SVAD: From Single Image to 3D Avatar via Synthetic Data Generation with Video Diffusion and Data Augmentation

Yonwoo Choi
SECERN AI
yonwoo.choi@secern.ai
https://yc4ny.github.io/SVAD/

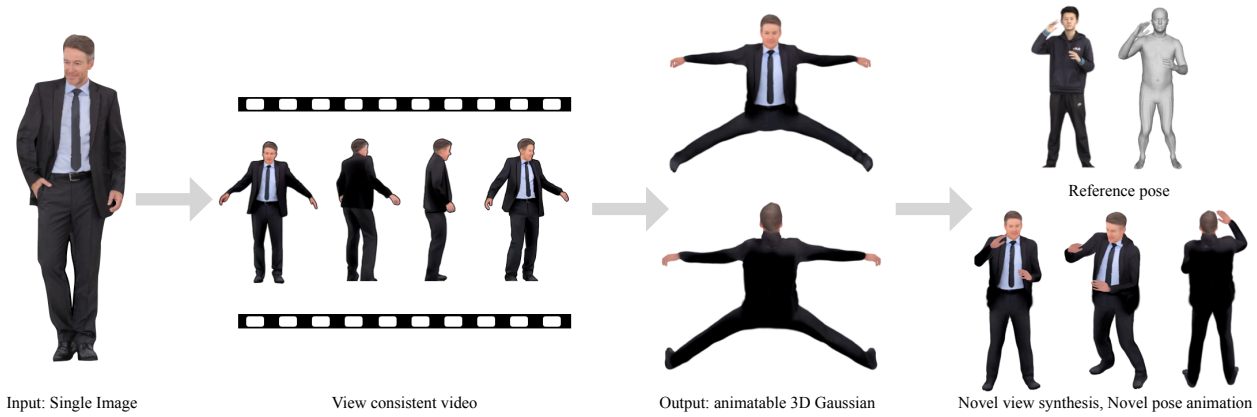


Figure 1. **SVAD.** Our method creates high-fidelity 3D avatars from a single image through synthetic data generation. We leverage video diffusion to generate pose-conditioned animations, enhance them with identity preservation and image restoration modules, then train a 3D Gaussian Splatting avatar. The resulting avatars maintain consistent identity across novel poses and viewpoints while enabling real-time rendering, outperforming state-of-the-art approaches.

Abstract

Creating high-quality animatable 3D human avatars from a single image remains a significant challenge in computer vision due to the inherent difficulty of reconstructing complete 3D information from a single viewpoint. Current approaches face a clear limitation: 3D Gaussian Splatting (3DGS) methods produce high-quality results but require multiple views or video sequences, while video diffusion models can generate animations from single images but struggle with consistency and identity preservation. We present SVAD, a novel approach that addresses these limitations by leveraging complementary strengths of existing techniques. Our method generates synthetic training data through video diffusion, enhances it with identity preservation and image restoration modules, and utilizes this refined data to train 3DGS avatars. Comprehensive evaluations demonstrate that SVAD outperforms state-of-the-art (SOTA) single-image methods in maintaining identity consistency and fine details across novel poses and viewpoints, while enabling real-time rendering capabilities. Through our data augmentation pipeline, we overcome the dependency on dense monocular or multi-view training data typically required by traditional 3DGS approaches. Extensive quantitative, qualitative comparisons show our method achieves superior performance across multiple metrics against baseline models. By effectively combining the generative power of diffusion models with both the high-quality results and rendering efficiency of 3DGS, our work establishes a new approach for high-fidelity avatar generation from a single image input.

1. Introduction

The ability to generate animatable 3D human avatars from minimal input data, such as a single-image, has significant potential across a range of applications. Traditional methods, particularly those based on 3DGS, have demonstrated considerable success in producing high-quality avatars [14, 34, 68, 69, 77, 86, 87, 93, 109, 120]. These methods rely on dense input data, typically monocular or multi-view

video [14, 34, 68, 77, 86, 93, 120], to achieve high fidelity across varied viewpoints and poses. This reliance on extensive video input complicates deployment in single-image scenarios, where ensuring viewpoint consistency and adaptability to novel poses becomes a key challenge.

Recent advancements in video diffusion models offer a potential solution by enabling animation generation from a single static image [33, 49, 97, 103, 121]. These models use certain conditions in diffusion processes to create video sequences, demonstrating the powerful generative capabilities of diffusion for single-image-driven animation. However, diffusion models often struggle to maintain temporal coherence, leading to inconsistent features and identity drift across frames [5, 19, 31, 90] . Additionally, their iterative denoising process for each frame introduces significant computational overhead, limiting their feasibility for real-time or interactive applications where rapid rendering across novel views is essential.

To overcome these challenges, we propose SVAD, a novel synthetic data generation and avatar creation pipeline that synergizes the generative flexibility of diffusion models with the efficient rendering capabilities of 3DGS avatars. Our approach leverages video diffusion model [94] to generate diverse pose-conditioned synthetic training data from a single-image. This synthetic data is refined through an identity-preservation module and an image restoration module to ensure that perceptual identity consistency and structural fidelity are preserved across diverse poses and temporal sequences. The resulting high-quality synthetic dataset is then used to train a 3DGS avatar model [68], which benefits from the rapid rendering capabilities inherent to 3DGS. By combining the generative strengths of diffusion for synthetic data creation with the efficiency of 3DGS for rendering, SVAD achieves consistent, high-quality 3D avatar animations from single-image input.

In summary, our main contributions are:

- We introduce a novel pipeline that generates high-quality synthetic training data from a single-image to create detailed, animatable 3D human avatars.
- We develop a comprehensive data augmentation approach that combines identity preservation and image restoration to ensure consistent identity and fine details across diverse poses.
- We demonstrate through extensive experiments that our synthetic data-driven approach significantly outperforms SOTA single-image avatar generation methods in identity preservation and novel pose adaptation while maintaining efficient real-time rendering.

2. Related Work

Diffusion Model for Human Image Animation The use of diffusion models has led to significant advancements in human image animation, enabling the generation of realistic and temporally consistent animations from static images [1, 6, 9, 22, 35, 41, 75, 79, 84, 88, 89, 106, 108, 113, 116]. Early methods, such as PIDM [4] and DreamPose [47], focused on improving texture fidelity by employing texture diffusion modules to align texture patterns between reference and target images. These methods, while enhancing detail preservation, still face challenges in maintaining temporal stability across frames.

Recent works, including DisCo [97] and Animate Anyone [33], have extended diffusion models to improve temporal consistency and fine-grained control in human animation tasks. DisCo leverages dual ControlNets [112] to separately control pose and background elements, providing more robust conditioning for complex motion sequences. Similarly, Animate Anyone integrates a ReferenceNet with temporal attention layers to ensure appearance consistency and smooth transitions across frames, thereby addressing flickering issues commonly observed in earlier models.

Dynamic 3D Gaussian based Avatars The concept of Gaussian splatting for 3D avatars has emerged recently as an innovative approach to explicit scene representation [48]. This technique models a scene as a collection of 3D Gaussian elements, each containing photometric and geometric properties. During rendering, these Gaussian splats are projected onto the image plane, creating the final rendered output. The efficiency of 3DGS has been demonstrated in both static [37, 44, 57] and dynamic [20, 46, 55, 58, 66] scenes, making it a versatile tool for various applications. Recent advancements [7, 8, 13, 18, 36, 39, 59, 76, 77, 96, 122] have explored the use of 3DGS to create photorealistic human avatars across different scenarios. These methods commonly rely on multi-view data [61, 72, 118] or monocular video [34, 39, 59, 68, 77] as input to achieve high-quality, consistent results. The advantage of 3DGS lies in its ability to produce temporally stable animated avatars with superior quantitative metrics.

3. Method

To generate high-quality human avatars from a single-image, facilitating free-viewpoint rendering and realistic animation, we integrate the generative capabilities of video diffusion models with the rendering efficiency of 3D Gaussian-based avatars. We start by leveraging a pretrained video diffusion model [94] for character animation to produce initial synthetic data, as described in Sec. 3.1. Directly using these frames to train a 3DGS avatar model [68], however, often yields poor results, with challenges in preserving facial identity, clothing details, and maintaining consistent multi-view coherence across side and back views. To address these issues and enhance avatar quality, we introduce a data augmentation pipeline in Sec. 3.2 comprising identity-preservation and image-restoration modules to refine the diffusion outputs. With the augmented synthetic

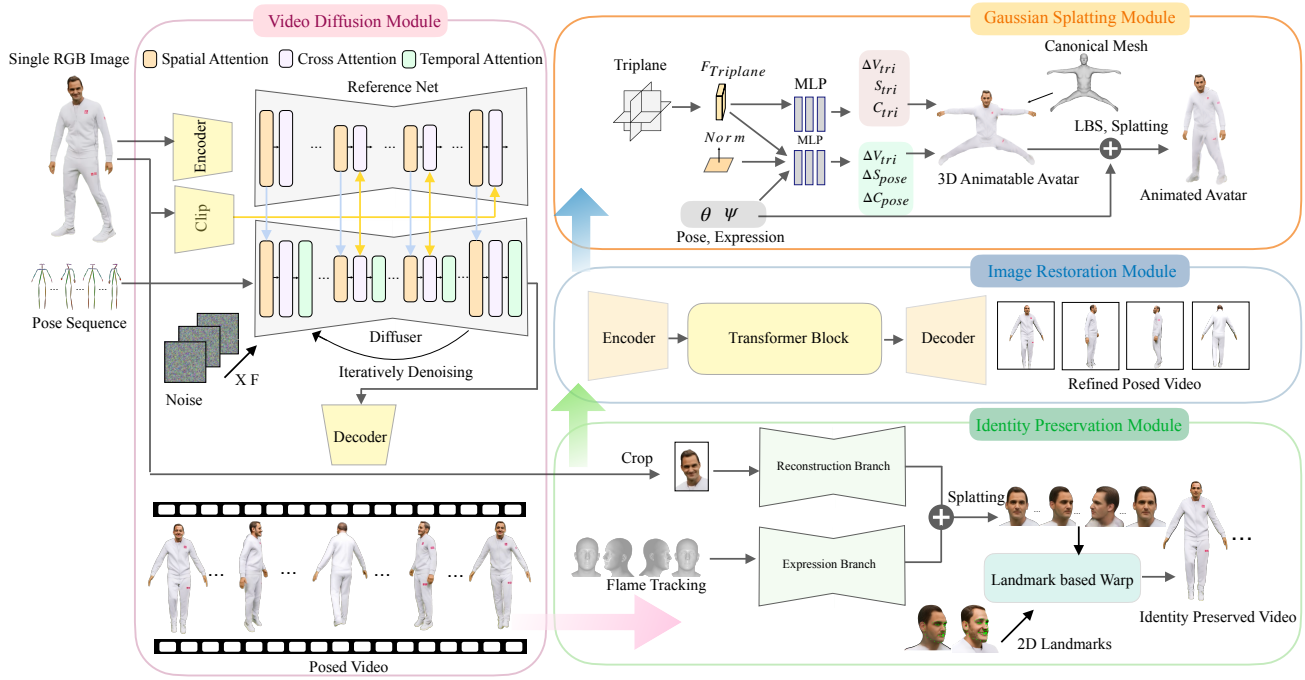


Figure 2. **Overall Pipeline of SVAD.** Starting from a single input image, the diffusion model generates pose-conditioned animations, which are refined using an identity preservation module and an image restoration module. The refined outputs are then used to train the 3DGS avatar, enabling high-fidelity, animatable 3D avatars with consistent details across poses and viewpoints.

data, we proceed to train a 3DGS avatar model, as outlined in Sec. 3.3. The following sections detail the technical methodologies employed in our approach.

3.1. Video Diffusion Module

To generate an animated character video V from a single input image I, we leverage MusePose [94], a finetuned variant of Animate Anyone [33], which is a SOTA video diffusion model designed for realistic human animation while maintaining temporal consistency and appearance fidelity. MusePose employs a U-Net [81]-based diffusion architecture with integrated pose and temporal controls, allowing for pose-guided animation across frames. For our pipeline, we utilize a pose sequence video from a sequence from the People Snapshot [2] dataset, which depicts a subject performing a full-body rotation with arms extended horizontally. This sequence results in 189 frames that serve as pose inputs to the MusePose video diffusion model.

The model architecture incorporates several key components for effective character animation. The denoising UNet is implemented as a 3D UNet [16] with motion modules for temporal coherence. Specifically, we use Vanilla motion modules [26, 27] with temporal self-attention blocks at resolutions of [1, 2, 4, 8] and in the mid-block. Each transformer [95] block contains 8 attention heads, with temporal position encoding enabling positional awareness across a sequence of up to 128 frames. To incorporate pose guidance, a lightweight Pose Guider encodes the motion control signal from the predefined 2D keypoints into a pose-aligned latent representation $P(p_t) \in \mathbb{R}^{H \times W \times C}$. For a pose feature $p_t \in \mathbb{R}^{J \times 2}$ at time t, where J is the number of keypoints, we align the encoding to ensure continuity between frames by adding this encoded pose signal to the noise latent z_t:

$$z_t = z_t + P(p_t) \tag{1}$$

For the diffusion process, we adopt a v-prediction [83] formulation with zero-SNR sampling [63], using a scaled linear beta schedule with $\beta_{\text{start}} = 0.00085$ and $\beta_{\text{end}} = 0.012$. The DDIM [91] sampler is configured for efficient inference with 20 sampling steps and a classifier-free guidance [30] scale of 3.5.

A critical challenge in character animation is ensuring anatomical consistency between the reference image and the motion poses. Direct application of pose control can result in unnatural animations due to mismatches in body proportions [9]. Therefore, we employ a comprehensive pose alignment procedure that adapts the source pose to match the reference character's physical characteristics.

Given a reference pose P_{ref} and a source pose P_{src} detected using DWpose [105], we compute scale parameters $\mathbf{S} = \{s_1, s_2, \ldots, s_{10}\}$ for ten distinct body regions: neck, face, shoulders, upper arms, lower arms, hands, torso, up-

Figure 3. **3D Avatars trained by SVAD.** SVAD generates high quality 3D avatars with just a single-image. The trained avatars can be rendered from any view point, in any pose.

per legs, and lower legs. For each body part i, we compute its scale factor s_i as the ratio between the corresponding keypoint distances. For body parts with bilateral symmetry (e.g., arms), we average the scales from both sides:

$$s_{arm\_upper} = \frac{1}{2}\left(\frac{\|p^2_{ref} - p^3_{ref}\|}{\|p^2_{src} - p^3_{src}\|} + \frac{\|p^5_{ref} - p^6_{ref}\|}{\|p^5_{src} - p^6_{src}\|}\right) \quad (2)$$

To apply these scales to the source pose, we use a rotation matrix transformation centered at anchor points specific to each body part:

$$p' = c_i + s_i \cdot (p - c_i) \quad (3)$$

where c_i is the anchor center for part i. This hierarchical approach ensures body proportions match the reference while maintaining the overall pose structure.

3.2. Data Augmentation Module

Training the 3DGS model using only outputs from the video diffusion model often results in low-fidelity avatars, particularly in terms of facial details and high-frequency features like hands and clothing. To address these challenges, we introduce a data augmentation module that enhances the quality of the training data. This module includes an identity preservation sub-module ensuring coherence in facial details across frames and a image restoration submodule which refines texture quality and high-frequency details, resulting in more realistic textures. This comprehensive data augmentation significantly improves the synthetic training data, enabling the 3DGS avatar model integrated in the future to generate more realistic and detailed 3D avatars.

Identity preservation sub-module. To ensure consistent and realistic facial details across frames, we implement an identity preservation module that combines 3D head reconstruction and facial fusion techniques. From a single input image, we first create a 3D Gaussian-based head avatar using a method inspired by Chu *et al*. [15], which employs a novel *dual-lifting* approach that predicts both forward and backward lifting distances.

Figure 4. **Qualitative Evaluation** on the People Snapshot dataset and of THuman dataset scan renderings. From a single-image input, SVAD generates high-quality, animatable 3D avatars.

Given an input image I_s, global and local features F_{local} are extracted using a frozen DINOv2 [71] backbone. These features are used to predict forward and backward lifting distances, positioning 3D Gaussians G_{pos} as follows:

$$G_{\text{pos}} = [\mathbf{p}_s + E_{\text{Conv0}}(F_{\text{local}}) \cdot \mathbf{n}_s, \mathbf{p}_s - E_{\text{Conv1}}(F_{\text{local}}) \cdot \mathbf{n}_s], \tag{4}$$

where $\mathbf{p}_s$ is the initial point plane, $\mathbf{n}_s$ is the normal vector, and E_{Conv} are convolutional layers predicting offsets. To capture expression variations, we bind 3DMM [60] features:

$$G_{\text{expr}} = \text{MLP}(F_{\text{3DMM}} + F_{\text{global}}). \tag{5}$$

To animate this 3D head avatar, we separately track FLAME [60] parameters $\Theta = \{\beta, \psi, \theta, \phi\}$ from our predefined pose sequence video (the same sequence used in the video diffusion module), where $\beta \in \mathbb{R}^{300}$ represents shape parameters, $\psi \in \mathbb{R}^{100}$ expression parameters, $\theta \in \mathbb{R}^6$ global pose parameters, and $\phi \in \mathbb{R}^6$ eye pose parameters. These tracked parameters serve as animation controls for the reconstructed 3D head. Using these tracked FLAME parameters, we render the 3D head avatar to generate a sequence of head images that match our predefined pose sequence. These renderings provide high-quality, identity-consistent facial details across different viewpoints. Since the quality of the renderings deteriorates for back-of-head views, we selectively apply the face fusion process only to frames where the head is front-facing (front and side views).

For the face fusion process, we detect facial landmarks [51] on both the diffusion-generated frame I_{orig} and the rendered head image I_{head}, compute an affine transformation for alignment, and use Poisson image editing [74] for seamless blending:

$$\min_I \int_\Omega \|\nabla I - \nabla I_{\text{warp}}\|^2 \, \mathrm{d}x \, \mathrm{d}y, \quad \text{subject to } I|_{\partial\Omega} = I_{\text{orig}}|_{\partial\Omega}, \tag{6}$$

where Ω is defined by the facial mask. This ensures temporally consistent facial details while preserving the original identity throughout the animation sequence.

Image restoration sub-module. Finally, to preserve quality of fine detailed regions, we employ an image restoration module based on the work of Chen *et al*. [12], specifically their diffusion-based image restoration method BFRffusion. This approach leverages the generative prior encapsulated in the pretrained Stable Diffusion [80] model to enhance image details through a comprehensive architecture that effectively extracts features from low-quality images and restores realistic facial details.

For our implementation, we set the super-resolution scale factor to $s = 1.5$, which our empirical analysis showed provides an optimal balance between detail enhancement and artifact suppression. We observed that scale factors $s < 1.5$ produce insufficient detail recovery, while factors $s > 2.0$ introduce perceptual artifacts (particularly in specular regions such as eyes) and significantly increase computational demands during avatar training. The diffusion process uses 50 DDIM sampling steps with:

$$z_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{z_t - \sqrt{1-\alpha_t}\epsilon_\theta(z_t)}{\sqrt{\alpha_t}} \right) + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(z_t) \tag{7}$$

where $\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$ and ϵ_θ is the denoising network. We utilize a classifier-free guidance scale of $w = 3.5$, with the guidance equation:

$$\hat{\epsilon}_\theta(z_t) = (1+w)\epsilon_\theta(z_t) - w\epsilon_\theta(z_t, \emptyset) \tag{8}$$

where $\epsilon_\theta(z_t, \emptyset)$ represents the unconditional prediction. This achieves an optimal balance between restoration quality and processing speed. For face regions, the method employs a face restoration helper with facial landmark detection to specifically enhance facial details, ensuring identity consistency across generated frames. Restored faces are blended with Poisson image editing.

This image restoration submodule significantly improves the fidelity and realism of our synthetic training data by restoring fine facial details, enhancing texture quality in clothing and accessories, and improving overall image coherence. The refined data enables the 3DGS avatar to learn more accurate representations with consistent high-frequency details that persist across poses and viewpoints.

3.3. 3D Human Gaussian Splatting Module

We apply the architecture of a 3DGS based avatar method introduced by Moon *et al*. [68], which integrates the SMPL-X [73] model with a 3D Gaussian-based representation to produce animatable human avatars. Each 3D Gaussian acts as a vertex connected by a pre-defined mesh topology following SMPL-X. This hybrid representation combines the expressive surface modeling of SMPL-X with the flexibility of a volumetric approach, allowing for smooth interpolation across the body surface essential for realistic animations.

Each Gaussian point is associated with positional data $\mathbf{V} \in \mathbb{R}^{N\times 3}$, RGB color values $\mathbf{C} \in \mathbb{R}^{N\times 3}$, and a scale parameter $\mathbf{S} \in \mathbb{R}^{N}$, where N is the number of Gaussians. The Gaussian splatting rendering equation is:

$$I = f(V, \exp(S), C, K, E), \tag{9}$$

where V represents positions, S denotes scale, C colors, and K and E camera parameters.

Pose-dependent deformations are applied through an MLP network, predicting offsets for each Gaussian based on SMPL-X pose parameters:

$$\mathbf{V}_{\text{pose}} = \mathbf{V} + \Delta\mathbf{V}_{\text{pose}} + \Delta\mathbf{V}_{\text{expr}}. \tag{10}$$

To maintain spatial coherence, a Laplacian regularizer [70, 92] minimizes the difference between the Laplacian of the canonical mesh and the deformed Gaussian points:

$$L_{\text{Lap}} = \left\| \Delta\mathbf{V}_{\text{canonical}} - \Delta\mathbf{V}_{\text{deformed}} \right\|^2. \tag{11}$$

This approach combined with our augmented synthetic data achieves highly realistic, animatable avatars capable of real-time rendering with smooth deformations across facial expressions, body movements, and hand gestures.

4. Experiments

4.1. Datasets and Metrics

People-Snapshot Dataset [2] We conduct our avatar evaluation on the People-Snapshot dataset, which features video recordings of subjects performing 360-degree rotations. Following both Anim-NeRF [102] and InstantAvatar [42], we address a known limitation in this dataset: the provided pose parameters often exhibit misalignment with the actual image content. Anim-NeRF addressed this by optimizing pose parameters for both training and test sequences. To ensure fair comparison with existing methods, we adopt these same optimized pose parameters and keep them frozen throughout our training process for fair comparison.

THuman Dataset [107] For evaluating single-image 3D human reconstruction, we employ the THuman dataset, adhering to the methodology established in Ultraman [11]. Our procedure involves randomly selecting 100 scans and generating renderings from four viewpoints (front, left, right, back). We then measure the similarity between our reconstructed outputs and the ground-truth scan renderings from these identical perspectives, facilitating objective comparison with other SOTA methods.

Evaluation Metrics Our evaluation framework uses four metrics to quantify reconstruction quality: PSNR [24], SSIM [99], LPIPS [114], and CLIP Similarity [78] (referred to as CLIP in our tables). This combination provides comprehensive assessment across different dimensions: PSNR for pixel accuracy, SSIM for structural coherence, LPIPS for perceptual alignment with human vision, and CLIP for semantic consistency at the feature level. The use of these metrics enables thorough evaluation of both fine-grained detail, and overall perceptual quality.

Method	Female-4-casual			Male-3-casual			Female-3-casual			Male-4-casual		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
HumanNeRF [102]	27.07	0.9615	0.0151	26.90	0.9605	0.0181	24.46	0.9516	0.0269	25.50	0.9397	0.0357
GaussianAvatar [34]	30.84	0.9771	0.0140	30.98	0.9790	0.0145	29.55	0.9762	0.0220	28.78	0.9755	0.0230
ExAvatar [68]	30.98	0.9789	0.0333	29.75	0.9628	0.0402	29.74	0.9678	0.0458	28.89	0.9666	0.0500
ExAvatar [68] (Single Image)	20.42	0.9427	0.0656	23.24	0.9448	0.0562	20.12	0.9492	0.0543	23.74	0.9497	0.0610
Ours (Single Image)	21.51	0.9442	0.0528	22.54	0.9467	0.0484	21.96	0.9609	0.0541	23.71	0.9570	0.0592

Table 1. **Quantitative Evaluation** on the People Snapshot [2] Dataset. Our approach demonstrates superior performance on *single-image* input, outperforming the baseline on most of the metrics. The top two results for *single-image* input are highlighted in first and second, with the overall best result highlighted in first. Note that methods that use monocular input utilize approximately 200 input frames.

4.2. Quantitative Evaluation

We quantitatively evaluate the quality of single-image 3D avatars generated by our method against SOTA 3D avatar generation methods [34, 68, 102]. While current 3D avatar models generally require a monocular video as input, we assess our model's performance using a single-image as input on ExAvatar [68]. Additionally, we report results using the original full training set of approximately 200 input frames for monocular input based avatar models for reference. As shown in Table 1, our model achieves highest scores on most of the metrics among single-image input methods. We further compare our approach with single-view 3D human reconstruction methods [11, 29, 38, 82, 115], many of which employ the SMPL model, allowing for animatability through mesh fitting and reposing techniques, such as those in Editable Humans [28]. We randomly sample 100 scans from the THuman dataset and report results. We repose our trained avatar using ground-truth SMPL-X parameters and compare with the ground-truth scan renderings from the same views. As presented in Table 2, our method surpasses all baselines, demonstrating superior quality in 3D human reconstruction tasks.

4.3. Qualitative Evaluation

Figure 4 shows the overall quality of our generated 3D avatars from single-images in the People Snapshot and the THuman dataset. Figure 5, Figure 6 shows that our method performs superior compared to current SiTH [29]. For single-image avatar generation, we evaluate on the People Snapshot dataset and compare against ExAvatar [68]. For fairness, we train ExAvatar for the same number (12,000) of iterations. Figure 7 shows that for single-image avatar generation, our method performs superior especially for the back and side views.

4.4. Ablation Study

In this section, we conduct ablation studies to validate each component of our methods. The average metrics over 4 sequences in the People Snapshot dataset are reported in Table 3. It shows that our methods modules are required

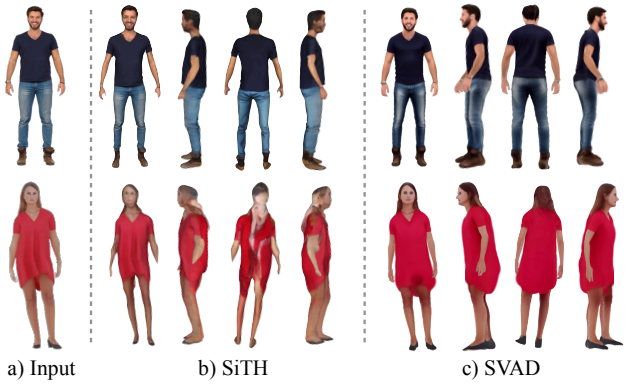


Figure 5. **Qualitative Evaluation** against SiTH [29]. Our approach better reconstructs complex contours and subtle features, resulting in a more lifelike and coherent side-view appearance.

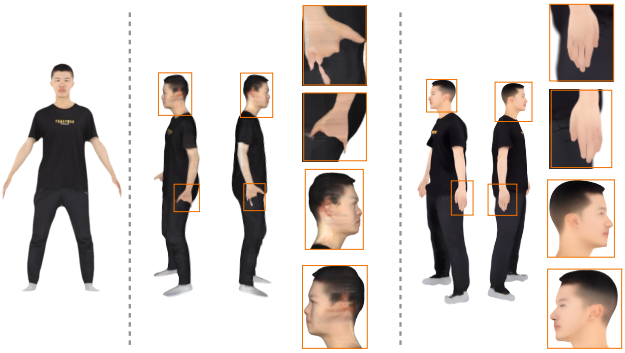
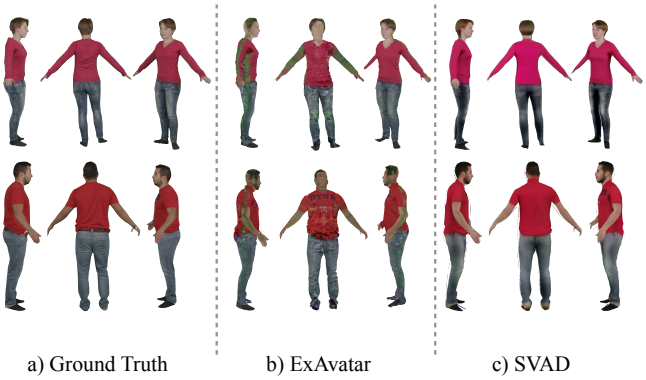


Figure 6. **Qualitative Evaluation** against SiTH [29]. Our method reconstructs fine detail (hands), while preserving original identity in facial regions.

to reach the optimal performance reflected by all the metrics. Using the THuman dataset, we apply the same evaluation technique as in our quantitative evaluation. Results show that our method performs the best in PSNR, SSIM and CLIP similarity and performs second best in LPIPS. Figure 8 shows visual results of the effect of the image restoration module. High-detailed regions such as clothing texture,



a) Ground Truth b) ExAvatar c) SVAD

Figure 7. **Qualitative Evaluation** against ExAvatar [68] in single-image to 3D avatar task. Our method generates more plausible back and side views with the generated synthetic dataset.

Method	PSNR↑	SSIM↑	LPIPS↓	CLIP↑
PIFu[82]	15.62	0.8921	0.1903	0.8612
TeCH[38]	15.85	0.8892	0.1667	0.8890
Ultraman[11]	18.13	0.9019	0.1334	0.9089
SIFU[115]	18.59	0.8591	0.1402	0.8873
SiTH[29]	19.98	0.9018	0.1294	0.9084
Ours	**20.92**	**0.9291**	**0.1124**	**0.9321**

Table 2. **Quantitative Evaluation** on single-image to 3D human reconstruction tasks on 100 scan renderings of the THuman [107] Dataset. Top two results are colored as first second .

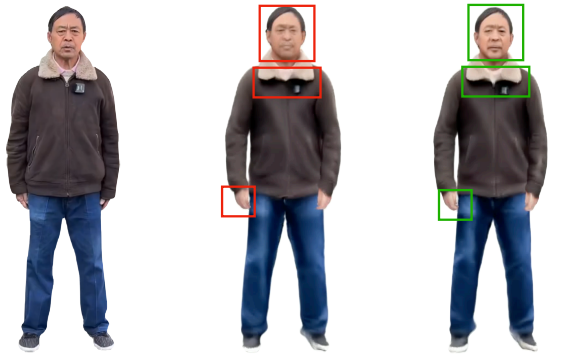
Method	PSNR↑	SSIM↑	LPIPS↓	CLIP↑
w/o Identity Preserve	22.19	0.9419	0.0623	0.9231
w/o Image Restoration	22.61	0.9298	0.0645	0.9239
Ours (Full)	22.79	0.9502	0.0594	0.9241

Table 3. **Ablation study** on the People Snapshot dataset. Our full model consistently outperforms variants with individual components removed across all metrics.

Method	PSNR↑	SSIM↑	LPIPS↓	CLIP↑
w/o Identity Preserve	20.12	0.9256	0.1294	0.9284
w/o Image Restoration	20.16	0.9212	0.0799	0.9201
Ours (Full)	20.92	0.9291	0.1124	0.9321

Table 4. **Ablation study** on the THuman dataset. The full model achieves superior performance in most metrics, demonstrating the importance of each component in our pipeline.

fingers, and facial details are better preserved when applying our module. Figure 9, shows the visual effect of the identity preservation module. We clearly show that original input's facial details are more preserved our module.



a) Input b) w/o image restoration c) Full model

Figure 8. **Ablation study** on the image restoration module. We show that applying the module into our pipeline recover fine details on the final avatar output.

a) Input (Cropped) b) w/o identity preserve c) Full Model

Figure 9. **Ablation study** on the identity preservation module. We show that with the module, the final avatar maintains facial details on the original input image.

5. Conclusion and Discussion

In this work, we introduced SVAD, a novel synthetic data generation approach for creating high-fidelity, animatable 3D human avatars from a single image. By combining the generative power of diffusion models with the rendering efficiency of 3D Gaussian Splatting, SVAD produces avatars that maintain consistent identity across varied poses and viewpoints. Through comprehensive experiments, we demonstrate that our method achieves SOTA performance.

Limitations and Future Work. Our method faces several limitations. First, inaccurate background segmentation of training frames produces floating artifacts. Second, our approach struggles with complex clothing textures and loose outfits due to limitations of the video diffusion model in generating detailed synthetic data. Finally, the computational requirements present practical challenges—the video diffusion step demands substantial resources, and the complete pipeline requires 5-6 hours per avatar generation. Future work will focus on improving handling of diverse clothing types and optimizing computational performance.

References

[1] Badour AlBahar, Shunsuke Saito, Hung-Yu Tseng, Changil Kim, Johannes Kopf, and Jia-Bin Huang. Single-image 3d human digitization with shape-guided diffusion. In *SIGGRAPH Asia*, 2023. 2

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, 2018. 3, 6, 7, 13, 20, 21

[3] C Bradford Barber, David P Dobkin, and Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM TOMS*, 1996. 14

[4] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *CVPR*, 2023. 2

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 2

[6] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K. Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In *CVPR*, 2024. 2, 17

[7] Hyunsoo Cha, Byungjun Kim, and Hanbyul Joo. Pegasus: Personalized generative 3d avatars with composable attributes. In *CVPR*, 2024. 2

[8] Hyunsoo Cha, Inhee Lee, and Hanbyul Joo. Perse: Personalized 3d generative avatars from a single portrait. In *CVPR*, 2025. 2

[9] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *ICCV*, 2019. 2, 3

[10] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022. 16

[11] Mingjin Chen, Junhao Chen, Xiaojun Ye, Huan-ang Gao, Xiaoxue Chen, Zhaoxin Fan, and Hao Zhao. Ultraman: Single image 3d human reconstruction with ultra speed and detail. *arXiv preprint arXiv:2403.12028*, 2024. 6, 7, 8

[12] Xiaoxu Chen, Jingfan Tan, Tao Wang, Kaihao Zhang, Wenhan Luo, and Xiaochun Cao. Towards real-world blind face restoration with generative diffusion prior. *IEEE TCSVT*, 2024. 6, 15

[13] Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. Monogaussianavatar: Monocular gaussian point-based head avatar. In *SIGGRAPH*, 2024. 2

[14] Yushuo Chen, Zerong Zheng, Zhe Li, Chao Xu, and Yebin Liu. Meshavatar: Learning high-quality triangular human avatars from multi-view videos. *ECCV*, 2024. 1, 2

[15] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. *NeurIPS*, 2024. 4, 14

[16] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *MICCAI*, 2016. 3

[17] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020. 15

[18] Helisa Dhamo, Yinyu Nie, Arthur Moreau, Jifei Song, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Headgas: Real-time animatable head avatars via 3d gaussian splatting. In *ECCV*, 2025. 2

[19] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *CVPR*, 2023. 2

[20] Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen. 4d gaussian splatting: Towards efficient novel view synthesis for dynamic scenes. *SIGGRAPH*, 2024. 2

[21] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *SIGGRAPH*, 2021. 15

[22] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *ECCV*, 2022. 2

[23] Wei Gao, Meihong Yang, Wei Zhang, and Libin Liu. Efficient ofdm channel estimation with rrdbnet. In *ISCC*, 2022. 15

[24] Rafael C Gonzalez and Richard E Woods. *Digital Image Processing*. Prentice Hall, 2008. 6

[25] Colin Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1991. 14

[26] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. *ECCV*, 2024. 3

[27] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *ICLR*, 2024. 3

[28] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In *CVPR*, 2023. 7

[29] I Ho, Jie Song, Otmar Hilliges, et al. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *CVPR*, 2024. 7, 8

[30] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3, 13

[31] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 2022. 2, 13

[32] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *SIGGRAPH*, 2022. 17

[33] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, 2024. 2, 3, 13

[34] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *CVPR*, 2024. 1, 2, 7

[35] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. In *ICCV*, 2023. 2

[36] Shoukang Hu, Tao Hu, and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *CVPR*, 2024. 2

[37] Letian Huang, Jiayang Bai, Jie Guo, Yuanqi Li, and Yanwen Guo. On the error analysis of 3d gaussian splatting and an optimal projection strategy. *CoRR*, 2024. 2

[38] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. Tech: Text-guided reconstruction of lifelike clothed humans. In *3DV*, 2024. 7, 8

[39] Rohit Jena, Ganesh Subramanian Iyer, Siddharth Choudhary, Brandon Smith, Pratik Chaudhari, and James Gee. Splatarmor: Articulated gaussian splatting for animatable humans from monocular rgb videos. *arXiv preprint arXiv:2311.10812*, 2023. 2

[40] Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control, 2023. 17

[41] Suyi Jiang, Haoran Jiang, Ziyu Wang, Haimin Luo, Wenzheng Chen, and Lan Xu. Humangen: Generating human radiance fields with explicit priors. In *CVPR*, 2023. 2

[42] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *CVPR*, 2023. 6

[43] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023. 15

[44] Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaoxiao Long, Wenping Wang, and Yuexin Ma. Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. In *CVPR*, 2024. 2

[45] Arlene John, Jishnu Sadasivan, and Chandra Sekhar Seelamantula. Adaptive savitzky-golay filtering in non-gaussian noise. *IEEE TSP*, 2021. 14, 16

[46] Brennan Jones, Yaying Zhang, Priscilla NY Wong, and Sean Rintel. Belonging there: Vroom-ing into the uncanny valley of xr telepresence. *ACM HCI*, 2021. 2

[47] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *ICCV*, 2023. 2

[48] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023. 2

[49] Taeksoo Kim and Hanbyul Joo. Target-aware video diffusion models. *arXiv preprint arXiv:2503.18950*, 2025. 2

[50] Taeksoo Kim, Byungjun Kim, Shunsuke Saito, and Hanbyul Joo. Gala: Generating animatable layered assets from a single scan. In *CVPR*, 2024. 17

[51] Davis King. Dlib-ml: A machine learning toolkit. http://dlib.net, 2009. 5, 14

[52] Diederik P Kingma. Adam: A method for stochastic optimization. *ICLR*, 2015. 16

[53] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 16, 18

[54] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *NeurIPS*, 2023. 17

[55] Agelos Kratimenos, Jiahui Lei, and Kostas Daniilidis. Dynmf: Neural motion factorization for real-time dynamic view synthesis with 3d gaussian splatting. In *ECCV*, 2025. 2

[56] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 17

[57] Byeonghyeon Lee, Howoong Lee, Xiangyu Sun, Usman Ali, and Eunbyung Park. Deblurring 3d gaussian splatting. *ECCV*, 2024. 2

[58] Inhee Lee, Byungjun Kim, and Hanbyul Joo. Guess the unseen: Dynamic 3d scene reconstruction from partial 2d glimpses. In *CVPR*, 2024. 2

[59] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *CVPR*, 2024. 2

[60] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *TOG*, 2017. 5, 13

[61] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *CVPR*, 2024. 2

[62] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J. Black. TADA! Text to Animatable Digital Avatars. In *International Conference on 3D Vision (3DV)*, 2024. 17

[63] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *WACV*, 2024. 3

[64] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 17

[65] Xiangyue Liu, Kunming Luo, Heng Li, Qi Zhang, Yuan Liu, Li Yi, and Ping Tan. Gaussianavatar-editor: Photorealistic animatable gaussian head avatar editor. *3DV*, 2025. 17

[66] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *3DV*, 2024. 2

[67] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *CVPR*, 2022. 15

[68] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. *ECCV*, 2024. 1, 2, 6, 7, 8, 16

[69] Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *CVPR*, 2024. 1

[70] Andrew Nealen, Takeo Igarashi, Olga Sorkine, and Marc Alexa. Laplacian mesh optimization. In *GRAPHITE*, 2006. 6

[71] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2023. 5

[72] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *CVPR*, 2024. 2

[73] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 6, 13, 16

[74] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. Association for Computing Machinery, 2023. 5, 14

[75] Sergey Prokudin, Michael J Black, and Javier Romero. Smplpix: Neural avatars from 3d human models. In *WACV*, 2021. 2

[76] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *CVPR*, 2024. 2

[77] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *CVPR*, 2024. 1, 2

[78] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 6, 13

[79] Yurui Ren, Ge Li, Shan Liu, and Thomas H Li. Deep spatial transformation for pose-guided person image generation and animation. *IEEE TIP*, 2020. 2

[80] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 6

[81] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3, 13

[82] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019. 7, 8

[83] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *ICLR*, 2022. 3

[84] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural re-rendering of humans from a single image. In *ECCV*, 2020. 2

[85] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns. *IEEE Sensors Journal*, 2020. 15

[86] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *CVPR*, 2024. 1, 2

[87] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-avatar: Expressive human avatars. In *CVPR*, 2023. 1

[88] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *NeurIPS*, 2019. 2

[89] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021. 2

[90] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[91] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ICLR*, 2020. 3, 13

[92] Olga Sorkine, Daniel Cohen-Or, Yaron Lipman, Marc Alexa, Christian Rössl, and H-P Seidel. Laplacian surface editing. In *Eurographics*, 2004. 6

[93] David Svitov, Pietro Morerio, Lourdes Agapito, and Alessio Del Bue. Haha: Highly articulated gaussian human avatars with textured mesh prior. *ACCV*, 2024. 1, 2

[94] Zhengyan Tong, Chao Li, Zhaokang Chen, Bin Wu, and Wenjiang Zhou. Musepose: a pose-driven image-to-video framework for virtual human generation. *arxiv*, 2024. 2, 3, 13

[95] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 3

[96] Jie Wang, Xianyan Li, Jiucheng Xie, Feng Xu, and Hao Gao. Gaussianhead: Impressive 3d gaussian-based head avatars with dynamic hybrid neural field. *arXiv e-prints*, pages arXiv–2312, 2023. 2

[97] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *CVPR*, 2024. 2

[98] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, 2021. 15

[99] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004. 6

[100] Zhou Wang, Alan C Bovik, and Hamid R Sheikh. Structural similarity based image quality assessment. In *Digital Video image quality and perceptual coding*. CRC Press, 2017. 14

[101] Eric W Weisstein. Affine transformation. *https://mathworld. wolfram. com/*, 2004. 14

[102] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, 2022. 6, 7

[103] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *CVPR*, 2024. 2

[104] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *NeurIPS*, 2024. 16

[105] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *ICCV*, 2023. 3, 13, 15

[106] Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. Pose-guided human animation from a single image in the wild. In *CVPR*, 2021. 2

[107] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *CVPR*, 2021. 6, 8, 22

[108] Wing-Yin Yu, Lai-Man Po, Ray CC Cheung, Yuzhi Zhao, Yu Xue, and Kun Li. Bidirectionally deformable motion modulation for video-based human pose transfer. In *ICCV*, 2023. 2

[109] Yifei Zeng, Yuanxun Lu, Xinya Ji, Yao Yao, Hao Zhu, and Xun Cao. Avatarbooth: High-quality and customizable 3d human avatar generation. *arXiv preprint arXiv:2306.09864*, 2023. 1

[110] Bowen Zhang, Yiji Cheng, Chunyu Wang, Ting Zhang, Jiaolong Yang, Yansong Tang, Feng Zhao, Dong Chen, and Baining Guo. Rodinhd: High-fidelity 3d avatar generation with diffusion models. *ECCV*, 2024. 17

[111] Hao Zhang, Yao Feng, Peter Kulits, Yandong Wen, Justus Thies, and Michael J Black. Teca: Text-guided generation and editing of compositional 3d avatars. In *3DV*, 2024. 17

[112] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2

[113] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *CVPR*, 2022. 2

[114] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[115] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *CVPR*, 2024. 7, 8

[116] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *CVPR*, 2022. 2

[117] Xin Zhao, Huan Zhao, Xiangfei Li, and Han Ding. Path smoothing for five-axis machine tools using dual quaternion approximation with dominant points. *IJPEM*, 2017. 14

[118] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *CVPR*, 2024. 2

[119] Yang Zheng, Qingqing Zhao, Guandao Yang, Wang Yifan, Donglai Xiang, Florian Dubost, Dmitry Lagun, Thabo Beeler, Federico Tombari, Leonidas Guibas, and Gordon Wetzstein. Physavatar: Learning the physics of dressed 3d avatars from visual observations. In *ECCV*, 2024. 18

[120] Yang Zheng, Qingqing Zhao, Guandao Yang, Wang Yifan, Donglai Xiang, Florian Dubost, Dmitry Lagun, Thabo Beeler, Federico Tombari, Leonidas Guibas, et al. Physavatar: Learning the physics of dressed 3d avatars from visual observations. *ECCV*, 2024. 1, 2

[121] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. *ECCV*, 2024. 2

[122] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. *3DV*, 2023. 2

A. Implementation Details

In this section, we provide comprehensive technical details of SVAD. We first describe the predefined pose sequences that serve as conditioning inputs for our video diffusion model. Next, we elaborate on the video diffusion module, the identity preservation module and image restoration module for enhancing facial fidelity and overall texture quality. Finally, we elaborate on the training process for our 3DGS avatar, including the SMPL-X [73] parameter fitting procedure and the optimization strategy for the 3D Gaussian representation.

A.1. Predefined Pose Sequences

To initialize frame generation for our pipeline, we rely on a predefined set of poses extracted from the People Snapshot [2] dataset. Specifically, we utilize the *male-4-casual* sequence, which depicts a subject performing a full-body rotation with arms extended horizontally. Using DW-Pose [105], we extract 2D keypoints $K \in \mathbb{R}^{J\times 2}$, where $J = 17$ is the number of keypoints, from this sequence to create a standardized pose template. This sequence serves as the conditioning input for the video diffusion model, resulting in 189 frames of pose-guided human animation, with a resolution of 1024×1024.

Our experiments revealed that inference with lower resolutions such as 512×512 produced animations with significantly degraded facial details, which adversely affected subsequent processing steps. Particularly, the landmark-based face fusion technique requires accurate facial landmark detection, which proved unreliable on low-resolution outputs. The absence of distinct facial features in 512×512 outputs led to inconsistent landmark detection, compromising the accuracy of 3D head rendering and warping operations. The higher 1024×1024 resolution preserves critical facial details, enabling robust landmark detection and consistent face fusion results across the generated sequence.

A.2. Video Diffusion Module

For our video diffusion module, we leverage MusePose [94], a modified variant of Animate Anyone [33], specifically designed for pose-guided video generation from a single image. The architecture follows a UNet-based [81] denoising diffusion model with temporal modeling capabilities, enabling coherent video generation while maintaining consistency with the reference image.

During inference, the video diffusion pipeline performs iterative denoising of random noise guided by the reference image and pose sequence. We configure the DDIM sampler [91] with 20 sampling steps and a classifier-free guidance [30] scale of 3.5 which keeps balance between generation quality and inference speed. The network architecture employs a 3D variant of the standard UNet architecture, where temporal layers enable information exchange across video frames. The reference image features are extracted using a CLIP vision encoder [78] and processed through a reference UNet. These features are transferred to the denoising UNet via a custom attention mechanism:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{12}$$

where Q represents queries from the denoising UNet features, while K and V are derived from the reference image features. This mechanism ensures that generated frames maintain the appearance details of the reference image.

The pose conditioning is handled by the PoseGuider module, which processes pose skeleton images through a series of convolutional layers to create pose feature embeddings. These embeddings are added to the latent noise to spatially align the generation with target poses:

$$z_t = z_t + P(p_t) \tag{13}$$

where z_t is the noise latent at timestep t, $p_t \in \mathbb{R}^{J\times 2}$ is the pose feature at time t, and $P(\cdot)$ represents the pose guider. The PoseGuider has an input convolutional layer, followed by blocks with increasing channel dimensions $(16, 32, 64, 128)$, and a zero-initialized output projection to the conditioning embedding channels.

For handling longer video sequences beyond the model's context window, we employ a sliding window [31] approach. The model processes frames in overlapping chunks of length $S = 48$ with an overlap of $O = 4$ frames. This enables the generation of arbitrarily long sequences while maintaining temporal consistency. The generative process for each video segment can be expressed as:

$$V_{i:i+S} = \mathcal{G}(I_{\text{ref}}, P_{i:i+S}, z) \tag{14}$$

where $V_{i:i+S}$ represents the generated video segment from frame i to $i + S$, $\mathcal{G}$ is our diffusion model, I_{ref} is the reference image, $P_{i:i+S}$ are the corresponding pose skeletons, and z is the random noise. By processing these overlapping segments and blending them at the boundaries, the final full-length human-animated video has smooth transitions.

A.3. Identity Preservation Module

Following the initial frame generation by the video diffusion model, we refine the facial regions to enhance identity consistency and detail preservation. Our identity preservation pipeline consists of three main components: FLAME [60] parameter tracking A.3.1, 3D head rendering A.3.2, and face fusion A.3.3. Each component plays a crucial role in generating high-quality, identity-consistent facial regions in our data augmentation pipeline.

A.3.1. FLAME Parameter Tracking

We begin by tracking FLAME parameters from our predefined pose sequence video to guide the animation of our 3D

head avatar. Using a tracking engine with focal length set to 12.0, we extract parameters $\Theta = \{\beta, \psi, \theta, \phi\}$, where $\beta \in \mathbb{R}^{300}$ represents shape parameters, $\psi \in \mathbb{R}^{100}$ expression parameters, $\theta \in \mathbb{R}^6$ global pose parameters, and $\phi \in \mathbb{R}^6$ eye pose parameters.

To ensure smooth parameter transitions across frames, we apply Savitzky-Golay [45] filtering with a window length of 9 frames and polynomial order of 2. For rotation parameters, we employ quaternion-based smoothing [117] with a continuity enforcement algorithm to handle sign flips:

$$\mathrm{q}'_{t+1} = \begin{cases} -\mathrm{q}_{t+1}, & \text{if } \mathrm{q}_t \cdot \mathrm{q}_{t+1} < 0 \\ \mathrm{q}_{t+1}, & \text{otherwise} \end{cases} \tag{15}$$

Different parameter types are smoothed with specific momentum coefficients: rotation matrices $\alpha = 0.6$, translation vectors $\alpha = 0.6$, and eye pose parameters $\alpha = 0.7$. This comprehensive smoothing strategy eliminates jitter and ensures temporal consistency in the final animation sequence.

A.3.2. 3D Head Rendering

Using GAGAvatar [15] as our 3D head modeling framework, we utilize the tracked FLAME parameters to render high-quality facial images that match our predefined pose sequence. We leverage this model to render the 3D head with precise control over pose and expression. The rendering process begins with the FLAME model, which generates 3D vertices based on the tracked shape, expression, pose, and eye parameters. We then employ a mesh renderer with a resolution of 512×512 pixels, using the FLAME topology for face modeling where focal length is set to 12.0. This approach enables us to generate precisely controlled facial renderings that maintain the identity of the source image while adopting the pose and expression parameters from the target sequence.

A.3.3. Face Fusion Process

We selectively apply face fusion only to frames when the head rendering is front-facing. We determine this by analyzing eye landmark detection - specifically, when at least one eye is clearly visible and properly detected in the facial landmark set. This approach ensures face fusion is only applied to frames with reliable facial orientation, as the quality of renderings deteriorates for back-of-head views where no eyes are visible. After filtering, we perform structural similarity assessment [100] and landmark-based warping [101] with careful parameter tuning to ensure seamless integration.

First, we detect 68 facial landmarks using dlib [51] on both the diffusion-generated frame I_{orig} and the rendered head image I_{head} from GAGAvatar. Before applying the transformation, we validate the structural compatibility by computing a Procrustes disparity measure [25] between the landmark sets:

$$d(L_{\text{orig}}, L_{\text{head}}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|L_{\text{orig},i} - L_{\text{head},i}\|^2} \tag{16}$$

where L_{orig} and L_{head} are the normalized landmark sets. We skip fusion when the disparity exceeds a threshold of 0.01, preserving the original frame in cases where the structural alignment would produce unnatural results. For valid frames, we compute an affine transformation matrix through corresponding landmarks using:

$$M = \arg\min_{M} \sum_{i=1}^{68} \|M \cdot L_{\text{head},i} - L_{\text{orig},i}\|^2 \tag{17}$$

where M is a 2×3 affine transformation matrix. This matrix is estimated using a partial affine model that preserves scale while allowing for rotation and translation, maintaining proportional facial features during transformation. The warped image is then computed by applying the transformation:

$$I_{\text{warp}} = T(I_{\text{head}}, M, (w, h)) \tag{18}$$

where T represents the affine warping function that maps pixels from the source to destination image according to transformation M.

We then create a facial mask Ω by computing the convex hull [3] of the landmarks to define the facial region:

$$\Omega = \text{convexHull}(L_{\text{orig}}) \tag{19}$$

Finally, we apply seamless cloning, a gradient-domain blending implementation of Poisson image editing [74], centered at the face centroid (c_x, c_y) with a blending factor $\alpha = 1.0$:

$$I_{\text{fused}} = \text{PoissonBlend}(I_{\text{warp}}, I_{\text{orig}}, \Omega, (c_x, c_y)) \tag{20}$$

This procedure solves the Poisson equation:

$$\min_{I} \int_{\Omega} \|\nabla I - \nabla I_{\text{warp}}\|^2 \, dx \, dy, \text{ subject to } I|_{\partial\Omega} = I_{\text{orig}}|_{\partial\Omega} \tag{21}$$

The gradient-domain blending preserves boundary conditions from the original image while replacing interior gradients with those from the warped image. This approach maintains lighting conditions and color consistency across the boundary by solving for pixel values that create a smooth transition while matching gradient fields. The complete face fusion pipeline significantly reduces visible artifacts at the transition between the rendered face and the original image, allowing consistent identity preservation even under challenging viewpoints.

A.4. Image Restoration Submodule

To enhance the quality of video diffusion outputs, particularly in facial regions, we integrate a hybrid restoration pipeline based on BFRffusion [12]. Our approach combines diffusion-based facial enhancement with background upsampling to improve overall visual fidelity while preserving identity-specific details.

The restoration workflow begins with face detection using RetinaFace [17], which accurately localizes facial regions in each frame. For aligned facial areas, we maintain a consistent face size of 512×512 with a $1:1$ crop ratio. When processing non-aligned faces, we employ a landmark-based alignment process using a five-point facial landmark detector with an eye distance threshold of 5 pixels to filter out low-quality detections.

Each detected face undergoes diffusion-based restoration using a latent diffusion model. The process follows a conditional diffusion sampling approach:

$$z_{t-1} = \frac{\sqrt{\alpha_{t-1}} z_t - \sqrt{1-\alpha_t}\epsilon_\theta(z_t)}{\sqrt{\alpha_t}} + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(z_t) \tag{22}$$

where $\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$ and ϵ_θ is the denoising network. We implement classifier-free guidance with a scale of $w = 3.5$:

$$\hat{\epsilon}_\theta(z_t) = (1+w)\epsilon_\theta(z_t) - w\epsilon_\theta(z_t, \emptyset) \tag{23}$$

where $\epsilon_\theta(z_t, \emptyset)$ represents the unconditional prediction.

The diffusion sampling process uses 50 DDIM steps with a latent shape of $\mathbb{R}^{4\times64\times64}$ for 512×512 input images. The input facial image is first encoded to a latent representation through a VAE encoder, and the diffusion model progressively refines this representation before decoding it back to pixel space.

For background regions, we employ Real-ESRGAN [98] with an RRDBNet [23] architecture and a $2\times$ upsampling scale. The background upsampler processes images in tiles of 400×400 pixels with 10-pixel padding to handle high-resolution inputs efficiently while maintaining consistent quality across tile boundaries.

After separate processing of facial and background regions, we integrate the enhanced components using inverse affine transformations computed from the original facial alignment process. This creates a seamless composite where facial details are preserved and enhanced while maintaining natural transitions to background areas:

$$I_{\text{final}} = M_{\text{face}} \odot T^{-1}(I_{\text{face}}) + (1 - M_{\text{face}}) \odot I_{\text{bg}} \tag{24}$$

where T^{-1} represents the inverse transformation that maps the restored face back to its original position, and M_{face} is the binary mask indicating facial regions.

This comprehensive image restoration approach significantly enhances the perceptual quality of generated frames, particularly improving fine facial details that may be lost or degraded during the initial video diffusion process. The integration of specialized facial and background processing ensures optimal quality across the entire frame while maintaining computational efficiency.

A.5. Gaussian Avatar Submodule

To transform our synthetic data into a high-quality, animatable 3D avatar, we employ a two-stage process: first, we fit an SMPL-X model to our synthetic data sequences, then we train a 3D Gaussian Splatting representation using the fitted parameters as guidance.

A.5.1. SMPL-X Fitting Process

Prior to training the 3DGS avatar, we employ a comprehensive fitting process to obtain accurate SMPL-X parameters from our synthetic data. This multi-stage process ensures that the avatar's geometry accurately reflects the subject's physical characteristics and articulation.

Keypoint Extraction. The fitting pipeline begins with pose and shape estimation. We utilize DWPose [105] to extract 2D whole-body keypoints from each frame of our synthetic sequence. These keypoints provide critical information about body articulation across the sequence. The keypoints are represented as $K \in \mathbb{R}^{J\times3}$, where $J = 133$ includes 17 body, 68 face, and 42 hand keypoints, with each keypoint having $(x, y, \text{confidence})$ values. We then employ MMPOSE [85] with the RTMPose-L [43] model for refinement, using a confidence threshold of 0.5 to filter reliable detections.

Initial Parameter Estimation. For facial geometry, we leverage DECA [21] to estimate initial FLAME parameters. The optimization uses perspective projection with focal length of 5000 pixels and 1024×1024 resolution textures. The FLAME parameters include shape coefficients $\beta \in \mathbb{R}^{10}$, expression parameters $\phi \in \mathbb{R}^{10}$, and pose parameters for jaw and eyes.

For body pose and shape, we incorporate Hand4Whole [67] with the configuration: focal length of 2000, principal point at image center, and input shape of 256×256. This process yields initial estimates for SMPL-X parameters: global orientation $\theta_{\text{root}} \in \mathbb{R}^3$, body pose $\theta_{\text{body}} \in \mathbb{R}^{21\times3}$, jaw pose $\theta_{\text{jaw}} \in \mathbb{R}^3$, hand poses $\theta_{\text{hands}} \in \mathbb{R}^{30\times3}$, and shape parameters $\beta_{\text{shape}} \in \mathbb{R}^{10}$.

Parameter Optimization. These initial parameters are refined through an optimization process with multiple objectives. The primary loss function combines reprojection error, parameter regularization, and temporal smoothness:

$$L_{\text{fit}} = \lambda_{\text{kpt}} L_{\text{kpt}} + \lambda_{\text{reg}} L_{\text{reg}} + \lambda_{\text{temp}} L_{\text{temp}} \tag{25}$$

The keypoint reprojection loss L_{kpt} measures the distance between projected model joints and detected 2D key-

points, weighted by detection confidence:

$$L_{\text{kpt}} = \sum_{i=1}^{J} c_i \|\Pi(J_i(\theta, \beta)) - K_i\|_2^2 \tag{26}$$

where Π is the perspective projection function, $J_i(\theta, \beta)$ is the 3D position of joint i, K_i is the corresponding 2D keypoint, and c_i is its confidence score.

The regularization term L_{reg} penalizes deviation from prior pose and shape distributions:

$$L_{\text{reg}} = \|\beta\|_2^2 + \sum_j \|\theta_j - \theta_{\text{mean}}\|_2^2 \tag{27}$$

The temporal consistency term L_{temp} enforces smooth transitions between frames:

$$L_{\text{temp}} = \sum_{t=1}^{T-1} \|\theta_t - \theta_{t+1}\|_2^2 + \|\beta_t - \beta_{t+1}\|_2^2 \tag{28}$$

The optimization uses the Adam optimizer [52] with learning rate 1×10^{-3} and loss weights $\lambda_{\text{kpt}} = 1.0$, $\lambda_{\text{reg}} = 0.001$, and $\lambda_{\text{temp}} = 0.1$. The optimization proceeds in two stages: first optimizing global position and orientation with 100 iterations, then refining all parameters with 200 iterations.

Parameter Smoothing. To ensure temporal consistency and reduce jitter, we apply the same smoothing approach as used in our FLAME parameter tracking process in Section A.3.1. Specifically, we employ Savitzky-Golay [45] filtering with a window length of 9 frames and polynomial order of 2. For rotation parameters, we utilize the identical quaternion-based smoothing procedure with continuity enforcement to handle sign flips.

Segmentation and Depth Estimation. We generate foreground masks using Segment Anything [53] with the ViT-H backbone. The model uses keypoint-based prompting with valid keypoints as point coordinates, and a bounding box computed from these keypoints with an extension ratio of 1.2. We also extract depth information using Depth Anything V2 [104] with the ViT-L backbone. The depth maps are normalized and aligned with the SMPL-X mesh using the following procedure:

$$\begin{aligned} \text{scale} &= \frac{\sigma(\text{depth}_{\text{pred,fg}})}{\sigma(\text{depth}_{\text{smplx,fg}})} \\ \text{depth}'_{\text{pred}} &= \frac{\text{depth}_{\text{pred}}}{\text{scale}} \\ \text{depth}'_{\text{pred}} &= \text{depth}'_{\text{pred}} - \mu(\text{depth}'_{\text{pred,fg}}) + \mu(\text{depth}_{\text{smplx,fg}}) \end{aligned} \tag{29}$$

where σ and μ represent standard deviation and mean of depth values, and fg indicates foreground regions.

The extracted SMPL-X parameters $\Phi = \theta, \beta$, together with corresponding image observations $I_t t = 1^T$, foreground masks $M_t t = 1^T$, and aligned depth maps $D_t{}_{t=1}^{T}$, constitute a multi-modal conditioning set that guides the optimization of our 3D Gaussian representation.

A.5.2. 3DGS Avatar Training Process

With the fitted SMPL-X parameters and processed synthetic data, we proceed to train the 3DGS-based avatar [68]. The training begins by initializing the triplane representation [10] $T \in \mathbb{R}^{32\times128\times128}$, encoding 3D features for both body and facial regions. Gaussian parameters, including positions $\mathbf{V} \in \mathbb{R}^{N\times3}$, colors $\mathbf{C} \in \mathbb{R}^{N\times3}$, and opacity $\mathbf{O} \in \mathbb{R}^{N}$, are optimized through backpropagation with the following multi-objective loss function:

$$L = \lambda_{\text{RGB}} L_{\text{RGB}} + \lambda_{\text{SSIM}} L_{\text{SSIM}} + \lambda_{\text{LPIPS}} L_{\text{LPIPS}}, \tag{30}$$

where $\lambda_{\text{RGB}} = 0.8$, $\lambda_{\text{SSIM}} = 0.2$, and $\lambda_{\text{LPIPS}} = 0.2$ are the weights for the RGB reconstruction, structural similarity, and perceptual loss, respectively. The model is trained for 5 epochs with a batch size of 1, as required by the Gaussian splatting renderer.

The optimization process proceeds in two stages. During the warmup stage, Gaussian positions $\mathbf{V}$ are updated using an adaptive learning rate:

$$\alpha_{\text{position}}(t) = \alpha_{\text{init}} \times \left(1 - \frac{t}{T_{\text{max}}}\right) + \alpha_{\text{final}} \times \frac{t}{T_{\text{max}}}, \tag{31}$$

where $\alpha_{\text{init}} = 1.6 \times 10^{-4}$, $\alpha_{\text{final}} = 1.6 \times 10^{-6}$, and $T_{\text{max}} = 30,000$ iterations. Additional parameters, including opacity $\mathbf{O}$, scale $\mathbf{S}$, and feature parameters, are optimized with learning rates $\alpha_{\text{opacity}} = 0.05$, $\alpha_{\text{scale}} = 0.005$, and $\alpha_{\text{feature}} = 0.0025$, respectively.

Densification of the Gaussian distribution occurs between iteration 500 and 15,000, at intervals of 100 iterations. Gaussians with opacity values below a threshold ($\mathbf{O} < 0.005$) are pruned, and dense regions are refined using gradient-based adjustments. The pruning mechanism ensures efficient representation while preserving fidelity:

$$\mathbf{V}_{\text{new}} = \mathbf{V}_{\text{old}} - \eta \frac{\partial L}{\partial \mathbf{V}}, \tag{32}$$

where η is the learning rate and $\frac{\partial L}{\partial \mathbf{V}}$ represents the gradient of the loss with respect to Gaussian positions.

A hierarchical learning approach progressively increases the spherical harmonic degree d_{sh} from 0 to 3 over the course of training. The training loop dynamically adjusts Gaussian parameters, leveraging an Adam optimizer with a learning rate of 1×10^{-3} for the overall framework and parameter-specific rates for finer control. For our experiments, we employ the male SMPL-X [73] model due to its superior performance in complex sequences. The entire pipeline runs on a single GPU, ensuring scalability and efficiency.

Figure 10. **Text to 3D Avatar.** Our method enables the generation of animatable 3D avatars from text prompts. We show results for various textual descriptions processed through Flux-1 Dev [56] for image generation, followed by our single-image to 3D avatar pipeline.

B. Applications

Our pipeline for 3D avatar generation from single images also serves as a basis for developing further creative applications. By integrating our core pipeline with contemporary text-to-image synthesis [56] and image editing methods [64], we expand its range of use. This section outlines two such applications: a text-to-3D avatar generation workflow in Section B.1, which enables the creation of animatable 3D characters from textual inputs, and a text-guided avatar editing application in Section B.2, which permits semantic modifications to an avatar's visual features using text prompts.

B.1. Text to 3D Avatar

The generation of 3D avatars directly from textual descriptions is an area of interest, and various approaches [32, 40, 54, 62, 110] have be explored. We first leverage the Flux-1 Dev [56] model for text-to-image generation. This generated image then serve as input to our single-image to 3D avatar pipeline, producing the 3D avatar that preserve the features described in the text, as shown in Figure 10. This integration extends the applicability of our framework to scenarios where photographic references are unavailable, thus broadening the scope of generative 3D human representation.

B.2. Text-Guided 3D Avatar Editing

Our pipeline enables text-guided 3D avatar editing [6, 50, 65, 111] as illustrated in Figure 11. By integrating Step1X-edit [64], a text-based image editing diffusion model, we enable semantic modifications to the input image. Given an input image and a textual editing prompt, the diffusion model generates the modified reference image that incorporate the requested prompts. This edited image then proceeds through our standard single-image to 3D avatar pipeline, generating the modified 3D avatar that reflect the text-specified edits. This approach allows customizing avatar appearances without manual image manipulation or 3D modeling.

C. Failure Cases

Despite the demonstrated effectiveness of SVAD, we observe several limitations that highlight opportunities for future research. Our analysis reveals three primary categories of failure cases that affect the quality and consistency of the generated avatars: issues stemming from segmentation artifacts in Section C.1, challenges in accurately modeling loose clothing deformation in Section C.2, and inconsistencies observed in back view synthesis Section C.3.

Figure 11. **Text-Guided 3D Avatar Editing.** Our framework enables semantic editing of 3D avatars using textual prompts. The resulting edited images are then processed by our single-image to 3D avatar pipeline, producing updated, animatable 3D avatars that reflect the specified textual edits.

C.1. Segmentation Artifacts

Limitations in the segmentation model [53], leveraged in preprocessing the generated synthetic data for 3D avatar training, can impact our pipeline's output, as illustrated in Fig. 12. Background portions can be included in the training data for the 3D Gaussian representation. Such inaccuracies are most evident in posterior views of the reconstructed avatar, where artifacts from the original background are observed in the volumetric representation. These background elements remain visible during novel viewpoint synthesis, degrading the quality of rendered results. This shows our method's reliance on background segmentation, particularly for images with ambiguous foreground-background boundaries or similar color distributions.

C.2. Loose Clothing Deformation

Modeling loose clothing, such as dresses and skirts, presents challenges for our method, as demonstrated in Fig. 13. This problem arises from the parametric body model, which represents the human form as a close-fitting mesh and lacks explicit mechanisms for loose garments. While the rendered appearance can be plausible when the avatar's legs are proximate, fidelity degrades upon leg separation. The Gaussian splats, conditioned to align with to the parametric body surface, follow the leg geometry instead of preserving the garment's structure, leading to unrealistic deformations. This limitation indicates the need for modeling loose clothing deformation separately from the body mesh [119], especially for garments with flow-dependent behaviors that deviate from standard body topology.

C.3. Back View Synthesis Inconsistency

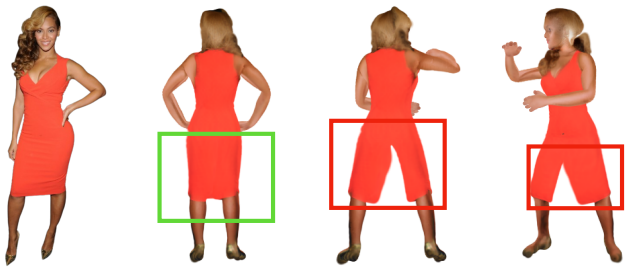
A fundamental limitation is the ill-posed problem of single-image to 3D generation, which particularly affects back view synthesis in our approach. As shown in Fig. 14, our method exhibits challenges in generating high-fidelity back views. While an input image provides strong frontal appearance cues, the synthesized back views can display reduced quality. Compared to ground truth renderings from similar viewpoints, our method often produce textures with lower fidelity, pattern inconsistencies, and blurred details. This behavior is attributed to two main factors: first, a potential bias in the video diffusion model towards frontal or near-frontal views during synthetic data generation, and second, the inherent ambiguity of inferring occluded geometry and appearance from a single perspective.

D. Societal Impact

SVAD generates animatable 3D avatars from single input images, enhancing accessibility to personal digital repre-

(a) Input (b) Avatar (c) Input (d) Avatar

Figure 12. **Failure case of segmentation artifacts.** When segmentation fails to properly separate the subject from the background, residual background elements become embedded in the avatar.



(a) Input (b) Legs Together (c) Legs Apart (d) Legs Apart

Figure 13. **Failure case of loose clothing deformation.** Our method struggles with modeling loose garments like dresses due to limitations of the underlying parametric body model."

(a) Input (b) Avatar (c) Ground Truth (d) Avatar (Zoom in)

Figure 14. **Failure case of back view synthesis.** The inherently ill-posed nature of single-image 3D generation results in degraded quality for unseen viewpoints."

sentation. As with other generative AI technologies addressing human likeness, this capability presents both positive and negative societal implications, contingent upon its application. Positively, SVAD allows broader access to 3D avatar creation, enabling users to readily produce personalized digital representations for VR/AR, gaming, and enhanced online communication without specialized expertise. This can foster more engaging virtual interactions and broaden creative expression, for example, through the rapid creation of avatars for virtual try-on or customized game characters. Negatively, the ease with which an avatar can be generated from any single image, potentially without consent, poses considerable risks. Unauthorized 3D likenesses could be exploited for deepfakes, impersonation, harassment, or privacy violations. While comprehensive technical safeguards against all misuse are challenging, we will emphasize responsible application and the critical need for consent in any public dissemination of our work. We advocate for a multi-faceted approach to mitigate these risks including detection technologies, ethical guidelines, and legal frameworks and hope our research prompts further discussion on these vital societal considerations for the responsible development and deployment of such technologies.

Figure 15. **3D Avatars from People Snapshot [2] dataset** Our method successfully generates high-fidelity avatars for various subjects from a single input image, demonstrating robust identity preservation and consistent appearance across novel poses and viewpoints. 🔍 **Zoom** in for more details.

Figure 16. **3D Avatars from People Snapshot [2] dataset.** SVAD enables creation of detailed and expressive avatars from a single image, accurately reproducing clothing details and facial features while maintaining realism in different poses. 🔍 **Zoom** in for more details.

Figure 17. **3D Avatars from the THuman [107] scan renderings.** Our approach generalizes well to the THuman dataset, producing realistic avatars with high geometric and texture fidelity. 🔍 **Zoom** in for more details.

Figure 18. **3D Avatars from Internet Images.** SVAD demonstrates strong generalization capability to in-the-wild images, successfully reconstructing recognizable 3D avatars of various celebrities from single unconstrained photographs. The method preserves distinctive appearance characteristics while enabling novel pose animation. 🔍 **Zoom** in for more details.