Supplementary Material for "LHM: Large Animatable Human Reconstruction Model for Single Image to 3D in Seconds"

Lingteng Qiu* Xiaodong Gu* Peihao Li* Qi Zuo*
Weichao Shen Junfei Zhang Kejie Qiu Weihao Yuan
Guanying Chen† Zilong Dong † Liefeng Bo
Tongyi Lab, Alibaba Group

Contents

1. Demo Video

Please kindly check the Demo Video for animation results of the reconstructed 3D avatar.

2. Details of Human Shape Reconstruction.

To enable accurate modeling of diverse human body shapes from single-view images, we adopt a two-stage approach. First, we sample points from the canonical SMPL-X mesh [4] to initialize Gaussian parameters. These parameters are then refined using blendshapes to capture variations in body morphology. This method reduces parameterization errors inherent in traditional shape parameterizations, while enabling the neural network to learn the transformation from a reference shape to target body configurations.

As shown in Fig. S1, our framework successfully reconstructs a wide range of human body types, including tall, average, stocky, and slender individuals.

*Equal contribution.
†Corresponding author.

Figure S1. Reconstruction from inputs with different shapes.

3. Details of Shape Regularizer

We apply the *as spherical as possible loss* to penalize excessive anisotropy in Gaussian primitives, following [7]:

$$\mathcal{L}_{\text{ASAP}} = \frac{1}{|P|} \sum_{p \in P} \max\left(\frac{\max(S_p)}{\min(S_p)}, r\right) - r \tag{12}$$

where S_p represents the scalings of 3D Gaussian at point p, and r is an empirical threshold value set to 3 in our implementation. The regularization effectively discouraging needle-like ellipsoids while preserving necessary shape variation.

4. Details of the Multimodal Transformer

Our Multimodal Body-Head Transformer (MBHT) is built on top of the recent Multimodal Transformers (MM-Transformer) [1].

The detailed architecture of MM-Transformer is summarized in Fig. S3. The 3D geometric body and head query tokens are fed as q and semantic image feature tokens are fed as h. MM-Transformer aggregates both features by attention mechanism with Adaptive Layer Normalization modulation guided by the extracted global context features.

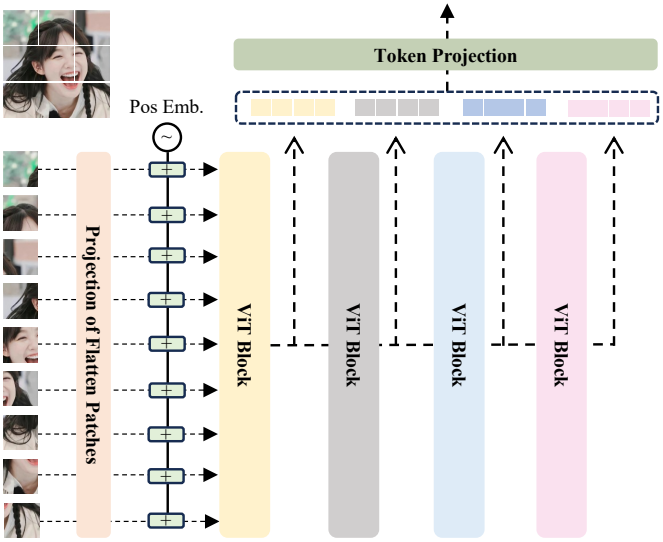

Figure S2. Architecture of our HFPE for multi-scale facial feature extraction

5. Details of Head Feature Pyramid Encoding

Given that the human head occupies a relatively small area within the input image and is subject to spatial downsampling during the encoding process, essential facial details are frequently lost. To address this challenge, we introduce a head feature pyramid encoding (HFPE) designed to aggregate multi-scale features of DINOv2 [3]. Figure S2 illustrates the architecture of HFPE.

6. Details of the Synthetic Training Dataset

To address viewpoint bias in natural videos, we supplement training with synthetic human scans from three sources: (1) 2K2K dataset [2] sampling 1,000 textured models, (2) Human4DiT [6] sampling 4,324 textured characters, and (3) 400 commercial assets from RenderPeople, culminating in 5,724 high-fidelity 3D human scans. Following AniGS [5]'s multi-view rendering protocol, we generate 30 azimuthal views per model with uniform angular spacing (12°intervals) under HDRI lighting conditions.

7. Effects of Canonical Space Regularization

We conduct an ablation study to assess the impact of the canonical space regularization design. Figure S4 shows that the *as spherical as possible* loss $\mathcal{L}_{ASAP}$ is effective in reducing semi-transparent boundary artifacts caused by Gaussians with distorted shapes.

Without the *as close as possible* loss $\mathcal{L}_{ACAP}$, the reconstruction results exhibit noticeable floating points around the human. These results clearly demonstrate the effectiveness of the proposed canonical space regularization losses.

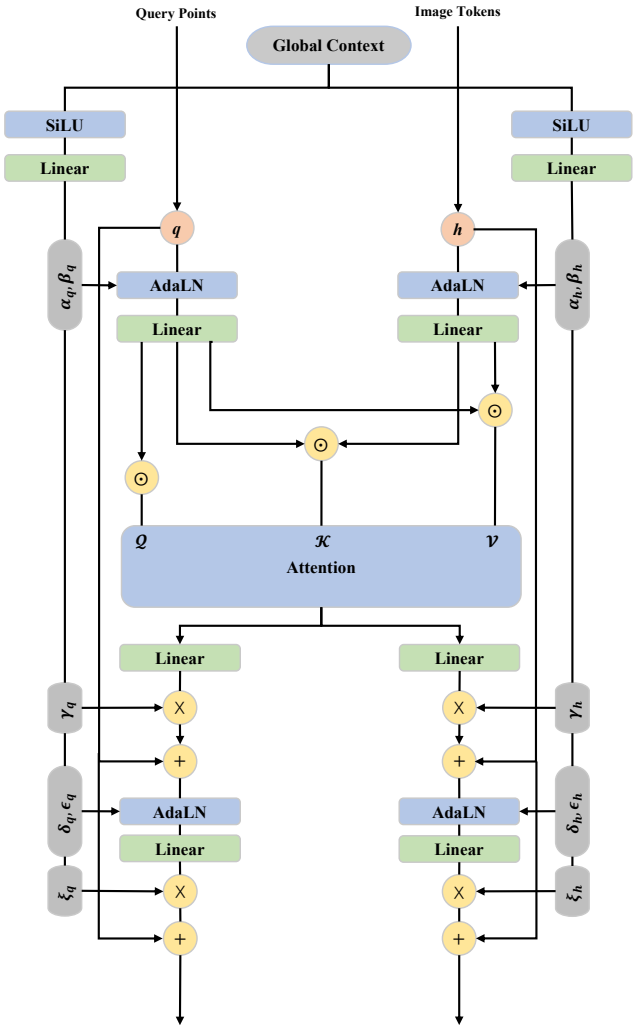

Figure S3. Detailed architecture of Multi-Modal Transformer [1].

Figure S4. Ablation for canonical space shape regularization.

8. More Results

Figure S5–Figure S6 showcase the reconstruction and animation results for input images featuring diverse appearances, clothing, and poses. Our method enables high-fidelity, animatable human avatar reconstruction in a single forward pass with photorealistic rendering, demonstrating

its strong generalization and effectiveness.

References

[1] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 1, 2

[2] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *CVPR*, 2023. 2

[3] 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. 2

[4] 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

[5] 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. 2

[6] Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. Human4dit: 360-degree human video generation with 4d diffusion transformer. *TOG*, 2024. 2

[7] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *CVPR*, 2024. 1

Figure S5. Visual results of 3D human reconstruction results from a single image (Part I). Best viewed with zoom-in.

Figure S6. Visual results of 3D human animation from a single image (Part II). Best viewed with zoom-in.