GIGA: Generalizable Sparse Image-driven Gaussian Avatars

Anton Zubekhin[1,2] Heming Zhu[1] Paulo Gotardo[3] Thabo Beeler[3]
Marc Habermann[1,2] Christian Theobalt[1,2]
[1] Max Planck Institute for Informatics, Saarland Informatics Campus
[2] Saarbrücken Research Center for Visual Computing, Interaction and AI [3] Google

Abstract

Driving a high-quality and photorealistic full-body human avatar, from only a few RGB cameras, is a challenging problem that has become increasingly relevant with emerging virtual reality technologies. To democratize such technology, a promising solution may be a generalizable method that takes sparse multi-view images of an unseen person and then generates photoreal free-view renderings of such identity. However, the current state of the art is not scalable to very large datasets and, thus, lacks in diversity and photorealism. To address this problem, we propose a novel, generalizable full-body model for rendering photoreal humans in free viewpoint, as driven by sparse multi-view video. For the first time in literature, our model can scale up training to thousands of subjects while maintaining high photorealism. At the core, we introduce a MultiHeadUNet architecture, which takes sparse multi-view images in texture space as input and predicts Gaussian primitives represented as 2D texels on top of a human body mesh. Importantly, we represent sparse-view image information, body shape, and the Gaussian parameters in 2D so that we can design a deep and scalable architecture entirely based on 2D convolutions and attention mechanisms. At test time, our method synthesizes an articulated 3D Gaussian-based avatar from as few as four input views and a tracked body template for unseen identities. Our method excels over prior works by a significant margin in terms of cross-subject generalization capability as well as photorealism.

1. Introduction

Driving your own full-body and photoreal virtual avatar from a scarce and affordable multi-view (i.e. four), camera setups has the potential to revolutionize communication, gaming, and remote collaboration. However, there are to date modeling challenges that remain unsolved: 1) Achieving photorealism and fidelity despite sensor scarcity and limited input data. 2) True generalization to unseen, novel identities. In this work, we attempt to jointly solve them by leveraging recent large-scale data capture efforts. Notably, this requires a generalizable method to synthesize digital humans at test time in a simple feed-forward manner, which is the subject of this work.

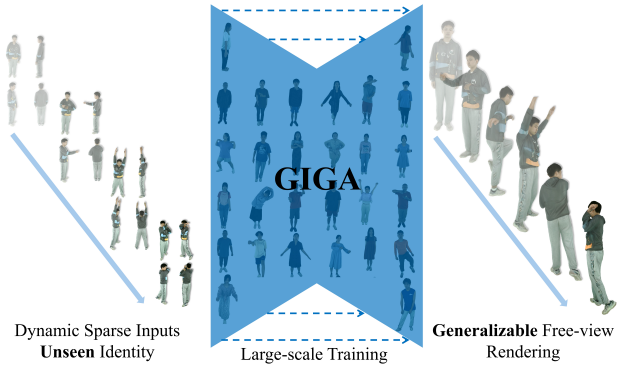


Figure 1. GIGA is trained efficiently on a large-scale human dataset. Given sparse views of an *unseen* identity, and respective skeletal poses, GIGA generates photorealistic dynamic renderings in free viewpoint.

Recent related works have focused on *person-specific* avatars, i.e., a learned representation that is trained per subject on dense dome-like camera setups. These representations may involve meshes [1, 5, 54], neural radiance fields [11, 33, 48], points [45], or volumetric primitives [36] such as 3D Gaussian splats [28, 31, 62]. Democratizing such technology is difficult since, for high-quality results, a dense camera dome is required *before* the character can be driven at inference time. Some methods aim to utilize simpler capture setups, e.g., monocular images [56, 57] or videos [8, 49]. However, due to the much scarcer input, their visual quality often falls short compared to multi-view methods. Single-image to 3D reconstruction methods [10, 40, 41, 56, 57] also primarily focus on adequate geometry reconstruction and produce a 3D asset, drivable with skeletal motion only, which typically lacks the skeletal motion-dependent geometry and appearance changes.

Designing a high-quality approach for lightweight human reconstruction and rendering, thus, remains an open problem: the model must accurately represent diverse human appearances, body types, and clothing configurations and it must correctly derive pose-dependent appearance changes from the scarce input signal. While previous works [15, 16, 30, 51] have shown promising first steps towards identity generalization, those methods were trained and evaluated on small-scale datasets [34] with few training subjects. Moreover, the quality and efficiency of these works are constrained by the limitations of implicit neural representations and non-scalable network architectures. More recently, large-scale datasets [3, 55] helped to pave the way for truly generalizable human avatar methods. However, training and evaluating such models at scale demands particular consideration. First, a generalizable model has to learn meaningful feature representations to scale, both, to the training set and outside of it. Second, the model architecture must be computationally and memory-efficient to capture fine detail and enable large-scale training. Third, at the core of such a method, the 3D representation has to yield high-quality reconstruction with minimal rendering time to maintain fast backpropagation during training and to ensure fast performance at test time.

To address these challenges, we propose GIGA, a feed-forward method to synthesize personalized virtual human avatars from sparse input views and a tracked body template at inference. Notably, for unseen subjects, GIGA *requires no personalized training on dense dome data*. At the core, we project sparse-view image information into the UV space of the SMPL-X model [32] while the digital human predicted by GIGA is represented as a texel-aligned Gaussian avatar [12, 16] greatly simplifying the task to a 2D-to-2D image translation. We propose a MultiHeadUNet, a UNet with multiple encoding and decoding heads, which takes the texture containing the projected image information as well as shape and motion codes and regresses per-texel Gaussian appearance and geometry parameters. In detail, we employ cross attention to inject motion information in the model, and skip-connections to propagate learning signal at different spatial scales inside of the network. Our architecture design choices ensure reliable learning of intrinsic feature statistics from the training data while maintaining person-specific information contained in the model inputs. The predicted Gaussian parameters can be posed into 3D space using the respective SMPL-X body pose, which allows rendering 2D images that can be compared to the ground truth during training. In summary, our contributions are:

- We introduce GIGA, a novel generalizable human avatar reconstruction and rendering approach operating in a 4-view setup in a feed-forward regime at test time.
- We design GIGA as a MultiHeadUNet with motion conditioning relying on attention mechanism to synthesize high-quality 3D Gaussian-based avatars with dynamic appearance changes.
- GIGA for the first time in literature can be effectively trained on thousands of human identities, and shows compelling results in large-scale generalization.

We demonstrate the generalization capabilities of GIGA through a comprehensive evaluation on large-scale datasets MVHumanNet [55], DNA-Rendering [3], and THuman2.0 [59]. Experimental results show that GIGA significantly outperforms prior works in visual quality as well as identity and pose generalization.

2. Related Work

Person-specific Human Capture and Rendering. Neural implicit approaches for novel view synthesis [27, 29, 47] were primarily designed for per-scene optimization. This paradigm naturally extended to human modeling through various implicit representations: radiance fields [11, 34, 44], signed distance functions [48, 52, 63], and occupancy fields [26, 40, 41]. These methods typically rely on parametric human body models [24, 32] as geometric proxies to initialize neural implicit representations. The inherent limitations of generic parametric templates in capturing person-specific geometry led to methods utilizing personalized mesh templates [5, 20, 36], which demonstrate superior reconstruction quality, mimicking pose-dependent appearance changes. Drivable Volumetric Avatars (DVA) [36] target telepresence applications by representing human avatars as mixtures of volumetric primitives [23], regressed from texel-aligned image and pose features extracted from 3 input views. Holoported Characters [42] combines a personalized mesh template with dynamic feature textures predicted from partial texture and normal maps, operating with 4 input views. This method achieves 4K resolution rendering through a superresolution network applied to rendered feature images. 3D Gaussian Splatting [13] marked a significant advancement in, both, rendering quality and computational efficiency. 3D Gaussians have been successfully adapted for free-viewpoint human avatar rendering in multi-view [12, 28, 31, 65] and monocular [7, 8, 35] setups. These approaches remain constrained to person-specific optimization and cannot be easily extended to cross-identity training.

Generalizable Human Capture and Rendering. Recent research has addressed the challenge of synthesizing free-viewpoint videos of human performances from sparse multi-view captures with generalization across subjects [6, 9, 64]. Neural Human Performer (NHP) [15] integrates pixel-aligned visual features with skeletal pose information extracted from multi-view video sequences through cross-attention. However, NHP's performance degrades in regions occluded from input viewpoints due to its di-

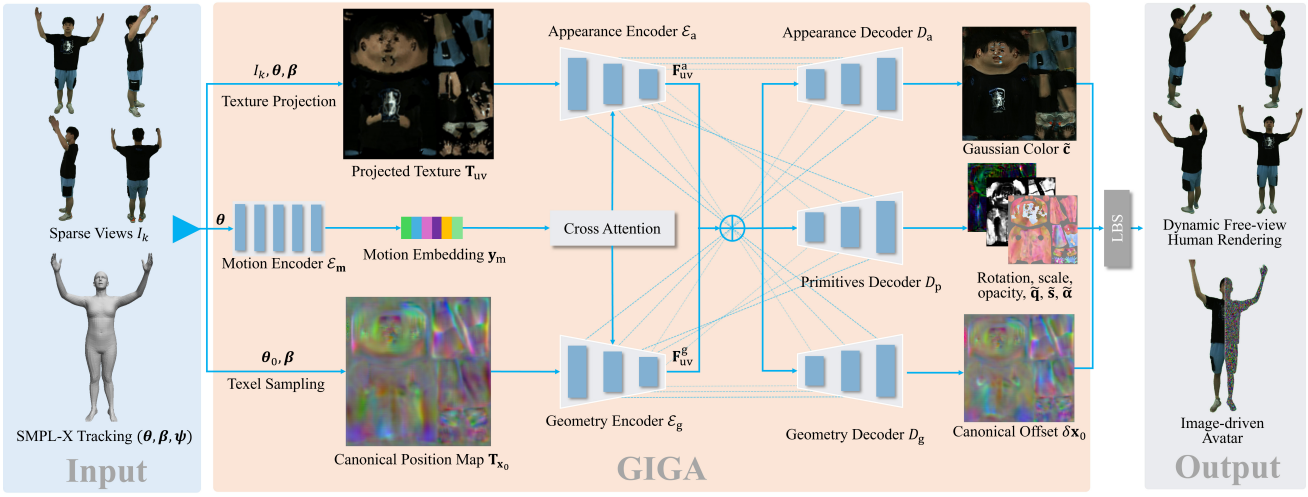


Figure 2. **Method Overview.** GIGA generates dynamic textures of 3D Gaussians for photoreal avatars from 4 input views $\mathcal{I}_k$ and a tracked body template $(\boldsymbol{\theta}, \boldsymbol{\beta})$. An initial RGB texture $\mathbf{T}_{\text{uv}}$ is gathered from the input images and passed to the appearance encoder $\mathcal{E}_{\text{a}}$ to extract appearance features $\mathbf{F}^{\text{a}}_{\text{uv}}$. A canonical position map $\mathbf{T}_{\mathbf{x}_0}$ is processed by the geometry encoder $\mathcal{E}_{\text{g}}$ into geometry features $\mathbf{F}^{\text{g}}_{\text{uv}}$. Both intermediate features are motion-dependent via cross-attention conditioning on the observed pose $\boldsymbol{\theta}$. Gaussian texel maps are regressed by separate decoders, each for an individual group of parameters. Skip-connections (dashed lines) propagate intermediate features from encoders to decoders. Output of GIGA is articulated with linear blend skinning.

rect operation in the observed pose space. TransHuman [30] addresses this limitation by learning to synthesize a NeRF-based representation in the canonical pose space, employing a transformer architecture that processes individual body parts as tokens. While NHP utilizes cross-attention to fuse temporal features first and frame-specific multi-view features next, TransHuman constructs an implicit representation through N-nearest neighbors' token aggregation. Neural Novel Actor (NNA) [51] disentangles appearance and geometry by processing spatial point features and SMPL surface features through a graph CNN, enabling separate prediction of person-specific appearance and pose-dependent deformation, as proposed by Liu et al. [21].

The identity generalization capabilities of these methods remain limited due to training on datasets with at most 10 subjects. Additionally, their reliance on implicit representations and network architecture design imposes significant computational costs during rendering, further increasing training time on larger datasets. In contrast, Generalizable Human Gaussians (GHG) [16] introduces an alternative approach for reconstructing static human models from sparse input views. By leveraging a large-scale dataset of textured human 3D scans [59], GHG demonstrates generalization to unseen subjects while achieving near real-time performance by representing avatars as multiple scaffolds of 3D Gaussians in the UV space. However, since GHG is supervised with ground-truth 3D data, its training cannot be easily extended to multi-view datasets.

3. Method

GIGA aims at mapping sparse image observations (4 views) of an *unseen* human to a 3D Gaussian-based avatar that can be rendered from a free viewpoint. This is achieved while relying only on an estimated parametric body model that captures the rough shape and pose of the person. Importantly, our method requires *no* personalized training.

This section first formalizes the problem setting and relevant background knowledge (Sec. 3.1). It then introduces our generalizable human representation (Sec. 3.2), our training strategy (Sec. 3.3), and implementation details (Sec. 3.4). Fig. 2 provides an overview.

3.1. Problem Setting and Background

Problem Setting. GIGA assumes a collection of multi-view videos of several subjects with per-frame subject segmentation masks for training. Each subject is captured by $\hat{K} = 16$ calibrated cameras, with $\pi_{\hat{k}}$ denoting the projection matrix for camera $\hat{k}$. Each video frame is annotated with SMPL-X [32] parameters released as additional labels in most multi-view human performance capture datasets [3, 55, 59]. At test time, GIGA takes as input sparse-view videos as driving signals and regresses 3D Gaussian splats that faithfully preserve the appearance and clothing dynamics from these driving videos, even for unseen identities. For training, we select $K = 4$ sparse camera views and use native SMPL-X parameters to construct the input for the encoder network. The images from the remaining $K' = 12$ views serve as ground truth to supervise the outputs de-

coded by the network.

3D Gaussian Splatting (3DGS). 3DGS [13] models a scene using 3D Gaussian primitives. Each primitive $\mathcal{G} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \mathbf{c}\}$ is paramterized by its position $\boldsymbol{\mu} \in \mathbb{R}^3$, covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times3}$, opacity $\alpha \in \mathbb{R}$, and RGB colors $\mathbf{c} \in \mathbb{R}^3$. The covariance matrix $\boldsymbol{\Sigma}$ is factorized as $\boldsymbol{\Sigma} = \mathbf{RSS}^T\mathbf{R}^T$, where the rotation $\mathbf{R}$ matrix is obtained from the quaternion $\mathbf{q} \in \mathbb{R}^4$ and the diagonal scaling matrix $\mathbf{S} = \mathrm{diag}\,(\mathbf{s})$, with per-axis scales $\mathbf{s} \in \mathbb{R}^3$. The Gaussians $\mathcal{G}$ is then rendered from the target camera π to an image I and an accumulated density image A using a Gaussian rasterizer

$$I, A = \mathcal{R}(\mathcal{G}, \pi). \tag{1}$$

SMPL-X Human Body Template. SMPL-X [32] is a parametric human body model with articulated limbs, hands, and an expressive face. SMPL-X has $N = 10475$ vertices, $J = 54$ joints, and can be articulated using linear blend skinning [19] to obtain a set of posed vertices $\mathcal{V} \in \mathbb{R}^{N\times3}$. Formally, SMPL-X is defined as a function $\mathcal{V} = \mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\psi})$, with $\boldsymbol{\theta} \in \mathbb{R}^{(J+1)\times3}$ describing J joint angles and a rigid root transformation, body shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$, and face expression $\boldsymbol{\psi} \in \mathbb{R}^{10}$. As we tested GIGA without expression tracking, we fixed $\boldsymbol{\psi}$ to the neutral expression.

3.2. Generalizable Human Representation

The original 3D Gaussian representation is a point cloud where each primitive is essentially independent from the others. A naive approach would optimize separate sets of 3D Gaussians for each training identity. Within such a set, each Gaussian is optimized almost independently from every other, leading to an absence of shared statistical information between neighboring Gaussians. Moreover, it would be hard to establish semantic correspondence between different sets of Gaussians.

Inspired by previous *personalized*, animatable Gaussian avatar methods [12, 31, 50], we represent virtual humans as texel-aligned 3D Gaussian maps within the texture space $\mathcal{M}_{\mathrm{uv}}$ of the SMPL-X template mesh. Anchoring 3D Gaussians to the texture space $\mathcal{M}_{\mathrm{uv}}$ solves the aforementioned problems: texel positions are fixed on the template surface, making 3D Gaussians assigned to same texels semantically similar across various characters; 2D textures can be efficiently processed by convolutional neural networks; and Gaussians from the same region of the texture space will be entangled thanks to the locality bias of convolutions. As such, GIGA can learn cross-subject information and predict the final texel-aligned Gaussian avatar with a single pass of a 2D convolutional neural network.

Input Appearance Encoding. We use information from the input views to capture identity-specific appearance with pose-dependent variations. To transform this information from image to texture space, we first compute partial textures $\mathbf{T}_{\mathrm{uv},k}$ for each view π_k by projecting each pixel onto the 3D surface of SMPL-X. This 3D point can be mapped into texture space using the UV mapping of SMPL-X such that the final image pixel color is projected onto the respective 2D location in the texture map. We then fuse those textures according to visibility resulting in the final texture $\mathbf{T}_{\mathrm{uv}} \in \mathbb{R}^{T\times T\times3}, T = 512$ in the UV space $\mathcal{M}_{\mathrm{uv}}$.

This texture is the input to our appearance encoder $\mathcal{E}_{\mathrm{a}}$:

$$\mathbf{F}^{\mathrm{a}}_{\mathrm{uv}}, \mathbf{H}^{\mathrm{a}} = \mathcal{E}_{\mathrm{a}}\left(\mathbf{T}_{\mathrm{uv}}; \mathbf{y}_m\right), \tag{2}$$

which extracts appearance features $\mathbf{F}^{a}_{\mathrm{uv}} \in \mathbb{R}^{T_f\times T_f\times d}$, that encode character-specific appearance and identity information. The encoder $\mathcal{E}_{\mathrm{a}}$ consist of 2D convolutional downsampling residual blocks. Feature maps $\mathbf{H}^{\mathrm{a}}$ from also taken from the downsampling levels for later usage in decoding.

$\mathcal{E}_{\mathrm{a}}$ is conditioned on the motion embedding $\mathbf{y}_{\mathrm{m}}$. Following Rombach et al. [37], this conditioning is implemented using cross-attention layers as final layers of the encoder.

Motion Embeddings. The input texture $\mathbf{T}_{\mathrm{uv}}$ contains only a localized pose-dependent learning signal from the template surface. We additionally use an MLP-based motion encoder $\mathcal{E}_{\mathrm{m}}$ to construct motion embeddings for SMPL-X poses $\boldsymbol{\theta}$, which adds global body motion awareness at the encoding stage:

$$\mathbf{y}_{\mathrm{m}} = \mathcal{E}_{\mathrm{m}}\left(\boldsymbol{\theta}\right). \tag{3}$$

Input Geometry Encoding. Even though appearance information is texel-aligned, it is insufficient to infer correct human shapes. To address this, we employ a geometry encoder $\mathcal{E}_{\mathrm{g}}$ to extract approximate geometry information from the body template. Starting with a T-posed SMPL-X mesh $\mathcal{V}(\boldsymbol{\theta}_0, \boldsymbol{\beta})$, we compute a canonical position map $\mathbf{T}_{\mathbf{x}_0} \in \mathbb{R}^{T\times T\times3}$ from canonical vertex coordinates $\mathbf{x}_0\left(\boldsymbol{\beta}\right) \in \mathbb{R}^{N\times3}$ using the UV map $\mathcal{M}_{\mathrm{uv}}$. The geometry encoder $\mathcal{E}_{\mathrm{g}}$ has the same architecture as the appearance encoder $\mathcal{E}_{\mathrm{a}}$ and produces geometry features $\mathbf{F}^{\mathrm{g}}_{\mathrm{uv}} \in \mathbb{R}^{T_f\times T_f\times d}$ with a stack of feature maps $\mathbf{H}^{\mathrm{g}}$:

$$\mathbf{F}^{\mathrm{g}}_{\mathrm{uv}}, \mathbf{H}^{\mathrm{g}} = \mathcal{E}_{\mathrm{g}}\left(\mathbf{T}_{\mathbf{x}_0}; \mathbf{y}_{\mathrm{m}}\right). \tag{4}$$

To handle dynamically changing details in the final shape, the geometry encoder $\mathcal{E}_{\mathrm{g}}$ is also conditioned on the motion embedding $\mathbf{y}_{\mathrm{m}}$.

Gaussian Primitives Regression. For the decoding stage, we design three separate decoders (Fig. 2): $\mathcal{D}_{\mathrm{a}}$ for the appearance, $\mathcal{D}_{\mathrm{p}}$ for Gaussian scales, quaternions and opacities, and $\mathcal{D}_{\mathrm{g}}$ for Gaussian offsets. All three share the same convolutional architecture and receive appearance and geometry features $\mathbf{F}^{\mathrm{a}}_{\mathrm{uv}}, \mathbf{F}^{\mathrm{g}}_{\mathrm{uv}}$ as inputs:

$$\mathcal{G}'_{\mathrm{uv}} = \mathcal{D}_{\{\mathrm{a}|\mathrm{p}|\mathrm{g}\}}\left([\mathbf{F}^{\mathrm{a}}_{\mathrm{uv}}, \mathbf{F}^{\mathrm{g}}_{\mathrm{uv}}]; [\mathbf{H}^{\mathrm{a}}, \mathbf{H}^{\mathrm{g}}]\right). \tag{5}$$

The output map $\mathcal{G}'_{\mathrm{uv}}$ contains RGB color channels $\mathbf{c}$, quaternions $\mathbf{q}_0$, normalized scales $\mathbf{s}'$, opacities α, and offsets $\delta\mathbf{x}_0$, which are defined in the canonical T-pose. To utilize

Figure 3. **Qualitative Comparison against Dynamic Methods.** We show results for identity generalization. GIGA (Ours) achieves significantly higher quality, while also being able to faithfully synthesize a virtual avatar after training on a large-scale dataset [55].

the representational power of the shared texel space and to propagate semantic information from encoders to decoders at different spatial scales, we use UNet-like [39] skip-connections, building a feature propagation bridge from each encoder to each decoder. We stack intermediate feature maps $\mathbf{H}^{\mathrm{a}}$ and $\mathbf{H}^{\mathrm{g}}$ along the feature dimension and feed them to every decoder at each corresponding upsampling layer. To convert a texel-aligned Gaussian map to the observed pose space $\boldsymbol{\theta}$, we use linear blend skinning applied to both offsets $\delta\mathbf{x}_0$ and quaternions $\mathbf{q}_0$: $\delta\mathbf{x} = \mathrm{lbs}\left(\delta\mathbf{x}_0, \boldsymbol{\theta}\right); \mathbf{q} = \mathrm{lbs}\left(\mathbf{q}_0, \boldsymbol{\theta}\right)$. We also multiply normalized Gaussian scales $\mathbf{s}'$ by a hyperparameter ρ: $\mathbf{s} = \rho\mathbf{s}'$, value $\rho = 5e^{-3}$ was empirically found adequate to provide sufficient coverage of the template with Gaussians.

The resulting 3D Gaussian map $\mathcal{G}_{\mathrm{uv}}$ is rendered for each camera view π_k using differentiable Gaussian rasterizer [58], yielding an RGB and accumulated opacity image pair, $I_k, A_k = \mathcal{R}\left(\mathcal{G}_{\mathrm{uv}}, \pi_k\right)$, as in Eq. (1).

3.3. Training

GIGA generates Gaussian texture maps at 512×512 resolution. As some texels in these textures are not associated with any triangle, they are discarded during rendering.

Training Objective. GIGA is trained to minimize a combination of the reconstruction $\mathcal{L}_{\mathrm{rec}}$ and geometry-related regularization terms $\mathcal{L}_{\mathrm{reg}}$:

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \mathcal{L}_{\mathrm{reg}}\,, \tag{6}$$

Reconstruction Term. Following Kerbl et al. [13], we compute the mean absolute error $\mathcal{L}_{\mathrm{L1}}$ and the structural similarity measure $\mathcal{L}_{\mathrm{ssim}}$ [53] between the rendered image I_k and the ground-truth image $I_{\mathrm{gt},k}$. We additionally evaluate the VGG-based [43] perceptual loss [61] $\mathcal{L}_{\mathrm{LPIPS}}$ on randomly sampled patches with centers within the ground truth segmentation mask $A_{\mathrm{gt},k}$. To ensure scale-invariance of the perceptual features, we sample 16 patches of varying sizes: 128×128, 256×256, and 512×512; then resize all patches to 256×256 before computing the loss, following Cao et al. [2]. To improve outlines of Gaussian primitives, we also compute $\mathcal{L}_{\mathrm{mask}}$, the mean squared error between the rendered opacity images A_k and ground truth character segmentation masks $A_{\mathrm{gt},k}$. The final reconstruction term is averaged over all images in the batch and defined as:

$$\mathcal{L}_{\mathrm{rec}} = \lambda_1\mathcal{L}_{\mathrm{L1}} + \lambda_2\mathcal{L}_{\mathrm{ssim}} + \lambda_3\mathcal{L}_{\mathrm{LPIPS}} + \lambda_4\mathcal{L}_{\mathrm{mask}}\,, \tag{7}$$

with $\lambda_1 = \lambda_2 = \lambda_3 = 0.5$ and $\lambda_4 = 0.1$ in all experiments.
Regularization Term. Modeling various dynamically changing geometrical shapes with Gaussian offsets $\delta\mathbf{x}_0$ is particularly challenging during the early stages of optimization. To this end, we introduce additional penalties to con-

Figure 4. **Qualitative Comparison against Static Method.** While sharing similar concepts with GHG [16], GIGA (Ours) outperforms it on THuman2.0 [59] dataset thanks to our design choices. GHG also cannot be straightforwardly applied to dynamic sequences, which is not the case for GIGA.

strain some of the predicted Gaussian parameters:

$$\begin{aligned}\mathcal{L}_{\text{reg}} &= \lambda_5 \mathcal{L}_{\delta \mathbf{x}_0} + \lambda_6 \left(\mathcal{L}_{\delta \mathbf{x}_0;\alpha} + \mathcal{L}_{\mathbf{s}';\alpha} \right) + \lambda_7 \mathcal{L}_{\alpha} \; = \\ &= \lambda_5 \left\| \delta \mathbf{x}_0 \right\|_2 + \lambda_6 \left(\left\| \mathbb{1}_{[\alpha<\epsilon]} \mathbf{x}_0 \right\|_2 + \left\| \mathbb{1}_{[\alpha<\varepsilon]} \mathbf{s}' \right\|_2 \right) + \\ &+ \lambda_7 \text{Beta}\left(\alpha\right) \; , \end{aligned} \tag{8}$$

with $\lambda_5 = 0.15, \lambda_6 = \lambda_7 = 0.1$. The offset penalty $\mathcal{L}_{\delta \mathbf{x}_0}$ prevents Gaussians from drifting far away from the mesh.

Due to UV mapping distortions, some mesh regions may be oversampled with Gaussians. The poor initial model state will encourage some Gaussians to become transparent and inadequately large and move away from the mesh. To reclaim those Gaussians for actual modeling, we additionally penalize offsets of low-opacity primitives with $\mathcal{L}_{\delta \mathbf{x}_0;\alpha}$, where the indicator function $\mathbb{1}_{[\alpha<\epsilon]}$ triggers the extra penalty only when Gaussian opacity α is below a threshold ϵ. Importantly, we disable gradient flow through the indicator function. We similarly penalize scales $\mathbf{s}$ when opacity is below ε with $\mathcal{L}_{\mathbf{s}';\alpha}$. Following Lombardi et al. [22], we encourage all Gaussians to be either completely opaque or completely transparent by computing negative log-likelihood of the beta distribution $\text{Beta}\left(0.5, 0.5\right)$ as $\mathcal{L}_{\alpha}$.

3.4. Implementation Details

We train GIGA for 500k iterations on a single NVIDIA H100 GPU using 1 subject per batch. All models are optimized with AdamW [25], with the learning rate being linearly increased from 0 to $1e^{-4}$ during the first 25k training steps. We also set the weight decay parameter to $1e^{-4}$.

At the warmup stage, we also apply linear annealing to the predicted offsets $\delta \mathbf{x}_0$ in addition to other penalties, Eq.(8). We empirically find it crucial, as it allows the model to focus on approximate appearance reconstruction before refining person-specific geometry, effectively preventing unstable offset predictions.

We render 4 random views for each input subject in the dataset image resolution to compute the training objective.

Dataset	MVHumanNet			THuman2.0	
Metric / Method	PSNR	SSIM	LPIPS	PSNR	LPIPS
NHP	17.72	0.5678	402.7	—	—
TH	17.46	0.5808	392.1	—	—
GHG	—	—	—	21.90	133.4
GIGA	**22.19**	**0.7526**	**70.2**	**24.70**	**51.6**

Table 1. **Quantitative Comparison.** We evaluate generalization to unseen identities of Neural Human Performer (NHP) and TransHuman (TH) on MVHumanNet, and Generalizable Human Gaussians (GHG) on THuman2.0.

Configuration	↑ PSNR	↑ SSIM	↓ LPIPS
I, Enc-Dec	21.93	0.7450	75.4
II, UNet	21.94	0.7449	73.4
III, 2Enc-3Dec	21.88	0.7441	77.1
IV (**GIGA**) MultuHeadUNet	**22.19**	**0.7526**	**70.2**
IVa, w/o CrossAttn	21.69	0.7377	76.5
IVb, w/o offset annealing	—-	—-	—-
IVc, w/o offset penalty	21.73	0.7454	71.0
IVd, w/o opacity penalties	21.78	0.7465	71.1

Table 2. **Ablation Results for GIGA.** We examine our design choices, proposed in Sec. 3, see detailed analysis in Sec. 4.4.

4. Results

We first explain the datasets and metrics (Sec. 4.1). Then, we show qualitative results (Sec. 4.2) and comparisons (Sec. 4.3). Lastly, we ablate our design choices (Sec. 4.4).

4.1. Dataset and Metrics

Datasets. To evaluate GIGA, we use the multi-view human performance capture datasets MVHumanNet [55] and DNA-Rendering [3], and also static 3D human scans dataset from THuman2.0 [59]. We split MVHumanNet as follows: IDs $100000 - 100999$ are used for training and IDs $101000 - 102386$ are held out for validation and testing. From the testing part, we choose 40 subjects for quantitative evaluation. For experiments with DNA-Rendering, we use the released Part-2 for training (400 subjects) and Part-1 for validation (40 subjects). In both cases, we select 12 cameras for supervision by picking every 4th camera from the

	Number of training characters			
Metric	100	250	500	970
↑ PSNR	21.02	21.37	21.18	22.19
↑ SSIM	0.7165	0.7291	0.7321	0.7526
↓ LPIPS	86.9	82.7	80.1	70.2

Table 3. **Effect of the Dataset Size.** Increasing number of subjects in the training dataset leads to a consistent improvement in terms of generalization to novel identities.

	Training set → Testing set	
Metric	DNA → DNA	MVH → DNA
↑ PSNR	19.63	19.49
↑ SSIM	0.7255	0.7138
↓ LPIPS	106.4	114.9

Table 4. **Generalization to Novel Data.** Training on MVHumanNet allows GIGA to generalize to DNA-Rendering with minor quantitaive performance drop in comparison with GIGA trained on DNA-Rendering.

provided camera calibration. From each subject sequence, we sample every 20th pose for training, but not more than 15 poses per character. We follow the GHG [16] training and validation split of THuman2.0 and render multi-view images from original 3D scans similarly to GHG.

Metrics. To evaluate rendering quality, we compute the following metrics: PSNR, SSIM [53], and the perceptual LPIPS [61] using AlexNet [14] features (scaled by 1000).

4.2. Qualitative Results

Fig. 3, 4, and 6 (and the supplementary video) present GIGA's novel view rendering results of unseen subjects performing unseen poses. Notably, GIGA achieves photorealistic and view-consistent rendering and effectively captures fine details such as clothing wrinkles and intricate textures. We highlight that for novel subjects, GIGA achieves a rendering quality comparable to that of the training subjects, demonstrating its generalization ability to novel identities.

4.3. Comparison

Competing Methods. We compare GIGA to other generalizable dynamic image-driven methods discussed in Sec. 2: NHP [15] and TransHuman (TH) [30]. We exclude NNA [51] from comparisons, as we could not reproduce the training results from the code provided by the authors. We train NHP and TH using the same training/validation split of MVHumanNet and follow the same training setup as for our method. We also adopt GHG [16] as a baseline and train GIGA on the THuman2.0 dataset [59].

Baselines. Tab. 1 provides quantitative comparisons of GIGA against state-of-the-art baselines for generalizable human rendering. In all cases, we evaluate the generalization ability to unseen identities from held-out validation sequences. Fig. 3 provides an overview of qualitative results. NHP relies on sparse 3D convolutions to process volumetric features in the observed pose space, thus, suffering from missing input signals due to occlusions and failing to generalize to unseen identities. Despite operating in a canonical template pose space and tokenizing the template for processing with a transformer-based network, TransHuman also cannot learn meaningful priors from the large data collection. Both NHP and TransHuman are extremely slow due to implicit representations at their cores, which significantly limits their generalization abilities. GIGA, on the other hand, utilizes the power of the shared texel space to a maximum degree: all feature representations for digital humans are defined in the same texel space, and intermediate features enhance the quality of the final prediction through skip-connections. MultiHeadUNet is also significantly more computationally efficient, which explains, both, qualitative and quantitative improvements over baselines.

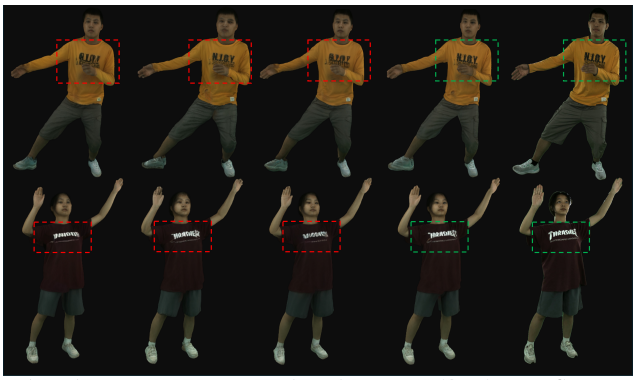
While being much more efficient than previous works, GHG targets only static reconstruction from sparse input views. GHG also models humans as a set of 3D Gaussian scaffolds in the observed pose space and cannot be easily extended to dynamic scenarios. GHG handles Gaussian color prediction by pretraining a separate texture inpainting network. GIGA learns to operate with appearance and geometry features simultaneously, yielding results of higher quality, both, visually (Fig. 4) and quantitatively (Tab. 1).

Cross-Dataset Validation. To demonstrate cross dataset generalization we also train two variants of GIGA where one is trained and tested on the DNA-Rendering dataset while the other variant is trained on MVHumanNet and tested on the DNA-Rendering. We follow train/test splits specified in Sec. 4.1. Tab. 4 and Fig. 6 demonstrate that our method is capable to generalize across datasets as the model trained on the large-scale MVHumanNet data performs comparable on DNA-Rendering, which clearly shows that our method effectively learns the prior from large datasets.

4.4. Ablation Study

Model Architecture. To benchmark MultiHeadUNet (**IV**) at the core of GIGA, we propose 3 alternative architectures with approximately the same number of trainable parameters ($\simeq$90M): a simple encoder-decoder model (**I**), a conventional UNet with skip-connections between corresponding up- and downsampling blocks of the encoder and decoder (**II**), and a model with 2 encoders $\mathcal{E}_{\mathrm{a}}, \mathcal{E}_{\mathrm{g}}$ and 3 decoders $\mathcal{D}_{\mathrm{a}}, \mathcal{D}_{\mathrm{p}}, \mathcal{D}_{\mathrm{g}}$, but without skip-connections (**III**). We observe, both, quantitatively (Tab. 2) and qualitatively (Fig. 5) that the configuration **IV** leads to a higher quality overall, being particularly helpful at preservation of fine appearance details, observed in the input signal.

Motion Conditioning. Configurations **I-IV** use an MLP-based motion encoder $\mathcal{E}_{\mathrm{m}}$ and cross-attention motion conditioning by default. We additionally remove cross atten-



I, 1Enc-1Dec **II**, UNet **III**, 2Enc-3Dec **IV (Ours)** GT

Figure 5. **Impact of Different GIGA Network Architectures.** While all proposed GIGA configurations produce reasonable results at coarse level, multiple encoding and decoding heads with additional skip-connections (**IV**, MultiHeadUNet) manages fine-grained appearance better.

Figure 6. **Cross-Dataset Examples.** After training on MVHumanNet, GIGA successfully generalizes to novel subjects from DNA-Rendering (MVH→DNA) and performs similarly to GIGA, specifically trained on DNA-Rendering (DNA→DNA).

tion blocks from **IV**, leaving intermediate features $\mathbf{F}^{\mathrm{a}}_{\mathrm{uv}}, \mathbf{F}^{\mathrm{g}}_{\mathrm{uv}}$ without motion conditioning (**IVa**). As it is evident from the quantitative evaluation (2), enabling motion conditioning via cross-attention improves the visual quality of GIGA's Gaussian avatars.

Regularizations. Additional geometry penalties also contribute to the quality of GIGA. Removing opacity-related penalties $\mathcal{L}_{\alpha}, \mathcal{L}_{\delta\mathbf{x}_0;\alpha}, \mathcal{L}_{\mathbf{s};\alpha}$ (**IVd**), and then the offset penalty $\mathcal{L}_{\delta\mathbf{x}_0}$ (**IVc**) from the complete training objective (7) leads to quality degradation (Tab. 2), as some of the Gaussians become underconstrained. If no offset annealing is used during the warmup stage (**IVb**), GIGA fails to converge.

Scaling Dataset Size. GIGA also benefits from training on large data collections, as reported in Tab. 3 (please see the supplemental materials for the qualitative comparisons). While scaling a digital human reconstruction model to hundreds of identities presents a challenge as evidenced by the fact that prior works failed to train on such large training corpora, it leads to an improved generalization overall.

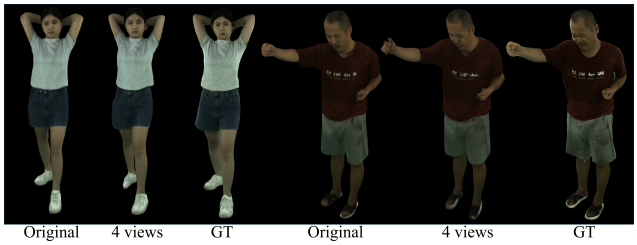


Figure 7. **Ablation on Sparsity of Tracking Cameras.** SMPL-X in MVHumanNet is estimated using tracking from 48 cameras. Retracking motion sequences from only 4 views can still yield reasonable SMPL-X templates for GIGA.

Tracking from Sparse Views. Original skeleton tracking and SMPL-X parameters in MVHumanNet are obtained from a dense set of views. To mimic a real-world scenario, we re-track a small number of sequences from a set of 4 views and fit SMPL-X models to estimated 3D keypoints. Qualitative results of GIGA with sparsely tracked SMPL-X template are shown in Fig. 7. If sparse tracking is mostly correctly aligned with the actual human actor, then GIGA can produce results, similar to dense view tracking case. For real-world use cases, a suitable sparse tracking solution should be applied, which is beyond the scope of this paper.

5. Limitations

GIGA shows unprecedented scalability that enables training on thousands of multi-view videos, thanks to our efficient representation and highly scalable architecture, and respective generalization without sacrificing rendering quality. However, it still faces some limitations that should be addressed in the future. While using SMPL-X as body template greatly facilitates generalization, it does not allow us to handle non-rigid dynamics (e.g., hair and loose clothing) properly without additional assumptions or physics-based priors. Here, a more advanced human shape prior that includes clothing geometry might alleviate some of these limitations [4, 38, 60]. Moreover, the dependency on parametric body models and motion tracking leads to quality degradation in case of template misalignment or inaccurate tracking. Future work could explore end-to-end optimization of body shape and pose parameters, which has already proven to be successful for face-only rendering approaches [46].

6. Conclusion

This work presented GIGA, a generalizable sparse image-drivable Gaussian avatar. Trained on a large-scale

multi-view dataset, GIGA synthesizes texel-aligned 3D Gaussian avatars from sparse input views in a feed-forward manner. Our approach achieves state-of-the-art generalization to unseen identities while preserving person-specific pose-dependent appearance changes thanks to our scalable architecture and efficient representation. We believe our proposed model could benefit future research in this domain and take another step towards enabling more accessible and immersive remote collaboration.

References

[1] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabián Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM Trans. Graph.*, 40(4), 2021. 1

[2] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason Saragih. Authentic volumetric avatars from a phone scan. *ACM Trans. Graph.*, 41(4), 2022. 5

[3] Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, Daxuan Ren, Lei Yang, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, Bo Dai, and Kwan-Yee Lin. DNA-Rendering: A Diverse Neural Actor Repository for High-Fidelity Human-centric Rendering. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19925–19936, 2023. 2, 3, 6

[4] Artur Grigorev, Bernhard Thomaszewski, Michael J. Black, and Otmar Hilliges. HOOD: Hierarchical graphs for generalized modelling of clothing dynamics. 2023. 8

[5] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Trans. Graph.*, 40(4), 2021. 1, 2

[6] 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. 2

[7] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. GaussianAvatar: Towards Realistic Human Avatar Modeling from a Single Video via Animatable 3D Gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 634–644, 2024. 2

[8] Shoukang Hu, Tao Hu, and Ziwei Liu. GauHuman: Articulated Gaussian Splatting from Monocular Human Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20418–20431, 2024. 1, 2

[9] Jiancheng Huang, Mingfu Yan, Songyan Chen, Yi Huang, and Shifeng Chen. Magicfight: Personalized martial arts combat video generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10833–10842, 2024. 2

[10] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-Guided Reconstruction of Lifelike Clothed Humans. In *2024 International Conference on 3D Vision (3DV)*, pages 1531–1542, 2024. 1

[11] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 1, 2

[12] Yujiao Jiang, Qingmin Liao, Xiaoyu Li, Li Ma, Qi Zhang, Chaopeng Zhang, Zongqing Lu, and Ying Shan. UV Gaussians: Joint Learning of Mesh Deformation and Gaussian Textures for Human Avatar Modeling. *arXiv preprint arXiv:2403.11589*, 2024. 2, 4

[13] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4), 2023. 2, 4, 5

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 2017. 7

[15] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 7, 12

[16] Youngjoong Kwon, Baole Fang, Yixing Lu, Haoye Dong, Cheng Zhang, Francisco Vicente Carrasco, Albert Mosella-Montoro, Jianjin Xu, Shingo Takagi, Daeil Kim, Aayush Prakash, and Fernando De la Torre. Generalizable Human Gaussians for Sparse View Synthesis. *European Conference on Computer Vision*, 2024. 2, 3, 6, 7, 13

[17] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular Primitives for High-Performance Differentiable Rendering. *ACM Transactions on Graphics*, 39(6), 2020. 12

[18] Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. xFormers: A modular and hackable Transformer modelling library. `https://github.com/facebookresearch/xformers`, 2022. 12

[19] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 811–818. 2023. 4

[20] Lingjie Liu, Weipeng Xu, Marc Habermann, Michael Zollhöfer, Florian Bernard, Hyeongwoo Kim, Wenping Wang, and Christian Theobalt. Neural Human Video Rendering by Learning Dynamic Textures and Rendering-to-Video Translation. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 2020. 2

[21] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural Actor: Neural Free-view Synthesis of Human Actors with Pose Control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021. 3

[22] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4), 2019. 6

[23] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 2

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2

[25] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2017. 6

[26] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10461–10471, 2021. 2

[27] 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. In *ECCV*, 2020. 2

[28] Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human Gaussian Splatting: Real-time Rendering of Animatable Avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 788–798, 2024. 1, 2

[29] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 2

[30] Xiao Pan, Zongxin Yang, Jianxin Ma, Chang Zhou, and Yi Yang. TransHuman: A Transformer-based Human Representation for Generalizable Neural Human Rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3544–3555, 2023. 2, 3, 7, 12

[31] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. ASH: Animatable Gaussian Splats for Efficient and Photoreal Human Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1165–1175, 2024. 1, 2, 4

[32] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 4

[33] 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 *ICCV*, 2021. 1

[34] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *CVPR*, 2021. 2

[35] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5020–5030, 2024. 2

[36] Edoardo Remelli, Timur Bagautdinov, Shunsuke Saito, Chenglei Wu, Tomas Simon, Shih-En Wei, Kaiwen Guo, Zhe Cao, Fabian Prada, Jason Saragih, and Yaser Sheikh. Drivable Volumetric Avatars using Texel-Aligned Features. In *ACM SIGGRAPH 2022 Conference Proceedings*, New York, NY, USA, 2022. Association for Computing Machinery. 1, 2

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022. 4

[38] Boxiang Rong, Artur Grigorev, Wenbo Wang, Michael J. Black, Bernhard Thomaszewski, Christina Tsalicoglou, and Otmar Hilliges. Gaussian Garments: Reconstructing Simulation-Ready Clothing with Photorealistic Appearance from Multi-View Video, 2024. 8

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 5

[40] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 1, 2

[41] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2

[42] Ashwath Shetty, Marc Habermann, Guoxing Sun, Diogo Luvizon, Vladislav Golyanik, and Christian Theobalt. Holoported Characters: Real-time Free-viewpoint Rendering of Humans from Sparse RGB Cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1206–1215, 2024. 2

[43] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014. 5

[44] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in neural information processing systems*, 34:12278–12291, 2021. 2

[45] Shih-Yang Su, Timur Bagautdinov, and Helge Rhodin. NPC: Neural Point Characters from Video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14795–14805, 2023. 1

[46] Kartik Teotia, Hyeongwoo Kim, Pablo Garrido, Marc Habermann, Mohamed Elgharib, and Christian Theobalt. Gaus-

sianHeads: End-to-End Learning of Drivable Gaussian Head Avatars from Coarse-to-fine Representations. *ACM Trans. Graph.*, 43(6), 2024. 8

[47] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. In *Advances in Neural Information Processing Systems*, pages 27171–27183. Curran Associates, Inc., 2021. 2

[48] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. ARAH: Animatable Volume Rendering of Articulated Human SDFs. In *European Conference on Computer Vision*, 2022. 1, 2

[49] Shaofei Wang, Bozidar Antic, Andreas Geiger, and Siyu Tang. IntrinsicAvatar: Physically Based Inverse Rendering of Dynamic Humans from Monocular Videos via Explicit Ray Tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1877–1888, 2024. 1

[50] Shaofei Wang, Tomas Simon, Igor Santesteban, Timur Bagautdinov, Junxuan Li, Vasu Agrawal, Fabian Prada, Shoou-I Yu, Pace Nalbone, Matt Gramlich, Roman Lubachersky, Chenglei Wu, Javier Romero, Jason Saragih, Michael Zollhoefer, Andreas Geiger, Siyu Tang, and Shunsuke Saito. Relightable full-body gaussian codec avatars. *arXiv.org*, 2501.14726, 2025. 4

[51] Yiming Wang, Qingzhe Gao, Libin Liu, Lingjie Liu, Christian Theobalt, and Bao Xin Chen. Neural Novel Actor: Learning a Generalized Animatable Neural Representation for Human Actors. *IEEE Transactions on Visualization and Computer Graphics*, 30:5719–5732, 2022. 2, 3, 7

[52] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3295–3306, 2023. 2

[53] Zhou Wang, Alan Bovik, Hamid Sheikh, and Eero Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *Image Processing, IEEE Transactions on*, 13:600 – 612, 2004. 5, 7

[54] Donglai Xiang, Timur Bagautdinov, Tuur Stuyck, Fabian Prada, Javier Romero, Weipeng Xu, Shunsuke Saito, Jingfan Guo, Breannan Smith, Takaaki Shiratori, et al. Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *ACM Transactions on Graphics (TOG)*, 41 (6):1–15, 2022. 1

[55] Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, et al. MVHumanNet: A Large-scale Dataset of Multi-view Daily Dressing Human Captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19801–19811, 2024. 2, 3, 5, 6, 13

[56] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, 2022. 1

[57] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[58] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An Open-Source Library for Gaussian Splatting. *arXiv preprint arXiv:2409.06765*, 2024. 5

[59] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, 2021. 2, 3, 6, 7

[60] Ilya Zakharkin, Kirill Mazur, Artur Grigorev, and Victor Lempitsky. Point-based modeling of human clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14718–14727, 2021. 8

[61] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*, 2018. 5, 7

[62] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. AvatarReX: Real-time Expressive Full-body Avatars. *ACM Trans. Graph.*, 42(4), 2023. 1

[63] Heming Zhu, Fangneng Zhan, Christian Theobalt, and Marc Habermann. Trihuman: a real-time and controllable triplane representation for detailed human geometry and appearance synthesis. *ACM Transactions on Graphics*, 44(1): 1–17, 2024. 2

[64] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, pages 145–162. Springer, 2024. 2

[65] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. *arXiv preprint arXiv:2311.08581*, 2023. 2

GIGA: Generalizable Sparse Image-driven Gaussian Avatars

Supplementary Material

7. Projected Texture Estimation

The input views consist of rich identity-specific appearance and pose-dependent variations. To aggregate the identity and pose-dependent information from input views, we adopt inverse texture mapping, which projects the imagery from multiple input views to the texel space of the template mesh.

Mesh Normalization. Before constructing inputs for GIGA, we normalize canonical mesh $\mathcal{V}(\boldsymbol{\theta}_0, \boldsymbol{\beta})$. We compute the scaling factor $\rho_{\text{body}} = \max(|\widetilde{\mathbf{x}}_0|)$ from canonical mesh vertices $\widetilde{\mathbf{x}}_0 \in \mathbb{R}^{N\times 3}$. Normalized vertices $\mathbf{x}_0 = \rho_{\text{body}}^{-1}\widetilde{\mathbf{x}}_0$ fit in the cubic region $[-1, 1]^3$. The same scaling procedure is applied to translation vectors $\pi_{k,\text{o}} \in \mathbb{R}^3$ of each of K cameras.

Partial Texture Computation. After articulating the normalized canonical mesh to the observed pose $\boldsymbol{\theta}_j$, we obtain posed vertex coordinates $\mathbf{x}_j$. The next step is an initialization of texel coordinates buffer $\mathbf{T}_{\mathbf{x}} \in \mathbb{R}^{T\times T\times 3}$. We set $\mathbf{x}_j$ as attributes to the mesh $\mathcal{V}(\boldsymbol{\theta}_j, \boldsymbol{\beta})$ vertices and perform texture sampling w.r.t UV parametrization $\mathcal{M}_{\text{uv}}$ to fill buffer $\mathbf{T}_{\mathbf{x}}$ with coordinates of posed texels. In the following, we will drop the pose index j, assuming that all operations are performed for the observed pose. For each input view k, pixel coordinates are calculated for each texel:

$$\mathbf{T}'_{\mathbf{x},k} = \text{Proj}_k(\mathbf{T}_{\mathbf{x}}), \tag{9}$$

where Proj_k denotes OpenGL-style projection to clip-space of view π_k.

The partial texture $\mathbf{T}_{\text{uv},k}$ is bilinearly sampled from the image I_k using pixel coordinates $\mathbf{X}'_k$:

$$\mathbf{T}_{\text{uv},k} = \text{GridSample}\left(I_k, \mathbf{T}'_{\mathbf{x},k}\right). \tag{10}$$

Visibility Check and Texture Aggregation. Not every texel is observed from the view π_k. Hence, we need to discard invisible texels from the partial texture. We first render a depth image of the body template $\mathcal{V}(\boldsymbol{\theta}, \boldsymbol{\beta})$ and retrieve vertex visibility buffer provided by the differentiable rasterizer [17]. After barycentric interpolating the visibility buffer $\mathcal{M}_{\text{uv}}$, we obtain the visibility mask $\mathbf{V}_{\text{uv},k} \in \mathbb{R}^{T\times T}$. Then, we compute angle visibility scores $\mathbf{V}^a_{\text{uv},k} \in \mathbb{R}^{T\times T}$:

$$\mathbf{V}^a_{\text{uv},k} = (\mathbf{N} \cdot \text{unit}(\pi_{k,\text{o}} - \mathbf{T}_{\mathbf{x}})), \tag{11}$$

where $\mathbf{N} \in \mathbb{R}^{T\times T\times 3}$ are per-texel normals obtained through barycentric interpolation of the vertex normals. unit denotes L2-normalization for per-texel viewing direction, and $(\cdot)$ denotes dot product between vectors. Next, we calculate indices $\overline{k}$ of partial texture with highest visibility scores

$$\overline{k} = \underset{k}{\text{argsort}}\left(\mathbf{V}^a_{\text{uv},k}\right). \tag{12}$$

Finally, the (body-) pose-dependent RGB texture map $\mathbf{T}_{\text{uv}}$ is computed as follows:

$$\mathbf{T}_{\text{uv}} = \text{gather}\left(\mathbf{T}_{\text{uv},k} \odot \mathbf{V}_{\text{uv},k}, \overline{k}\right) \tag{13}$$

where $\odot$ stands for Hadamard product and gather performs aggregation of individual texels specified by indices $\overline{k}$.

Figure 8. **Qualitative Results from Front View.** More qualitative results produced by GIGA. All subjects are from the test split. GT stands for ground truth.

8. Baselines.

We use the open-source implementations of both NHP [15] and TransHuman [30] for training. For the sake of efficiency, we replace original attention layers in baselines with their more efficient analogs [18]. Moreover, TransHuman requires clusterization of body template vertices. Therefore, we follow the original codebase and cluster SMPL-X

Figure 9. **Ablation on Number of Subjects for Training.** We provide qualitative results of GIGA, trained on smaller subsets of MVHumanNet [55]

vertices using K-Means with $K = 300$ clusters. For evaluation against GHG [16], we use validation data released by authors. GHG metric numbers are taken from the original paper.

9. More Qualitative Results

Here we show additional qualitative examples from GIGA results. Fig. 8 presents additional qualitative results for subjects from the testing split. We also present free-viewpoint renderings in Fig. 10 for subjects used neither for training nor testing.

Figure 10. **Free-viewpoint Rendering.** Views from a circular trajectory around unseen characters.