arXiv:2604.02320v1 [cs.CV] 2 Apr 2026

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

Junxuan Li*, Rawal Khirodkar*, Chengan He, Zhongshi Jiang, Giljoo Nam, Lingchen Yang, Jihyun Lee, Egor Zakharov, Zhaoen Su, Rinat Abdrashitov, Yuan Dong, Julieta Martinez, Kai Li, Qingyang Tan, Takaaki Shiratori, Matthew Hu, Peihong Guo, Xuhua Huang, Ariyan Zarei, Marco Pesavento, Yichen Xu, He Wen, Teng Deng, Wyatt Borsos, Anjali Thakrar, Jean-Charles Bazin, Carsten Stoll, Ginés Hidalgo, James Booth, Lucy Wang, Xiaowen Ma, Yu Rong, Sairanjith Thalanki, Chen Cao, Christian Häne, Abhishek Kar, Sofien Bouaziz, Jason Saragih, Yaser Sheikh, Shunsuke Saito†

Codec Avatars Lab, Meta

https://junxuan-li.github.io/lca



Figure 1. **Large-scale Codec Avatars (LCA).** (*Left*) We generate avatars from a handful of images in seconds. LCA follows the pre/post-training paradigm, achieving broad generalization with high-fidelity reconstruction over pretraining alone. (*Middle*) The resulting avatars are highly detailed with faithful 3D structures, fully animatable with expression, gaze, and body pose, even for out-of-domain samples. (*Right*) LCA further supports *loose garments* and *relighting* while retaining generalizability by only modifying the post-training.

Abstract

High-quality 3D avatar modeling faces a critical trade-off between fidelity and generalization. On the one hand, multi-view studio data enables high-fidelity modeling of humans with precise control over expressions and poses, but it struggles to generalize to real-world data due to limited scale and the domain gap between the studio environment and the real world. On the other hand, recent large-scale avatar models trained on millions of in-the-wild samples show promise for generalization across a wide range of identities, yet the resulting avatars are often of low-quality due to inherent 3D ambiguities. To address this, we present Large-Scale Codec Avatars (LCA), a high-fidelity, full-body 3D avatar model that generalizes to world-scale populations in a feedforward manner, enabling efficient inference. Inspired by the success of large language models and vision foundation models, we present, for the first time, a pre/post-training paradigm for 3D avatar modeling at scale: we pre-train on 1M in-the-wild videos to learn broad priors over appearance and geometry, then post-train on high-quality curated data to enhance expressivity and fidelity. LCA generalizes across hair styles, clothing, and demographics while providing precise, fine-grained facial expressions and finger-level articulation control, with strong identity preservation. Notably, we observe emergent generalization to relightability and loose garment support to unconstrained inputs, and zero-shot robustness to stylized imagery, despite the absence of direct supervision.

*Core contributors †Project lead

1. Introduction

Photorealistic human avatars [35, 36, 49, 53, 66, 75, 77] present the opportunity to transform how humans communicate [72], but broad adoption requires systems that work robustly for everyone. We present an approach that, given a handful of images of a subject, produces an identity-preserving 3D avatar that can be accurately driven by subtle facial expressions, full-body motion, and fine-grained hand poses. Achieving this goal requires preserving two key properties that trade-off against each other: (1) *generalization* across clothing, hairstyles, accessories, demographics, and environments, and (2) *fidelity* preserving precise motion and 3D consistent authenticity. Together, these properties are necessary for enabling a true communication service for everyone, as identity must be preserved no matter whom you interact with and the signal embedded in behavior and appearance must be preserved no matter what they do.

Most existing approaches trade generalization and fidelity against each other. One line of work uses **high-quality studio data**, with the number of identities typically in the thousands at most [10, 22, 35, 39, 53, 70], as they use expensive optimization pipelines for personalization. These systems deliver authentic and expressive avatars but they generalize poorly beyond the captured domains. Another line trains on diverse corpora [25, 50, 51] from **in-the-wild data**. These models generalize more broadly in a feedforward manner but often produce distortions from unobserved views, blur in body parts, and limited expressivity.

Recently, large-scale pre/post-training has achieved remarkable success in resolving the aforementioned trade-off in language modeling [1, 60, 62], vision models [6, 32, 56] and video generation [4, 34, 63]. Pretraining learns broad priors for generalization from million-to-billion-scale training data, and the model is post-trained with high-quality curated data to align the learned representation with a target task. Inspired by the success in adjacent domains, we present Large-scale Codec Avatars (LCA), a pre/post-train framework for human avatar creation. LCA first *pre-trains* on millions of in-the-wild videos to learn human priors over appearance and geometry for generalization. It then *post-trains* on high-resolution, multi-view studio captures [42] spanning thousands of identities to specialize for precise control and photorealism. We show, for the first time, that this two-stage approach at scale breaks the generalization-fidelity trade-off, yielding avatars that are fully expressive, identity-preserving, and robustly generated under real-world conditions. Figure 1 qualitatively compares the two stages: the pretrained model generalizes across ethnicity, clothing, and hairstyles, but exhibits muted expressions with distorted 3D shapes, whereas the post-training produces more expressive facial animation with faithful 3D structure while preserving the identity.

Extending the pre/post-training paradigm to human avatar modeling poses a unique challenge: the architecture needs to be scalable, expressive, and efficient. To unify training with studio and in-the-wild data, we adopt a scalable architecture that implicitly associates one or more reference images of a subject with animatable 3D Gaussians [50]. Unlike prior methods based on studio data [22, 35], our approach does not require high-quality conditioning data, such as geometry and texture maps, allowing seamless support of pre/post-training. To capture full-body expressivity, LCA uses a two-branch design: one outputs canonical appearance and geometry, and the other decodes correctives to the canonical output driven by body/hand poses of an expressive body model [47] and facial expression latent codes learned in a self-supervised manner similar to [69].

Specifically, we form two token streams: image tokens from off-the-shelf visual extractors [32] and geometric tokens from a template body mesh in a canonical pose. A large transformer [54] backbone fuses these tokens. To support a variable number of input images, we adopt a hybrid attention scheme that alternates global attention over all tokens with per-image self-attention blocks [64]. A canonical MLP branch decodes the output tokens into per-Gaussian canonical attributes (center, rotation, scale, opacity, and color). A corrective MLP branch predicts per-Gaussian attribute offsets conditioned on the output tokens and the driving signals. The attributes with correctives are transformed to the target pose via linear blend skinning (LBS), and then rendered with differentiable 3D Gaussian splatting [31].

The most remarkable characteristic of LCA is its strong generalizability. LCA faithfully reconstructs clothing, hairstyles, and accessories that do not exist in the post-training data (*e.g.*, eyewear, headwear). We also show that LCA can easily incorporate additional features such as loose-garment handling and relighting [66] while retaining its generalizability to unconstrained inputs by only modifying the post-training stage. Moreover, LCA generalizes to stylized or fictional characters despite explicitly filtering them out from both pre/post-training. Our experiments show that LCA sets a new state-of-the-art in avatar modeling and faithfully captures subtle facial expressions and whole-body motion, including finger-level articulation. Finally, its modular design enables avatar creation in seconds and real-time animation: the pose-dependent residual head is lightweight and runs per frame, while the transformer inference is executed only once during generation.

In summary, our contributions are as follows:

- We are the first to show that million-scale pre/post-training simultaneously yields broad generalization with high-fidelity outputs for animatable avatar creation.
- We propose a new architecture that supports flexible identity conditioning while supporting faithful facial and whole-body animation.

- Our core design is versatile and efficient: LCA extends to additional features, including loose garment handling and relighting, with minimal modifications, and the avatars can be animated in real-time.

2. Related Work

Studio-Based 3D Avatars. 3D human avatar modeling has been actively studied over the past decades [65], and the data available for avatar creation has been a key factor affecting the fidelity of the resulting avatars. The line of work achieving the highest quality typically relies on multi-view studio data captured in calibrated, highly controlled environments [2, 3, 12, 35–37, 41, 42, 49, 53, 58, 75, 77]. Such setups provide dense observations of the identity across diverse viewpoints, appearances, and motions, enabling the effective learning of 3D avatar representations (e.g., NeRF [19, 43, 48], 3DGS [49, 49, 53]) to achieve high authenticity and expressiveness. Relightability is another core capability that can be effectively learned in studio settings equipped with light stages [5, 23, 24, 35, 53, 66, 71], which are essential for achieving photorealistic appearance under varying illumination. Despite their remarkable quality, acquiring calibrated multi-view captures is impractical for users, and these methods often perform poorly when directly generalized to in-the-wild inputs due to domain gap.

In-the-Wild 3D Avatars. Unlike studio-based avatars, approaches that create avatars from in-the-wild (ITW) data reflect more practical real-world scenarios. Most existing methods either (1) learn a feedforward 3D avatar reconstruction model from large-scale, casually captured images or videos [50, 67, 76], or (2) optimize 3D avatar representations directly from ITW captures [20, 28, 29, 44, 55]. However, the avatar creation problem in this setting remains highly under-constrained for achieving high *3D* fidelity, as ITW captures typically provide only sparse and monocular observations. To mitigate this, recent methods attempt to reduce the 3D ambiguity by (1) leveraging image generative models to augment ITW observations [55, 67], or (2) learning a universal prior model additionally trained on multi-view data [22, 35]. Nevertheless, these approaches still require expensive test-time fine-tuning on ITW captures to achieve reasonable quality – falling short of the fidelity and authenticity attained by avatars created from studio-captured data. In summary, studio-based avatars achieve high fidelity but lack generalizability, whereas in-the-wild avatars exhibit the opposite trade-off. To bridge this gap, we incorporate pre/post-training for 3D avatar modeling that jointly leverages the advantages of both data regimes.

Large-Scale Pre/Post-Training. Beyond the 3D avatar domain, recent large language models (LLMs) [1, 7, 15, 27, 59–62] and image or video generative models [4, 8, 34, 63] have demonstrated remarkable performance, achieving both high fidelity and strong generalization. This success largely stems from a two-stage learning paradigm comprising *pre-training* and *post-training*. In the *pretraining stage*, models are trained on massive, diverse datasets to learn comprehensive inductive priors without focusing on specific downstream objectives. While this stage provides robust generalization, it often yields suboptimal fidelity due to noisy, heterogeneous data. In the subsequent *post-training stage*, the model is fine-tuned on smaller, high-quality data to enhance fidelity, alignment, and controllability. For example, recent LLMs [1, 7, 15, 27, 59–62] are pre-trained on trillions of internet-scale text tokens and then post-trained to align with human preferences (e.g., RLHF [45], DPO [52]). Similarly, modern image and video generative models [4, 8, 34, 63] are first pre-trained on large-scale visual data and later post-trained for higher fidelity or controllability [13, 26, 33]. Despite its demonstrated effectiveness in other domains, large-scale pre- and post-training have not yet been explored for 3D avatar modeling—a direction we argue is crucial for achieving fidelity and generalization.

3. Large-scale Codec Avatars

In this section, we detail the architecture (Section 3.1), objective (Section 3.2), data preparation and pre/post-training setup(Section 3.3), and feature extensions (Section 3.4).

3.1. Architecture

Tokenization. Given N images of the full-body $\mathbf{I}_i^{\text{body}} = \{\mathbf{I}_i \in \mathbb{R}^{H\times W\times 3}\}_{i=1}^N$ and face close-up $\mathbf{I}_i^{\text{face}}$, we compute their image features with Sapiens [32], denoted $\mathcal{E}_{\text{sap}}$, and mapped to a D-dimensional token space by a shared single-layer MLP $\mathcal{F}_{\text{proj}}$:

$$\mathbf{T}_i^{\text{body}} = \mathcal{F}_{\text{proj}}(\mathcal{E}_{\text{sap}}(\mathbf{I}_i^{\text{body}})), \tag{1}$$

$$\mathbf{T}_i^{\text{face}} = \mathcal{F}_{\text{proj}}(\mathcal{E}_{\text{sap}}(\mathbf{I}_i^{\text{face}})), \tag{2}$$

where $\mathbf{T}_i^{\text{body}}, \mathbf{T}_i^{\text{face}} \in \mathbb{R}^{P\times D}$ with P denoting the number of patches. In addition, we sample G anchor points from a template 3D human mesh with positions $\mathbf{X} \in \mathbb{R}^{G\times 3}$, following LHM [50]. These are encoded by a positional encoder $\mathcal{F}_{\text{PE}}$ and projected to the same hidden dimension via $\mathcal{F}_{\text{proj-gs}}$ to give $\mathbf{T}^{\text{gs}} \in \mathbb{R}^{G\times D}$:

$$\mathbf{T}^{\text{gs}} = \mathcal{F}_{\text{proj-gs}}(\mathcal{F}_{\text{PE}}(\mathbf{X})). \tag{3}$$

Together, these form two token streams: (i) image tokens $\{\mathbf{T}_i^{\text{body}}, \mathbf{T}_i^{\text{face}}\}_{i=1}^N$ and (ii) geometric tokens $\mathbf{T}^{\text{gs}}$.

Transformer. To efficiently share information across the two token streams, each LCA encoder layer consists of three stages: (i) image attention—self-attention among the image tokens (ii) geometric attention—self-attention among the geometric tokens and (iii) multimodal attention—self-attention over the concatenated image and

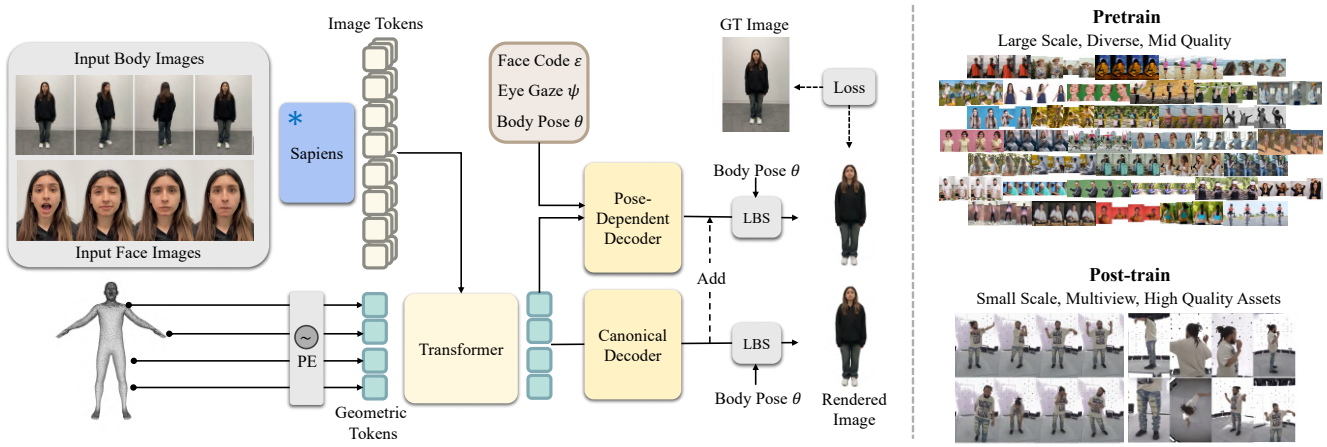


Figure 2. (*Left*) **Overview.** Given multiple images of a subject, we extract *image tokens* from full-body images and face crops, and *geometric tokens* from a template mesh. The LCA encoder alternates image-only, geometry-only, and multimodal attention to fuse information across streams. Our decoders, canonical and pose-dependent, predict Gaussian attributes, which are skinned via linear blend skinning (LBS) and rendered to novel views. Training uses photometric reconstruction losses. (*Right*) **Pretraining vs. Post-Training.** LCA pretrains on large-scale, unconstrained monocular videos of single subjects with mixed (mid/low) quality, then post-trains on high-quality, multi-view studio captures. Pretraining drives broad generalization whereas post-training improves fidelity and 3D completeness.

geometric tokens to fuse body, face, and geometric cues. For each view i,

$$\mathbf{T}_i^{\text{body}} = \mathcal{A}_{\text{image}}(\mathbf{T}_i^{\text{body}}), \tag{4}$$

$$\mathbf{T}_i^{\text{face}} = \mathcal{A}_{\text{image}}(\mathbf{T}_i^{\text{face}}), \tag{5}$$

where $\mathcal{A}_{\text{image}}$ applies self-attention to the image tokens independently, together with standard operations such as LayerNorm [68], residual connections, and MLPs. Similarly, we apply geometric attention using $\mathcal{A}_{\text{geometry}}$ on

$$\mathbf{T}^{\text{gs}} = \mathcal{A}_{\text{geometry}}(\mathbf{T}^{\text{gs}}). \tag{6}$$

We then concatenate per-view outputs, $\mathbf{T}^{\text{body}} = [\mathbf{T}_1^{\text{body}}, \ldots, \mathbf{T}_N^{\text{body}}]$, $\mathbf{T}^{\text{face}} = [\mathbf{T}_1^{\text{face}}, \ldots, \mathbf{T}_N^{\text{face}}]$ and perform the multimodal attention using $\mathcal{A}_{\text{multimodal}}$,

$$\mathbf{T}^{\text{gs}}, \mathbf{T}^{\text{body}}, \mathbf{T}^{\text{face}} = \mathcal{A}_{\text{multimodal}}(\mathbf{T}^{\text{gs}}, \mathbf{T}^{\text{body}}, \mathbf{T}^{\text{face}}). \tag{7}$$

$\mathcal{A}_{\text{multimodal}}$ architecturally resembles the body–face MMDiT [16] block of LHM [50], which uses masked attention so that only face geometric tokens attend to face image tokens. Together $\mathcal{A}_{\text{image}}, \mathcal{A}_{\text{geometry}}, \mathcal{A}_{\text{multimodal}}$ operations constitute a single layer among the L layers of our encoder. This design supports an arbitrary number of input views [64] and enables bidirectional information exchange between image and geometric tokens.

Gaussian Decoder. We decode the geometric tokens into 3D Gaussian attributes – position, rotation, scale, opacity, and color. Our decoder has two heads: a canonical head to capture static features, and a pose-dependent head to model pose-driven effects such as facial expressions, eye gaze, hand pose and clothing deformations.

Canonical: The encoded geometric tokens $\mathbf{T}^{\text{gs}}$ are decoded into canonical 3D Gaussians using an MLP head $\mathcal{H}_{\text{cano}}$,

$$\boldsymbol{c}, \boldsymbol{p}, \boldsymbol{o}, \boldsymbol{q}, \boldsymbol{s} = \mathcal{H}_{\text{cano}}(\mathbf{T}^{\text{gs}}), \tag{8}$$

where $\boldsymbol{c} \in \mathbb{R}^{kG\times 3}, \boldsymbol{p} \in \mathbb{R}^{kG\times 3}, \boldsymbol{o} \in \mathbb{R}^{kG}, \boldsymbol{q} \in \mathbb{R}^{kG\times 4}, \boldsymbol{s} \in \mathbb{R}^{kG\times 3}$ denote the color, position, opacity, quaternion rotation, and scale of each 3D Gaussian, respectively. Note that k is a Gaussian-to-token ratio: each geometric token is expanded into k distinct Gaussians by $\mathcal{H}_{\text{cano}}$.

Pose-Dependent: Given body pose [47] $\boldsymbol{\theta} \in \mathbb{R}^{138}$, face expression code $\boldsymbol{\varepsilon} \in \mathbb{R}^{128}$, and gaze direction $\boldsymbol{\psi} \in \mathbb{R}^{6}$, we concatenate these driving signals with the geometric tokens $\mathbf{T}^{\text{gs}}$ and pass them to an MLP head $\mathcal{H}_{\text{pose}}$ to predict the pose- and expression-dependent deltas of the Gaussian attributes:

$$\Delta\boldsymbol{c}, \Delta\boldsymbol{p}, \Delta\boldsymbol{q}, \Delta\boldsymbol{s} = \mathcal{H}_{\text{pose}}(\boldsymbol{T}^{\text{gs}}, \boldsymbol{\theta}, \boldsymbol{\varepsilon}, \boldsymbol{\psi}). \tag{9}$$

We apply these deltas to the canonical attributes $(\boldsymbol{c}, \boldsymbol{p}, \boldsymbol{o}, \boldsymbol{q}, \boldsymbol{s})$ to obtain pose- and expression-aware Gaussians, while keeping opacities $\boldsymbol{o}$ fixed during animation to promote stability across poses and expressions. Figure 2 illustrates our LCA architecture overview.

3.2. Loss

Our training objective combines a photometric rendering loss with Gaussian regularizations. We transform the canonical and pose-dependent Gaussian attributes to the target view using linear blend skinning (LBS) and render the canonical image $\hat{\boldsymbol{I}}_{\text{cano}}$ and the pose-dependent image $\hat{\boldsymbol{I}}_{\text{pose}}$. We supervise both renderings with ℓ_1 and LPIPS [74] loss,

$$\mathcal{L}_{\text{img}}(\boldsymbol{I}, \hat{\boldsymbol{I}}) = \mathcal{L}_{\ell_1}(\boldsymbol{I}, \hat{\boldsymbol{I}}) + \mathcal{L}_{\text{LPIPS}}(\boldsymbol{I}, \hat{\boldsymbol{I}}). \tag{10}$$

We regularize Gaussian positions $\boldsymbol{p}$ and scales $\boldsymbol{s}$ as,

$$\mathcal{L}_{\text{reg}}(\boldsymbol{p}, \boldsymbol{s}) = \mathcal{L}_{\text{ACAP}}(\boldsymbol{p}) + \mathcal{L}_{\text{ASAP}}(\boldsymbol{s}), \tag{11}$$

where $\mathcal{L}_{\text{ACAP}}$ and $\mathcal{L}_{\text{ASAP}}$ are position- and scale-regularizers [50]. The total loss per training sample is,

$$\mathcal{L} = \mathcal{L}_{\text{img}}\left(\boldsymbol{I}, \hat{\boldsymbol{I}}_{\text{cano}}\right) + \mathcal{L}_{\text{img}}\left(\boldsymbol{I}, \hat{\boldsymbol{I}}_{\text{pose}}\right) + \lambda \mathcal{L}_{\text{reg}}(\boldsymbol{p}, \boldsymbol{s}), \tag{12}$$

where λ is the regularization weight. We observe that adding the photometric rendering loss explicitly against $\hat{\boldsymbol{I}}_{\text{cano}}$ leads to faster convergence.

3.3. Pretraining and Post-Training

We use distinct data sources for LCA's two-stage training. Fig. 2 (*Right*) contrasts the data sources used in each stage.

Pretraining. We curated an in-the-wild dataset of 1 million monocular, human-centric videos. Each video contains a single subject and has a minimum diagonal resolution of 256 pixels. In addition, we collect 40,000 upper-body videos with diverse facial expressions, and full-body videos from $\sim$ 1,000 subjects performing a broad range of motions, captured from diverse viewpoints. During pretraining, both the encoder and decoder are randomly initialized and trained with a higher learning rate. This stage builds broad generalization to diverse inputs.

Post-Training. We use the term *post-training* to refer to supervised fine-tuning on a small, high-quality dataset to improve avatar animatability and visual fidelity. We use a multi-view capture system [42] to record dynamic human performances. The setup uses 200 calibrated, synchronized cameras capturing 4K images. Participants perform casual motions, yielding on average $\sim$ 5,000 frames per subject. In total, we collect recordings from 2,737 participants for model training. During post-training, we start from the pre-trained checkpoint and apply layer-wise learning-rate decay to preserve knowledge acquired during pretraining. Training on this high-quality multi-view data refines the model, improving 3D completeness and fine-grained details.

3.4. Post-Training Extensions

LCA architecture is versatile and extends to multiple applications with minimal modifications in the post-training.

Loose Garment Support. Methods that use predefined skinning weights often produce garment-splitting artifacts when animating loose garments (*e.g.* skirts) [50, 51]. Nevertheless, we adopt such a conventional approach during pre-training for scalability. Specifically, given a predefined skinning weight field $\mathcal{W}$, we deform each Gaussian as

$$\hat{\boldsymbol{p}} = \text{LBS}(\boldsymbol{\theta}, \boldsymbol{p} + \Delta\boldsymbol{p}; \mathcal{W}). \tag{13}$$

Since such fixed skinning weights cannot account for the large variation introduced by clothed humans, we enable loose-garment support in post-training stage. Inspired by [14, 21, 46, 57], we introduce a two-level (coarse-to-fine) learnable deformation module. We define a set of intermediate nodes $\boldsymbol{n} \in \mathbb{R}^{N_{\text{node}} \times 3}$ to encode a low-dimensional deformation subspace articulated by $\boldsymbol{\theta}$ through node-level skinning weights $\mathcal{W}'$. These nodes drive full-Gaussian deformation via embedded deformation weights $\mathcal{W}''$ [57] using 4-nearest-neighbors (Figure 3). To enable subject-specific variation, we parameterize this subspace using learnable correctives applied on the spatial canonical weights [38]:

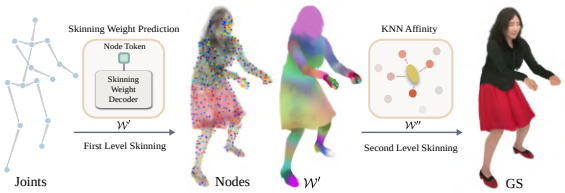


Figure 3. **Node-Based Deformation Model.** We use a flexible two-level learnable deformation model to adapt skinning weights learning for post-training.

$$\mathcal{W}' = \mathcal{W}(\boldsymbol{n}) + \mathcal{H}_{\text{skin}}(\mathbf{T}^{\text{node}}), \tag{14}$$

where $\mathcal{H}_{\text{skin}}$ is an MLP head. For simplicity, node locations $\boldsymbol{n}$ and tokens $\mathbf{T}^{\text{node}} \in \mathbb{R}^{N_{\text{node}} \times D}$ are uniformly sub-sampled from the canonical Gaussians. To learn $\mathcal{H}_{\text{skin}}$, we add regularizations to encourages smooth yet sparse correctives:

$$\mathcal{L}_{\text{skin}} = \mathcal{L}_{\text{ARAP}}(\hat{\boldsymbol{p}}, \boldsymbol{p}) + \lambda_{\text{skw}} \mathcal{L}_{\ell_1}(\mathcal{H}_{\text{skin}}), \tag{15}$$

where $\mathcal{L}_{\text{ARAP}}$ is the As-Rigid-As-Possible loss [30], which regularizes the deformation induced by the learned skinning, and $\mathcal{L}_{\ell_1}$ promotes sparsity. Despite post-training on only a handful of loose garments data, the model generalizes well across unseen garments and identities (Figure 7).

Relighting. To support relighting of the reconstructed avatars, we use learnable radiance transfer proposed in [66]. Adding relightability to LCA only requires the replacement of MLP heads for the canonical decoder and pose-dependent decoder. The model is post-trained with time-multiplexed light stage data [5, 42].

4. Experiments

4.1. Implementation Details

Preprocessing. We process each training dataset to compute (i) foreground and face segmentation, (ii) a pixel-aligned body mesh, (iii) facial-expression estimates, and (iv) eye-gaze estimates. We use fine-tuned models for all these tasks based on Sapiens [32] for both pretraining and post-training data. We primarily use body pose, expressions and eye-gaze to model pose-dependent behaviors, these signals are highest quality in the post-training data.

Model. LCA transformer consists of $L = 8$ layers, each token is $D = 1024$ dimensional. This results in a model size

Train-Source	Capture-Studio			In-the-Wild		
	L1↓	LPIPS↓	PSNR↑	L1↓	LPIPS↓	PSNR↑
In-the-Wild	0.0117	0.0904	27.464	0.0099	0.0507	27.669
Capture-Studio	0.0091	0.0766	29.741	0.0128	0.0706	26.632
Mixed	0.0087	0.0724	30.025	0.0099	0.0519	27.998
Pre→Post (Ours)	0.0082	0.0688	30.514	0.0096	0.0491	28.175

Table 1. Effect of training schemes evaluated across domains.

of 800M parameters. We process all images at a resolution of 1024×768. During training, we randomly sample the number of images N from $[1, 4]$. At evaluation, we fix $N = 4$. We set the Gaussian-to-token ratio $k = 8$ and $G = 8192$, resulting in 65,536 Gaussians per avatar.

Training. We use the AdamW optimizer [40] with a learning rate of 4×10^{-4} and weight decay of 0.05. The learning rate follows a cosine annealing schedule, preceded by a brief linear warm-up. We use mixed-precision training and gradient clipping with a maximum norm of $||\nabla||_2 = 1.0$. Standard image augmentations like random cropping, scaling, flipping, and photometric distortions are used. For post-training, following [32], we use differential learning rates to maintain generalization, applying lower rates to earlier layers and progressively higher rates to later layers. Specifically, we set the layer-wise decay to 0.65. The learning rates for both decoders are unchanged.

Evaluation. We evaluate on two test sets: capture-studio, and in-the-wild. The capture-studio set contains randomly sampled views from a multi-view setup for 100 subjects. The in-the-wild set contains monocular videos of 1000 fully held-out subjects captured under unconstrained conditions. We report L1, LPIPS [74], and PSNR [17], computed exclusively on human pixels using segmentation masks.

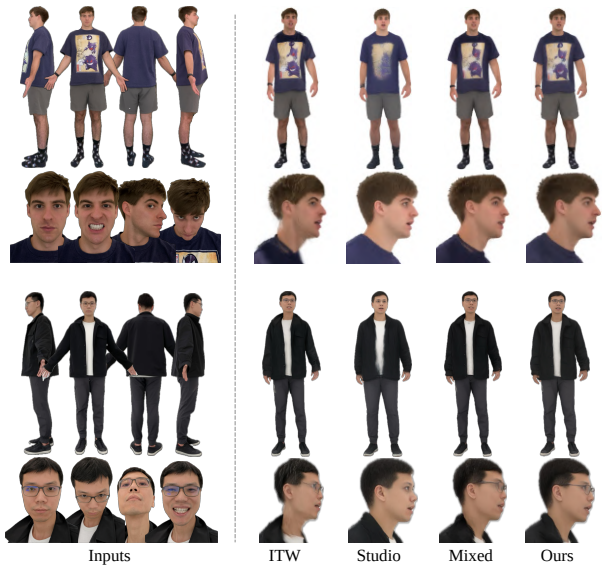


Figure 4. **Pretraining vs. Post-Training.** Qualitative comparison of models trained on multiple data sources and training strategies.

Method	Capture-Studio			In-the-Wild		
	L1↓	LPIPS↓	PSNR↑	L1↓	LPIPS↓	PSNR↑
(a) Multi-view						
ExAvatar [44]	0.011	0.065	23.925	0.033	0.120	18.010
UP2You [9]	0.021	0.085	19.585	0.025	0.086	19.062
MV-LHM* [50]	0.008	0.043	26.606	0.009	0.040	25.796
LCA (Ours)	0.007	0.041	27.483	0.008	0.029	27.802
(b) Single-view						
LHM [50]	0.016	0.072	21.897	0.036	0.109	18.322
LCA (Ours)	0.008	0.043	26.878	0.008	0.030	27.685

Table 2. Quantitative comparison with state-of-the-art 3D avatar methods. * denotes methods trained by us for multi-view inputs.

4.2. Pretraining vs. Post-Training

We study the effectiveness of large-scale pretraining and post-training across diverse data sources. Table 1 compares models trained only on capture-studio data, only on in-the-wild data, on their mixture, and our proposed pretrain→post-train scheme. Surprisingly, the studio-only model does not perform best even on the studio domain, suggesting that it overfits to identity appearance and viewpoints. In contrast, a model trained on mixed data provides a strong baseline, achieving about 30.0 PSNR on studio and 28.0 PSNR on in-the-wild test set, which mirrors the most common training recipe in existing methods [11, 50, 55]. Our pretrain→post-train approach, however, outperforms this mixed strategy across domains – reaching 30.5 PSNR on studio data highlighting improved realism and 28.2 PSNR on the in-the-wild set showcasing stronger generalization.

Figure 4 qualitatively compares feed-forward avatar creation from eight input images for all methods. We show that the in-the-wild-only model produces blurry avatar with incomplete 3D geometry, whereas the studio-only model is sharper and more 3D-complete but fails to generalize to unseen clothing and accessories such as jackets and eyewear. The mixed-data model behaves similarly to the in-the-wild model with slightly better fidelity. Our approach yields sharper, more 3D-complete avatars that generalize better to diverse clothing and accessories.

4.3. Comparison with State-of-the-Art Methods

Table 2 compares LCA with existing state-of-the-art methods [9, 44, 50] using their recommended evaluation protocols, in both multi-view and monocular setups. For fairness, we extend monocular baselines to multi-view setting using our data, *e.g.*, MV-LHM denotes our multi-view extension of LHM [50] trained with the mixture of our pretraining and post-training data. LCA consistently outperforms prior methods across setups and data distributions. In the multi-view studio setting, it surpasses ExAvatar [44] by 3.56 dB PSNR, and by 9.8 dB in the in-the-wild setting. Because

ExAvatar [44] is optimization-based, it does not generalize well to unseen viewpoints during avatar fitting, leading to lower scores overall. We observe the largest gains on faces and extremities such as fingers and legs. When repurposed for monocular (single-view) input, LCA outperforms LHM [50] by 5.0 dB PSNR on studio data and by 9.3 dB PSNR in the in-the-wild setting. Overall, our method produces sharper local details (*e.g.* mouth, fingers) and more faithful articulation than the baselines.

Figure 5 provides qualitative comparisons. LCA avatars produce faithful geometry and appearance, whereas other methods often exhibit 3D distortions, especially around the nose and hips. Our avatar also preserves fine appearance details such as shoelaces, benefiting from high-resolution supervision. Despite never being post-trained on examples with eyewear or headwear, LCA generalizes to these, as well as to diverse hairstyles, clothing, ethnicities, and even stylized characters (see Figure 1). Overall, LCA produces more expressive facial animations with better preservation of subtle changes such as eye gaze, cheek motion, and inner-mouth details. We additionally compare with Wan-Animate [13] (2D video diffusion) and GUAVA [73] (upper-body 3D avatar) in the supplementary material.

# Videos	Capture-Studio			In-the-Wild		
	L1↓	LPIPS↓	PSNR↑	L1↓	LPIPS↓	PSNR↑
10K	0.0083	0.0681	30.393	0.0109	0.0536	27.665
100K	0.0083	0.0677	30.457	0.0108	0.0533	27.755
1M	0.0082	0.0688	30.514	0.0096	0.0491	28.175

Table 3. Effect of scaling pretraining data on downstream performance across data distributions.

4.4. Discussion

Effect on Attention Maps. We show that LCA implicitly learns semantic correspondences between mesh vertices and input images. Figure 6 visualizes the post-training change in last-layer attention between selected geometric token and image patches, shown as heatmaps. For instance, a vertex on the back of the head attends to image regions corresponding to the head, while a hand vertex attends to pixels around the left hand. Compared to pretraining, post-training reduces the noise in the attention maps, yielding cleaner correspondences with human pixels.

Scaling Pretraining Data. We study how the pretraining data scale affects performance. To this end, we subsample the data to 10K, 100K, and 1M videos while keeping all other details fixed. All models are pretrained and post-trained using the same hyperparameters. Table 3 shows that

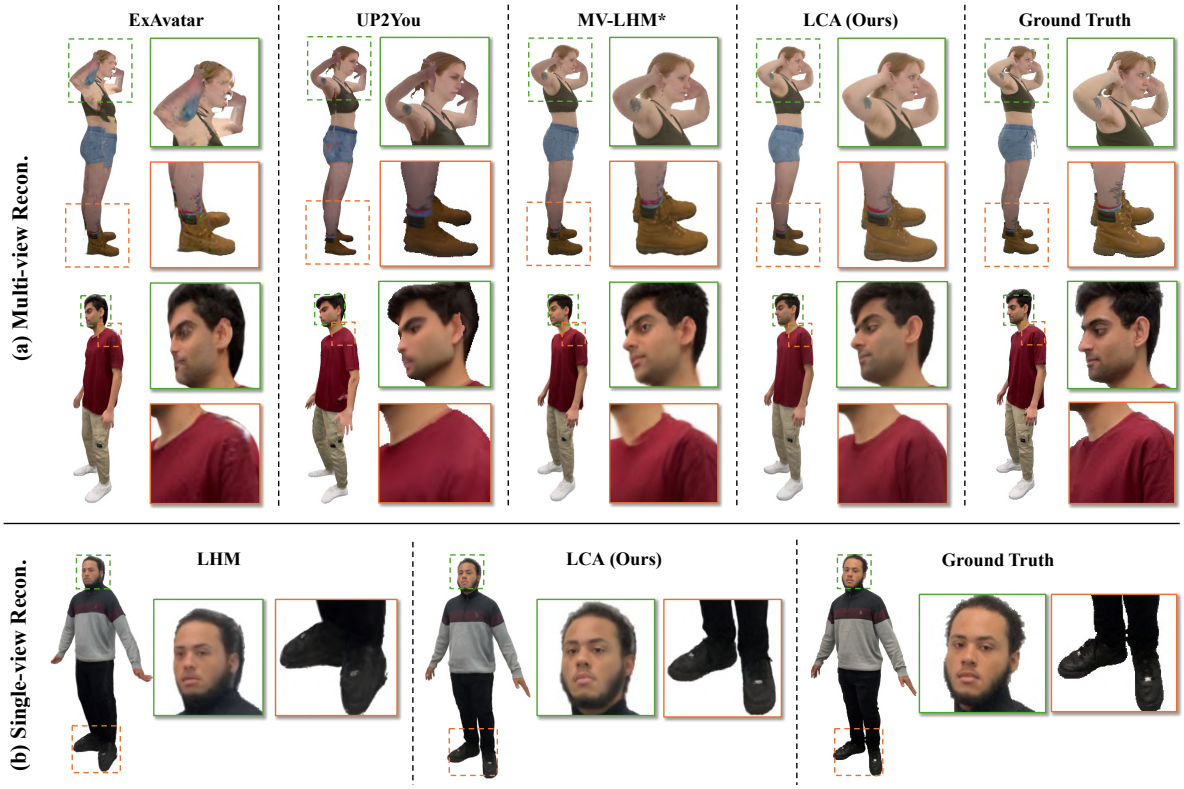


Figure 5. **Qualitative Comparison with State-of-the-Art Methods.** LCA outperforms in both multi-view and monocular settings.

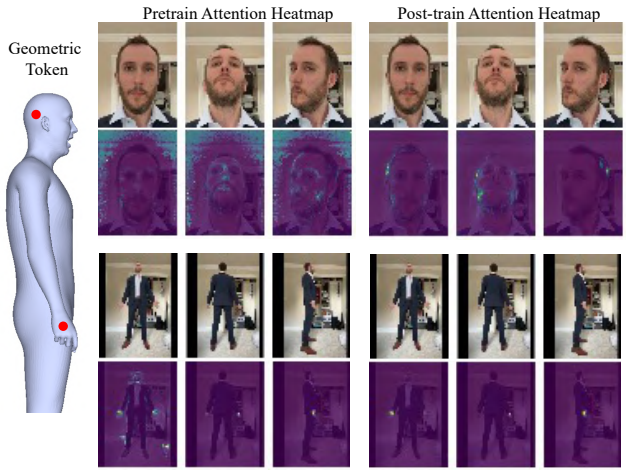


Figure 6. **Attention Map Visualization.** Post-training yields cleaner semantic correspondences in last-layer attention maps between geometric and image tokens. The selected geometric tokens on the mesh are shown in red.

on the studio benchmark, all three settings achieve similar performance, suggesting that, when combined with multi-view post-training, even a relatively small amount of in-the-wild pretraining suffices for the studio domain. In contrast, performance on the in-the-wild test set depends strongly on pretraining scale: increasing from 10K to 100K to 1M videos significantly reduces reconstruction error.

Loose Garment Support. We post-train the LCA model with loose garment support on 147 subjects with loose garments in addition to our existing post-train dataset. As shown in Figure 7, the original post-trained LCA suffers from splitting leg artifacts due to the predefined skinning weights derived from a minimally clothed body template. In contrast, the modified model more faithfully deforms loose garments without splitting inside the garments even for unseen identities casually captured with a mobile phone.

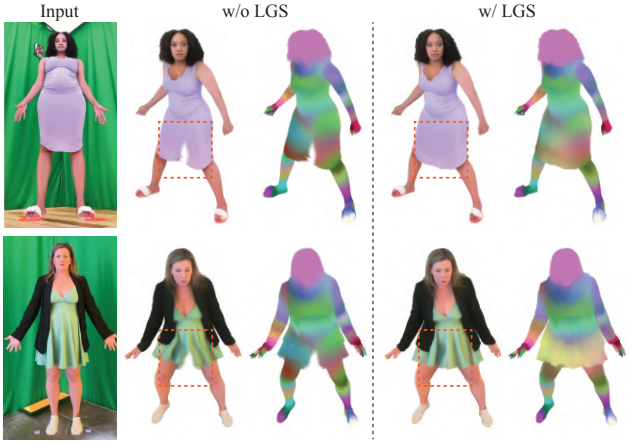


Figure 7. **Loose Garment Support (LGS).** (*Left*) Frontal view of the input condition. (*Middle*) Post-trained LCA avatar without loose garment support, while the general shape is recovered, skirts behave like pants when moving. (*Right*) LCA with loose garment support produces plausible animations without splitting garments.

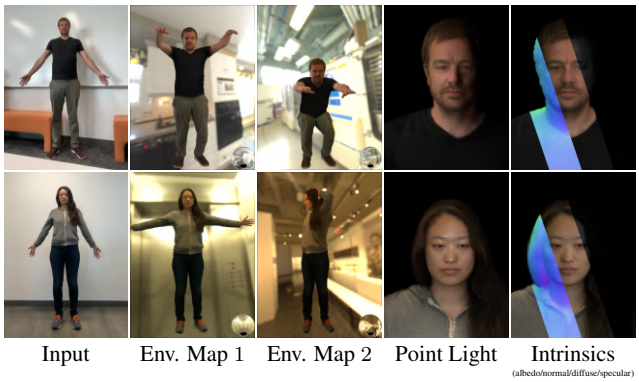


Figure 8. **Relightable LCA.** We demonstrate relighting under HDRI environment maps and point lights, alongside the recovered intrinsic properties of the avatar.

Relighting. We post-train our relightable LCA extension on multi-view captures with time-multiplex light patterns including fully-lit and partially-lit frames. As shown in Figure 8, LCA recovers plausible reflectance and produces consistent, photorealistic relighting across diverse lighting environments, despite being conditioned only on unconstrained phone captures at test time. We also visualize albedo and normal recovered by LCA.

Real-Time Drivability. Our residual design with separate canonical and pose-dependent decoders enables efficient inference. We process input images once with the transformer and canonical decoder, and driving thereafter uses only the lightweight pose-dependent decoder. This yields real-time performance at 586 FPS on an A100 GPU.

Limitations. Finer details, such as clothing with embroidery and intricate textures, remain challenging. Moreover, future work should address secondary motion dynamics such as hair motion and movement of accessories like handbags. The model does not handle heavy occlusions or fast motion blur, which can degrade reconstruction quality.

5. Conclusion

We introduce LCA, a pre/post-training framework for modeling full-body avatars with high-fidelity animation and faithful 3D details. Our experiments show that it is possible to achieve generalization and high-fidelity generation by effectively leveraging large-scale in-the-wild data and high-quality studio data with the proposed two-stage training strategy, similar to frontier language and vision models. Our design is versatile, enabling feature extensions non-trivial to incorporate with large-scale data such as relighting and loose garment handling, and the resulting avatars run in real time. The multi-stage training paves the way for truly scalable high-fidelity avatar creation for everyone.

Acknowledgments

We thank Amol Agrawal, Hernan Badino, Chun-Wei Chan, Yueh-Tung Chen, Shen-Chi Chen, Yuhua Chen, Carol Cheng, Tingfang Du, Marco Dal Farra, Ryan Frazier, Sidi Fu, Ke Gao, Liuhao Ge, Chen Guo, Aaqib Habib, Ish Habib, Andrew Hou, Yuta Inoue, Ethan James, Austin James, Fei Jiang, Sam Johnson, Justin Joseph, Anjani Josyula, Song Ju, Kevin Kane, Kai Kang, Thomas Keady, Taylor Koska, Sanjeev Kumar, Jess Kuts, Jianchao Li, Steven Longay, Alex Ma, Kevyn McPhail, Sergiu Munteanu, Conor O'Hollaren, Sam Pepose, Albert Parra Pozo, Wei Pu, David Rogers, Javier Romero, Igor Santesteban, Jake Simmons, Tomas Simon, Nir Sopher, Sam Sussman, 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, Chengxiang Yin, Jiu Xu, Bo Yang, Shoou-I Yu, and Junchen Zhang for data processing and discussion.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3

[2] Oleg Alexander, Mike Rogers, William Lambeth, Jen-Yuan Chiang, Wan-Chun Ma, Chuan-Chang Wang, and Paul Debevec. The digital emily project: Achieving a photorealistic digital actor. *IEEE Computer Graphics and Applications*, 30 (4):20–31, 2010. 3

[3] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabian Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM Transactions on Graphics (TOG)*, 40(4):1–17, 2021. 3

[4] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2, 3

[5] Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. Deep relightable appearance models for animatable faces. *ACM Transactions on Graphics (ToG)*, 40(4):1–15, 2021. 3, 5

[6] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025. 2

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[8] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. 3

[9] Zeyu Cai, Ziyang Li, Xiaoben Li, Boqian Li, Zeyu Wang, Zhenyu Zhang, and Yuliang Xiu. Up2you: Fast reconstruction of yourself from unconstrained photo collections. *arXiv preprint arXiv:2509.24817*, 2025. 6

[10] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shunsuke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, et al. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022. 2

[11] Hyunsoo Cha, Inhee Lee, and Hanbyul Joo. PERSE: personalized 3d generative avatars from A single portrait. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 15953–15962. Computer Vision Foundation / IEEE, 2025. 6

[12] Yushuo Chen, Zerong Zheng, Zhe Li, Chao Xu, and Yebin Liu. Meshavatar: Learning high-quality triangular human avatars from multi-view videos. In *European Conference on Computer Vision*, pages 250–269. Springer, 2024. 3

[13] 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. 3, 7, 2

[14] Kevin G Der, Robert W Sumner, and Jovan Popović. Inverse kinematics for reduced deformable models. *ACM Transactions on graphics (TOG)*, 25(3):1174–1179, 2006. 5

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[16] 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 *Forty-first international conference on machine learning*, 2024. 4

[17] Fernando A Fardo, Victor H Conforto, Francisco C De Oliveira, and Paulo S Rodrigues. A formal evaluation of psnr as quality measurement parameter for image segmentation algorithms. *arXiv preprint arXiv:1605.07116*, 2016. 6

[18] 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

[19] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*, pages 8649–8658. Computer Vision Foundation / IEEE, 2021. 3

[20] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos

in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12858–12868, 2023. 3

[21] Chen Guo, Tianjian Jiang, Manuel Kaufmann, Chengwei Zheng, Julien Valentin, Jie Song, and Otmar Hilliges. Reloo: Reconstructing humans dressed in loose garments from monocular video in the wild. In *European Conference on Computer Vision*, pages 21–38. Springer, 2024. 5

[22] Chen Guo, Junxuan Li, Yash Kant, Yaser Sheikh, Shunsuke Saito, and Chen Cao. Vid2avatar-pro: Authentic avatar from videos in the wild via universal prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5559–5570, 2025. 2, 3

[23] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 38(6):1–19, 2019. 3

[24] Mingming He, Pascal Clausen, Ahmet Levent Taşel, Li Ma, Oliver Pilarski, Wenqi Xian, Laszlo Rikker, Xueming Yu, Ryan Burgert, Ning Yu, et al. Diffrelight: Diffusion-based facial performance relighting. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 3

[25] Yisheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Weihao Yuan, Zilong Dong, and Liefeng Bo. Lam: Large avatar model for one-shot animatable gaussian head. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–13, 2025. 2

[26] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 3

[27] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 3

[28] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16922–16932, 2023. 3

[29] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, pages 402–418. Springer, 2022. 3

[30] Yuheng Jiang, Zhehao Shen, Yu Hong, Chengcheng Guo, Yize Wu, Yingliang Zhang, Jingyi Yu, and Lan Xu. Robust dual gaussian splatting for immersive human-centric volumetric videos. *ACM Transactions on Graphics (TOG)*, 43 (6):1–15, 2024. 5

[31] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4):1–14, 2023. 2

[32] 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. 2, 3, 5, 6, 1

[33] Boeun Kim, Hea In Jeong, JungHoon Sung, Yihua Cheng, Jeongmin Lee, Ju Yong Chang, Sang-Il Choi, Younggeun Choi, Saim Shin, Jungho Kim, et al. Personabooth: Personalized text-to-motion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22756–22765, 2025. 3

[34] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2, 3

[35] Junxuan Li, Chen Cao, Gabriel Schwartz, Rawal Khirodkar, Christian Richardt, Tomas Simon, Yaser Sheikh, and Shunsuke Saito. Uravatar: Universal relightable gaussian codec avatars. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2, 3

[36] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision*, pages 419–436. Springer, 2022. 2

[37] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19711–19722, 2024. 3

[38] Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu. Learning implicit templates for point-based clothed human modeling. In *European Conference on Computer Vision*, pages 210–228. Springer, 2022. 5

[39] Di Liu, Teng Deng, Giljoo Nam, Yu Rong, Stanislav Pidhorskyi, Junxuan Li, Jason Saragih, Dimitris N Metaxas, and Chen Cao. Lucas: Layered universal codec avatars. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21127–21137, 2025. 2

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[41] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 64–73, 2021. 3

[42] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, et al. Codec avatar studio: Paired human captures for complete, driveable, and generalizable avatars. *Advances in Neural Information Processing Systems*, 37:83008–83023, 2024. 2, 3, 5

[43] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[44] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. 3, 6, 7

[45] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 3

[46] Xiaoyu Pan, Jiaming Mai, Xinwei Jiang, Dongxue Tang, Jingxiang Li, Tianjia Shao, Kun Zhou, Xiaogang Jin, and Dinesh Manocha. Predicting loose-fitting garment deformations using bone-driven motion networks. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 5

[47] Jinhyung Park, Javier Romero, Shunsuke Saito, Fabian Prada, Takaaki Shiratori, Yichen Xu, Federica Bogo, Shoou-I Yu, Kris Kitani, and Rawal Khirodkar. Atlas: Decoupling skeletal and shape parameters for expressive parametric human modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6508–6518, 2025. 2, 4

[48] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 3

[49] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 2, 3

[50] 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. 2, 3, 4, 5, 6, 7, 1

[51] 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. 2, 5

[52] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023. 3

[53] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 130–141, 2024. 2, 3

[54] Joonghyuk Shin, Alchan Hwang, Yujin Kim, Daneul Kim, and Jaesik Park. Exploring multimodal diffusion transformers for enhanced prompt-based image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19492–19502, 2025. 2

[55] Geonhee Sim and Gyeongsik Moon. Persona: Personalized whole-body 3d avatar with pose-driven deformations from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12670–12680, 2025. 3, 6

[56] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 2

[57] Robert W Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. In *ACM siggraph 2007 papers*, pages 80–es. 2007. 5

[58] Tiancheng Sun, Zexiang Xu, Xiuming Zhang, Sean Fanello, Christoph Rhemann, Paul Debevec, Yun-Ta Tsai, Jonathan T Barron, and Ravi Ramamoorthi. Light stage super-resolution: continuous high-frequency relighting. *ACM Transactions on Graphics (TOG)*, 39(6):1–12, 2020. 3

[59] Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. Scale efficiently: Insights from pre-training and fine-tuning transformers. *arXiv preprint arXiv:2109.10686*, 2021. 3

[60] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2

[61] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[62] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2, 3

[63] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 3

[64] 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. 2, 4, 1

[65] Ruihe Wang, Yukang Cao, Kai Han, and Kwan-Yee K Wong. A survey on 3d human avatar modeling–from reconstruction to generation. *arXiv preprint arXiv:2406.04253*, 2024. 3

[66] Shaofei Wang, Tomas Simon, Igor Santesteban, Timur Bagautdinov, Junxuan Li, Vasu Agrawal, Fabian Prada, Shoou-I Yu, Pace Nalbone, Matt Gramlich, et al. Relightable full-body gaussian codec avatars. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025. 2, 3, 5

[67] Zhenzhen Weng, Jingyuan Liu, Hao Tan, Zhan Xu, Yang Zhou, Serena Yeung-Levy, and Jimei Yang. Template-free single-view 3d human digitalization with diffusion-guided lrm. *arXiv preprint arXiv:2401.12175*, 2024. 3

[68] Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and improving layer normalization. *Advances in neural information processing systems*, 32, 2019. 4

[69] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *Advances in Neural Information Processing Systems*, 37:660–684, 2024. 2

[70] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2024. 2

[71] Haotian Yang, Mingwu Zheng, Wanquan Feng, Haibin Huang, Yu-Kun Lai, Pengfei Wan, Zhongyuan Wang, and Chongyang Ma. Towards practical capture of high-fidelity relightable avatars. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 3

[72] Kevin Yu, Gleb Gorbachev, Ulrich Eck, Frieder Pankratz, Nassir Navab, and Daniel Roth. Avatars for teleconsultation: Effects of avatar embodiment techniques on user perception in 3d asymmetric telepresence. *IEEE Transactions on Visualization and Computer Graphics*, 27(11):4129–4139, 2021. 2

[73] 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. 7, 2, 3

[74] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4, 6

[75] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive full-body avatars. *ACM Transactions on Graphics (TOG)*, 42(4): 1–19, 2023. 2, 3

[76] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26308–26319, 2025. 3

[77] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. In *2025 International Conference on 3D Vision (3DV)*, pages 979–990. IEEE, 2025. 2, 3

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

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 [32] 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 [18], with half sampled on the face region and half on the body to provide higher resolution in the face area. The positional encoder $\mathcal{F}_{\text{PE}}$ uses fixed Fourier features: given a 3D point, we compute $\sin$ and $\cos$ at 6 logarithmically spaced frequency bands $(2^0, 2^1, \ldots, 2^5)$ 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 ($\mathcal{F}_{\text{proj-gs}}$) 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 $\mathcal{A}_{\text{image}}$, we follow VGGT [64] 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 $\mathcal{A}_{\text{multimodal}}$, we adopt a two-stage design inspired by the Body–Face MM-T block in LHM [50]. 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,

$$\mathbf{T}^{\text{gs-face}}, \mathbf{T}^{\text{gs-body}} = \text{split}(\mathbf{T}^{\text{gs}}), \tag{16}$$

$$\mathbf{T}^{\text{global}} = \mathcal{F}_{\text{proj}}(\text{AvgPool}(\mathbf{T}^{\text{body}})), \tag{17}$$

$$\mathbf{T}^{\text{gs-face}}, \mathbf{T}^{\text{face}} = \mathcal{A}_{\text{MM-T}}(\mathbf{T}^{\text{gs-face}}, \mathbf{T}^{\text{face}}; \mathbf{T}^{\text{global}}), \tag{18}$$

$$\mathbf{T}^{\text{gs}} = \text{concat}(\mathbf{T}^{\text{gs-face}}, \mathbf{T}^{\text{gs-body}}), \tag{19}$$

$$\mathbf{T}^{\text{gs}}, \mathbf{T}^{\text{body}} = \mathcal{A}_{\text{MM-T}}(\mathbf{T}^{\text{gs}}, \mathbf{T}^{\text{body}}; \mathbf{T}^{\text{global}}). \tag{20}$$

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

B.3. Gaussian Decoder

Both the canonical decoder $\mathcal{H}_{\text{cano}}$ and the pose-dependent decoder $\mathcal{H}_{\text{pose}}$ 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 runtime 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, $\mathcal{H}_{\text{pose}}$, 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

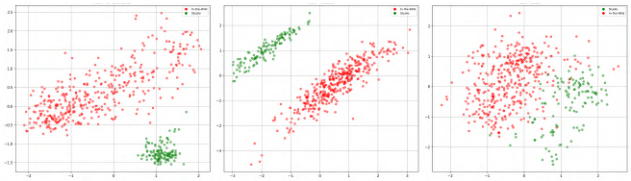


Figure S1. **PCA of Geometric Token Features.** Visualization of the feature space distributions produced by models trained with different strategies. Green points denote studio-captured subjects, while red points denote in-the-wild subjects.

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, γ=0.00 (no decay, all layers trained at the same rate) severely degrades studio metrics, indicating catastrophic forgetting of pretraining knowledge. Both γ=0.30 and γ=0.65 perform well; we use γ=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× 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.

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.

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 γ=0.00 (no decay) severely degrading studio performance.

Train Data Size	Capture-Studio			In-the-Wild		
Num. Identities	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× 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.

F. Comparison with Alternative Paradigms

We qualitatively compare LCA with Wan-Animate [13], a 2D video diffusion method, and GUAVA [73], 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.

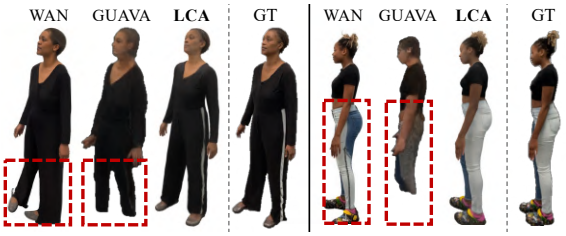


Figure S2. **Qualitative Comparison with Alternative Paradigms.** Comparison with Wan-Animate [13] (2D video diffusion) and GUAVA [73] (upper-body 3D Gaussian avatar).

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 Borsos, 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, Liuhao 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, Shoou-I Yu, and Junchen Zhang.
Evaluation Data: Andrew Hou, Austin James, Fei Jiang, Alex Ma, and Conor O'Hollaren.