Supplementary Material for "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

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 [7], MimicMotion [6], StableAnimator [5], and our PERSONA. Fig. S2 provides an example from the study, with (a), (b), (c), and (d) corresponding to MimicMotion [6], Champ [7], our PERSONA, and StableAnimator [5], respectively.

Qualitative comparisons. Fig.S3 compares our PERSONA with 3D-based state-of-the-art methods[2, 3]. PERSONA achieves more accurate pose-driven deformations with more stable and consistent renderings. Fig.S4 compares PERSONA with diffusion-based methods[5–7], 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 [7]	0.88
MimicMotion [6]	0.36
StableAnimator [5]	0.24
PERSONA (Ours)	**25.56**

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

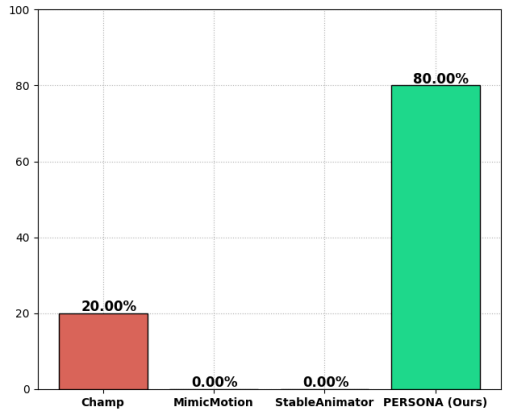


Figure S1. User preference study results from 40 participants.

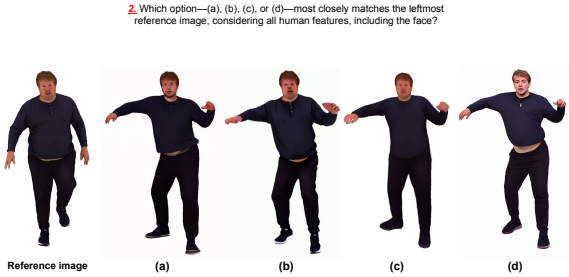


Figure S2. An example of our user study.

Figure S3. Comparison of state-of-the-art 3D-based methods [2, 3] and our PERSONA.

Figure S4. Comparison of state-of-the-art diffusion-based methods [5–7] and our PERSONA.

S2. Avatars in canonical space

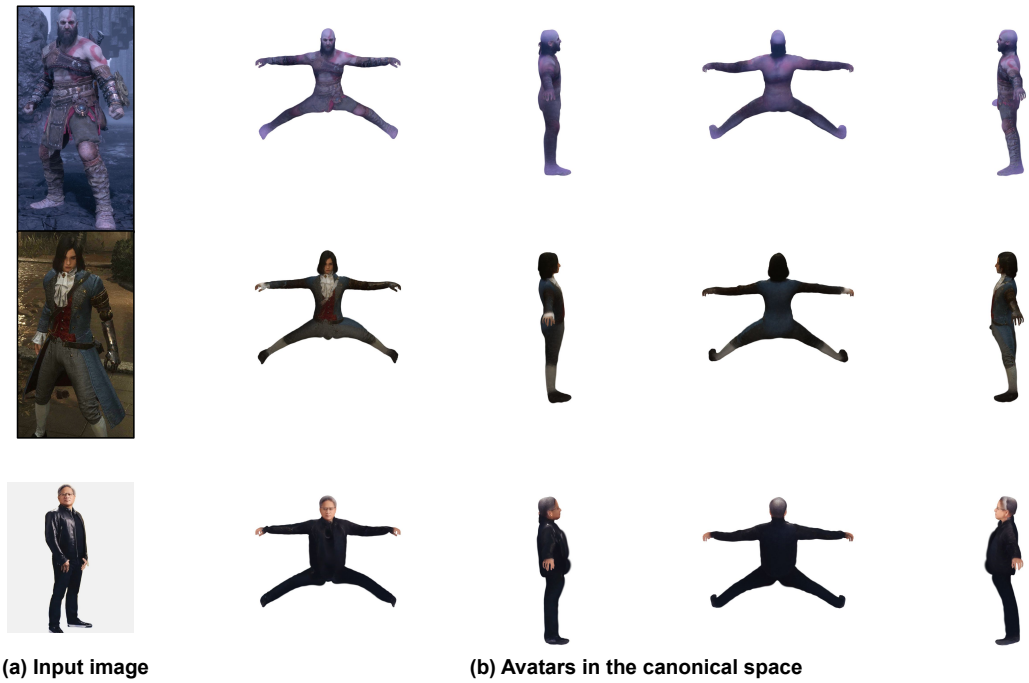
Fig. S5, S6, and S7 showcase various avatars created from a single input image. These avatars are rendered in canonical space without applying our pose-driven deformations. Despite being constructed from just a single image, the avatars achieve high-quality renderings from multiple viewpoints, including 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 of the main manuscript) 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 of the main manuscript) and geometry estimators (Sec. 5.2 of the main manuscript) 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 [1] 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.

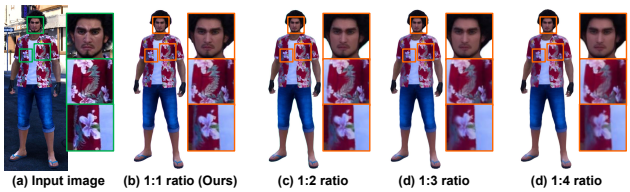


Figure S8. Rendering comparisons with different input image-to-generated frame ratios in balanced sampling. For a clearer comparison, avatars are rendered using the viewpoint and pose of the input image.

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 [4] 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.

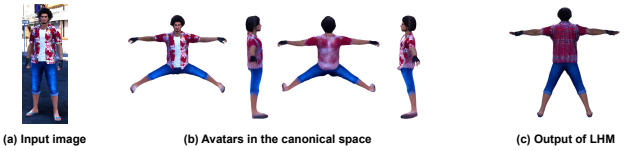


Figure S9. Limitation of PERSONA. Due to texture inconsistencies of generated frames, used to train our PERSONA, complex patterns in invisible regions are challenging to render sharply. Even very recent feed-forward method [2] fail to generate plausible textures.

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 ren-

der sharply due to inconsistencies in the generated frames used to train PERSONA. We observe that even recent feed-forward methods [2] 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.

References

[1] 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. 6

[2] 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, 6, 7

[3] 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

[4] 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

[5] 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. 1, 3

[6] 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

[7] 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. 1, 3