning (2023) 1, 4
18. He, Y., Gu, X., Ye, X., Xu, C., Zhao, Z., Dong, Y., Yuan, W., Dong, Z., Bo, L.: Lam: Large avatar model for one-shot animatable gaussian head. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–13 (2025) 4
19. Ho, I., Song, J., Hilliges, O., et al.: Sith: Single-view textured human reconstruction with image-conditioned diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 538–549 (2024) 3
20. Hong, Y., Zhang, J., Jiang, B., Guo, Y., Liu, L., Bao, H.: Stereopifu: Depth aware clothed human digitization via stereo vision. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 3
21. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. In: International Conference on Learning Representations. vol. 2024, pp. 50678–50702 (2024) 2, 4
22. Hu, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8153–8163 (2024) 1, 4
23. Hu, S., Hong, F., Pan, L., Mei, H., Yang, L., Liu, Z.: Sherf: Generalizable human nerf from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9352–9364 (2023) 4
24. Huang, Y., Yi, H., Liu, W., Wang, H., Wu, B., Wang, W., Lin, B., Zhang, D., Cai, D.: One-shot implicit animatable avatars with model-based priors. In: IEEE Conference on Computer Vision (ICCV) (2023) 4
25. Huang, Y., Yi, H., Xiu, Y., Liao, T., Tang, J., Cai, D., Thies, J.: Tech: Text-guided reconstruction of lifelike clothed humans. In: 2024 International Conference on 3D Vision (3DV). pp. 1531–1542. IEEE (2024) 4
26. Jiang, J., Zeng, W., Zheng, Z., Yang, J., Liang, C., Liao, W., Liang, H., Zhang, Y., Gao, M.: Omnihuman-1.5: Instilling an active mind in avatars via cognitive simulation. arXiv preprint arXiv:2508.19209 (2025) 1, 4

27. Jiang, S., Jiang, H., Wang, Z., Luo, H., Chen, W., Xu, L.: Humangen: Generating human radiance fields with explicit priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12543–12554 (2023) 3
28. Jiang, W., Yi, K.M., Samei, G., Tuzel, O., Ranjan, A.: Neuman: Neural human radiance field from a single video. In: Proceedings of the European conference on computer vision (ECCV) (2022) 11, 12
29. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023) 2
30. Khirodkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models. arXiv preprint arXiv:2408.12569 (2024) 3, 5
31. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 10
32. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024) 2, 8
33. Li, H., Li, Y., Yang, Y., Cao, J., Zhu, Z., Cheng, X., Chen, L.: Dispose: Disentangling pose guidance for controllable human image animation. arXiv preprint arXiv:2412.09349 (2024) 1, 4
34. Li, P., Zheng, W., Liu, Y., Yu, T., Li, Y., Qi, X., Chi, X., Xia, S., Cao, Y.P., Xue, W., et al.: Pshuman: Photorealistic single-image 3d human reconstruction using cross-scale multiview diffusion and explicit remeshing. In: Proceedings of the computer vision and pattern recognition conference. pp. 16008–16018 (2025) 2, 3
35. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6), 194:1–194:17 (2017), https://doi.org/10.1145/3130800.3130813 2
36. Liu, Z., Dong, H., Chharia, A., Wu, H.: Human-vdm: Learning single-image 3d human gaussian splatting from video diffusion models. arXiv preprint arXiv:2409.02851 (2024) 3
37. Lu, Y., Dong, J., Kwon, Y., Zhao, Q., Dai, B., De la Torre, F.: Gas: Generative avatar synthesis from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12883–12893 (2025) 4
38. Luo, Y., Rong, Z., Wang, L., Zhang, L., Hu, T.: Dreamactor-m1: Holistic, expressive and robust human image animation with hybrid guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11036–11046 (2025) 4
39. Men, Y., Lei, B., Yao, Y., Cui, M., Lian, Z., Xie, X.: En3d: An enhanced generative model for sculpting 3d humans from 2d synthetic data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9981–9991 (2024) 3
40. Men, Y., Yao, Y., Cui, M., Bo, L.: Mimo: Controllable character video synthesis with spatial decomposed modeling. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21181–21191 (2025) 1, 4
41. Meng, R., Wang, Y., Wu, W., Zheng, R., Li, Y., Ma, C.: Echomimicv3: 1.3 b parameters are all you need for unified multi-modal and multi-task human animation. arXiv preprint arXiv:2507.03905 (2025) 4
42. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020) 2
43. Pan, P., Su, Z., Lin, C., Fan, Z., Zhang, Y., Li, Z., Shen, T., Mu, Y., Liu, Y.: Humansplat: Generalizable single-image human gaussian splatting with structure

priors. Advances in Neural Information Processing Systems **37**, 74383–74410 (2024) 2, 3, 4
44. Pang, H., Zhu, H., Kortylewski, A., Theobalt, C., Habermann, M.: Ash: Animatable gaussian splats for efficient and photoreal human rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1165–1175 (2024) 5
45. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019) 1, 2
46. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023) 1, 4
47. Peng, B., Wang, J., Zhang, Y., Li, W., Yang, M.C., Jia, J.: Controlnext: Powerful and efficient control for image and video generation. arXiv preprint arXiv:2408.06070 (2024) 4
48. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022) 3
49. Qiu, L., Gu, X., Li, P., Zuo, Q., Shen, W., Zhang, J., Qiu, K., Yuan, W., Chen, G., Dong, Z., et al.: Lhm: Large animatable human reconstruction model for single image to 3d in seconds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14184–14194 (2025) 2, 4, 10, 11, 12
50. Qiu, L., Zhu, S., Zuo, Q., Gu, X., Dong, Y., Zhang, J., Xu, C., Li, Z., Yuan, W., Bo, L., et al.: Anigs: Animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21148–21158 (2025) 2, 4
51. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 1
52. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2304–2314 (2019) 3
53. Saito, S., Simon, T., Saragih, J., Joo, H.: 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. pp. 84–93 (2020) 3
54. Sim, G., Moon, G.: Persona: Personalized whole-body 3d avatar with pose-driven deformations from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12670–12680 (2025) 2, 4, 7
55. Tan, S., Gong, B., Wang, X., Zhang, S., Zheng, D., Zheng, R., Zheng, K., Chen, J., Yang, M.: Animate-x: Universal character image animation with enhanced motion representation. ICLR 2025 (2025) 4
56. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In: European Conference on Computer Vision. pp. 1–18. Springer (2024) 2, 4
57. Tu, S., Xing, Z., Han, X., Cheng, Z.Q., Dai, Q., Luo, C., Wu, Z.: Stableanimator: High-quality identity-preserving human image animation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21096–21106 (2025) 1, 2, 4

58. Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10039–10049 (2021) 4
59. Wang, X., Zhang, S., Tang, L., Zhang, Y., Gao, C., Wang, Y., Sang, N.: Unianimate-dit: Human image animation with large-scale video diffusion transformer. arXiv preprint arXiv:2504.11289 (2025) 4
60. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 5
61. Weng, Z., Liu, J., Tan, H., Xu, Z., Zhou, Y., Yeung-Levy, S., Yang, J.: Template-free single-view 3d human digitalization with diffusion-guided lrm. arXiv preprint arXiv:2401.12175 (2024) 4
62. Wu, Y., Chen, X., Li, W., Jia, S., Wei, H., Feng, K., Chen, J., Li, Y., He, A., Zhang, W., et al.: Sings: Animatable single-image human gaussian splats with kinematic priors. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5571–5580 (2025) 4
63. Xiang, J., Gao, X., Guo, Y., Zhang, J.: Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1802–1812 (2024) 5
64. Xiang, J., Guo, Y., Hu, L., Guo, B., Yuan, Y., Zhang, J.: Expressive talking human from single-image with imperfect priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10398–10409 (2025) 2, 4, 7
65. Xiong, Z., Kang, D., Jin, D., Chen, W., Bao, L., Cui, S., Han, X.: Get3dhuman: Lifting stylegan-human into a 3d generative model using pixel-aligned reconstruction priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023) 3
66. Xiu, Y., Yang, J., Cao, X., Tzionas, D., Black, M.J.: Econ: Explicit clothed humans optimized via normal integration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 512–523 (2023) 3
67. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: Icon: Implicit clothed humans obtained from normals. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13286–13296. IEEE (2022) 3
68. Xiu, Y., Ye, Y., Liu, Z., Tzionas, D., Black, M.J.: Puzzleavatar: Assembling 3d avatars from personal albums. ACM Transactions on Graphics (TOG) (2024) 4
69. Xu, Z., Zhang, J., Liew, J.H., Yan, H., Liu, J.W., Zhang, C., Feng, J., Shou, M.Z.: Magicanimate: Temporally consistent human image animation using diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1481–1490 (2024) 4
70. Xu, Z., Yu, Z., Zhou, Z., Zhou, J., Jin, X., Hong, F.T., Ji, X., Zhu, J., Cai, C., Tang, S., et al.: Hunyuanportrait: Implicit condition control for enhanced portrait animation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15909–15919 (2025) 4
71. Yang, X., Chen, X., Gao, D., Wang, S., Han, X., Wang, B.: Have-fun: Human avatar reconstruction from few-shot unconstrained images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 742–752 (2024) 4
72. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4210–4220 (2023) 1

73. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5746–5756 (2021) 2, 9, 11, 12
74. Zhang, D., Liu, Y., Lin, L., Zhu, Y., Li, Y., Qin, M., Li, Y., Wang, H.: Guava: Generalizable upper body 3d gaussian avatar. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14205–14217 (2025) 2, 4, 6, 10, 11, 12
75. Zhang, J., Li, X., Zhang, Q., Cao, Y., Shan, Y., Liao, J.: Humanref: Single image to 3d human generation via reference-guided diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1844–1854 (2024) 4
76. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023) 1
77. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 8
78. Zhang, W., Yan, Y., Liu, Y., Sheng, X., Yang, X.: E 3gen: Efficient, expressive and editable avatars generation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 6860–6869 (2024) 2, 4
79. Zhang, Y., Gu, J., Wang, L.W., Wang, H., Cheng, J., Zhu, Y., Zou, F.: Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. In: International Conference on Machine Learning (2025) 2, 4, 7
80. Zhang, Z., Sun, L., Yang, Z., Chen, L., Yang, Y.: Global-correlated 3d-decoupling transformer for clothed avatar reconstruction. In: Advances in Neural Information Processing Systems (NeurIPS) (2023) 3
81. Zhang, Z., Yang, Z., Yang, Y.: Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9936–9947 (June 2024) 2, 3
82. Zhao, X., Xu, H., Song, G., Xie, Y., Zhang, C., Li, X., Luo, L., Suo, J., Liu, Y.: X-nemo: Expressive neural motion reenactment via disentangled latent attention. arXiv preprint arXiv:2507.23143 (2025) 4
83. Zheng, Z., Yu, T., Liu, Y., Dai, Q.: Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. IEEE transactions on pattern analysis and machine intelligence (2021) 3
84. Zhu, S., Chen, J.L., Dai, Z., Su, Q., Xu, Y., Cao, X., Yao, Y., Zhu, H., Zhu, S.: Champ: Controllable and consistent human image animation with 3d parametric guidance. arXiv preprint arXiv:2403.14781 (2024) 4
85. Zhuang, Y., Lv, J., Wen, H., Shuai, Q., Zeng, A., Zhu, H., Chen, S., Yang, Y., Cao, X., Liu, W.: Idol: Instant photorealistic 3d human creation from a single image. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26308–26319 (2025) 2, 4, 6, 8, 9, 10, 11, 12, 13

SplatCtrlA: Generalizable Single Image to Fully Controllable 3D Avatar

Supplementary Material

A Video Demo

We highly recommend readers watch the video demo in the supplementary materials for a more comprehensive and objective evaluation. In the video, we provide a concise overview of the problem addressed in this work and our proposed pipeline, followed by demonstrations of diverse novel-view synthesis and free-pose animation results. In addition, we compare our method with representative approaches across three key tasks: novel-view synthesis (NVS), half-body talking reenactment, and large-pose full-body animation. Finally, we present visual results from our ablation studies and showcase additional examples across various styles. These results demonstrate that our method enables one-shot reconstruction of photorealistic avatars with whole-body control, outperforming existing works in terms of human dynamic and appearance details across various poses.

B Additional Results

Robustness to challenging inputs. Fig. 8 demonstrates our robustness to challenging inputs, including large occlusions, articulations, and even seated poses.

Fig. 8: Our method shows great robustness to challenging inputs.

Comparison with Wan2.2-Animate. We compare with Wan2.2-Animate in Fig. 9. Despite good motion simulation, Wan2.2-Animate exhibits 2D inconsistencies (e.g., texture artifacts, missing details) and struggles with complex 3D geometries like hands under challenging driving, which our 3D-aware approach avoids.

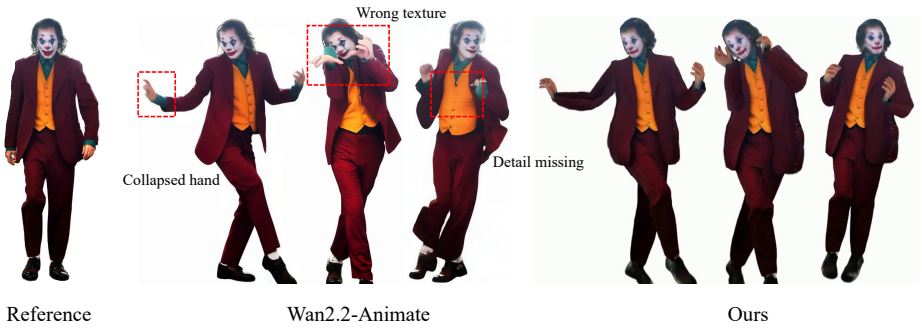

Fig. 9: Comparison between our method and Wan2.2-Animate.

C Implementation Details

C.1 Network Architecture

Image UV Encoder. We utilize the backbone of the frozen Sapiens-1B model for image feature encoding. For the UV Transformer, we adopt a ViT-based architecture similar to IDOL, consisting of $D = 16$ transformer layers with a width of 1536, which produces a 96-dimensional UV feature map. The subsequent upsampling network comprises three transposed convolutional layers (channels {512, 512, 256}, kernel 4), each doubling the spatial resolution, followed by three convolutional layers with output channels $\{128, 128, 32\}$ to further refine the features, yielding the final feature map $\mathbf{F}_{\text{uv}} \in \mathbb{R}^{768\times768\times32}$.

Input-aware Decoder. The geometry decoder and the color decoder are implemented as two separate convolutional networks, which transform the 768×768 inverse UV RGB and the UV feature map $\mathbf{F}_{\text{uv}}$ into their corresponding Gaussian attribute maps. In addition, for color decoding, the inverse UV image I_{uv} is first processed by a StyleUNet, which paints the invisible regions, before being concatenated with $\mathbf{F}_{\text{uv}}$ and passed to the color decoder.

Gaussian Blending Network. The Gaussian blending network consists of a StyleUNet followed by a "zero convolution", namely a 1×1 convolution layer whose weights and biases are both initialized to zero. The schematic illustration of the facial enhancement module can be found in Fig. 10.

D Broader Impact

Our work enables the reconstruction of fully controllable 3D avatars from a single photo, allowing for realistic animations with expressive body gestures and natural expression changes. We consider this a significant advancement in the research and practical applications of multimodal digital humans. However, this technology carries the risk of misuse, such as generating fake videos of individuals

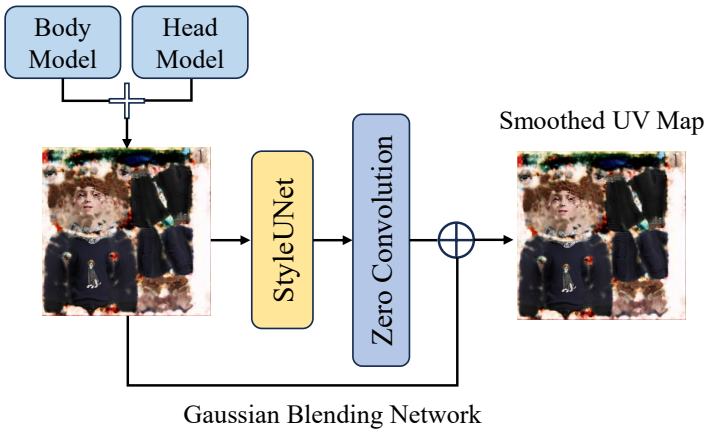

Fig. 10: Visualization of the facial enhancement module.

to spread false information or harm reputations. We strongly condemn such unethical applications. While it may not be possible to entirely prevent malicious use, we believe that conducting research in an open and transparent manner can help raise public awareness of potential risks. Additionally, we hope our work can inspire further advancements in forgery detection technologies.