Supplemantary Material for "Zero-Shot Reconstruction of Animatable 3D Avatars with Cloth Dynamics from a Single Image"

Joohyun Kwon Geonhee Sim Gyeongsik Moon
Korea University
{juheanqueen, kh6362, mks0601}@korea.ac.kr
https://juhyeon-kwon.github.io/DynaAvatar.github.io/

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: Comparisons to related works
- Sec. S2: Dataset reannotation comparisons
- Sec. S3: Ablation studies
- Sec. S4: Architecture details
- Sec. S5: Implementation details

S1. Comparisons to related works

We compare DynaAvatar with state-of-the-art methods, physics-based approaches, and diffusion-based approaches to show its advantages. Please refer to the accompanying supplementary video for the full animation results.

S1.1. Comparison to state-of-the-art methods

S1.1.1. Qualitative comparisons

Fig. S1 shows comparisons of our DynaAvatar and previous state-of-the-art methods [9, 11, 14] from in-the-wild input image. Our method successfully reconstructs and animates avatars with high-fidelity cloth dynamics. For instance, when the subject raises their arms, the upper garment naturally lifts upward, exhibiting physically plausible motion-dependent dynamics.

In contrast, baseline methods generate animations without incorporating motion-dependent dynamics. Consequently, the resulting animations often lack realism, as the garments remain static regardless of the body's movement. Our method effectively overcomes this limitation by leveraging the Dynamic Transformer, resulting in superior visual realism.

Note that PF-LHM [10] is excluded due to code unavailability. Nevertheless, as it takes only the pose cues without motion information (*i.e.*, a sequence of poses), similar to PERSONA [11], it is expected to lack the capability to represent motion-dependent dynamics, thereby underperforming compared to our motion-aware approach.

Table S1. Comparison of face consistency (FC) on DNA-Rendering.

Methods	FC ↑
IDOL [60]	0.625
LHM [37]	0.697
DynaAvatar (Ours)	**0.712**

Table S2. Comparison of computational costs.

Methods	Zero-shot	Cloth dynamics	Time	# of params.
PERSONA [43]	✗	✓	3.11h	4M
LHM-500M [37]	✓	✗	1.08s	356M
LHM-1B [37]	✓	✗	3.00s	1.1B
DynaAvatar (Ours)	✓	✓	1.82s	719M

S1.1.2. Quantitative comparisons

Face consistency. Table. S1 on DNA-Rendering shows that DynaAvatar achieves superior image consistency, as shown by the higher face consistency (FC) compared to baseline methods. FC is measured via cosine similarity in the ArcFace [2] embedding space.

Inference Latency. As shown in Table. S2, DynaAvatar ensures fast inference via its zero-shot architecture, avoiding the lengthy per-subject optimization of PERSONA. While our Dynamic Transformer adds moderate overhead compared to LHM-500M, it remains more efficient than LHM-1B in both time and parameters. This cost is essential for capturing dynamic deformations, offering a superior trade-off for motion-dependent cloth dynamics that LHM lacks. All metrics are measured on a RTX pro 6000 GPU.

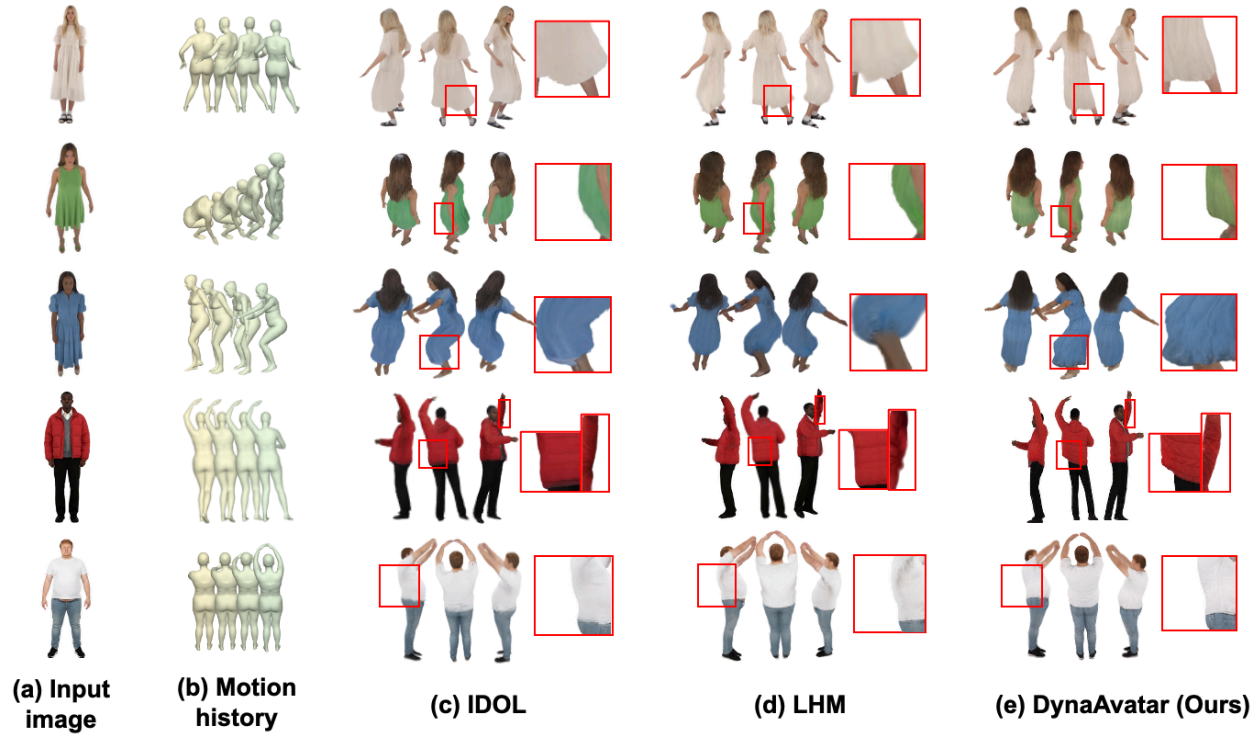

Figure S1. Comparison between DynaAvatar and previous single-image–based state-of-the-art methods [9, 14] on in-the-wild images.

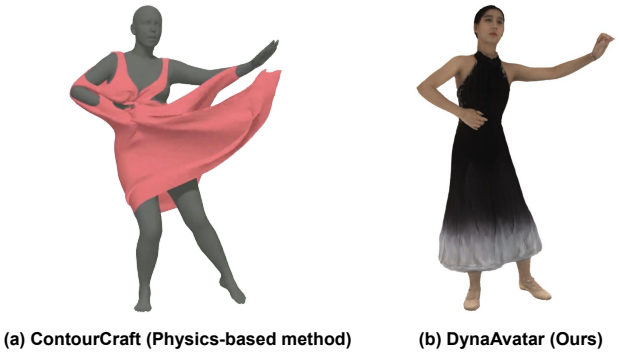

Figure S2. Comparison between physics-based method [4] and DynaAvatar.

S1.2. Comparison to physics-based approaches

Fig. S2 compares the physics-based method [3, 4] with DynaAvatar, highlighting the instability of the former under in-the-wild scenarios. Note that we used a garment template of a similar type to the input image for the physics-based simulation. As shown in the Fig. S2 left, applying physics simulation to in-the-wild sequences often leads to catastrophic failures where the cloth unrealistically flies away or drifts. This instability stems from the imperfect in-the-wild pose estimation. Specifically, pose errors frequently cause the body mesh to penetrate the garment mesh, creating invalid collision constraints. These interpenetrations trigger erroneous inputs in the simulation, causing garment to become unstable. Moreover, these methods primarily focus on geometric garment deformation, often lacking the capability to model photorealistic, full-body appearance.

In contrast, DynaAvatar robustly synthesizes both motion-dependent cloth dynamics and high-fidelity appearance, even when driven by in-the-wild motion sequences. These results validate DynaAvatar as a robust and practical method for animating avatars from single images.

S1.3. Comparison to diffusion-based approaches

Fig. S3 compares the state-of-the-art diffusion-based method [12] with DynaAvatar, highlighting the limitations of diffusion models. These models fundamentally require pixel-level alignment between the reference image and the target pose. When this constraint is violated due to large global motion, the generated results suffer from severe degradation, exhibiting noticeable artifacts and hallucinations.

Moreover, since diffusion-based methods predominantly center the target pose along the y-axis within a fixed output resolution, body parts such as arms are frequently cropped. Furthermore, significant movement along the x-axis often

Figure S3. Comparison between diffusion-based method [12] and DynaAvatar.

Table S3. Effectiveness of our dataset reannotations on 4D-Dress.

Settings	PSNR ↑	SSIM ↑	LPIPS ↓
(1)	20.72	0.952	0.085
(2)	20.91	0.953	0.084
(3) (Ours)	**23.74**	**0.960**	**0.064**

causes the subject to move out of frame, cutting off parts of the animation.

In contrast, DynaAvatar is free from these alignment constraints and robustly handles large global motions. This capability stems from our Dynamic Transformer, which effectively incorporates motion features via attention mechanisms without relying on explicit spatial alignment.

S2. Dataset Reannotation comparisons

Fig. S4 provides additional comparisons between (b) the original SMPL-X fittings and (c) our reannotated results. Our reannotations produce more accurate and visually plausible poses, whereas the original annotations often contain noisy predictions and noticeable temporal jitter. Such instability in the original annotations hinders learning a reliable and practical relationship between human motion and cloth deformation. In contrast, our reannotated sequences exhibit significantly improved temporal consistency and pose accuracy. As a result, our reannotated datasets are directly usable for training motion-dependent deformation models.

S3. Ablation studies

We provide additional ablation study results to validate our design choices.

S3.1. Dataset reannotations

Table S3 on 4D-Dress shows the value of our reannotations by fixing the architecture while varying training annotations. We compare three settings: (1) the original annotations, (2) reannotation of the originally available frames, and (3) our fully reannotated dataset (Sec. 4). Results show that our reannotations (3) yield far superior results compared to the original annotations (1), which suffers from 80% missing frames. Fig. 4 and Sec. S2 additionally show the value of our reannotations.

S4. Architecture details

Fig. 2 and Sec. 3.1 of the main manuscript show architecture of the proposed DynaAvatar. We provide detailed descriptions of each component.

S4.1. Static Transformer

The Static Transformer takes two distinct image tokens: body tokens and head tokens, extracted via Sapiens [6] and DINOv2 [7], respectively. It consists of several layers, each of which is composed of a Body Transformer block and a Head Transformer block. The Body Transformer block utilizes the body tokens as key and value to update the input query tokens, whereas the Head Transformer block utilizes the head tokens. Additionally, we compute the global average of the body tokens and inject this feature into the Static Transformer through adaptive Layer Normalization (AdaLN) [8].

S4.2. Motion encoder

The Motion encoder is designed as a simple MLP that takes the motion history as input and outputs motion tokens. We construct a motion history representation from the pose sequence, which consists of $K = 22$ body joints. For a pose sequence with T frames, we concatenate the 3D joint linear velocities, 6D rotation-parameterized [13] pose, pose velocity, and pose acceleration, resulting in a 21-dimensional motion vector per joint. This motion history is flattened to form a tensor of shape $\mathbb{R}^{T\times(K\cdot 21)}$, which is subsequently mapped to T motion tokens via the motion encoder.

S4.3. Dynamic Transformer

The Dynamic Transformer utilizes the encoded motion tokens to update the query features output by the Static Transformer. Unlike the Static Transformer layer, which comprises two distinct blocks, Dynamic Transformer layer is implemented as a single block where motion tokens act as keys and values. Furthermore, the last element of the motion tokens is injected via AdaLN to explicitly provide the target pose context.

S5. Implementation details

We observed that constructing a batch with a single subject across multiple timeframes and views yields better convergence than stacking multiple subjects. Accordingly, for the training of DynaAvatar, we configure the batch with $F = 4$ temporal frames and $V = 4$ views per subject, resulting in a batch size of 16. Our model is trained using the AdamW

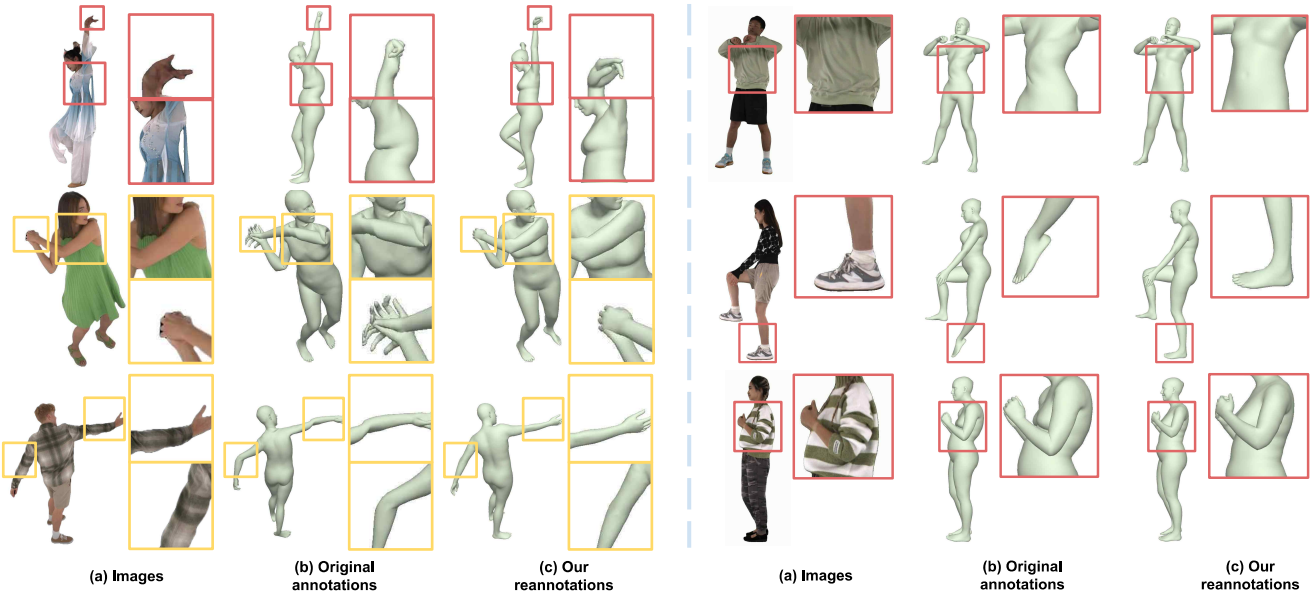

Figure S4. Comparison between (b) the original annotations and (c) our reannotations. The bounding box colors indicate the source datasets: Red denotes DNA-Rendering [1], and Yellow denotes Actors-HQ [5].

optimizer with an initial learning rate of 4×10^{-4} and gradient clipping set to 0.1. We apply LoRA to all linear layers in the Static Transformer with a rank $r = 32$, scaling alpha $\alpha = 64$, and a dropout rate of 0.1. The training is conducted on 8 NVIDIA RTX Pro 6000 GPUs for a total of 40K iterations, taking approximately 90 hours. The DynaFlow loss is activated after 20K iterations to ensure that a coarse geometry is established.

References

[1] Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, et al. DNA-Rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *ICCV*, 2023. 4

[2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 1

[3] Artur Grigorev, Michael J Black, and Otmar Hilliges. HOOD: Hierarchical graphs for generalized modelling of clothing dynamics. In *CVPR*, 2023. 2

[4] Artur Grigorev, Giorgio Becherini, Michael Black, Otmar Hilliges, and Bernhard Thomaszewski. ContourCraft: Learning to resolve intersections in neural multi-garment simulations. *ACM TOG*, 2024. 2

[5] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. HumanRF: High-fidelity neural radiance fields for humans in motion. *ACM TOG*, 2023. 4

[6] 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. 3

[7] 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. 3

[8] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 3

[9] 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

[10] Lingteng Qiu, Peihao Li, Qi Zuo, Xiaodong Gu, Yuan Dong, Weihao Yuan, Siyu Zhu, Xiaoguang Han, Guanying Chen, and Zilong Dong. PF-LHM: 3D animatable avatar reconstruction from pose-free articulated human images. *arXiv preprint arXiv:2506.13766*, 2025. 1

[11] Geonhee Sim and Gyeongsik Moon. PERSONA: Personalized whole-body 3D avatar with pose-driven deformations from a single image. In *ICCV*, 2025. 1

[12] 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 *ICML*, 2025. 2, 3

[13] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 3

[14] 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. In *CVPR*, 2025. 1, 2