CVPR Workshop SyntaGen #16

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

Anonymous CVPR Workshop SyntaGen submission

Paper ID 16

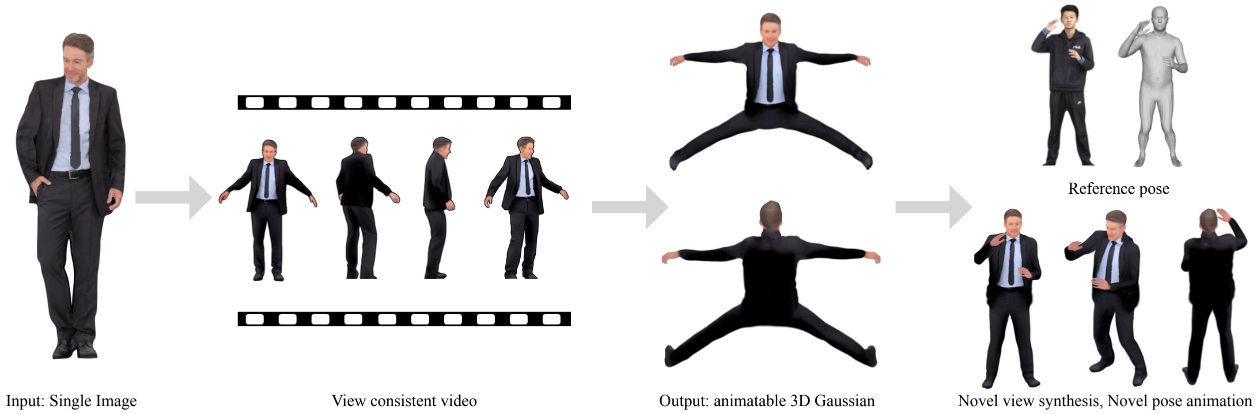

Figure 1. **SVAD.** We present a novel pipeline that leverages video diffusion models and data augmentation methods to generate high-quality synthetic training data from a single human image. This synthetic data generation approach enables us to train 3D Gaussian Splatting avatars with significantly improved fidelity, outperforming state-of-the-art single-image avatar creation methods while preserving identity and fine details across novel poses and viewpoints.

Abstract

Creating animatable 3D human avatars from a single image remains a significant challenge with applications in virtual reality and human-centered AI. Traditional 3D Gaussian Splatting (3DGS) methods produce high-quality avatars but require monocular video sequences or multi-view inputs, while video diffusion models can animate from static images but struggle with temporal coherence and identity preservation. We present SVAD, a novel framework for synthetic data generation and avatar creation that addresses these limitations. SVAD leverages video diffusion models to generate an initial set of synthetic pose-conditioned animations from a single image, then enhances this synthetic data through identity preservation and image restoration modules. This high-quality synthetic dataset enables training of 3DGS avatar models that maintain subject fidelity and fine details across diverse poses and viewpoints. Our approach combines the generative capabilities of diffusion models with the rendering efficiency of 3DGS, resulting in state-of-the-art performance in single-image avatar creation. Experiments demonstrate that SVAD's synthetic data generation pipeline significantly improves temporal stability and identity consistency compared to existing methods, while enabling real-time rendering for interactive applications.

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 [9, 19, 37, 38, 44, 51, 52, 56, 67, 74]. These methods rely on dense input data, typically monocular or multi-view video [9, 19, 37, 44, 51, 56, 74], 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 [18, 59, 62, 75]. These models use pose-conditioned 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. 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 models 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-based avatar model [37], 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.

Our contributions can be summarized as follows:

- We introduce SVAD, a novel pipeline for generating high-quality synthetic training data from a single image, enabling the creation of detailed and animatable 3D human avatars.
- We incorporate an identity-preservation module and an image restoration technique to refine diffusion-generated synthetic data, ensuring consistency in identity and fine details across diverse poses and viewpoints.
- We demonstrate that our synthetic data generation approach significantly improves the quality of 3DGS avatars compared to state-of-the-art single-image methods, while maintaining efficient real-time rendering capabilities.
- We provide extensive experiments and evaluations showing that SVAD's synthetic data-driven approach achieves superior performance in novel pose adaptation and identity preservation for single-image avatar creation.

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, 4, 5, 14, 20, 25, 42, 46, 50, 53, 54, 64, 66, 69, 72]. Early methods, such as PIDM [3] and DreamPose [29], 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 [59] and Animate Anyone [18], have extended diffusion models to improve temporal consistency and fine-grained control in human animation tasks. DisCo leverages dual ControlNets [68] 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 [30]. 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 3D Gaussian splatting has been demonstrated in both static [22, 27, 32] and dynamic [13, 28, 31, 36] scenes, making it a versatile tool for various applications. Recent advancements [8, 12, 21, 24, 33, 43, 44, 58, 76] have explored the use of 3DGS to create photorealistic human avatars across different scenarios. These methods commonly rely on multi-view data [34, 40, 73] or monocular video [19, 24, 33, 37, 44] 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 just 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 for character animation to produce initial synthetic data, as described in Sec. 3.1. Directly using these frames to train a 3DGS avatar model, however, often yields poor results, with challenges in preserving facial identity, clothing details, and maintaining consistent multiview 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 data, we proceed to train a 3DGS avatar model, as out-

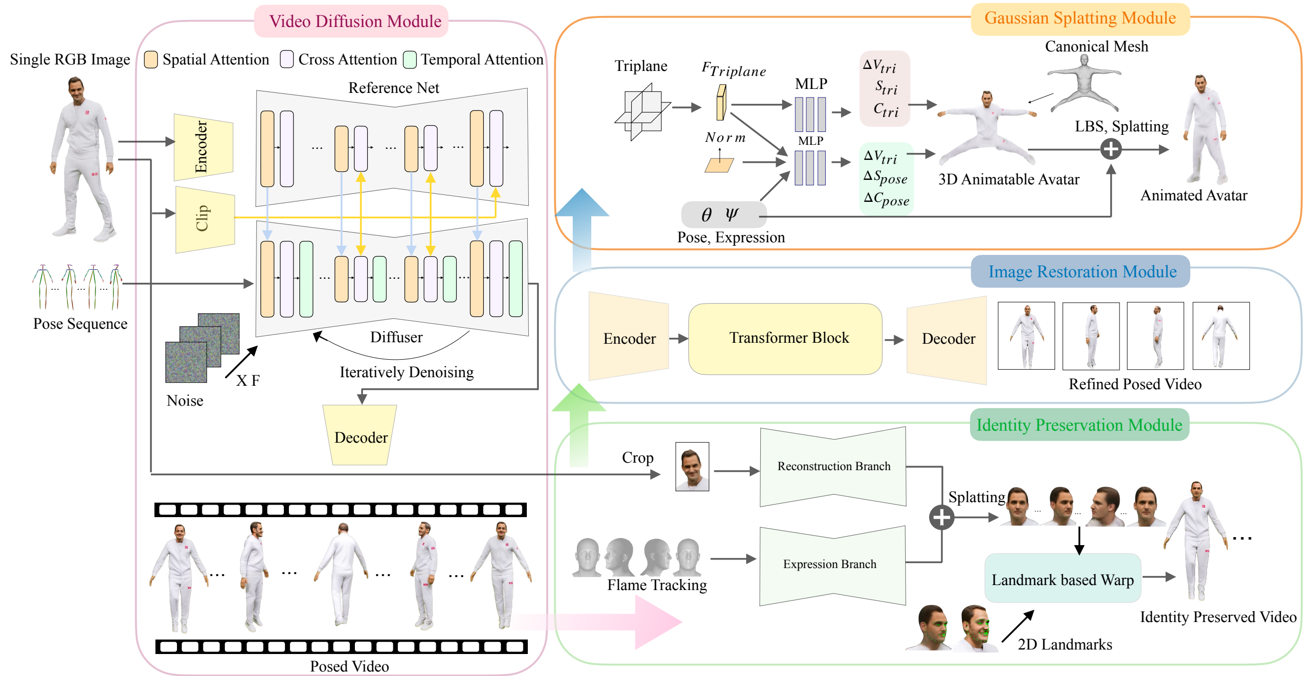

Figure 2. **Overall Pipeline of SVAD.** Starting from a single input image, the diffusion model generates pose-conditioned animations, which are refined using an identity preservation module and an image restoration module. The refined outputs are then used to train the 3DGS avatar, enabling high-fidelity,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

lined 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 [57], a finetuned variant of Animate Anyone [18], which is a state-of-the-art video diffusion model designed for realistic human animation while maintaining temporal consistency and appearance fidelity. MusePose employs a U-Net-based diffusion architecture with integrated pose and temporal controls, allowing for pose-guided animation across frames.

The model architecture incorporates several key components for effective character animation. The denoising UNet is implemented as a 3D UNet [11] with motion modules for temporal coherence. Specifically, we use Vanilla motion modules with temporal self-attention blocks at resolutions of [1, 2, 4, 8] and in the mid-block. Each transformer 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 [49] formulation with zero-SNR sampling [35], using a scaled linear beta schedule with $\beta_{\text{start}} = 0.00085$ and $\beta_{\text{end}} = 0.012$. The DDIM sampler [55] is configured for efficient inference with 20 sampling steps and a classifier-free guidance scale of 3.5. The temporal consistency loss L_{temp} minimizes discrepancies across successive frames by enforcing coherence in appearance and pose:

$$L_{\text{temp}} = \sum_{t=1}^{T-1} \|F_t - F_{t+1}\|^2 \tag{2}$$

where $F_t \in \mathbb{R}^{H\times W\times 3}$ represents the RGB frame output at time t.

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. 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 [63], we compute scale parameters $\mathbf{S} = \{s_1, s_2, \ldots, s_{10}\}$ for ten distinct body regions: neck,

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.

face, shoulders, upper arms, lower arms, hands, torso, upper legs, and lower legs. For each body part i, we compute its scale factor s_i as the ratio between the corresponding keypoint distances:

$$s_i = \frac{d_i^{ref}}{d_i^{src}} \tag{3}$$

where d_i^{ref} and d_i^{src} represent the Euclidean distances between keypoints. For body parts with bilateral symmetry (e.g., arms), we average the scales from both sides:

$$s_{arm\_upper} = \frac{1}{2}\left(\frac{\|p^2_{ref} - p^3_{ref}\|}{\|p^2_{src} - p^3_{src}\|} + \frac{\|p^5_{ref} - p^6_{ref}\|}{\|p^5_{src} - p^6_{src}\|}\right) \tag{4}$$

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

$$p' = c_i + s_i \cdot (p - c_i) \tag{5}$$

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 head reconstruction and facial fusion techniques. This module leverages a 3DGS-based head reconstruction method inspired by Chu *et al.* [10] to create a 3D Gaussian-based head avatar from a single input image using a novel *dual-lifting* approach that predicts both forward and backward lifting distances.

Given an input image I_s, global and local features F_{local} are extracted using a frozen DINOv2 [39] backbone. These features are used to predict forward and backward lifting distances, positioning 3D Gaussians G_{pos} as follows:

$$G_{\text{pos}} = [\mathbf{p}_s + E_{\text{Conv0}}(F_{\text{local}}) \cdot \mathbf{n}_s, \mathbf{p}_s - E_{\text{Conv1}}(F_{\text{local}}) \cdot \mathbf{n}_s], \quad (6)$$

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 features:

$$G_{\text{expr}} = \text{MLP}(F_{\text{3DMM}} + F_{\text{global}}). \quad (7)$$

After generating the head avatar renderings, we detect facial landmarks on both the original frame I_{orig} and the generated head image I_{head}, compute an affine transformation for alignment, and use Poisson image editing [41] for seamless fusion:

$$\min_I \int_\Omega \|\nabla I - \nabla I_{\text{warp}}\|^2 \, dx \, dy, \quad \text{subject to } I|_{\partial\Omega} = I_{\text{orig}}|_{\partial\Omega}, \quad (8)$$

where Ω is defined by the facial mask, ensuring temporally consistent facial details throughout the animation.

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.* [7], specifically their diffusion-based image restoration method. BFRffusion leverages the generative prior encapsulated in the pretrained Stable Diffusion model [47] to enhance image details through a comprehensive architecture that effectively extracts features from low-quality images and restores realistic facial details.

In our method, we set the super-resolution scale factor to 2, enhancing input frames while maintaining computational efficiency. The process uses 50 DDIM sampling steps with a classifier-free guidance scale of 3.5, achieving a 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.

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 synthetic 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.* [37], which integrates the SMPL-X 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 $\mathbf{V} \in \mathbb{R}^{N\times3}$, RGB color values $\mathbf{C} \in \mathbb{R}^{N\times3}$, and a scale parameter $\mathbf{S} \in \mathbb{R}^{N}$, where N is the number of Gaussians. The Gaussian splatting rendering equation is:

$$I = f(V, \exp(S), C, K, E), \quad (9)$$

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:

$$\mathbf{V}_{\text{pose}} = \mathbf{V} + \Delta\mathbf{V}_{\text{pose}} + \Delta\mathbf{V}_{\text{expr}}. \quad (10)$$

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

$$L_{\text{Lap}} = \|\Delta\mathbf{V}_{\text{canonical}} - \Delta\mathbf{V}_{\text{deformed}}\|^2. \quad (11)$$

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 We use the People-Snapshot dataset[2], which contains videos of individuals rotating in front of a stationary camera. For consistency and fair comparison, we adhere to the evaluation protocol established by InstantAvatar[26] .

THuman The PeopleSnapshot dataset has limited pose variations. To evaluate the performance on more challenging test poses, we utilize the THuman dataset[65]. This dataset includes a diverse set of poses that require higher flexibility and adaptability.

Evaluation Metrics. We consider four metrics, PSNR[15], SSIM[60], LPIPS[70], and Clip Similarity[45](denoted as CLIP in tables) to assess the reconstruction quality on three datasets.

CVPR Workshop SyntaGen #16

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 [61]	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 [19]	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 [37]	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 [37] (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 without the *Single Image* utilize approximately 200 input frames.

Figure 4. **Synthetic Data Generated by SVAD.** Comparison of the original input image (left), unprocessed synthetic data from video diffusion (middle), and our enhanced synthetic data after applying identity preservation and image restoration modules (right). Note how our data augmentation approach preserves facial identity and significantly improves detail quality in the synthetic training data.

4.2. Quantitative Evaluation

We quantitatively evaluate the quality of single-image 3D avatars generated by our method against SOTA 3D avatar generation methods [19, 37, 61]. 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 [37]. 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 [6, 17, 23, 48, 71], many of which employ the SMPL model, allowing for animatability through mesh fitting and reposing techniques, such as those in Editable Humans [16]. We randomly sample 100 scans from the THuman dataset and report results. We repose our trained avatar using ground-truth SMPL-X parameters, then render four viewpoints (front, left, right, back) for each model and compare with the ground-truth scan renderings from the same views. As presented in Table 4, our method surpasses all other baselines, demonstrating superior quality in single-image 3D human reconstruction tasks.

Method	PSNR↑	SSIM↑	LPIPS↓	CLIP↑
PIFu[48]	15.62	0.8921	0.1903	0.8612
TeCH[23]	15.85	0.8892	0.1667	0.8890
Ultraman[6]	18.13	0.9019	0.1334	0.9089
SIFU[71]	18.59	0.8591	0.1402	0.8873
SITH[17]	19.98	0.9018	0.1294	0.9084
Ours	**20.92**	**0.9291**	**0.1124**	**0.9321**

Table 2. **Quantitative Evaluation.** SVAD compared to recent single-image based 3D human generation SOTAs. Top two results are colored as first second.

4.3. Qualitative Evaluation

Figure 5 shows the overall quality of our generated 3D avatars from single images in the People Snapshot and the THuman dataset. Figure 6, Figure 7 shows that our method performs superior compared to current single-view human reconstruction SOTA [17]. For single image avatar generation, we evaluate on the People Snapshot dataset and compare against the 3D-GS based avatar SOTA [37]. For fairness, we train the SOTA with the single input image for the same amount of iterations (12k iter) as our method. Figure 8 shows that for single image avatar generation, our method performs superior especially for the back and side views.

Figure 5. **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.

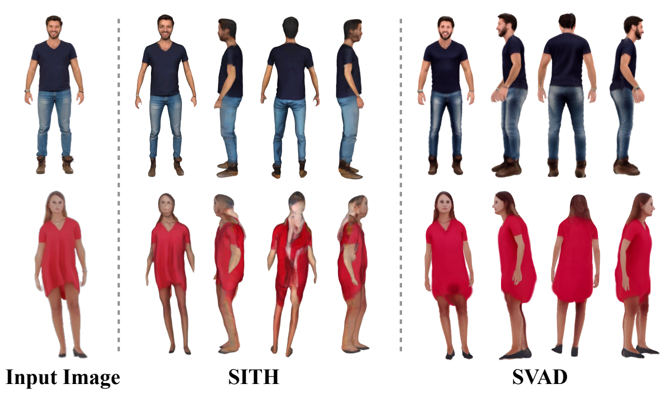

Figure 6. **Qualitative Comparison with SITH [17].** Our approach better reconstructs complex contours and subtle features, resulting in a more lifelike and coherent side-view appearance.

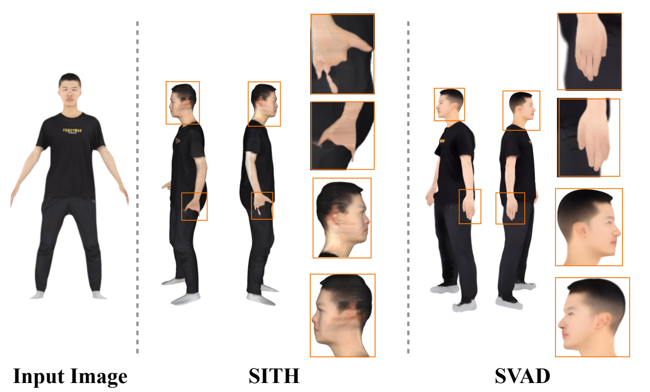

Figure 7. **Qualitative Evaluation against SITH [17].** Our method reconstructs fine detail(hands), while preserving original identity in facial regions.

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 to reach the optimal performance reflected by all the metrics. Using the THuman dataset, we apply the same eval-

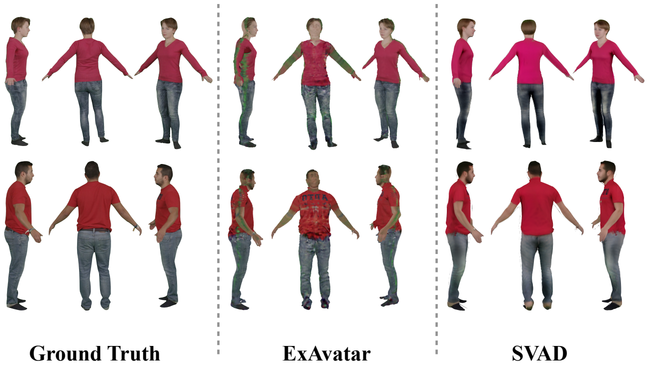

Figure 8. **Qualitative Evaluation against ExAvatar in single image input.**

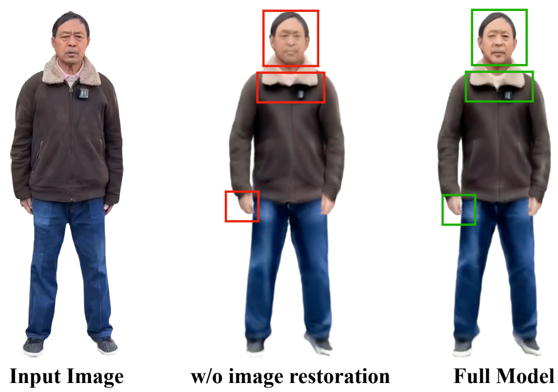

Figure 9. **Ablation study on the image restoration module.** We show that applying the module into our pipeline recover fine details on the final avatar output.

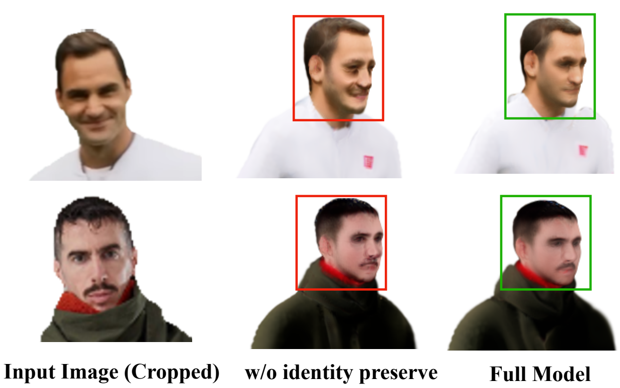

Figure 10. **Ablation study on the identity preservation module.** We show that with the module, the final avatar maintains facial details on the original input image.

uation 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 9 shows visual results of the effect of the image restoration module. High-detailed regions such as clothing texture,

Method	PSNR↑	SSIM↑	LPIPS↓	CLIP↑
w/o Identity Preserve	25.19	0.9419	0.0623	0.9231
w/o Image Restoration	25.61	0.9298	0.0645	0.9239
Ours (Full)	25.79	0.9502	0.0594	0.9241

Table 3. **Ablation study on the People Snapshot dataset.**

Method	PSNR↑	SSIM↑	LPIPS↓	CLIP↑
w/o Identity Preserve	21.12	0.9256	0.0898	0.9284
w/o Image Restoration	21.16	0.9212	0.0799	0.9201
Ours (Full)	21.95	0.9291	0.0824	0.9321

Table 4. **Ablation study on the THuman dataset.**

fingers, and facial details are better preserved when applying our module. Figure 10, shows the visual effect of the identity preservation module. We clearly show that original input's facial details are more preserved with the presence of our module.

5. Conclusion and Future Work

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 leveraging video diffusion models to generate pose-conditioned synthetic training data, and enhancing this data through identity preservation and image restoration, SVAD successfully addresses key challenges in single-image avatar creation. Our method demonstrates how carefully refined synthetic data can overcome limitations in traditional approaches, enabling stable, visually consistent avatars that retain the original subject's identity and details across varied poses and viewpoints. Through comprehensive evaluations, SVAD achieves state-of-the-art performance compared to existing methods while maintaining the rendering efficiency of 3DGS.

Limitations and Future Work. Our method has three primary limitations. First, the synthetic data generation process relies on computationally intensive video diffusion models, creating a bottleneck in the pipeline. Second, while effective for diverse poses represented in the synthetic training data, SVAD struggles with extreme or unconventional poses that fall outside this distribution. Third, the method does not handle object interactions, limiting its applicability in dynamic environments.

In future work, we plan to explore more efficient synthetic data generation techniques to reduce computational requirements while maintaining quality. We aim to expand the diversity of synthetic training data to include extreme poses and object interactions, and investigate how synthetic data generation could benefit multi-subject scenarios and clothing variation to extend SVAD's capabilities beyond single-subject avatars.

CVPR Workshop SyntaGen #16

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 Conference Papers*, pages 1–11, 2023. 2

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, pages 8387–8397, 2018. 5, 6

[3] 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*, pages 5968–5976, 2023. 2

[4] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In *CVPR*, pages 958–968, 2024. 2

[5] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *ICCV*, pages 5933–5942, 2019. 2

[6] 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] 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. 5

[8] Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. Monogaussianavatar: Monocular gaussian point-based head avatar. In *SIGGRAPH*, pages 1–9, 2024. 2

[9] Yushuo Chen, Zerong Zheng, Zhe Li, Chao Xu, and Yebin Liu. Meshavatar: Learning high-quality triangular human avatars from multi-view videos. *arXiv preprint arXiv:2407.08414*, 2024. 1

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

[11] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 3

[12] 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*, pages 459–476, 2025. 2

[13] Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen. 4d gaussian splatting: Towards efficient novel view synthesis for dynamic scenes. *arXiv preprint arXiv:2402.03307*, 2024. 2

[14] 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*, pages 1–19. Springer, 2022. 2

[15] Rafael C Gonzalez and Richard E Woods. *Digital Image Processing*. Prentice Hall, 2008. 5

[16] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In *CVPR*, pages 21024–21035, 2023. 6

[17] I Ho, Jie Song, Otmar Hilliges, et al. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *CVPR*, pages 538–549, 2024. 6, 7

[18] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, pages 8153–8163, 2024. 2, 3

[19] 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*, pages 634–644, 2024. 1, 2, 6

[20] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. In *ICCV*, pages 9352–9364, 2023. 2

[21] Shoukang Hu, Tao Hu, and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *CVPR*, pages 20418–20431, 2024. 2

[22] 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

[23] 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*, pages 1531–1542, 2024. 6

[24] 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

[25] Suyi Jiang, Haoran Jiang, Ziyu Wang, Haimin Luo, Wenzheng Chen, and Lan Xu. Humangen: Generating human radiance fields with explicit priors. In *CVPR*, pages 12543–12554, 2023. 2

[26] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *CVPR*, pages 16922–16932, 2023. 5

[27] 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*, pages 5322–5332, 2024. 2

[28] Brennan Jones, Yaying Zhang, Priscilla NY Wong, and Sean Rintel. Belonging there: Vroom-ing into the uncanny valley of xr telepresence. *ACM HCI*, 5(CSCW1):1–31, 2021. 2

[29] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *ICCV*, pages 22623–22633, 2023. 2

[30] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4):139–1, 2023. 2

CVPR Workshop SyntaGen #16

[31] Agelos Kratimenos, Jiahui Lei, and Kostas Daniilidis. Dynmf: Neural motion factorization for real-time dynamic view synthesis with 3d gaussian splatting. In *ECCV*, pages 252–269, 2025. 2

[32] Byeonghyeon Lee, Howoong Lee, Xiangyu Sun, Usman Ali, and Eunbyung Park. Deblurring 3d gaussian splatting. *arXiv preprint arXiv:2401.00834*, 2024. 2

[33] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *CVPR*, pages 19876–19887, 2024. 2

[34] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *CVPR*, pages 19711–19722, 2024. 2

[35] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5404–5411, 2024. 3

[36] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713*, 2023. 2

[37] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. *ECCV*, 2024. 1, 2, 5, 6

[38] 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*, pages 788–798, 2024. 1

[39] 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. *arXiv preprint arXiv:2304.07193*, 2023. 5

[40] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *CVPR*, pages 1165–1175, 2024. 2

[41] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 577–582. 2023. 5

[42] Sergey Prokudin, Michael J Black, and Javier Romero. Smplpix: Neural avatars from 3d human models. In *WACV*, pages 1810–1819, 2021. 2

[43] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *CVPR*, pages 20299–20309, 2024. 2

[44] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *CVPR*, pages 5020–5030, 2024. 1, 2

[45] 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. 5

[46] Yurui Ren, Ge Li, Shan Liu, and Thomas H Li. Deep spatial transformation for pose-guided person image generation and animation. *IEEE TIP*, 29:8622–8635, 2020. 2

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 5

[48] 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*, pages 2304–2314, 2019. 6

[49] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 3

[50] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural re-rendering of humans from a single image. In *ECCV*, pages 596–613, 2020. 2

[51] 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*, pages 1606–1616, 2024. 1

[52] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-avatar: Expressive human avatars. In *CVPR*, pages 16911–16921, 2023. 1

[53] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *NeurIPS*, 32, 2019. 2

[54] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, pages 13653–13662, 2021. 2

[55] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[56] David Svitov, Pietro Morerio, Lourdes Agapito, and Alessio Del Bue. Haha: Highly articulated gaussian human avatars with textured mesh prior. *arXiv preprint arXiv:2404.01053*, 2024. 1

[57] 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. 3

[58] 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

[59] 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*, pages 9326–9336, 2024. 2

[60] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 5

[61] 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*, pages 16210–16220, 2022. 6

[62] 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*, pages 1481–1490, 2024. 2

[63] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 3

[64] 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*, pages 15039–15048, 2021. 2

[65] 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. 5

[66] 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*, pages 7502–7512, 2023. 2

[67] 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

[68] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2

[69] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *CVPR*, pages 7713–7722, 2022. 2

[70] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 5

[71] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *CVPR*, pages 9936–9947, 2024. 6

[72] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *CVPR*, pages 3657–3666, 2022. 2

[73] 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*, pages 19680–19690, 2024. 2

[74] 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. *arXiv preprint arXiv:2404.04421*, 2024. 1

[75] 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

[76] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. *arXiv preprint arXiv:2311.08581*, 2023. 2