Supplementary Material for *HRM^{2}Avatar: High-Fidelity Real-Time Mobile Avatars from Monocular Phone Scans*

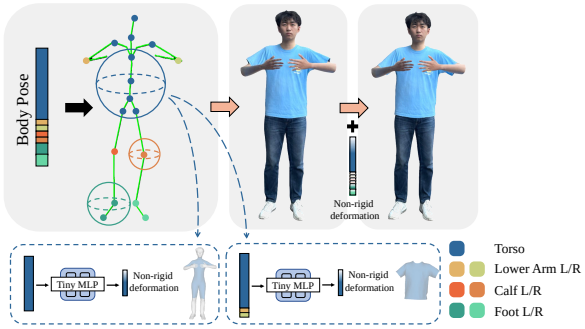
CHAO SHI*, Alibaba Group, China
SHENGHAO JIA*, Shanghai Jiao Tong University, China and Alibaba Group, China
JINHUI LIU, Alibaba Group, China
YONG ZHANG†, Alibaba Group, China
LIANGCHAO ZHU, Alibaba Group, China
ZHONGLEI YANG‡, Alibaba Group, China
JINZE MA, Alibaba Group, China
CHAOYUE NIU, Shanghai Jiao Tong University, China
CHENGFEI LV†, Alibaba Group, China

A Regional MLPs

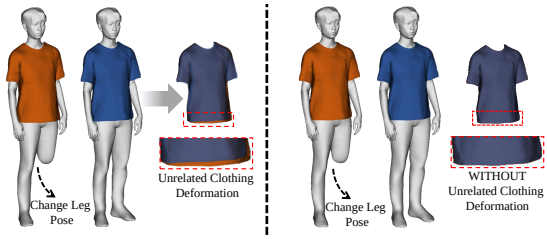
As discussed in the paper, pose-dependent non-rigid deformation and illumination are regressed via lightweight MLPs. To reduce parameter redundancy and improve generalization, we adopts a regional MLP architecture for the pose-dependent deformation. Our key insight stems from the regional nature of deformation effects. As shown in Fig. 1, we partition the body joints into seven distinct regions and assign vertices to these regions based on skinning weights, with each region governed by a lightweight MLP. Each MLP exclusively processes joint poses from its assigned anatomical region to predict localized vertex deformations. For garment, we manually select joints that directly influence garment deformation as MLP inputs. We employ PCA-based low-dimensional projection to suppress high-frequency pose components, thereby mitigating interference from unstable signal fluctuations.

B Double-Side Clothing Modeling

Different from body mesh, clothing mesh exhibits dual-sided characteristics. For instance, the outer side of a sleeve typically reflects the color of the fabric, while the inner side appears black due to the absence of light. To accurately model this phenomenon, we introduce a pre-defined discrete visibility label l for the Gaussians attached to clothing mesh, indicating whether the Gaussian is located on the front or back side of the surface. As shown in Fig. 2, we determine the visibility of Gaussians by computing the dot product between the view direction and the surface normal. During rendering, visibility-based culling of occluded Gaussian components induces heterogeneous optical properties on individual triangle.



(a) Regional MLPs. The input of clothing MLPs are manually adjusted according to specific clothing characteristics.



(b) The ablation study of Regional MLPs. Compared with a single large MLP, the proposed RMLP framework effectively isolates joint movements, preventing their influence from propagating to unrelated regions; for instance, leg joint motions do not interfere with clothing deformation.

Fig. 1. An illustration on Regional MLP.

B.1 Stick Gaussian

In non-hair regions, we constrain Gaussians to two dimensions on the mesh surface by setting their normal-direction scale and offset to zero, permitting only rotation about the normal, referred to as stick Gaussian. As shown in Fig. 4, this constraint reduces

*Both authors contributed equally to this research.
†Corresponding Author.
‡Project Leader.

SA Conference Papers '25, Hong Kong, Hong Kong

ACM ISBN 979-8-4007-2137-3/2025/12
https://doi.org/10.1145/3757377.3763894

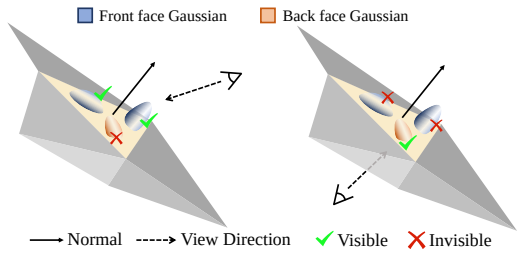


Fig. 2. Gaussian Visibility.

spike introduced during animation without compromising clarity for non-hair surfaces.

C Regularization Losses

Our regularization term is decomposed as:

$$\mathcal{L}_{\text{reg}} = \lambda_{\mathbf{s}}\mathcal{L}_{\mathbf{s}} + \lambda_{\text{uv}}\mathcal{L}_{\text{uv}} + \lambda_{\text{sh}}\mathcal{L}_{\text{sh}} + \lambda_{\text{vLap}}\mathcal{L}_{\text{vLap}} + \lambda_{\text{aoLap}}\mathcal{L}_{\text{aoLap}} + \lambda_{\text{area}}\mathcal{L}_{\text{area}}, \tag{1}$$

where $\mathcal{L}_{\mathbf{s}}$ constrains Gaussian scale $\mathbf{s}$ within a threshold, $\mathcal{L}_{\text{uv}}$ constrains Gaussian center stay inside the triangle, $\mathcal{L}_{\text{sh}}$ applies regularization to higher-order components of $\mathbf{sh}$, $\mathcal{L}_{\text{vLap}}$ and $\mathcal{L}_{\text{aoLap}}$ is Laplacian regularizer on vertex offsets and shadow, $\mathcal{L}_{\text{area}}$ encourages Gaussian splats to completely cover each triangle within the hand mesh. We set the hyper-parameters $\lambda_{\mathbf{s}} = 100$, $\lambda_{\text{uv}} = 100$, $\lambda_{\text{sh}} = 0.01$, $\lambda_{\text{vLap}} = 5 \times 10^{-4}$, $\lambda_{\text{aoLap}} = 5 \times 10^{-7}$, $\lambda_{\text{area}} = 0.01$.

We detail the computation of $\mathcal{L}_{\text{area}}$ in the following. For each Gaussian i, the local area is computed as:

$$A_{\text{local},i} = s_{x,i} \times s_{z,i} \times o_i$$

In this expression, $s_{x,i}$ and $s_{z,i}$ represent the scaling factors of the Gaussian along two axes that lie parallel to the plane of the triangle to which the Gaussian is associated. Subsequently, for each triangle t, the total area covered by associated Gaussians is accumulated:

$$A_{\text{total},t} = \sum_{i \in T_t} A_{\text{local},i},$$

where T_t represents the set of Gaussians linked to triangle t.

The regularization imposed by $\mathcal{L}_{\text{area}}$ ensures adequate coverage through the following expression:

$$\mathcal{L}_{\text{area}} = \sum_{t} \max(0, \epsilon_{\text{area}} - A_{\text{total},t})$$

This loss function is designed to promote complete coverage of each triangle by its associated Gaussians, specifically addressing issues such as holes in the avatar's hands, as illustrated in Fig. 3. The hand's texture is particularly challenging due to significant self-occlusion and constraints from limited resolution and viewpoints in training images, making this targeted approach essential for mitigating the resulting artifacts in generated model.

D Collision Loss

Prior collision losses [Bertiche et al. 2021, 2022; Lin et al. 2024] occasionally fail in our scenario due to initial alignment errors, especially when dealing with collisions of long sleeves, as shown in Fig. 5. Our collision loss $\mathcal{L}_{\text{collision}}$ enforces bidirectional normal consistency between body and clothing surfaces through:

$$\mathcal{L}_{\text{collision}} = \sum_{(b,c) \in \mathcal{N}} \min((c - b) \cdot \mathbf{n}_b - \epsilon, 0)^2 \cdot H(\mathbf{n}_c \cdot \mathbf{n}_b), \tag{2}$$

where $\mathcal{N}$ represents the set of nearest neighbor correspondences vertex pair (b, c) between body and clothing, $\mathbf{n}_*$ represents the vertex normal of vertex $*$, ϵ is a small positive threshold to increase robustness, and $H(\cdot)$ is the Heaviside step function.

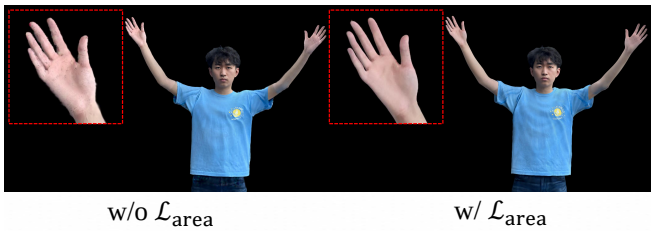


Fig. 3. Ablation on $\mathcal{L}_{\text{area}}$.

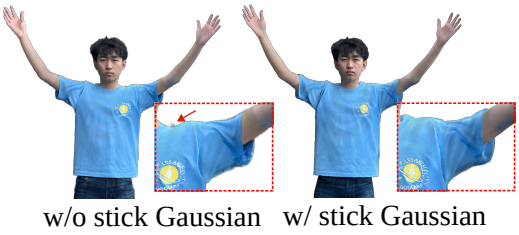


Fig. 4. Ablation on stick Gaussian. Restricting the Gaussian outside of hair to the mesh surface can effectively prevent Gaussian spikes.

E Implementation Details

The background masks of the captured images are segmented by BiRefNet [Zheng et al. 2024], and the SMPL-X parameters are estimated using GVHMR [Shen et al. 2024], HAMER [Pavlakos et al. 2024] and Hand4Whole [Moon et al. 2022]. During optimization, body pose $\boldsymbol{\theta}$ are optimized for all training images to compensate registration errors. The gradients of body pose $\boldsymbol{\theta}$ parameters are broadcast within *StaticSequence* in the first 50k steps. Gaussian attributes (u, v, w), $\mathbf{s}$, o are fixed in first 100k steps to ensure correct geometry optimization.

F Image Acquisition Setup

All images were captured in manual mode with fixed camera parameters throughout the acquisition process to ensure consistent color and exposure across the entire dataset. This approach minimizes inter-shot variations in color representation that might otherwise arise from automatic exposure adjustments, thereby enhancing the reliability of cross-view comparisons. To reduce image noise and motion blur, the ISO sensitivity was set to the lowest feasible value, and the exposure time was kept as short as possible while maintaining sufficient illumination. This configuration effectively suppressed

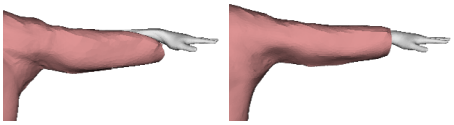


Fig. 5. Ablation on $\mathcal{L}_{\text{collision}}$. Due to geometric misalignment between initial clothing and body geometry, PBNS's collision loss may erroneously converge at wrist regions, causing the body to penetrate through the clothing. By introducing the normal consistency constraint, we achieve correct convergence of the clothing-body relationship.

sensor noise induced by high ISO settings and minimized motion artifacts caused by prolonged exposure, thus preserving image quality and facilitating accurate 3D reconstruction.

G Runtime Memory

We measured the runtime memory consumption (in MB) of three models with sizes of 530K, 1M, and 2M splats, respectively. By applying chunk-based compression, we achieved an average asset compression ratio of 10×, reducing the runtime memory usage to approximately one-tenth of that of the uncompressed data.

Table 1. Runtime memory (MB) of three avatars with different splat count.

Splats	Uncompressed	Compressed	Ratio
530K	130.82	14.35	9.10x
1M	265.97	28.42	9.35x
2M	496.23	42.90	11.5x

H More Comparison with Baseline Methods

Due to the requirement for full-body imagery in ExAvatar [Moon et al. 2025] and GaussianAvatar [Hu et al. 2024], all comparative avatars are reconstructed using full-body frames $I_{sg} \cup I_d$. Considering data fairness, we conducted an additional comparative experiment. We remove all close-up images from the training set to form an ablation version of our methods, as shown in Tab. 2 and Fig. 7. Our method outperforms the baseline in terms of metrics and quality.

Table 2. The metric comparisons between baseline methods and our ablation version on our dataset.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
GaussianAvatar [Hu et al. 2024]	19.78	0.931	0.075
ExAvatar [Moon et al. 2025]	24.43	0.948	0.051
ours w/o close-up	**25.83**	**0.954**	**0.042**

I Failure Case

Our framework handles most daily motions well but may exhibit mesh interpenetration during motions outside the training distribution. For example, squatting may cause the top hem to penetrate through the thigh regions, as shown in Fig. 6.

Fig. 6. A failure case.

References

Hugo Bertiche, Meysam Madadi, and Sergio Escalera. 2021. PBNS: Physically Based Neural Simulation for Unsupervised Garment Pose Space Deformation. *ACM Trans. Graph.* 40, 6, Article 198 (dec 2021), 14 pages. doi:10.1145/3478513.3480479

Hugo Bertiche, Meysam Madadi, and Sergio Escalera. 2022. Neural Cloth Simulation. *ACM Trans. Graph.* 41, 6, Article 220 (Nov. 2022), 14 pages. doi:10.1145/3550454.3555491

Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. 2024. GaussianAvatar: Towards Realistic Human Avatar Modeling from a Single Video via Animatable 3D Gaussians. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, USA, 634–644. doi:10.1109/CVPR52733.2024.00067

Siyou Lin, Zhe Li, Zhaoqi Su, Zerong Zheng, Hongwen Zhang, and Yebin Liu. 2024. LayGA: Layered Gaussian Avatars for Animatable Clothing Transfer. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA) *(SIGGRAPH '24)*. Association for Computing Machinery, New York, NY, USA, Article 37, 11 pages. doi:10.1145/3641519.3657501

Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. 2022. Accurate 3D Hand Pose Estimation for Whole-Body 3D Human Mesh Estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, New Orleans, Louisiana, 2307–2316. doi:10.1109/CVPRW56347.2022.00257

Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. 2025. Expressive Whole-Body 3D Gaussian Avatar. In *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer Nature Switzerland, Cham, 19–35.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. 2024. Reconstructing Hands in 3D with Transformers. In *CVPR*. IEEE, Seattle, USA.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. 2024. World-Grounded Human Motion Recovery via Gravity-View Coordinates. In *SIGGRAPH Asia 2024 Conference Papers* (Tokyo, Japan) *(SA '24)*. Association for Computing Machinery, New York, NY, USA, Article 144, 11 pages. doi:10.1145/3680528.3687565

Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. 2024. Bilateral Reference for High-Resolution Dichotomous Image Segmentation. *CAAI Artificial Intelligence Research* 3 (2024), 9150038.

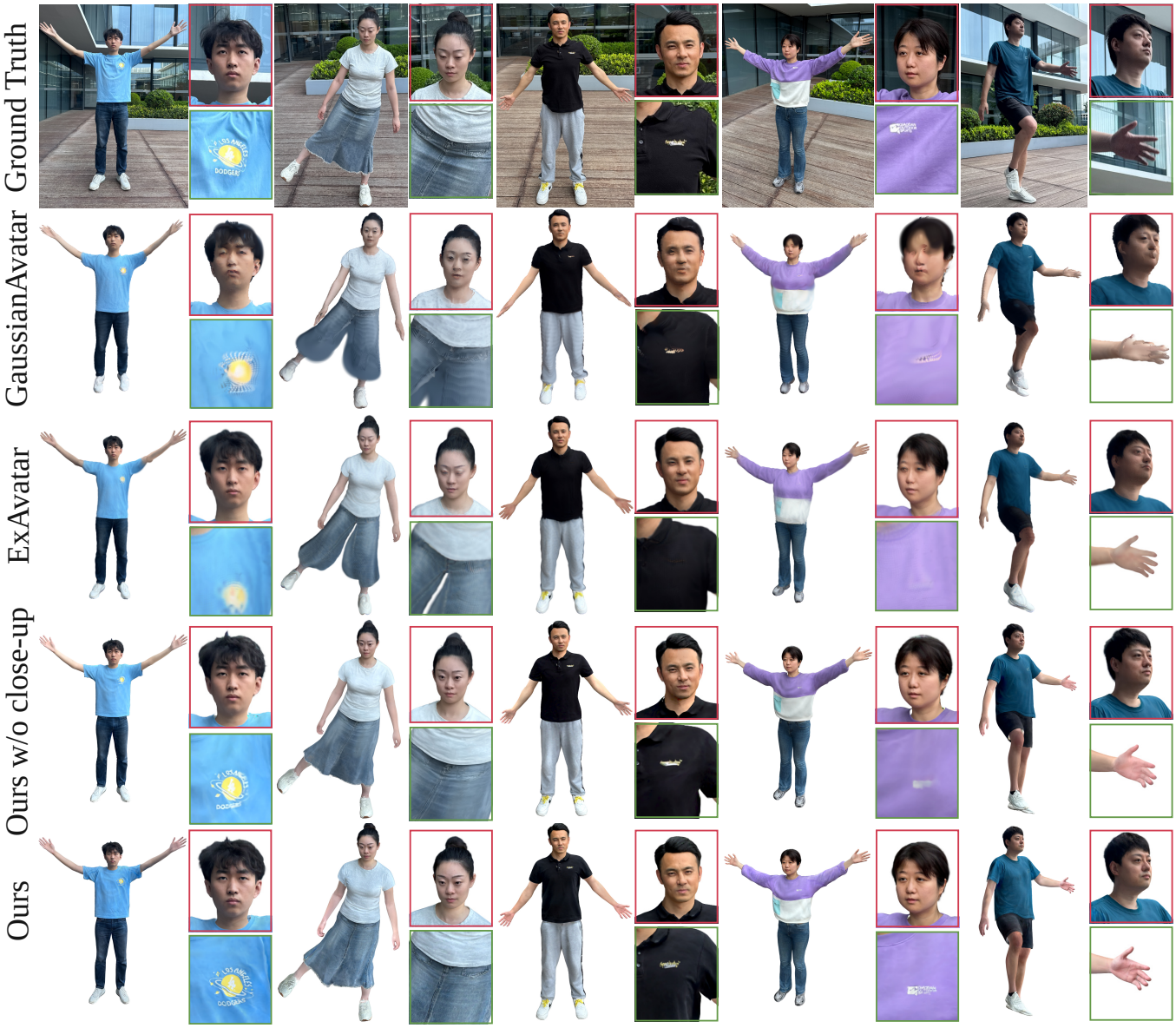


Fig. 7. Self-driven animation comparzisons between monocular avatar methods and our ablation version.