

Large-scale Codec Avatars: The Unreasonable Effectiveness of Large-scale Avatar Pretraining

Supplementary Material

Supplementary Material

A. Additional Qualitative Results

Please see supplementary video for additional qualitative results.

B. Network Architecture

In this section, we describe the design of our network architecture in detail.

B.1. Tokenization Details

We use Sapiens-1B [3] as our image feature extractor. Face crops are obtained by detecting face keypoints using Sapiens, computing a bounding box from the keypoints, and cropping and resizing to the target resolution. The $G=8,192$ geometric tokens are sampled from the surface of a template mesh from MHR [2], with half sampled on the face region and half on the body to provide higher resolution in the face area. The positional encoder \mathcal{F}_{PE} uses fixed Fourier features: given a 3D point, we compute \sin and \cos at 6 logarithmically spaced frequency bands ($2^0, 2^1, \dots, 2^5$) and concatenate the result with the raw input, yielding a 39-dimensional encoding per point. This encoding is then projected to the token dimension $D=1024$ via a single MLP layer ($\mathcal{F}_{proj-gs}$) for use in the subsequent attention layers.

B.2. Transformer

As described in Section 3.1 of the main paper, each LCA Transformer layer comprises three components: (1) self-attention over image tokens, (2) self-attention over geometry tokens, and (3) multimodal attention. All attention modules use 16 heads.

For the image self-attention module \mathcal{A}_{image} , we follow VGGT [5] by augmenting the image token sequence with four additional learned registry tokens, which are discarded after the layer. We also apply 2D Rotary Positional Encoding (2D-RoPE) to preserve spatial information within each image.

For the multimodal attention module $\mathcal{A}_{multimodal}$, we adopt a two-stage design inspired by the Body-Face MM-T block in LHM [4]. Specifically, face image tokens attend to face geometry tokens first. The resulting face geometry features are concatenated with body geometry tokens, and this combined set is then attended by body image tokens,

enabling bidirectional cross-modal interaction. Formally,

$$\mathbf{T}^{gs-face}, \mathbf{T}^{gs-body} = \text{split}(\mathbf{T}^{gs}), \quad (1)$$

$$\mathbf{T}^{global} = \mathcal{F}_{proj}(\text{AvgPool}(\mathbf{T}^{body})), \quad (2)$$

$$\mathbf{T}^{gs-face}, \mathbf{T}^{face} = \mathcal{A}_{MM-T}(\mathbf{T}^{gs-face}, \mathbf{T}^{face}; \mathbf{T}^{global}), \quad (3)$$

$$\mathbf{T}^{gs} = \text{concat}(\mathbf{T}^{gs-face}, \mathbf{T}^{gs-body}), \quad (4)$$

$$\mathbf{T}^{gs}, \mathbf{T}^{body} = \mathcal{A}_{MM-T}(\mathbf{T}^{gs}, \mathbf{T}^{body}; \mathbf{T}^{global}). \quad (5)$$

We compute the global feature $\mathbf{T}^{global} \in \mathbb{R}^{1 \times D}$ by first averaging the body image tokens across all spatial locations, followed by a learnable projection through an MLP \mathcal{F}_{proj} .

B.3. Gaussian Decoder

Both the canonical decoder \mathcal{H}_{cano} and the pose-dependent decoder \mathcal{H}_{pose} are lightweight networks, each composed of four fully connected layers. We use LeakyReLU activation functions between layers to improve stability and gradient flow. The hidden dimensionality of all intermediate layers is set to 128. This compact design enables fast inference while maintaining sufficient representational capacity for high-quality Gaussian decoding.

B.4. Inference Efficiency

To achieve real-time performance, we decouple the inference process into a one-time initialization stage and a run-time animation stage. The computationally intensive transformer encoder and canonical decoder are executed once per subject to generate the canonical Gaussian parameters and geometry tokens. This initialization step takes approximately 2.1 seconds on a single NVIDIA A100 GPU.

For subsequent animation frames, only the lightweight pose-dependent decoder, \mathcal{H}_{pose} , is evaluated. This component is highly efficient, requiring approximately 1.7 ms per forward pass. This design ensures that our method allows for high-fidelity, interactive applications such as VR/AR telepresence and real-time character control.

C. Training Parameters

We train our models using 64 NVIDIA A100 GPUs (80GB) via Distributed Data Parallel (DDP). For both training stages, we set the per-GPU batch size to 1, resulting in a total effective batch size of 64. The pretraining stage is conducted for 1×10^5 iterations, taking approximately 100 hours. Subsequently, the post-training stage is fine-tuned for 1×10^4 iterations, which requires approximately

10 hours to converge. For the loss weights, we set the ℓ_1 and LPIPS coefficients to 0.1 each, and the regularization weight $\lambda = 1.0$.

D. Analysis of Latent Feature Distribution

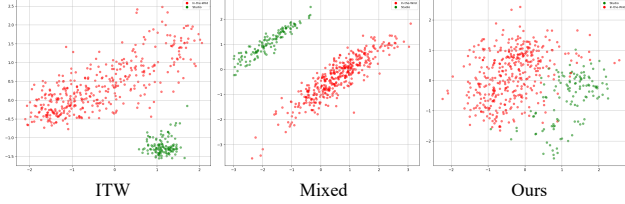


Figure S1. **PCA of Geometric Token Features.** Visualization of the feature space distributions produced by models trained with different strategies. Green points denote studio-captured subjects, while red points denote in-the-wild subjects.

We analyze the distribution of the learned geometric token features, T^{gs} , to understand how different training strategies handle the domain gap between datasets. We extract features for unseen subjects from both the studio-capture and in-the-wild test sets. For each subject, we compute a global feature vector by averaging the geometric tokens and projecting them into 2D space using Principal Component Analysis (PCA). As shown in Figure S1, models trained on ITW-only or Mixed data form distinct clusters for studio (green) and in-the-wild (red) samples, limiting synergetic improvement of both fidelity and generalization. In contrast, our proposed pre/post-training strategy effectively aligns these distributions, treating inputs from both domains consistently. This suggests that our approach learns a robust, high-fidelity, and domain-agnostic representation of human geometry, effectively leveraging the two distinctive training data sources.

E. Additional Ablation Studies

We ablate the decoder architecture and post-training learning rate decay (γ) in Tab. S1. Our dual-branch residual design outperforms the single-branch non-residual variant, likely due to improved pose-dependent decoupling. For the learning rate decay, $\gamma=0.00$ (no decay, all layers trained at the same rate) severely degrades studio metrics, indicating catastrophic forgetting of pretraining knowledge. Both $\gamma=0.30$ and $\gamma=0.65$ perform well; we use $\gamma=0.65$ in our final model as it offers a good balance across both domains.

We also study the effect of data scale on the pre/post-training paradigm (Tab. S2). Training LCA at $10\times$ smaller scale (100K pretraining identities and 500 post-training identities) still yields improvements over baselines, confirming that the benefits of our two-stage approach are not solely attributable to data scale.

Configuration	Capture-Studio			In-the-Wild		
	L1↓	LPIPS↓	PSNR↑	L1↓	LPIPS↓	PSNR↑
Decoder Architecture						
Single-Branch	0.0088	0.0742	30.370	0.0103	0.0516	27.936
Dual-Branch (Ours)	0.0082	0.0688	30.514	0.0096	0.0491	28.175
Post-Training Learning Rate Decay (γ)						
$\gamma = 0.00$	0.0117	0.0904	27.464	0.0096	0.0507	27.669
$\gamma = 0.30$	0.0076	0.0614	30.609	0.0096	0.0482	28.252
$\gamma = 0.65$ (Ours)	0.0082	0.0688	30.514	0.0096	0.0491	28.175
$\gamma = 1.00$	0.0085	0.0694	30.464	0.0114	0.0512	27.478

Table S1. **Ablation study on decoder architecture and post-training learning rate decay.** Our dual-branch residual design outperforms a single-branch variant. The learning rate decay is critical for preserving pretraining knowledge, with $\gamma=0.00$ (no decay) severely degrading studio performance.

Train Data Size Num. Identities	Capture-Studio			In-the-Wild		
	L1↓	LPIPS↓	PSNR↑	L1↓	LPIPS↓	PSNR↑
Pre-100K + Post-500	0.0088	0.0723	30.398	0.0130	0.0606	26.754
Pre-1M + Post-2.7K (Ours)	0.0082	0.0688	30.514	0.0096	0.0491	28.175

Table S2. **Effect of training data scale.** Pre/post-training benefits persist even at $10\times$ smaller scale (100K pretraining identities, 500 post-training identities), with consistent trends across both domains.

Method	L1↓	LPIPS↓	PSNR↑
No Deformer	0.0304	0.2452	24.538
LCA (Ours)	0.0238	0.2189	26.013

Table S3. **Loose garment deformer ablation.** Quantitative evaluation on loose-garment sequences. The full model with the deformer improves all metrics and reduces splitting artifacts.

We quantitatively evaluate the effect of the deformer module on loose-garment sequences in Tab. S3. The full model with the deformer improves all metrics and significantly reduces splitting artifacts. We find the deformer benefits from multi-view post-training data to effectively constrain cloth deformation.

F. Comparison with Alternative Paradigms

We qualitatively compare LCA with Wan-Animate [1], a 2D video diffusion method, and GUAVA [6], an upper-body 3D Gaussian avatar method, in Fig. S2. Relative to video diffusion approaches, LCA is more compute-efficient, enables real-time on-device animation, and exhibits stronger holistic 3D consistency, while diffusion baselines can hallucinate and struggle with long-form generation. Compared to GUAVA, which only supports upper-body reconstruction, LCA produces full-body avatars with higher fidelity.

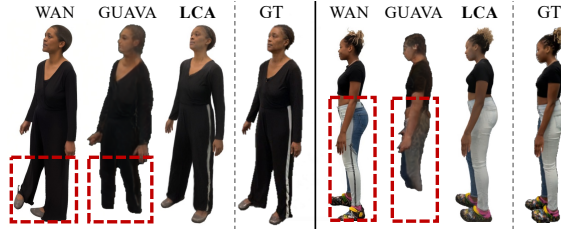


Figure S2. **Qualitative Comparison with Alternative Paradigms.** Comparison with Wan-Animate [1] (2D video diffusion) and GUAVA [6] (upper-body 3D Gaussian avatar).

G. Author Contributions

Project Lead. Shunsuke Saito.

Core Contributors. Junxuan Li and Rawal Khirodkar.

Loose Garment Support. Zhongshi Jiang, Lingchen Yang, and Rinat Abdrashitov.

Relighting. Giljoo Nam and Chengan He.

Evaluation & Benchmarking. Jihyun Lee.

Research Discussions. Egor Zakharov, Abhishek Kar, Christian Häne, Sofien Bouaziz, Jason Saragih, Yaser Sheikh, and Chen Guo.

Data Pipeline & Processing. Contributors listed alphabetically by last name.

Pretraining: Jean-Charles Bazin, James Booth, Wyatt Borso, Yuan Dong, Peihong Guo, Ginés Hidalgo, Matthew Hu, Xiaowen Ma, Julieta Martinez, Marco Pesavento, Yu Rong, Takaaki Shiratori, Carsten Stoll, Zhaoen Su, Anjali Thakrar, Sairanjith Thalanki, Lucy Wang, He Wen, Yichen Xu, and Ariyan Zarei.

Post-Training: Amol Agrawal, Hernan Badino, Chen Cao, Chun-Wei Chan, Yueh-Tung Chen, Shen-Chi Chen, Yuhua Chen, Carol Cheng, Teng Deng, Tingfang Du, Marco Dal Farra, Ryan Frazier, Sidi Fu, Ke Gao, Lihao Ge, Aaqib Habib, Ish Habib, Xuhua Huang, Yuta Inoue, Ethan James, Sam Johnson, Justin Joseph, Anjani Josyula, Song Ju, Kevin Kane, Kai Kang, Thomas Keady, Taylor Koska, Sanjeev Kumar, Jess Kuts, Jianchao Li, Kai Li, Steven Longay, Kevyn McPhail, Sergiu Munteanu, Sam Pepose, Albert Parra Pozo, Wei Pu, David Rogers, Javier Romero, Igor Santesteban, Jake Simmons, Tomas Simon, Nir Sopher, Sam Sussman, Qingyang Tan, Autumn Trimble, Harshita Tupili, Julien Valentin, Carlos Vallespi-Gonzalez, Kiran Vekaria, Kishore Venkateshan, Simon Venshtain, Harsh Vora, Yimu Wang, Yuzhi Wang, Michael Wu, Longhua Wu,

Jiu Xu, Bo Yang, Chengxiang Yin, Shou-I Yu, and Junchen Zhang.

Evaluation Data: Andrew Hou, Austin James, Fei Jiang, Alex Ma, and Conor O'Hollaren.

References

- [1] Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025. 2, 3
- [2] Aaron Ferguson, Ahmed AA Osman, Berta Bescos, Carsten Stoll, Chris Twigg, Christoph Lassner, David Otte, Eric Vignola, Fabian Prada, Federica Bogo, et al. Mhr: Momentum human rig. *arXiv preprint arXiv:2511.15586*, 2025. 1
- [3] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2024. 1
- [4] 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. *CoRR*, abs/2503.10625, 2025. 1
- [5] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 1
- [6] Dongbin Zhang, Yunfei Liu, Lijian Lin, Ye Zhu, Yang Li, Minghan Qin, Yu Li, and Haoqian Wang. Guava: Generalizable upper body 3d gaussian avatar. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14205–14217, 2025. 2, 3