

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

Supplementary Material

001 Supplementary Material

002 This supplementary material provides additional details to
003 complement the main paper. In Sec. 1, we elaborate on the
004 implementation details of our pipeline, covering the pre-
005 defined pose sequences utilized in our pipeline, the image
006 restoration module for enhancing facial fidelity and overall
007 texture quality, and the fitting and training methodology for
008 the Gaussian avatar module. Sec. 2 presents results on the
009 data augmentation methods, highlighting their impact on
010 improving the quality of training data for the 3DGS-based
011 avatar model, with a focus on identity preservation and im-
012 age restoration. In Sec. 3, we demonstrate the robustness
013 of our method in handling challenging poses, including ex-
014 treme body movements, occlusions, and non-frontal facial
015 orientations. Sec. 4 discusses failure cases, identifying key
016 limitations, while Sec. 5 outlines potential directions for fu-
017 ture work aimed at addressing these challenges and further
018 enhancing the robustness and realism of our approach.

019 1. Implementation Details

020 1.1. Predefined Pose Sequences

021 To initialize frame generation for our pipeline, we rely on a
022 predefined set of poses extracted from the People Snapshot
023 dataset. Specifically, we utilize the *male-4-casual* sequence,
024 which depicts a subject performing a full-body rotation with
025 arms extended horizontally. Using DWPose [15], we extract
026 2D keypoints $K \in \mathbb{R}^{J \times 2}$, where $J = 17$ is the number of
027 keypoints, from this sequence to create a standardized pose
028 template. This sequence serves as the conditioning input
029 for all video diffusion model generations, resulting in 187
030 frames per sequence, each with a resolution of 1024×1024
031 pixels.

032 Following the initial frame generation by the video dif-
033 fusion model, we refine the facial regions to enhance iden-
034 tity consistency and detail preservation. To achieve this,
035 we leverage GAGAvatar [3] to generate a 3D head avatar
036 from the single input image. The generated 3D head, repre-
037 sented as a set of 3D Gaussians with parameters $G = \{V \in$
038 $\mathbb{R}^{N \times 3}, C \in \mathbb{R}^{N \times 3}, S \in \mathbb{R}^N\}$, where N is the number of
039 Gaussians, is fused into the raw video diffusion output to
040 replace the initial low-fidelity facial regions.

041 To ensure accurate alignment between the generated
042 3D head and the original diffusion output, we extract
043 FLAME [8] parameters $\theta \in \mathbb{R}^{|\theta|}$, which encode expres-
044 sion and pose, and apply them to guide the rendering of
045 the 3D head. The FLAME parameters include facial shape

$\beta \in \mathbb{R}^{10}$, pose $\psi \in \mathbb{R}^6$, and expression $\phi \in \mathbb{R}^{10}$.

For all experiments, the body pose sequence and
FLAME parameters obtained above remain fixed, provid-
ing a consistent reference for pose-guided video generation
and refinement.

051 1.2. Image Restoration Submodule

052 To provide more details on the image restoration submod-
053 ule, we leverage the work by Chen *et al.* [2] and apply
054 super-resolution to enhance the quality of our training data.
055 Specifically, we use a hybrid restoration pipeline that in-
056 tegrates Real-ESRGAN [13] as the background upsampler
057 and a diffusion-based face restoration method to ensure both
058 global fidelity and local detail preservation.

059 The restoration process begins with face detection and
060 alignment using RetinaFace [4]. The detected facial regions
061 are then passed through a diffusion model, guided by con-
062 ditional embeddings generated from the low-resolution in-
063 put. The model iteratively refines the high-resolution details
064 while maintaining consistency with the original identity.

065 For background regions, Real-ESRGAN [13] is applied
066 to upscale non-facial areas without introducing artifacts.
067 The restored facial regions are seamlessly integrated into
068 the upscaled background using a face restoration helper
069 module [2]. This ensures that the enhanced facial details
070 blend naturally with the surrounding context.

071 1.3. Gaussian Avatar Submodule

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

077 1.3.1 SMPL-X Fitting Process

078 Prior to training the Gaussian avatar, we employ a compre-
079 hensive fitting process to obtain accurate SMPL-X param-
080 eters from our synthetic data. This multi-stage process en-
081 sures that the avatar’s geometry accurately reflects the sub-
082 ject’s physical characteristics and articulation.

083 **Keypoint Extraction.** The fitting pipeline begins with pose
084 and shape estimation. We utilize DWPose [15] to extract 2D
085 whole-body keypoints from each frame of our synthetic se-
086 quence. These keypoints provide critical information about
087 body articulation across the sequence. The keypoints are
088 represented as $K \in \mathbb{R}^{J \times 3}$, where $J = 133$ includes 17

body, 68 face, and 42 hand keypoints, with each keypoint having $(x, y, \text{confidence})$ values. We then employ MM-POSE [12] with the RTMPose-L model for refinement, using a confidence threshold of 0.5 to filter reliable detections.

Initial Parameter Estimation. For facial geometry, we leverage DECA [5] 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 [9] 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 $\theta_{\text{body}} \in \mathbb{R}^{21 \times 3}$, jaw pose $\theta_{\text{jaw}} \in \mathbb{R}^3$, hand poses $\theta_{\text{hands}} \in \mathbb{R}^{30 \times 3}$, 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:

$$L_{\text{fit}} = \lambda_{\text{kpt}} L_{\text{kpt}} + \lambda_{\text{reg}} L_{\text{reg}} + \lambda_{\text{temp}} L_{\text{temp}} \quad (1)$$

The keypoint reprojection loss L_{kpt} measures the distance between projected model joints and detected 2D keypoints, weighted by detection confidence:

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

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:

$$L_{\text{reg}} = \|\beta\|_2^2 + \sum_j \|\theta_j - \theta_{\text{mean}}\|_2^2 \quad (3)$$

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

$$L_{\text{temp}} = \sum_{t=1}^{T-1} \|\theta_t - \theta_{t+1}\|_2^2 + \|\beta_t - \beta_{t+1}\|_2^2 \quad (4)$$

The optimization uses the Adam optimizer 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 Savitzky-Golay [6] filtering with

a window length of 9 frames and polynomial order of 2. For rotation parameters, we employ a quaternion-based smoothing procedure. The quaternion smoothing incorporates a continuity enforcement algorithm to handle sign flips:

$$q'_{t+1} = \begin{cases} -q_{t+1}, & \text{if } q_t \cdot q_{t+1} < 0 \\ q_{t+1}, & \text{otherwise} \end{cases} \quad (5)$$

Segmentation and Depth Estimation. We generate foreground masks using the Segment Anything Model [7] 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 [14] with the ViT-L backbone. The depth maps are normalized and aligned with the SMPL-X mesh using the following procedure:

$$\text{scale} = \frac{\sigma(\text{depth}_{\text{pred}, \text{fg}})}{\sigma(\text{depth}_{\text{smplx}, \text{fg}})}$$

$$\text{depth}'_{\text{pred}} = \frac{\text{depth}_{\text{pred}}}{\text{scale}}$$

$$\text{depth}'_{\text{pred}} = \text{depth}'_{\text{pred}} - \mu(\text{depth}'_{\text{pred}, \text{fg}}) + \mu(\text{depth}_{\text{smplx}, \text{fg}}) \quad (6)$$

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

These processes provide a comprehensive set of parameters and auxiliary data that serve as the foundation for the subsequent Gaussian avatar training.

1.3.2 3DGS Avatar Training Process

With the fitted SMPL-X parameters and processed synthetic data, we proceed to train the 3DGS-based avatar [10]. The training begins by initializing the triplane representation $T \in \mathbb{R}^{32 \times 128 \times 128}$, encoding 3D features for both body and facial regions. Gaussian parameters, including positions $\mathbf{V} \in \mathbb{R}^{N \times 3}$, colors $\mathbf{C} \in \mathbb{R}^{N \times 3}$, and opacity $\mathbf{O} \in \mathbb{R}^N$, are optimized through backpropagation with the following multi-objective loss function:

$$L = \lambda_{\text{RGB}} L_{\text{RGB}} + \lambda_{\text{SSIM}} L_{\text{SSIM}} + \lambda_{\text{LPIPS}} L_{\text{LPIPS}}, \quad (7)$$

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:

$$\alpha_{\text{position}}(t) = \alpha_{\text{init}} \times \left(1 - \frac{t}{T_{\text{max}}}\right) + \alpha_{\text{final}} \times \frac{t}{T_{\text{max}}}, \quad (8)$$

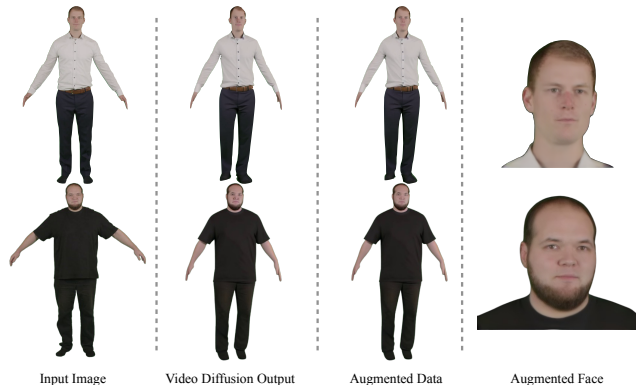


Figure 1. **Data Augmentation Results.** This figure highlights the effectiveness of our data augmentation pipeline, showcasing enhanced facial regions and overall image quality improvements achieved through the identity preservation and image restoration sub-modules

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:

$$\mathbf{V}_{\text{new}} = \mathbf{V}_{\text{old}} - \eta \frac{\partial L}{\partial \mathbf{V}}, \quad (9)$$

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 [11] model due to its superior performance in complex sequences. The entire pipeline runs on a single GPU, ensuring scalability and efficiency.

2. Details on Generated Synthetic Data

We show our augmented data from the sequences of the People Snapshot [1] dataset. As shown in Fig. 1, applying our data augmentation module consisted of the identity preservation and the image restoration sub-module enhance the overall quality of the data, especially the facial regions.



Figure 2. **Challenging Poses.** The figure illustrates the robustness of our method in handling extreme poses, including non-frontal views and dynamic motion scenarios, while maintaining high fidelity and consistency in the generated avatars.



Figure 3. **Failure cases of SVAD.** Examples include noisy back and side views, inconsistent clothing textures, and artifacts in non-frontal regions.

3. Challenging Poses

As shown in Fig. 2, our method demonstrates exceptional robustness to challenging poses, including extreme body movements, occlusions, and non-frontal facial orientations. This robustness is achieved through the integration of pose-guided video diffusion models and the 3D Gaussian splatting framework, which together enable high-fidelity avatar generation that remains consistent across a wide range of poses and motions. The ability to handle such diverse and dynamic poses is critical for applications requiring realistic and adaptable avatar rendering. The capability to handle such challenging poses and motion scenarios establishes the robustness and versatility of our method, making it well-suited for applications in gaming, virtual reality, and animation. Future enhancements, such as incorporating additional motion datasets and refining pose-guidance mechanisms, could further extend this capability to even more complex and dynamic scenarios.

4. Failure Cases

In this section, we analyze several failure cases observed in SVAD, revealing limitations in specific scenarios that highlight areas for potential improvement. As in Fig. 3 one of the primary challenges lies in the generation of back and side views. Despite using a pretrained video diffusion

227 model trained on 3D scan data, the inherent bias towards
228 frontal views within diffusion models often results in noisy
229 or inaccurate reconstructions of non-frontal regions. These
230 inconsistencies are particularly evident in textured areas,
231 such as clothing and hair, where fine details are difficult to
232 maintain without multi-view constraints.

233 Another issue arises in maintaining consistent textures
234 and lighting across different viewpoints. Artifacts such as
235 abrupt transitions in lighting or shading irregularities ap-
236 pear, particularly in side or back views. These imperfec-
237 tions likely stem from limitations in the data augmentation
238 process, as synthesized views may not fully capture the di-
239 versity of real-world lighting conditions and texture varia-
240 tions. These inconsistencies affect the overall visual fidelity
241 and reduce the photorealism of the rendered avatars.

242 Additionally, while the 3D Gaussian splatting represen-
243 tation is effective for free-viewpoint rendering, its reliance
244 on isotropic Gaussians can lead to oversmoothing in high-
245 frequency regions such as hands and facial features. This
246 limitation occasionally causes a loss of sharpness and detail
247 in regions where fine textures are crucial for realism. gauss

248 5. Future Work

249 Addressing these limitations in Sec. 4 requires several
250 enhancements. Improvements to the data augmentation
251 pipeline, such as introducing more realistic texture and
252 lighting variations, could help mitigate shading and tex-
253 ture artifacts. Regularization techniques could enforce more
254 consistent geometry and appearance across views, while
255 hybrid volumetric representations or pose-dependent defor-
256 mation fields could improve accuracy in challenging poses.
257 These advancements would help SVAD achieve greater ro-
258 bustness and fidelity across diverse scenarios.

References

- [1] 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. 3
- [2] 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. 1
- [3] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. *NeurIPS*, 2024. 1
- [4] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, pages 5203–5212, 2020. 1
- [5] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4): 1–13, 2021. 2
- [6] Arlene John, Jishnu Sadasivan, and Chandra Sekhar Seelamantula. Adaptive savitzky-golay filtering in non-gaussian noise. *IEEE Transactions on Signal Processing*, 69:5021–5036, 2021. 2
- [7] 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*, pages 4015–4026, 2023. 2
- [8] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *TOG*, 36(6):194–1, 2017. 1
- [9] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *CVPR*, pages 2308–2317, 2022. 2
- [10] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. *ECCV*, 2024. 2
- [11] 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*, pages 10975–10985, 2019. 3
- [12] 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*, 20(17):10032–10044, 2020. 2
- [13] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, pages 1905–1914, 2021. 1
- [14] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xianggang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 2
- [15] 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. 1