PERSONA: Personalized Whole-Body 3D Avatar with Pose-Driven Deformations from a Single Image

Geonhee Sim Gyeongsik Moon
Dept. of CSE, Korea University
{kh6362,mks0601}@korea.ac.kr
https://mks0601.github.io/PERSONA

Abstract

Two major approaches exist for creating animatable human avatars. The first, a 3D-based approach, optimizes a NeRF- or 3DGS-based avatar from videos of a single person, achieving personalization through a disentangled identity representation. However, modeling pose-driven deformations, such as non-rigid cloth deformations, requires numerous pose-rich videos, which are costly and impractical to capture in daily life. The second, a diffusion-based approach, learns pose-driven deformations from large-scale in-the-wild videos but struggles with identity preservation and pose-dependent identity entanglement. We present PERSONA, a framework that combines the strengths of both approaches to obtain a personalized 3D human avatar with pose-driven deformations from a single image. PERSONA leverages a diffusion-based approach to generate pose-rich videos from the input image and optimizes a 3D avatar based on them. To ensure high authenticity and sharp renderings across diverse poses, we introduce balanced sampling and geometry-weighted optimization. Balanced sampling oversamples the input image to mitigate identity shifts in diffusion-generated training videos. Geometry-weighted optimization prioritizes geometry constraints over image loss, preserving rendering quality in diverse poses.

1. Introduction

Creating an animatable human avatar is a long-standing challenge in computer vision and graphics. An animatable human avatar is a representation that (1) can be driven by novel whole-body poses and facial expressions and (2) can be rendered from any viewpoint. Early approaches [7, 27, 41, 42] relied on multi-view video data and accurately tracked 3D poses. While these methods achieved impressive results in controlled environments, their practical applicability was limited due to the difficulty of acquiring such data in everyday scenarios. Recent methods have significantly lowered data requirements, allowing avatars to be built from casually captured short monocular videos [9, 14, 18, 19, 24, 37, 43] or even a single image [15, 33, 63, 66], eliminating the need for multi-view recordings and precisely tracked 3D poses.

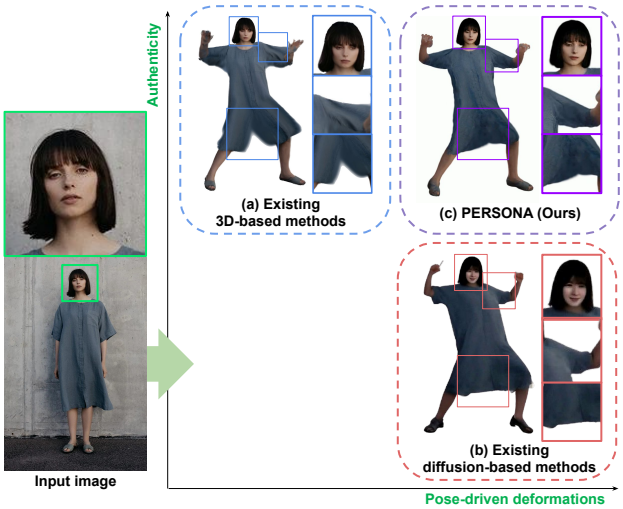


Figure 1. Comparison of (a) existing 3D-based method [44], (b) existing diffusion-based method [66], and (c) our PERSONA. PERSONA integrates the strengths of both approaches to achieve a personalized whole-body 3D avatar with pose-driven deformations.

Fig.1 illustrates the two dominant approaches for creating animatable human avatars. The first, a 3D-based approach[9, 14, 18, 19, 24, 37, 43–45, 70] (Fig.1 (a)), combines neural rendering techniques (*e.g.*, NeRF[34], 3DGS [21]) with 3D human parametric models (*e.g.*, SMPL [32], SMPL-X [40]). Neural rendering captures appearance and geometry, while parametric models enable animation. Early methods [9, 14, 18, 19, 24, 37, 43] reconstruct avatars from short monocular videos of subjects rotating in an A-pose. More recent approaches [44, 45, 70] directly generate animatable 3D Gaussian avatars in a feed-forward manner. This framework allows for clear dis-

entanglement of identity and pose, enabling personalized avatar creation and faithful identity preservation. However, as these methods animate avatars via 3D parametric models that mainly support rigid deformations, capturing pose-driven deformations such as non-rigid clothing motion requires large-scale pose-diverse datasets. Acquiring such data per subject is costly and impractical, leading most 3D methods to rely on static or simple-pose datasets. As a result, these avatars often lack expressiveness in handling complex, pose-dependent clothing deformations.

The second approach, a diffusion-based method (Fig. 1 (b)) [15, 33, 58, 63, 66, 69], generates animated human videos directly from a conditional 2D pose sequence using diffusion-based generative models [4, 12, 46, 54–57], without relying on neural rendering or parametric models. Trained on large-scale video datasets, these models effectively capture pose-driven deformations. However, they face significant challenges in identity preservation. They struggle to 1) retain the identity of the person in the input image and 2) maintain identity consistency when animating avatars, resulting in limited personalization capability. This limitation arises because identity representation is not fully disentangled from pose, often causing pose-dependent identity variations that distort the subject's original appearance.

We present PERSONA (Fig. 1 (c)), a framework for creating personalized 3D avatars with pose-driven deformations from a single image by leveraging diffusion-generated training videos. This eliminates the need for extensive per-individual data capture, making the pipeline highly scalable. PERSONA combines the strengths of 3D-based and diffusion-based approaches—3D-based methods effectively preserve identity but struggle with pose-driven non-rigid deformations in casually captured data (*e.g.*, a short monocular video of a subject rotating in an A-pose). In contrast, diffusion-based methods capture pose-dependent deformations but lack personalization. Our approach bridges this gap, achieving both identity preservation and pose-driven deformations in a scalable and efficient manner.

We introduce two key components to address the main challenges. First, we propose balanced sampling to ensure high authenticity in personalization. Diffusion-generated training videos often fail to fully preserve identity, leading to inconsistencies across poses. To mitigate identity shifts, our balanced sampling oversamples the input image during avatar optimization. In addition, we prevent baked-in artifacts such as shadows and pose-dependent geometry (*e.g.*, cloth wrinkles) of the input image. This approach enhances identity preservation while minimizing baked-in artifacts in novel poses.

Second, we propose geometry-weighted optimization to maintain sharp renderings across diverse poses. Diffusion-generated videos often contain inconsistent or artifact-prone textures, and directly optimizing the avatar on such frames degrades visual quality. Balanced sampling alone is insufficient for preserving sharp renderings in poses different from the input image, as the pose-driven deformation modeling module differentiates between the input image and generated frames based on their pose information, causing the model to adapt to the low-quality outputs of generated frames. Simply detaching textures from pose-driven deformation modeling is ineffective, as image loss still encourages geometry to replicate artifacts from generated frames, leading to degraded renderings. To address this, geometry-weighted optimization assigns low image loss weights and high geometry loss weights. Since geometry (*e.g.*, binary mask, depth maps, normal maps, and part segmentations) remains stable despite texture inconsistencies, it serves as a reliable constraint for non-rigid deformations. Additionally, omitting scale offsets in pose-driven deformation modeling prevents blurriness, significantly contributing to sharp renderings across diverse poses.

Methods	Pose-invariant ID	Pose-driven deform.	Single img.
GaussianAvatar [14]	✓	✓	✗
ExAvatar [37]	✓	✓	✗
IDOL [70]	✓	✗	✓
AniGS [45]	✓	✗	✓
LHM [44]	✓	✗	✓
Animate Anyone [15]	✗	✓	✓
MimicMotion [66]	✗	✓	✓
Champ [69]	✗	✓	✓
StableAnimator [58]	✗	✓	✓
PERSONA (Ours)	✓	✓	✓

Table 1. Comparison of existing human avatar creation methods and the proposed PERSONA. The first block [14, 37, 44, 45, 70] represents 3D-based approaches, while the second block [15, 58, 66, 69] corresponds to diffusion-based approaches. Each column indicates whether the avatar's identity representation is pose-invariant, whether it supports pose-driven deformations (*e.g.*, non-rigid cloth deformations), and whether it can be created from a single image.

Despite the complementary strengths of 3D-based and diffusion-based methods in personalization and pose-driven deformations, few studies have effectively integrated them. We hope our work provides valuable insights for both research directions. Our key contributions are as follows:

- We propose PERSONA, a framework for creating personalized 3D avatars with pose-driven deformations from a single image by leveraging diffusion-generated pose-rich training videos, eliminating the need for extensive per-individual video capture.
- We introduce balanced sampling to ensure authentic identity consistency. It mitigates identity shifts in diffusion-generated videos while preventing baked-in artifacts such as shadows and pose-dependent geometry.
- We propose geometry-weighted optimization, which prioritizes geometry constraints over image loss, ensuring

sharp renderings across diverse poses.

2. Related works

Tab. 1 compares existing 3D-based and diffusion-based human avatar creation methods and the proposed PERSONA.

3D-based human avatars. Early works [1, 3, 27–29, 38, 39, 41, 42, 68] required accurate 3D pose tracking with multi-view images. Since accurate 3D pose tracking is rarely available in daily life, recent works focus on creating 3D avatars from casually captured monocular videos. Jiang *et al*. [19] introduced NeuMan, an in-the-wild dataset with a NeRF-based baseline. Guo *et al*. [9] proposed self-supervised scene-human decomposition, while Jiang *et al*. [18] developed a fast 3D avatar pipeline. Kocabas *et al*. [24] and Hu *et al*. [14] leveraged 3DGS to improve representation and regression from a posed SMPL [32] mesh. Liu *et al*. [31] introduced a whole-body avatar without facial expression animation, and Deng *et al*. [8] applied image-to-image translation for rendering. Recent 3DGS-based methods further refine avatars, including Gaussian-mesh associations [51], hybrid surface-mesh representations [37], and gradient-based optimization refinements [13]. Xiu *et al*. [62] generated avatars from personal photo collections. Concurrently, Zhuang *et al*. [70] and Qiu *et al*. [45] proposed single-image 3D avatar frameworks trained on static 3D scans with multi-view renderings. However, they do not support non-rigid deformations, as capturing diverse poses and clothing behaviors is difficult, limiting the availability of suitable training data. We address this by leveraging diffusion-generated videos, avoiding the need for extensive 3D data capture.

Diffusion-based human avatars. With the success of recent generative models for image [12, 46, 54–57] and video generation [4], dedicated generative models have been developed to animate a human from a single reference image using target 2D pose sequences. Unlike 3D-based avatar creation methods, which rely on casually captured short monocular videos, these generative models leverage large-scale datasets to learn the prior distribution of human motion. Xu *et al*. [63] developed a video diffusion model to encode temporal dynamics, while Hu *et al*. [15] incorporated spatial attention for enhanced detail preservation. Zhang *et al*. [66] introduced confidence-aware pose guidance, and Men *et al*. [33] designed compact spatial encodings that account for the 3D nature of videos. Tu *et al*. [58] proposed an ID Adapter for identity preservation, and Zhu *et al*. [69] used SMPL renderings [32] as a conditioning signal for human video generation.

3D human recovery from a single image. 3D human pose estimation methods [6, 20, 25, 30, 35, 36, 47] regress 3D joint angles and shape parameters of 3D human models [32, 40] from a single image. While these representations are animatable, they lack high-fidelity detail due to their reliance on simplified naked body geometry without textures. 3D reconstruction methods [2, 10, 16, 17, 26, 48, 49, 53, 60, 61, 67] recover posed 3D human geometry from a single image, with some approaches [2, 16, 17, 26, 48, 53, 67] also reconstructing texture. Although these methods achieve high-fidelity reconstruction, their animation capability remains limited, as posed 3D scans are inherently difficult to animate [50].

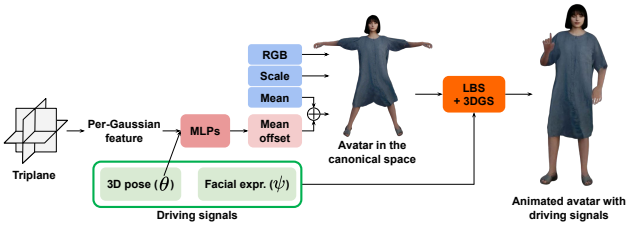


Figure 2. The pipeline of PERSONA. The mean offset from MLPs is used to represent pose-driven deformations.

3. Pipeline of PERSONA

Fig. 2 shows the pipeline of PERSONA.

Representation. We design PERSONA by combining the SMPL-X [40] parametric model with 3D Gaussian Splatting (3DGS) [21, 64]. SMPL-X enables whole-body animation, while 3DGS supports texture and geometry modeling along with rendering. Following ExAvatar [37], we adopt a hybrid representation of surface mesh and 3D Gaussians. Each vertex of the SMPL-X template mesh is modeled as a 3D Gaussian, with connectivity inherited from the SMPL-X triangle topology. To enhance generalization to novel views, we use isotropic Gaussians with constant opacity set to one.

Architecture. With the optimizable Gaussian features (*i.e.*, means, scales, and RGB colors), we introduce mean offsets to model pose-driven deformations. These offsets are predicted by multi-layer perceptrons (MLPs), which take as input the triplane features of each Gaussian point in the canonical space along with the 3D poses. To enhance generalization, the MLPs utilize only the 3D poses of 4-ring neighboring joints while setting non-neighboring joints to zero, following Saito *et al*. [50]. The final mean offsets, combined with facial expression-dependent vertex offsets from SMPL-X, are applied to the means in the canonical space.

Animation and rendering. The 3D human avatar is constructed in a canonical space and animated using SMPL-X 3D pose θ and facial expression parameter ψ. Each body Gaussian is assigned the average skinning weight of its 16 nearest SMPL-X template vertices, while original SMPL-X weights are retained for hands and face. The 3D Gaussians are animated using linear blend skinning (LBS), and the final rendering is performed with Mip-Splatting [64].

4. Generating training videos from an image

4.1. Pose-rich video generation

As shown in Fig.3, we generate pose-rich training videos using the diffusion-based human animation method MimicMotion [66]. The generated videos and the input image are used to construct our final avatar. These videos effectively compensate for the limited pose and deformation information available in a single image. In particular, they reveal how clothing deforms across different poses, which is essential for modeling pose-driven (*i.e.*, non-rigid) deformations. We generate various motion types, including dance sequences with peak poses, and rotating, light punching, and kicking actions that contain milder pose variations. We adopt MimicMotion for the training video generation as it outperforms other open-source diffusion-based methods [15, 58] for our task.

4.2. Geometric ID-preserving video generation

One of the key challenges in diffusion-based video generation is preserving the geometric identity of the person in the input image, such as bone lengths. To address this, we combine identity-related SMPL-X parameters (*e.g.*, shape) from the input image with target 3D poses that define the motion for animation. These target motions include a diverse range of actions, from mild (*e.g.*, rotation and light gestures) to strong (*e.g.*, dancing and kicking), and are extracted in advance from public videos using the ExAvatar [37] fitting process. We project the resulting SMPL-X keypoints to 2D space to create driving pose videos, which are used as input to diffusion-based animation methods [58, 66]. By using identity-related SMPL-X parameters from the input image, we better preserve the subject's identity compared to prior methods [58, 66], which rely on aligning a small number of vertical keypoints in the 2D space. Moreover, pre-computed 3D poses eliminate the need for the costly fitting process required by prior video-based methods [14, 37].

5. Personalize with pose-driven deformations

5.1. Balanced sampling

Balanced sampling. Fig. 4 (a) illustrates how balanced sampling alternates between the input image and generated video frames during training, effectively oversampling the input image. This helps prevent authenticity loss, as diffusion-generated videos often distort the subject's identity, especially in the face region. By using the input image more frequently, our approach maintains identity consistency in visible areas.

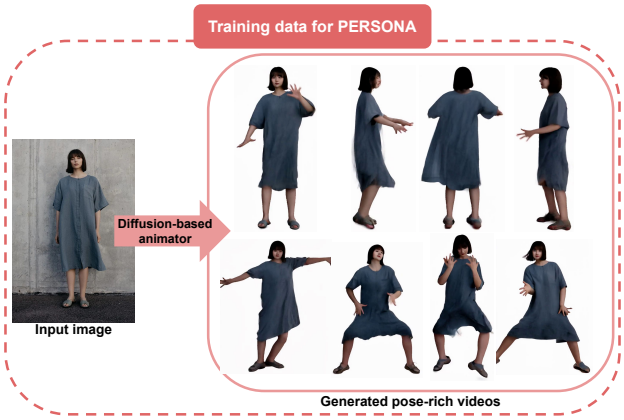


Figure 3. We use a diffusion-based animator [66] to generate pose-rich training videos from a single image. The input image and the generated videos together form our training set.

Reducing baked-in artifacts. Oversampling can introduce baked-in artifacts from the input image, including shadows in textures and seam artifacts between visible and invisible regions. These artifacts become embedded in the avatar's canonical space, causing issues when animating in novel poses, as they do not naturally adapt to new viewpoints or poses. To mitigate artifacts, we apply two strategies when supervising with the input image.

First, to reduce seam artifacts between visible and invisible regions, we identify seam boundaries by applying a Sobel filter to rendered positional maps from the canonical space, as shown in Fig. 5. The positional map is obtained by encoding the normalized 3D coordinates of Gaussians in the canonical space as RGB and rendering them in the posed screen space. Since the canonical space (*i.e.*, A-pose) spatially separates body parts, abrupt changes in the rendered positional map typically indicate transitions between unrelated regions—approximating seam boundaries. Unlike simple foreground masks, this approach can also detect internal boundaries between different body parts. We regularize these regions using separate RGBs supervised only on generated videos, which are free from the oversampling artifacts of the input image. Second, to avoid baked-in shadows, we use albedo images from Careaga *et al*. [5] as additional supervision, which contain minimal shading and help prevent shadow artifacts in the texture.

5.2. Geometry-weighted optimization

Geometry-weighted optimization. Fig. 4 (b) illustrates that geometry-weighted optimization applies low image loss weights and high geometry loss weights when optimizing MLPs for pose-driven deformation modeling. In this way, we mitigate rendering degradation caused by inconsistent and artifact-prone textures in generated frames. This approach enhances the robustness of the optimiza-

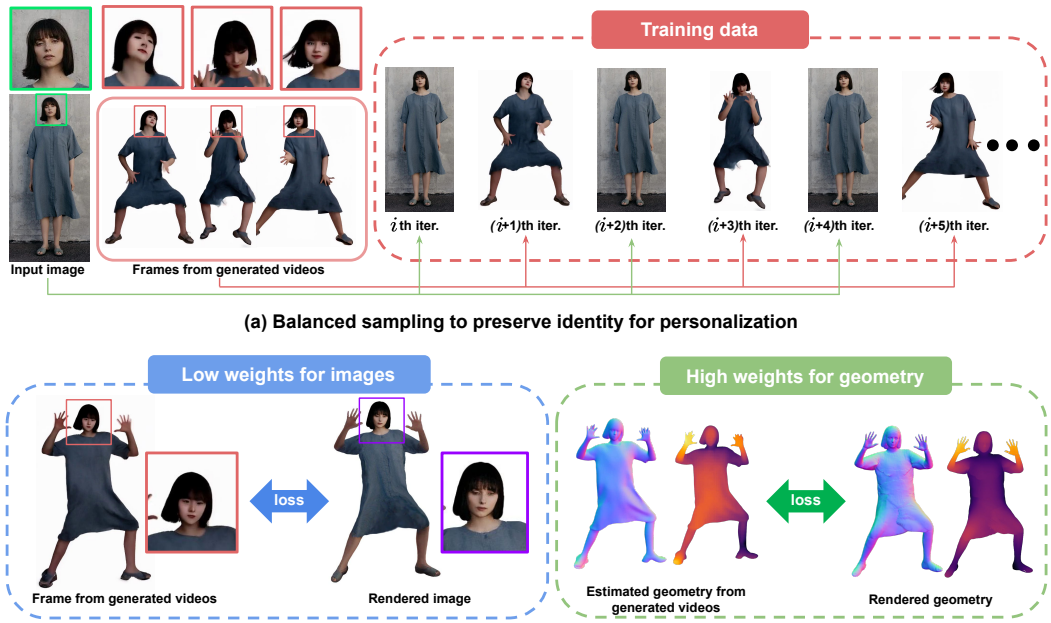


Figure 4. Two core components of PERSONA: balanced sampling for identity preservation and geometry-weighted optimization for sharp renderings in pose-driven deformations.

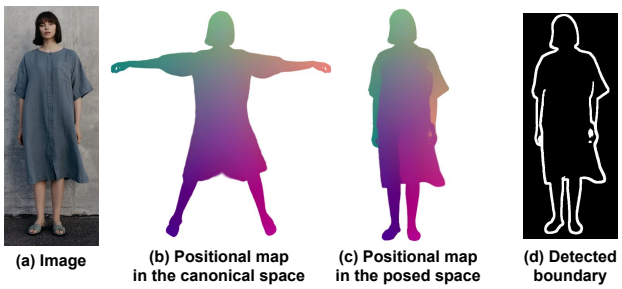


Figure 5. Seam boundary detection from positional maps. We obtain (c) the positional map by applying the pose to (b), then apply a Sobel filter to (c) to detect boundaries between different body parts as well as between foreground and background.

tion pipeline against texture artifacts in diffusion-generated frames, ensuring sharper renderings in poses different from the input image. Since per-frame pose information is used for pose-driven deformation modeling, balanced sampling alone (Sec. 5.1) is insufficient to maintain rendering quality in poses different from the input image. The pose-driven deformation modeling module differentiates between the input image and generated frames based on their poses, leading to sharp renderings when the pose matches the input image, but degraded quality in novel poses as the model adapts to artifacts in generated frames. Simply detaching textures from pose-driven deformation modeling is ineffective, as image loss still encourages geometry to replicate blur and artifacts of the generative videos.

Geometry-weighted optimization utilizes binary masks, depth maps, normal maps, and part segmentations extracted from diffusion-generated videos using SAM [23] and Sapiens [22] and compares them with the rendered outputs. Since geometry (*i.e.*, binary masks, depth maps, normal maps, and part segmentations) remains stable despite variations in texture quality, it serves as a reliable foundation for modeling pose-driven deformations while preserving rendering sharpness. The binary masks are rendered with a color value of one. To render depth maps, the depth value of each Gaussian is treated as a color attribute. Normal maps are computed per Gaussian using normal vectors, leveraging the hybrid representation of ExAvatar [37], and are similarly treated as colors. Finally, part segmentations are represented as RGB images, with colors assigned based on the Sapiens palette.

Preserving sharp renderings with mean offsets. To model pose-driven deformations, we apply only mean offsets to isotropic Gaussians. This approach shifts each Gaussian's position while keeping its shape and appearance fixed, allowing the avatar to deform without blurring textures. Such position-only deformation preserves texture sharpness, similar in spirit to mesh-based animation where vertex displacements alone maintain high-frequency texture details. In contrast, using scale offsets leads to blurry results, as Gaussians grow or shrink instead of shifting, and RGB offsets risk copying unreliable colors from generated frames. Although mean offsets could introduce gaps, dense high-resolution Gaussians and their overlapping nature ensure seamless renderings, even during large deformations

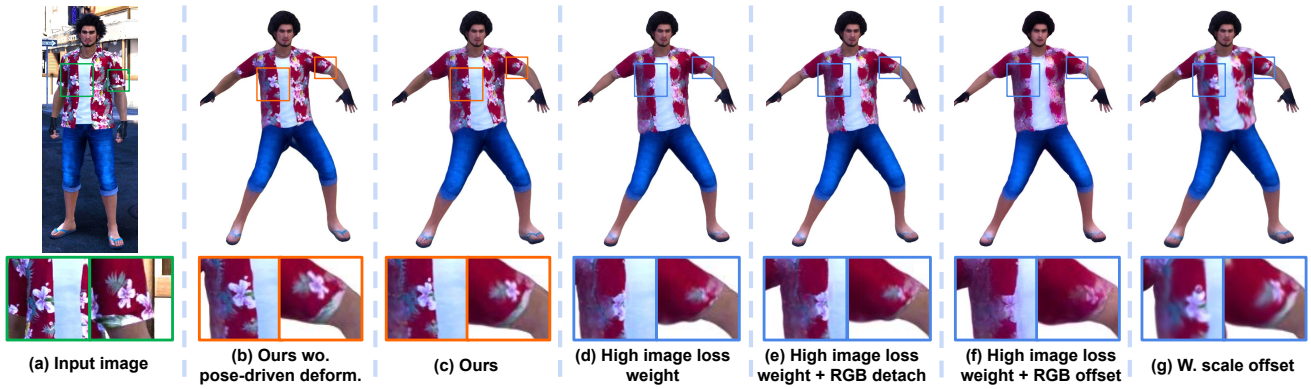


Figure 6. Comparison of various pose-driven deformation modeling strategies. Our geometry-weighted optimization is essential for maintaining authentic and sharp renderings.

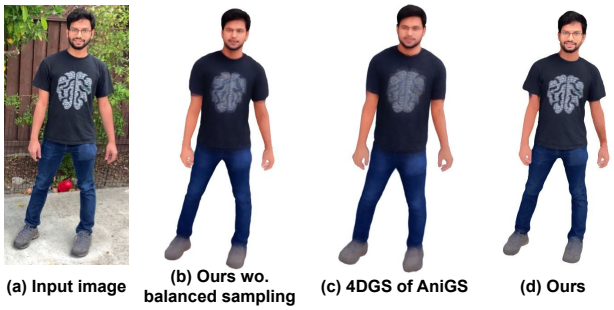


Figure 7. Effectiveness of our balanced sampling. Without it, the avatar loses the identity of the input image. The 4DGS approach of AniGS [45] also fails to preserve the subject's identity.

like dance motions.

6. Optimization

We optimize 3D Gaussian features (*i.e.*, means, scales, and RGB colors), triplane, MLP weights, and per-frame SMPL-X parameters. For supervision, we use standard image reconstruction loss functions, including $L1$, SSIM, and LPIPS [65], along with geometry regularizers such as Laplacian regularization, following Moon *et al.* [37]. In geometry-weighted optimization, we minimize the $L1$ distance between the rendered outputs and target geometry maps. Diffusion-generated videos often produce implausible hand shapes. To enforce geometric plausibility, we minimize the $L1$ distance between hand masks rendered with 3DGS and those rendered with SMPL-X meshes (not Gaussian points) using a standard mesh renderer, ensuring that the hand shape closely resembles SMPL-X hands.

7. Experiments

7.1. Protocol

We conduct quantitative comparisons against state-of-the-art methods using NeuMan [19] and X-Humans [52] datasets, following the evaluation protocols of previous works [14, 37, 52]. We measure rendering quality with PSNR, SSIM, and LPIPS [65] metrics. NeuMan provides short monocular videos captured in the wild, while X-Humans offers a diverse range of whole-body motions, including various body poses, hand gestures, and facial expressions. For qualitative comparisons, we evaluate animation capability using in-the-wild videos featuring intense dance performances, *different from our training set.*

Figure 8. Effectiveness of our detected boundary in balanced sampling. Without it, colors from one body part can leak into others, leading to noticeable artifacts.

7.2. Ablation studies

Balanced sampling. Fig. 7 demonstrates the importance of balanced sampling in preserving authenticity. Without it, as shown in Fig. 7 (b), the input image is underused during training, resulting in an avatar with a different identity and blurry textures due to inconsistencies in the generated frames. Fig. 7 (c) shows that 4DGS-based method [45] still suffers from identity loss and blurry renderings due to the severe imbalance between the input image and the generated frames. Their 4DGS treats inconsistencies across frames as a temporal sequence and trains MLPs with spatio-temporal features to differentiate the input image from generated frames. Fig. 8 shows that our boundary detection is necessary to prevent color leaking between different body parts.

Geometry-weighted optimization. Fig. 6 highlights the

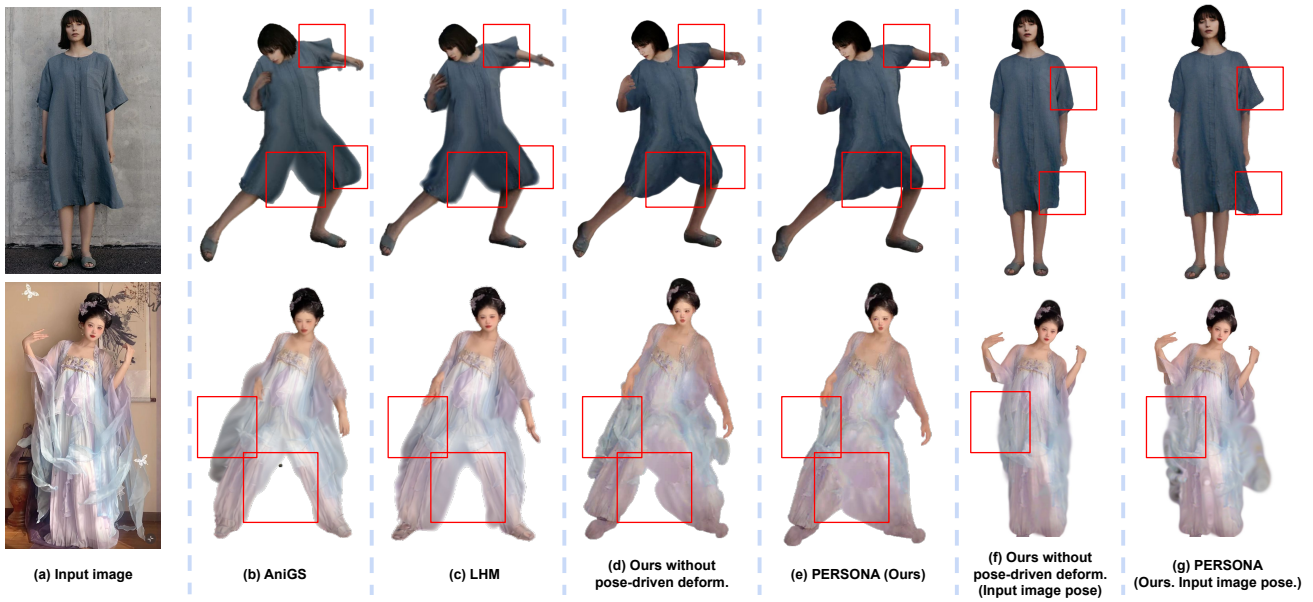


Figure 9. Effectiveness of our pose-driven deformations. (b,c): Previous 3D-based methods [44, 45] embed input-image-specific deformations (highlighted in red) into the avatar, leading to baked-in artifacts when animated to new poses. (d-g): Our method, PERSONA, mitigates this issue by explicitly modeling pose-driven deformations.

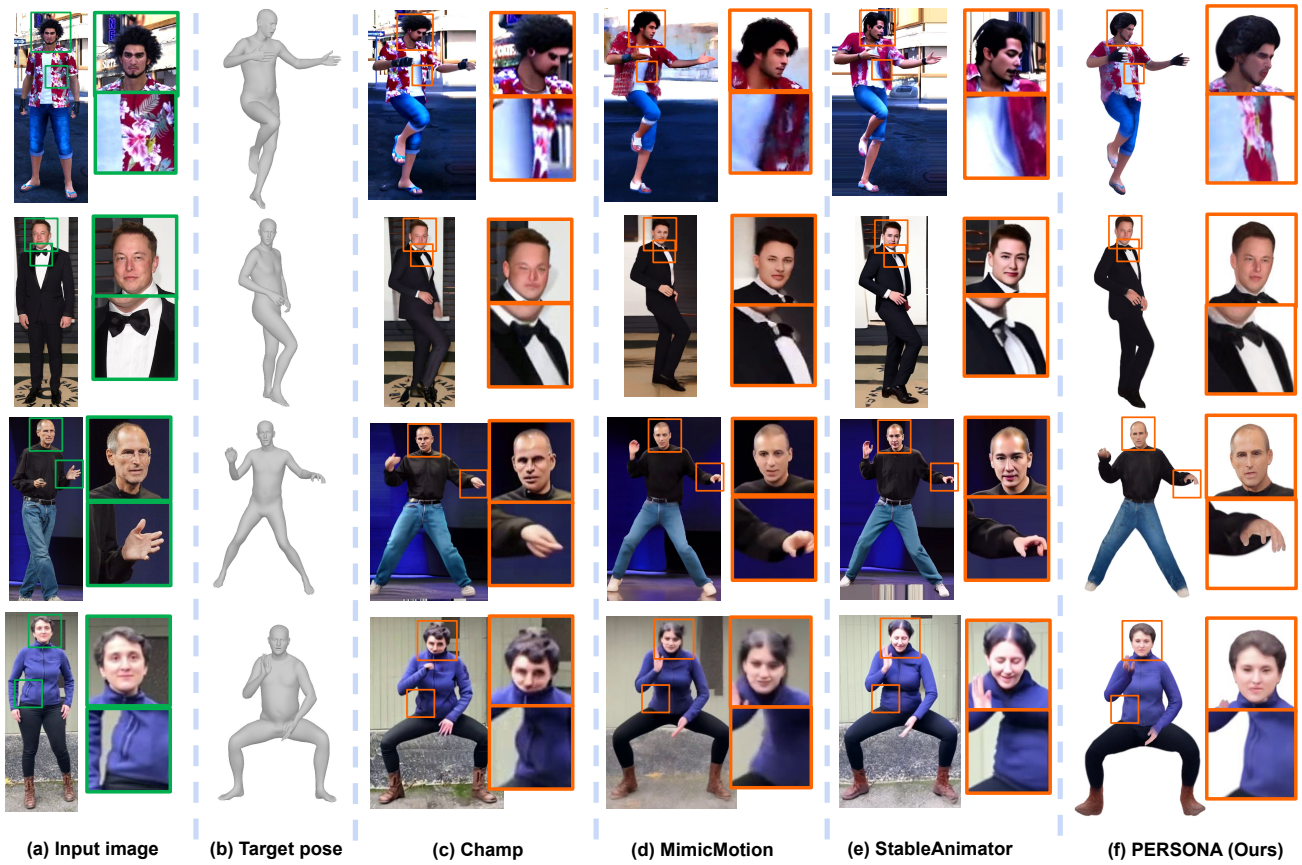


Figure 10. Comparison of diffusion-based state-of-the-art methods and our PERSONA.

importance of geometry-weighted optimization in preserving authentic and sharp renderings for pose-driven deforma-

Methods	*00028*			*00034*			*00087*		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
*** Available data: 10 videos and 3D registrations from multi-view images**									
X-Avatar [52]	28.57	0.976	0.026	28.05	0.965	0.035	30.89	0.970	0.030
ExAvatar [37]	**30.58**	**0.981**	**0.018**	**28.75**	**0.966**	**0.029**	**32.01**	**0.972**	**0.025**
*** Available data: a single image**									
ExAvatar [37]	21.83	0.962	0.058	23.04	0.951	0.066	24.58	0.958	0.061
TeCH [16]	20.28	0.950	0.059	22.26	0.938	0.055	24.74	0.945	0.052
SiTH [11]	15.00	0.934	0.110	18.28	0.935	0.098	21.65	0.950	0.087
Champ [69]	18.96	0.953	0.070	20.17	0.949	0.082	25.19	0.960	0.057
MimicMotion [66]	21.02	0.959	0.053	23.49	0.953	0.051	27.67	0.963	0.038
StableAnimator [58]	21.66	0.962	0.050	23.53	0.952	0.051	28.04	0.963	0.039
AniGS [45]	23.96	0.965	**0.040**	24.80	0.955	0.052	27.46	0.964	0.043
Ours wo. pose-driven deform.	23.15	0.968	0.049	26.32	0.960	0.047	29.25	0.966	0.038
PERSONA (Ours)	**24.76**	**0.972**	**0.040**	**27.60**	**0.963**	**0.042**	**29.79**	**0.968**	**0.035**

Table 2. Comparisons with previous works on the test set of X-Humans [52].

tions. Without it, inconsistent textures in generated frames cause identity shifts and blurriness. Fig. 6 (c) demonstrates effective pose-driven deformation, where raising the arms causes the clothing to move further away from the waist, reflecting a natural interaction between the body and the fabric. In contrast, Fig. 6 (b) shows unnatural adherence, with the clothing remaining tightly fitted to the body despite the raised arms due to the lack of deformation modeling. Fig. 6 (d) and (e) illustrate how high image loss weights degrade visual quality, forcing geometry to replicate texture inconsistencies and resulting in blurry renderings, even when RGB is detached. Finally, Fig. 6 (f) and (g) validate our choice to use only mean offsets, as scale offsets lead to excessive blurring, as discussed in Sec. 5.2. All variants are evaluated under the same balanced sampling and geometry loss settings for a fair comparison.

Pose-driven deformations. Fig.9 illustrates the effectiveness of our pose-driven deformations. Existing methods[44, 45] embed deformations present in the input image, which become baked-in artifacts when the avatar is animated to novel poses. In the first row, for example, the left arm and thigh should fall naturally due to gravity, but remain unnaturally bent. In the second row, the right side of the body similarly fails to respond naturally to gravity. In both cases, long skirts are often misinterpreted as pants, resulting in incorrect deformation behavior. In contrast, PERSONA explicitly models pose-driven deformations, allowing avatars to respond naturally to novel poses without inheriting input-specific artifacts. This leads to significantly improved visual fidelity under diverse poses.

7.3. Comparisons to state-of-the-art methods

Fig.9, Fig.10, Tab.2, and Tab.3 compare PERSONA with state-of-the-art methods [44, 45, 58, 66, 69]. Fig.9 shows that compared to existing 3D-based approaches [44, 45], PERSONA captures natural pose-driven clothing deformations more effectively, thanks to our geometry-weighted optimization. In addition, Fig.10 shows that compared to diffusion-based approaches [58, 66, 69], PERSONA better preserves identity, accurately maintaining facial features and clothing patterns through balanced sampling. Tab.2 and Tab.3 show that PERSONA outperforms all single-image-based methods, demonstrating the necessity and effectiveness of pose-driven deformations. All comparisons exclude background pixels and use official implementations.

Methods	PSNR↑	SSIM↑	LPIPS↓
*** Available data: a video**			
HumanNeRF [59]	27.06	0.967	0.019
InstantAvatar [18]	28.47	0.972	0.028
NeuMan [19]	29.32	0.972	0.014
Vid2Avatar [9]	30.70	0.980	0.014
GaussianAvatar [14]	29.94	0.980	0.012
3DGS-Avatar [43]	28.99	0.974	0.016
ExAvatar [37]	**34.80**	**0.984**	**0.009**
*** Available data: a single image**			
ExAvatar [37]	24.95	0.963	0.031
TeCH [16]	22.82	0.953	0.039
SiTH [11]	23.96	0.957	0.031
Champ [69]	27.27	0.968	**0.021**
MimicMotion [66]	26.12	0.970	0.029
StableAnimator [58]	26.58	0.968	0.025
AniGS [45]	28.27	0.969	0.027
LHM [44]	26.22	0.967	0.025
Ours wo. pose-driven deform.	28.02	0.972	0.025
PERSONA (Ours)	**29.20**	**0.974**	**0.021**

Table 3. Comparisons of previous works on the test set of NeuMan [19].

8. Conclusion

We introduce PERSONA, a framework that creates personalized 3D avatars with pose-driven deformations from a single image using diffusion-generated training videos. By combining 3D- and diffusion-based approaches, PERSONA ensures identity preservation and natural deformations. To address authenticity loss and rendering artifacts, we propose balanced sampling and geometry-weighted optimization. Our results show that PERSONA outperforms existing methods, providing a scalable solution for high-quality avatar creation.

Supplementary Material *for*

"PERSONA: Personalized Whole-Body 3D Avatar with Pose-Driven Deformations from a Single Image"

In this supplementary material, we provide more experiments, discussions, and other details that could not be included in the main text due to the lack of pages. The contents are summarized below:

- Sec. S1: More comparisons to state-of-the-art methods.
- Sec. S2: Rendered avatars in the canonical space.
- Sec. S3: More ablation studies.
- Sec. S4: Limitations of the proposed PERSONA.

S1. Comparisons to state-of-the-art methods

Running time comparison. Tab. S1 further highlights that PERSONA achieves real-time rendering speeds, whereas existing diffusion-based methods suffer from slow inference. All running times were measured under the same hardware setup using a single RTX A6000.

User study. Fig.S1 presents results from our user study, where participants strongly preferred our approach over existing diffusion-based methods. We conducted the study with 40 participants, each answering 10 questions in which they selected the image that best matched the input single image. The compared methods included Champ [69], MimicMotion [66], StableAnimator [58], and our PERSONA. Fig. S2 provides an example from the study, with (a), (b), (c), and (d) corresponding to MimicMotion [66], Champ [69], our PERSONA, and StableAnimator [58], respectively.

Qualitative comparisons. Fig.S3 compares our PERSONA with 3D-based state-of-the-art methods[44, 45]. PERSONA achieves more accurate pose-driven deformations with more stable and consistent renderings. Fig.S4 compares PERSONA with diffusion-based methods[58, 66, 69], where our method better preserves the subject's identity from the input image, resulting in more authentic avatars while still accurately modeling pose-driven deformations.

Methods	Frames per second
Champ [69]	0.88
MimicMotion [66]	0.36
StableAnimator [58]	0.24
PERSONA (Ours)	**25.56**

Table S1. Frames per second comparisons of various human animation methods.

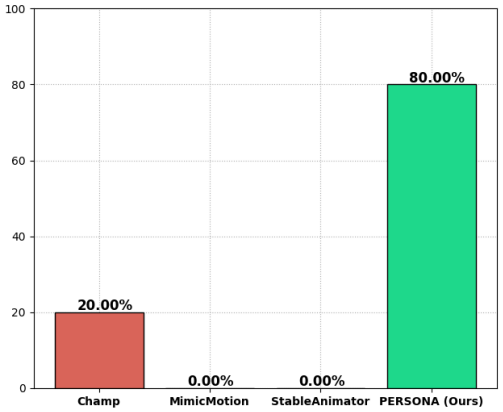


Figure S1. User preference study results from 40 participants.

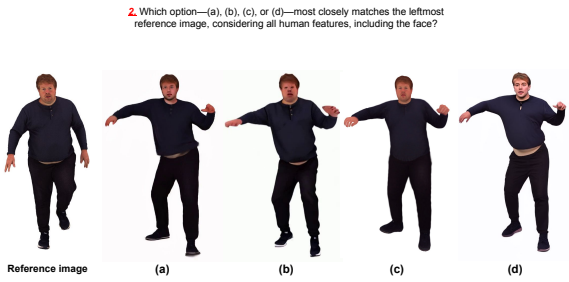


Figure S2. An example of our user study.

Figure S3. Comparison of state-of-the-art 3D-based methods [44, 45] and our PERSONA.

Figure S4. Comparison of state-of-the-art diffusion-based methods [58, 66, 69] and our PERSONA.

S2. Avatars in canonical space

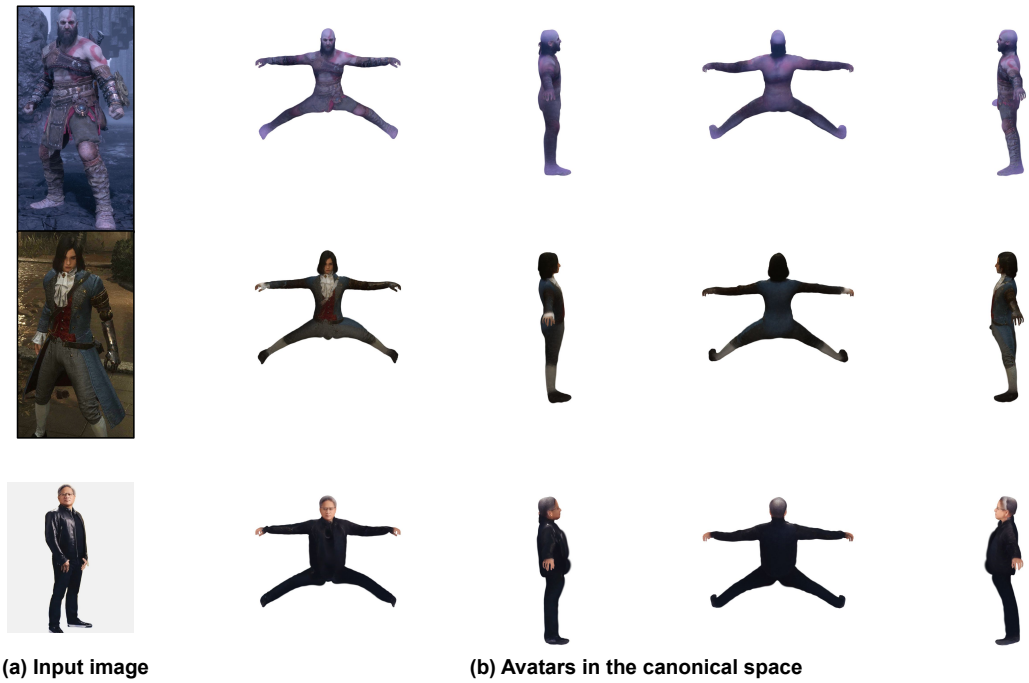
Fig. S5, S6, and S7 showcase various avatars created from a single input image. These avatars are rendered in canonical space without applying our pose-driven deformations. Despite being constructed from just a single image, the avatars achieve high-quality renderings from multiple viewpoints, including fully invisible regions, without noticeable artifacts. These results highlight the effectiveness of our avatar creation pipeline.

(a) Input image **(b) Avatars in the canonical space**

Figure S5. The input image and rendered avatars in the canonical space from multiple viewpoints.

(a) Input image (b) Avatars in the canonical space

Figure S6. The input image and rendered avatars in the canonical space from multiple viewpoints.



(a) Input image (b) Avatars in the canonical space

Figure S7. The input image and rendered avatars in the canonical space from multiple viewpoints.

S3. Ablation studies

Balanced sampling. Fig. S8 demonstrates the effectiveness of our 1:1 ratio between the input image and generated frames in balanced sampling. Reducing the use of the input image leads to a loss of authenticity and sharpness in the rendering, which is expected due to the inconsistent textures in the generated frames.

Loss weights for geometry-weighted optimization. Tab. S2 shows that using a high image loss weight (first row) significantly degrades rendering quality. This issue is mitigated by lowering the image loss weight (second row). However, further reducing it slightly harms rendering quality (third row), indicating the need for a balanced trade-off.

Pose-driven deformations. Tab. S3 demonstrates that our pose-driven deformation not only improves photometric metrics (as shown in Tab. 2 and 3) but also enhances geometry quality. Mask, depth, and normal metrics are measured as intersection-over-union, $L1$ distance between rendered and ground truth depth maps after aligning global translation, and the angular difference between rendered and ground truth normal maps, respectively.

Variants in pre-processing stages. Tab. S4 and Tab. S5 show how different training video generators (Sec. 4) and geometry estimators (Sec. 5.2) affect the final rendering quality. As shown in the tables, the choice of generator or the use of lighter geometry estimators has only a marginal impact on rendering quality. In particular, since we use enough number of generated frames (approximately 1K) for optimizing PERSONA, the geometric estimation errors from lighter models such as Sapiens [22] do not significantly degrade the final output.

S4. Limitations

Lack of dynamics. Despite its ability to represent pose-driven deformations, PERSONA cannot capture motion-dependent dynamics, which rely on velocity and acceleration. These dynamics are crucial for modeling complex deformations in loose-fitting clothing and hair. While we attempted to incorporate velocity and acceleration as additional inputs, our 3D avatar representation lacks separate layers for garments and hair, leading to unsatisfactory results. We believe that designing separate layers for garments and hair could be an interesting direction for future research.

Lack of fine-grained cloth wrinkles. Additionally, PERSONA struggles to capture fine, pose-dependent wrinkles in clothing, likely due to the lack of 3D consistency in diffusion-generated videos, which hinders accurate geometry and texture tracking and results in oversmoothed surfaces.

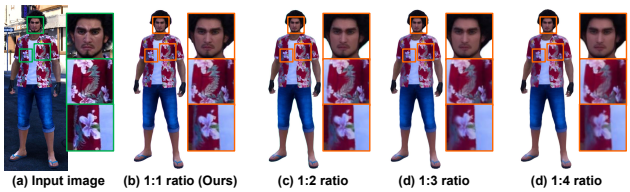


Figure S8. Rendering comparisons with different input image-to-generated frame ratios in balanced sampling. For a clearer comparison, avatars are rendered using the viewpoint and pose of the input image.

Geo. weight	Img. weight	PSNR↑	SSIM↑	LPIPS↓
1	1	28.18	0.969	0.030
1	**0.1**	**29.20**	**0.974**	**0.021**
1	0.01	29.00	0.970	0.023

Table S2. Effect of loss weights in our geometry-weighted optimization on the NeuMan test set. The second row (in bold) is ours.

Settings	Mask↑	Depth↓	Normal↓
Wo. pose-driven deform.	88.60	47.17	22.07
W. pose-driven deform. (Ours)	**90.06**	**46.13**	**21.73**

Table S3. Effectiveness of our pose-driven deformations on the X-Humans [52] test set. Units for mask, depth, and normal are %, mm, and degrees, respectively.

Generator	PSNR↑	SSIM↑	LPIPS↓
Champ	29.13	0.972	**0.019**
StableAnimator	28.98	0.970	0.024
MimicMotion (Ours)	**29.20**	**0.974**	0.021

Table S4. Effect of different training video generators on the NeuMan test set.

Sapiens models	PSNR↑	SSIM↑	LPIPS↓
0.3B (Smallest one)	28.98	0.971	0.023
1B (Ours)	**29.20**	**0.974**	**0.021**

Table S5. Effect of different geometry estimators on the NeuMan test set.

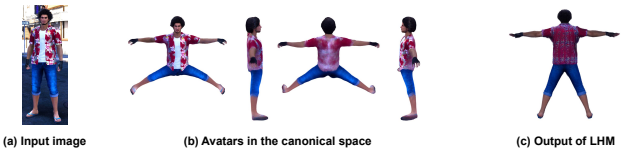


Figure S9. Limitation of PERSONA. Due to texture inconsistencies of generated frames, used to train our PERSONA, complex patterns in invisible regions are challenging to render sharply. Even very recent feed-forward method [44] fail to generate plausible textures.

Blurry rendering for complex patterns in invisible regions. Fig. S9 illustrates that our pipeline struggles to achieve sharp renderings in invisible regions when complex patterns are present. While our method produces plausible geometry and textures for these areas, as seen in Fig. S5 and Fig. S6, intricate patterns remain difficult to render sharply due to inconsistencies in the generated frames used to train PERSONA. We observe that even recent feed-

forward methods [44] fail to generate plausible textures. We believe this limitation could be addressed by incorporating more advanced image or video generative models.

Lack of relighting capability. Lastly, omitting RGB offsets in pose-driven deformation modeling prevents our method from handling relighting effects, such as natural shadows and reflections in novel environments. Addressing these challenges remains an avenue for future work.

Long pre-processing time. Generating training videos with diffusion-based animators requires significant pre-processing time due to their slow inference speed. It takes approximately one hour to generate training videos, whereas avatar training itself additionally takes 30 minutes. Exploring strategies to optimize data generation for a more efficient avatar creation pipeline presents an interesting direction for future research.

Acknowledgments

This work was partly supported by the ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2025-RS-2020-II201819). It was also supported by IITP grant funded by the MSIT (RS-2025-02653113, High-Performance Research AI Computing Infrastructure Support at the 2 PFLOPS Scale). Additional support was provided by Korea University research grants. It was also supported by the Artificial Intelligence Industrial Convergence Cluster Development Project funded by the MSIT and Gwangju Metropolitan City. This research was supported by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2025 (Project Name: Development of AI-based image expansion and service technology for high-resolution (8K/16K) service of performance contents, Project Number: RS-2024-00395886, Contribution Rate: 20%). This work was supported by the Technology Innovation Program (RS-2025-02653087, Development of a Motion Data Collection System and Dynamic Persona Modeling Technology) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea). This work was supported by the IITP grant funded by the MSIT (No. RS-2025-25441838, Development of a human foundation model for human-centric universal artificial intelligence and training of personnel).

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3D people models. In *CVPR*, 2018. 3

[2] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3D reconstruction of humans wearing clothing. In *CVPR*, 2022. 3

[3] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabian Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM TOG*, 2021. 3

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable Video Diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 3

[5] Chris Careaga and Yağız Aksoy. Intrinsic image decomposition via ordinal shading. *ACM TOG*, 2023. 4

[6] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In *ECCV*, 2020. 3

[7] Hongsuk Choi, Gyeongsik Moon, Matthieu Armando, Vincent Leroy, Kyoung Mu Lee, and Grégory Rogez. MonoNHR: Monocular neural human renderer. In *3DV*, 2022. 1

[8] Xiang Deng, Zerong Zheng, Yuxiang Zhang, Jingxiang Sun, Chao Xu, Xiaodong Yang, Lizhen Wang, and Yebin Liu. RAM-Avatar: Real-time photo-realistic avatar from monocular videos with full-body control. In *CVPR*, 2024. 3

[9] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2Avatar: 3D avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *CVPR*, 2023. 1, 3, 8

[10] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. DeepCap: Monocular human performance capture using weak supervision. In *CVPR*, 2020. 3

[11] Hsuan-I Ho, Jie Song, and Otmar Hilliges. SiTH: Single-view textured human reconstruction with image-conditioned diffusion. In *CVPR*, 2024. 8

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3

[13] Hezhen Hu, Zhiwen Fan, Tianhao Wu, Yihan Xi, Seoyoung Lee, Georgios Pavlakos, and Zhangyang Wang. Expressive gaussian human avatars from monocular rgb video. *arXiv preprint arXiv:2407.03204*, 2024. 3

[14] 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. *arXiv preprint arXiv:2312.02134*, 2023. 1, 2, 3, 4, 6, 8

[15] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate Anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, 2024. 1, 2, 3, 4

[16] 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. 3, 8

[17] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable reconstruction of clothed humans. In *CVPR*, 2020. 3

[18] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. InstantAvatar: Learning avatars from monocular video in 60 seconds. In *CVPR*, 2023. 1, 3, 8

[19] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. NeuMan: Neural human radiance field from a single video. In *ECCV*, 2022. 1, 3, 6, 8

[20] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 3

[21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023. 1, 3

[22] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *ECCV*, 2024. 5, 14

[23] 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. 5

[24] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. HUGS: Human gaussian splatting. In *CVPR*, 2024. 1, 3

[25] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 3

[26] Nikos Kolotouros, Thiemo Alldieck, Enric Corona, Eduard Gabriel Bazavan, and Cristian Sminchisescu. Instant 3d human avatar generation using image diffusion models. In *ECCV*, 2024. 3

[27] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural Human Performer: Learning generalizable radiance fields for human performance rendering. *NeurIPS*, 2021. 1, 3

[28] Youngjoong Kwon, Lingjie Liu, Henry Fuchs, Marc Habermann, and Christian Theobalt. DELIFFAS: Deformable light fields for fast avatar synthesis. *NeurIPS*, 2024.

[29] 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. 3

[30] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 3

[31] Xinqi Liu, Chenming Wu, Xing Liu, Jialun Liu, Jinbo Wu, Chen Zhao, Haocheng Feng, Errui Ding, and Jingdong Wang. GEA: Reconstructing expressive 3D gaussian avatar from monocular video. *arXiv preprint arXiv:2402.16607*, 2024. 3

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 2015. 1, 3

[33] Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. MIMO: Controllable character video synthesis with spatial decomposed modeling. *arXiv preprint arXiv:2409.16160*, 2024. 1, 2, 3

[34] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021. 1

[35] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In *ECCV*, 2020. 3

[36] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3D hand pose estimation for whole-body 3D human mesh estimation. In *CVPRW*, 2022. 3

[37] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3D gaussian avatar. In *ECCV*, 2024. 1, 2, 3, 4, 5, 6, 8

[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*, 2024. 3

[39] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. ASH: Animatable gaussian splats for efficient and photoreal human rendering. In *CVPR*, 2024. 3

[40] 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. 1, 3

[41] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021. 1, 3

[42] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural Body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 1, 3

[43] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3DGS-Avatar: Animatable avatars via deformable 3d gaussian splatting. In *CVPR*, 2024. 1, 8

[44] Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, and Liefeng Bo. LHM: Large animatable human reconstruction model from a single image in seconds. In *ICCV*, 2025. 1, 2, 7, 8, 9, 10, 14, 15

[45] Lingteng Qiu, Shenhao Zhu, Qi Zuo, Xiaodong Gu, Yuan Dong, Junfei Zhang, Chao Xu, Zhe Li, Weihao Yuan, Liefeng Bo, et al. AniGS: Animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. In *CVPR*, 2025. 1, 2, 3, 6, 7, 8, 9, 10

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3

[47] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3D whole-body pose estimation system via regression and integration. In *ICCVW*, 2021. 3

[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*, 2019. 3

[49] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *CVPR*, 2020. 3

[50] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *CVPR*, 2021. 3

[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*, 2024. 3

[52] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-Avatar: Expressive human avatars. In *CVPR*, 2023. 6, 8, 14

[53] Jisu Shin, Junmyeong Lee, Seongmin Lee, Min-Gyu Park, Ju-Mi Kang, Ju Hong Yoon, and Hae-Gon Jeon. CanonicalFusion: Generating drivable 3D human avatars from multiple images. In *ECCV*, 2024. 3

[54] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2, 3

[55] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *NeurIPS*, 2020.

[56] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

[57] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023. 2, 3

[58] Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. StableAnimator: High-quality identity-preserving human image animation. *arXiv preprint arXiv:2411.17697*, 2024. 2, 3, 4, 8, 9, 11

[59] 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. 8

[60] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. ICON: Implicit clothed humans obtained from normals. In *CVPR*, 2022. 3

[61] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. ECON: Explicit clothed humans optimized via normal integration. In *CVPR*, 2023. 3

[62] Yuliang Xiu, Yufei Ye, Zhen Liu, Dimitrios Tzionas, and Michael J Black. PuzzleAvatar: Assembling 3D avatars from personal albums. *ACM TOG*, 2024. 3

[63] 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. 1, 2, 3

[64] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-Splatting: Alias-free 3D gaussian splatting. In *CVPR*, 2024. 3

[65] 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

[66] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. In *International Conference on Machine Learning*, 2025. 1, 2, 3, 4, 8, 9, 11

[67] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction. *TPAMI*, 2021. 3

[68] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. AvatarRex: Real-time expressive full-body avatars. *ACM TOG*, 2023. 3

[69] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3D parametric guidance. In *ECCV*, 2024. 2, 3, 8, 9, 11

[70] Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. IDOL: Instant photorealistic 3D human creation from a single image. *arXiv preprint arXiv:2412.14963*, 2024. 1, 2, 3