Sequential Gaussian Avatars with Hierarchical Motion Context

Wangze Xu[1]* Yifan Zhan[1,2] Zhihang Zhong[1]† Xiao Sun[1]†
[1]Shanghai Artificial Intelligence Laboratory [2]The University of Tokyo

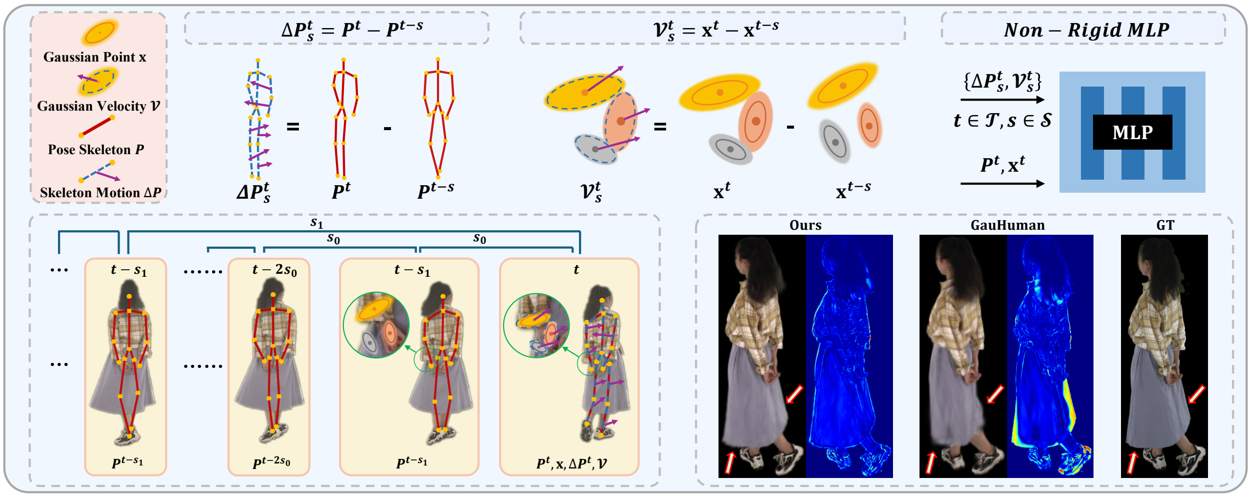
Gaussian Point x
Gaussian Velocity $\mathcal{V}$
Pose Skeleton P
Skeleton Motion ΔP
$\Delta P_s^t = P^t - P^{t-s}$
ΔP_s^t = P^t - P^{t-s}
$\mathcal{V}_s^t = \mathbf{x}^t - \mathbf{x}^{t-s}$
$\mathcal{V}_s^t$ = $\mathbf{x}^t$ - $\mathbf{x}^{t-s}$
$Non-Rigid\ MLP$
$\{\Delta P_s^t, \mathcal{V}_s^t\}$
$t \in \mathcal{T}, s \in \mathcal{S}$
$P^t, \mathbf{x}^t$
MLP
s_1
s_0
s_0
...
$t - s_1$
......
$t - 2s_0$
$t - s_1$
t
P^{t-s_1}
P^{t-2s_0}
P^{t-s_1}
$P^t, \mathbf{x}, \Delta P^t, \mathcal{V}$
Ours
GauHuman
GT

Figure 1. **Illustration of the Hierarchical Motion Context Design.** We model the motion-dependent appearance variations with both coarse skeleton condition ΔP and fine-grained point-wise velocity condition $\mathcal{V}$. ΔP describes the overall human skeleton motion, which is derived from the difference between the human poses at adjacent frames. $\mathcal{V}$ is the point-wise velocity that indicates finer-grained motion in local regions. To achieve robust deformation, we propose Spatio-temporal Multi-scale Sampling, which samples the overall motion trend and inter-frame details via diverse time intervals s for ΔP and $\mathcal{V}$.

Abstract

The emergence of neural rendering has significantly advanced the rendering quality of 3D human avatars, with the recently popular 3DGS technique enabling real-time performance. However, SMPL-driven 3DGS human avatars still struggle to capture fine appearance details due to the complex mapping from pose to appearance during fitting. In this paper, we propose SeqAvatar, which excavates the explicit 3DGS representation to better model human avatars based on a hierarchical motion context. Specifically, we utilize a coarse-to-fine motion conditions that incorporate both the overall human skeleton and fine-grained vertex motions for non-rigid deformation. To enhance the robustness of the proposed motion conditions, we adopt a spatio-temporal multi-scale sampling strategy to hierarchically integrate more motion clues to model human avatars. Extensive experiments demonstrate that our method significantly outperforms 3DGS-based approaches and renders human avatars orders of magnitude faster than the latest NeRF-based models that incorporate temporal context, all while delivering performance that is at least comparable or even superior. Project page: `https://zezeaaa.github.io/projects/SeqAvatar/`

1. Introduction

Recent research on digital humans has highlighted the efficiency of 3D Gaussian Splatting (3DGS) [27], demonstrating its capability for high-quality and real-time rendering. By defining T-pose human Gaussians in a canonical space and employing specific warping techniques, we can render human avatars from diverse perspectives and in any pose. The canonical Gaussian primitives and warping weights are then jointly optimized under supervision from video inputs.

Animatable human avatar reconstruction is primarily hindered by limited modeling of non-rigid warping, forcing a trade-off between rendering quality and animation capability. Current methods [22, 56] incorporate SMPL(-X) pose priors [43, 51] to guide motion for each observation

*This work was done during the author's internship at Shanghai Artificial Intelligence Laboratory. † denotes co-corresponding authors.

and rely on Linear Blend Skinning (LBS) for warping. Although these approaches are effective for pose-driven animation, they often struggle with per-frame non-rigid warping in scenarios involving complex garments.

We experimentally find that current human pose conditions [22, 56] do not fully capture the complex one-to-many mapping from pose to appearance, particularly for 3DGS-based approaches [27]. These methods [56, 68] rely on the spatial information of a template body in each frame's human pose to predict non-rigid deformations. However, they often overlook local details, such as garment deformations far from the skeleton, and cannot resolve cases where the same pose corresponds to different appearances during complex motions. Furthermore, although previous NeRF-based [44] method [8] has attempted to model motion sequences using human pose residuals, the inherently global nature of the pose sequence limits the ability to capture finer motion details. Naive pose sequence modeling does not fully account for the explicit characteristics of 3DGS.

In this paper, we introduce a hierarchical motion context condition for 3DGS-based human avatar modeling to address the limitations of relying solely on human pose, which provides only limited global skeletal information. Our approach improves the ability of 3DGS-based methods to accurately capture the complex relationship between human pose and appearance in challenging scenarios. Specifically, we design a coarse-to-fine motion condition that incorporates both overall skeletal movements and fine-grained point-wise motions. Leveraging the explicit nature of Gaussian primitives, this condition seamlessly integrates into 3DGS-based methods, enabling more precise predictions of complex non-rigid deformations. To further enhance robustness, we propose a spatio-temporal multi-scale sampling strategy for constructing the motion context with a larger receptive field. Spatially, we consider the motion states of neighboring points within the local region of each Gaussian primitive to obtain more stable motion embeddings. Temporally, we capture human motion patterns across multiple time scales, combining long-term trends with fine-grained inter-frame motion details. This improves the model's generalization to complex human movements.

To summarize, our key contributions are as follows:

1) We propose a novel hierarchical motion condition that integrates coarse-to-fine human motion, combining global skeletal poses with localized vertex residuals to enhance non-rigid deformation prediction.
2) We introduce a spatiotemporal multiscale sampling strategy that expands the receptive field of hierarchical motion context, improving generalization to complex motions.
3) Experiments on the I3D-Human [8], DNA-Rendering [9], and ZJU-MoCap [53] datasets show the effectiveness of the proposed **SeqAvatar**, which is capable of modeling details of the human body in complex motions.

2. Related Work

Neural Rendering. Neural rendering techniques have brought significant progress to human reconstruction and rendering. In particular, Neural Radiance Fields (NeRF) [44] introduces an implicit scene representation that models color and density using multilayer perceptrons, delivering photorealistic rendering results. More recently, point-based 3D Gaussian Splatting [27] utilizes 3D Gaussians to represent scenes explicitly, achieving real-time high-quality rendering. Building upon NeRF and 3DGS, subsequent works have advanced neural rendering across various dimensions, *e.g.*, enhancing visual fidelity [1, 2], improving sparse-view reconstruction quality [10, 52, 63, 71, 80], enabling pose-free optimization [37, 58, 61, 67], modeling dynamic scenes [13, 34, 49, 55, 69, 74], and accelerating training and inference [4, 12, 18, 39, 46, 60]. These developments also benefit human avatar modeling, which demands high-quality and efficient rendering.

SMPL(-X)-Based Neural Human Modeling. The SMPL(-X) family [43, 51] provides a parametric representation of the human body by decomposing it into pose-related and shape-related components using 3D mesh scans and principal component analysis (PCA). Its *pose blend shapes* enables body deformation through joint-wise pose blending, offering an efficient and compact way for animation. SMPL(-X) has thus become a cornerstone in human body modeling and animation, with many methods [11, 59] estimating its parameters directly from 2D inputs. Recent neural human reconstruction methods integrate SMPL(-X) with implicit representations such as NeRF [44] and 3DGS [27] to enable animatable avatars with high rendering quality. Different from purely 2D image animation methods [48, 62, 65, 73], these methods typically register the input data to a canonical T-pose space and use linear blend skinning (LBS) to transform 3D points into observation space based on SMPL(-X) poses. NeRF-based approaches [5–8, 14, 17, 19, 20, 30, 53, 64, 68, 75] focus on monocular or multi-view reconstruction. More recently, 3DGS-based methods [21–23, 25, 26, 29, 31–33, 36, 40, 41, 45, 47, 56, 76, 78, 79, 81] have gained popularity due to their real-time rendering and high fidelity. Some works [8, 23] further incorporate pose sequences as temporal context, but they suffer from slow rendering and lack fine details of motion caused by simply integrating coarse human pose embedding with NeRF.

Human Rendering with Temporal Embeddings The earliest modeling of dynamic radiance field [3, 13, 15, 16, 24, 34, 35, 42, 50, 57, 70] can be traced back to temporal embeddings in general scenes, where each frame's observation is obtained by constructing a canonical space and a time-conditioned deformation field. Motivated by these, a stream of research [38, 72] focuses on pure rendering quality, employing temporal embeddings instead of human pose to encode each frame of human videos. Although these methods achieve high-quality, temporally continuous rendering, they struggle with the lack of geometric constraints from human pose, making parameterized animation unfeasible.

3. Preliminary

SMPL(-X) Series [43, 51] adopt a parameterized framework to represent human bodies across diverse shapes and poses. For each frame, the human mesh is derived by deforming a canonical template mesh based on shape and pose parameters. Specifically, a 3D point $\mathbf{x}$ on the canonical mesh is warped to the corresponding point on the deformed mesh in the observation space, following

$$\mathbf{LBS}(\mathbf{x}, \mathbf{B}_k, \omega_k(\mathbf{x})) = \sum_{k=1}^{K} \omega_k(\mathbf{x}) \mathbf{B}_k \mathbf{x}, \tag{1}$$

where K is the total bone number and $\mathbf{B}_k$ is the transformation matrix for each bone. Specifically, $\mathbf{B}_k$ consists of a rotation and translation matrix $\mathbf{R}_k$ and $\mathbf{b}_k$. Here, $\mathbf{R}_k$ represents the global rotation of each joint, influenced by the local joint rotations $\boldsymbol{r}$, while $\mathbf{b}_k$ denotes each joint's translation, determined by the joint positions $\boldsymbol{j}$ and the human shape parameters β. The linear blending weight ω^k depends on $\mathbf{x}$ and is regressed from comprehensive human meshes.

3D Gaussian Splatting [27] utilizes a set of point-based Gaussian primitives to explicitly represent scenes, enabling real-time and high-quality rendering. Each Gaussian primitive is defined by its center position $\mathbf{x} \in \mathbb{R}^3$ and covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{3\times 3}$. To ensure positive semi-definiteness and simplify optimization, the covariance matrix Σ is decomposed into a rotation matrix $\mathbf{R}$ and a scaling matrix $\mathbf{S}$:

$$\mathbf{\Sigma} = \mathbf{R}\mathbf{S}\mathbf{S}^T\mathbf{R}^T, \tag{2}$$

where $\mathbf{R}$ and $\mathbf{S}$ are derived by a scaling vector $\mathbf{s} \in \mathbb{R}^3$ and a quaternion vector $\mathbf{r} \in \mathbb{R}^4$ in practice. Additionally, each Gaussian primitive is also assigned a color feature $sh \in \mathbb{R}^k$ represented by spherical harmonics (SH) and an opacity α for rendering. During the rendering process, a splatting technique [27] is used to project each Gaussian primitive onto the 2D image space with a viewing transform $\mathbf{W}$ and the Jacobian $\mathbf{J}$ of the projective transformation's affine approximation. The transformed covariance $\mathbf{\Sigma}'$ in the camera's coordinates is

$$\mathbf{\Sigma}' = \mathbf{J}\mathbf{W}\mathbf{\Sigma}\mathbf{W}^T\mathbf{J}^T. \tag{3}$$

After projection, the pixel color C is computed by blending N ordered Gaussian primitives overlapping at the pixel:

$$C = \sum_{i\in N} c_i \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j), \tag{4}$$

where c_i is computed from SH feature sh and α'_i is the product of α_i and probability density of i-th 2D Gaussian.

For optimization, a photometric loss is defined by a combination of L_1 and SSIM [66] losses:

$$L_{photo} = \lambda L_1(\hat{I}, I) + (1-\lambda)(1 - SSIM(\hat{I}, I)), \tag{5}$$

where $\hat{I}$ and I denote the rendered and ground-truth images, and λ controls the balance between the two terms.

4. Method

Fig. 2 shows the framework of our method. We use the explicit point-based 3DGS as the representation of the human body. Given a collection of input cameras and images, we optimize a set of Gaussian primitives $\{\mathcal{G}_i\}_{i=1}^{i=n}$ to fit the body's shape and appearance. Each Gaussian primitive $\mathcal{G}_i$ includes the center position $\mathbf{x}$, scaling vector $\mathbf{s}$, rotation quaternion vector $\mathbf{r}$, color feature sh and opacity α, where $\mathbf{x}$ is initialized from the SMPL template vertices $\{\mathbf{T}_i\}_{i=1}^{i=N}$. During the optimization process, we first apply a coarse-to-fine motion context to capture more accurate and fine-grained details (Sec. 4.1). To mitigate overfitting, we propose Spatio-Temporal Multi-Scale Sampling to obtain more robust point-wise Gaussian motion, which serves as the embedding for non-rigid deformation (Sec.4.2). Next, we adopt Linear Blend Skinning (LBS) to map canonical Gaussian primitives $\mathcal{G}$ to observation space and render images via differentiable splitting (Sec.4.3).

4.1. Non-Rigid Deformation with Coarse-to-fine Motion Context

Coarse Skeleton Motion. As illustrated in Fig. 1, we introduce a coarse-to-fine motion context to excavate additional conditions for non-rigid deformation. Typically, most human 3D avatar methods use the current frame pose as a condition to predict non-rigid deformation [22, 56]. While this provides spatial information to distinguish the motion of different body parts, it fails to capture temporal motion changes, as discussed in [8]. Therefore, given a frame at time t, we consider a sequence of regularly interval-sampled frames $\mathcal{T}$ to model inter-frame body motion variations:

$$\mathcal{T} = \{t-s, t-2s, ..., t-Ls\}, \tag{6}$$

where L is the sequence length and s is the time interval. The coarse motion of body skeletons at each time step $t \in \mathcal{T}$ can be derived by calculating the difference of poses between adjacent frames as illustrated in Fig. 1, following

$$\Delta\mathcal{P} = \{\Delta P^t = \delta(P^t, P^{t-s}) | t \in \mathcal{T}\}, \tag{7}$$

where $P \in \mathbb{R}^{K\times 3}$ is the body pose and δ denotes the difference $\Delta\mathcal{P}$ between two poses in axis-angle form. We employ an MLP $\boldsymbol{\mathcal{E}}_{\Delta\mathcal{P}}$ to encode the sequential skeleton mo-

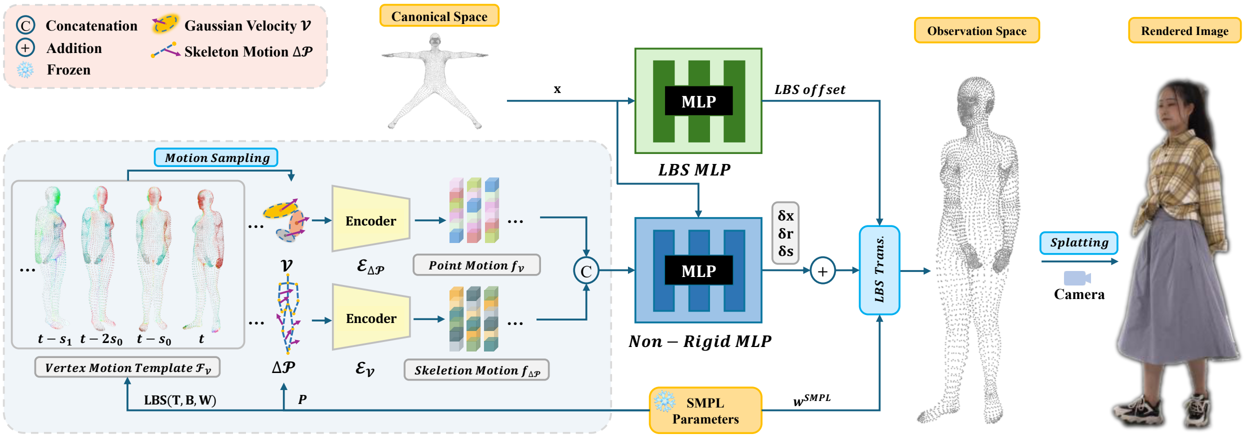

Figure 2. **Overview of the proposed method.** We first initialize canonical Gaussian positions $\mathbf{x}$ with SMPL template vertexes. For each Gaussian, we derive both coarse skeleton motion condition $f_{\Delta\mathcal{P}}$ and fine-grained vertex motion condition $f_{\mathcal{V}}$ that sampled from the vertex motion template $\mathcal{F}_{\mathcal{V}}$ (points in different colors represent different motions). Based on such hierarchical motion information, we utilize an MLP $\mathcal{E}_{non-rigid}$ to better predict each Gaussian's non-rigid deformation prediction. The non-rigid deformed Gaussians are then warped into observation space via the standard LBS transformation for rendering.

tion $\Delta\mathcal{P} \in \mathbb{R}^{L\times K\times 3}$, flattened along the temporal dimension, into a skeleton motion embedding $f_{\Delta\mathcal{P}} \in \mathbb{R}^{32}$,

$$f_{\Delta\mathcal{P}} = \mathcal{E}_{\Delta\mathcal{P}}(\Delta\mathcal{P}), \tag{8}$$

which serves as a condition for the subsequent non-rigid deformation. Further details are provided in the supplementary material.

Fine Vertex Motion. Compared to previous implicit NeRF-based methods, the point-based 3DGS representation enables us to explore more fine-grained temporal motion information. We derive a point-wise velocity vector v_i for each Gaussian primitive $\mathcal{G}_i$ to model the fine-grained localized body motion, which is beyond the scope of pose skeleton motion $\Delta\mathcal{P}$. By measuring the position variations across adjacent time steps, we calculate per-frame velocity

$$v^t = \frac{\mathbf{x_o}^t - \mathbf{x_o}^{t-s}}{s}, \tag{9}$$

where $\mathbf{x_o}$ denotes the warped coordinates in observation space. However, since the positions of Gaussian primitives are updated during optimization, it is not stable to obtain v_i based on such a dynamic variable. Moreover, transforming the canonical Gaussians into the observation space requires non-rigid deformation first, which leads to a circular dependency and conflicts with the current velocity design. To this end, we introduce a motion template field $\boldsymbol{\mathcal{F}}_{\mathbf{V}} = \{\mathbf{V}_i\}_{i=1}^{i=N}$ that stores the velocity of each SMPL vertex, and v_i for each Gaussian primitive is sampled from $\boldsymbol{\mathcal{F}}_{\mathbf{V}}$ based on the distance between query points and vertexes in this template. Specifically, given a time step t and the corresponding body pose $P^t = \{(\mathbf{R}_k^t, \mathbf{b}_k^t)|k \in K\}$, the SMPL template vertexes in $\mathbf{T}$ can be warped into observation space with the template skinning weights $\mathbf{W}$ and standard linear blend skinning, following

$$\mathbf{T_o}^t = \mathbf{LBS}(\mathbf{T}, \mathbf{B}^t, \mathbf{W}), \tag{10}$$

where $\mathbf{B}^t$ is the transformation matrix derived from rotation and translation matrix $\mathbf{R}^t$ and $\mathbf{b}^t$. $\mathbf{W}$ is the template SMPL LBS weights. The velocity $\mathbf{V}$ of each template vertexes $\mathbf{T}$ can be further derived as

$$\mathbf{V}^t = \frac{\mathbf{T_o}^t - \mathbf{T_o}^{t-s}}{s}. \tag{11}$$

We then sample each Gaussian's velocity from this motion template field $\boldsymbol{\mathcal{F}}_{\mathbf{V}}$, which is discussed in Sec. 4.2.

4.2. Spatio-temporal Multi-scale Sampling (STMS)

To enhance the robustness of human motion conditions, we propose modeling both local region motion information of the human body and motion patterns across different temporal windows to mitigate overfitting and improve generalization by capturing more comprehensive motion dynamics.

In the spatial dimension, we sample the τ nearest template to model the body's local region motion more robustly. Specifically, for each canonical Gaussian primitive $\mathcal{G}_i$, the τ nearest vertexes' velocities are sampled as input to an MLP $\mathcal{E}_{knn}$ to learn a local point-wise motion embedding

$$e_i^t = \mathcal{E}_{knn}(\{\mathbf{V}_j^t\}), \quad j \in \mathbf{KNN}(\mathbf{T}, \mathbf{x}_i), \tag{12}$$

where $\mathbf{KNN}(\mathbf{T}, \mathbf{x}_i)$ denotes the τ nearest SMPL template vertexes of the canonical Gaussian position $\mathbf{x}_i$. Similar to Eq. 7, we then apply an MLP $\boldsymbol{\mathcal{E}}_{\boldsymbol{\mathcal{V}}}$ to encode each Gaussian primitive's sequential motion embedding $\mathcal{V} = \{e^t|t \in \mathcal{T}\}$ into a point-wise sequential condition $f_{\mathcal{V}} \in \mathbb{R}^{96}$

$$f_{\mathcal{V}} = \mathcal{E}_{\mathcal{V}}(\mathcal{V}). \tag{13}$$

In the temporal dimension, to capture both the overall motion trend and the inter-frame motion details, we adopt a multi-scale sequence sampling strategy. In detail, we sample a series of sequences at several progressively increasing intervals to get human motion information across different temporal windows

$$\mathcal{S} = \{s = s_0 + i\Delta s\}_{i=0}^{i=m}, \tag{14}$$

where Δs is the increasing rate of sampling interval s. We then input the multi-scale sampled sequence motions into the skeleton motion encoder $\boldsymbol{\mathcal{E}_{\Delta\mathcal{P}}}$ and the point-wise motion encoder $\boldsymbol{\mathcal{E}_{\mathcal{V}}}$ to obtain hierarchical temporal motion embeddings. Therefore, Eq. 7 and Eq. 13 can be revised as

$$f_{\Delta\mathcal{P}} = \mathcal{E}_{\Delta\mathcal{P}}(\{\Delta\mathcal{P}_s\}), \quad s \in \mathcal{S} \tag{15}$$

$$f_{\mathcal{V}} = \mathcal{E}_{\mathcal{V}}(\{\mathcal{V}_s\}), \quad s \in \mathcal{S}. \tag{16}$$

In practice, we concatenate all skeleton motion conditions, $\Delta\mathcal{P}_s$ for $s \in \mathcal{S}$, and localized vertex motion conditions, $\mathcal{V}_s$ for $s \in \mathcal{S}$, across different sampling scales. These are then input to $\boldsymbol{\mathcal{E}_{\Delta\mathcal{P}}}$ and $\boldsymbol{\mathcal{E}_{\mathcal{V}}}$, respectively. More details are shown in the supplementary material.

Non-Rigid Deformation. Given the coarse skeleton motion $f_{\Delta\mathcal{P}} = \mathcal{E}_{\Delta\mathcal{P}}(\{\Delta\mathcal{P}_s\})$ condition and fine point-wise velocity condition $f_{\mathcal{V}} = \mathcal{E}_{\mathcal{V}}(\{\mathcal{V}_s\})$, we utilize an MLP to predict each Gaussian's non-rigid deformation as

$$\delta\mathbf{x}, \delta\mathbf{s}, \delta\mathbf{r} = \mathcal{E}_{non-rigid}(\mathbf{x}, P, f_{\Delta\mathcal{P}}, f_v). \tag{17}$$

The deformed canonical Gaussian $\mathcal{G}'$ is

$$\mathbf{x}' = \mathbf{x} + \delta\mathbf{x}, \tag{18}$$

$$\mathbf{s}' = \mathbf{s} + \delta\mathbf{s}, \tag{19}$$

$$\mathbf{r}' = \mathbf{r} \cdot \delta\mathbf{r}, \tag{20}$$

where $\cdot$ denotes the multiplication of two quaternions.

4.3. Optimization

Rigid Deformation. We utilize the standard $\mathbf{LBS}$ operation to map the non-rigid deformed Gaussians $\mathcal{G}_o$ into the observation space, following

$$\mathbf{x_o} = \mathbf{LBS}(\mathbf{x}', \mathbf{B}, \omega), \tag{21}$$

$$\mathbf{R_o} = \sum_{k=1}^{K} \omega_k(\mathbf{x}')\mathbf{B}_k\mathbf{R}', \tag{22}$$

where $\mathbf{R}'$ is the rotation matrix derived from the non-rigid deformed Gaussian's rotation quaternion $\mathbf{r}'$, and $\mathbf{R_o}$ represents the rotation matrix in observation space. Following the previous method [22], we utilize an MLP $\mathcal{E}_{lbs}$ to predict the LBS weight offsets for each query canonical point and update the sampled weights from the nearest SMPL vertex as

$$\omega_k(\mathbf{x}) = \omega_k^{SMPL}(\mathbf{x}) + \mathcal{E}_{lbs}(\mathbf{x}). \tag{23}$$

Similar to [8, 22, 68], we introduce a pose refinement MLP $\mathcal{E}_{pose}$ to refine the pose estimate from SMPL for a better fit to the human body.

Loss Function. With the transformed Gaussian primitives $\mathcal{G}_o$ in observation space, we apply the standard splitting [27] to render images as

$$I = \mathbf{Splatting}(\mathbf{x_o}, \mathbf{R_o}, s', \alpha, sh). \tag{24}$$

During the optimization process, we employ a combination of loss functions as supervision, summarized as

$$\mathcal{L} = \lambda_1\mathcal{L}_{color} + \lambda_2\mathcal{L}_{ssim} + \lambda_3\mathcal{L}_{lpips} + \mathcal{L}_{mask}, \tag{25}$$

where λ is used to balance the weight of different losses. $\mathcal{L}_{mask}$ [22] is an L_2 loss between the rendered α and human body mask. Similar to [56], we apply $\mathcal{L}_{isopos}$ and $\mathcal{L}_{isocov}$ to control the Gaussian primitive's position and covariance. Details are shown in the supplementary material.

5. Experiments

5.1. Datasets

DNA-Rendering [9] is a challenging human-centric rendering dataset. It contains diverse scenes from everyday life to professional occasions. We use 6 sequences with loose-fitting garments (1_0206_04, 2_0007_04, 2_0019_10, 2_0044_11, 2_0051_09, 2_0813_05) for experiments. We adopt 24 views for training and 6 views for testing.

I3D-Human [8] contains scenes closer to daily life, which comprises multi-view video frames of humans with loose clothing and complex movements. We conduct experiments on 4 sequences and use 4-5 views for training and the rest for testing. All training and testing data are retained to be the same in comparisons following Dyco's [8] data split.

5.2. Baselines and Metrics

We compare against the state-of-the-art human modeling methods including Dyco [8], 3DGS-Avatar [56], GART [31], and GauHuman [22]. Dyco is an implicit NeRF-based method that adopts pose sequence information for temporal modeling. We compare performance with Dyco to showcase the better efficacy of our fine-grained motion condition design. 3DGS-Avatar, GART, and GauHuman are explicit 3DGS-based methods, and all of them are originally designed for monocular inputs. For fairness, we extend them to multi-view input under the same settings. We report three key metrics: peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) [66], and learned perceptual image patch similarity (LPIPS) [77].

5.3. Comparison

Comparison on DNA-Rendering. Tab. 1 shows the quantitative results on DNA-Rendering dataset. The proposed

Table 1. **Quantitative Results on DNA-Rendering Dataset.** We report the performance of novel view rendering. Our method outperforms previous state-of-the-art human modeling methods. We mark the **best** and the **second best** methods in cells. LPIPS*=LPIPS ×10^3.

Subject:	1_0206_04			2_0007_04			2_0019_10			2_0044_11			2_0051_09			2_0813_05		
Metric:	PSNR↑	SSIM↑	LPIPS*↓	PSNR↑	SSIM↑	LPIPS*↓	PSNR↑	SSIM↑	LPIPS*↓	PSNR↑	SSIM↑	LPIPS*↓	PSNR↑	SSIM↑	LPIPS*↓	PSNR↑	SSIM↑	LPIPS*↓
3DGS-Avatar [56]	26.64	0.9439	50.09	28.06	0.9492	56.70	32.06	0.9710	31.12	28.23	0.9512	37.15	25.04	0.9534	42.22	31.75	0.9701	31.30
GART [31]	27.19	0.9454	56.47	28.22	0.9525	56.46	31.13	0.9686	35.16	28.90	0.9561	41.27	26.47	0.9628	42.21	32.06	0.9729	35.74
GauHuman [22]	27.96	0.9500	50.42	27.93	0.9496	56.93	31.63	0.9698	32.82	30.38	0.9641	34.19	25.99	0.9599	42.00	33.43	0.9764	29.42
Ours	30.81	0.9649	39.47	29.63	0.9566	43.72	34.97	0.9777	23.53	32.62	0.9747	23.83	28.45	0.9685	34.06	35.81	0.9843	20.83

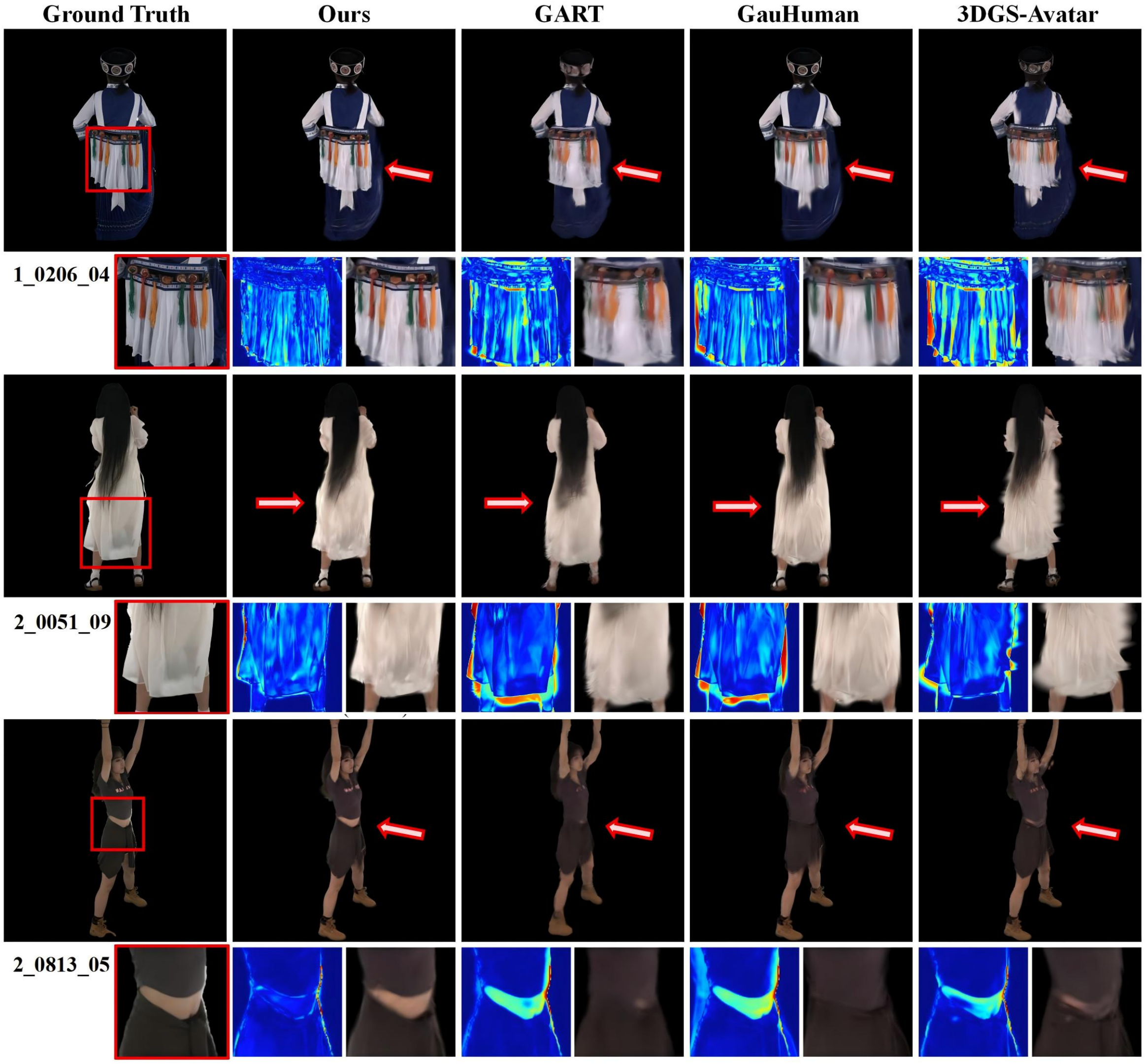

Figure 3. **Novel View Qualitative Results on DNA-Rendering.** We zoom into the local region and compute the error maps compared with ground truth images. The results show that our method achieves competitive results on both the overall and local region qualities.

method outperforms previous state-of-the-art human modeling methods across all metrics. To demonstrate the visual improvement, we compare the quality of rendering images in Fig. 3. Previous methods, which lack the temporal motion information for non-rigid deformations, are unable to predict the movement of Gaussian primitives properly, resulting in blurred renderings in regions with complex textures (scene 1_0206_04 in Fig. 3). Besides, the results in Fig. 3 show that the compared methods fail to capture the proper shape variation caused by movement (e.g., the white dress in scene 2_0051_09). Although these methods achieve a smooth appearance in these areas, the error maps show that they fail to match the actual shape closely. We leverage the hierarchical motion context to predict Gaussian defor-

Table 2. **Novel View Rendering Quantitative Results on I3D-Human.**

Subject:	ID1_1			ID1_2			ID2_1			ID3_1		
Metric:	PSNR↑	SSIM↑	LPIPS*↓	PSNR↑	SSIM↑	LPIPS*↓	PSNR↑	SSIM↑	LPIPS*↓	PSNR↑	SSIM↑	LPIPS*↓
3DGS-Avatar [56]	29.78	0.9600	31.40	31.29	0.9627	28.09	29.63	0.9604	37.33	32.73	0.9602	39.47
Dyco [8]	30.81	0.9617	28.29	31.45	0.9628	25.75	29.41	0.9618	33.26	32.50	0.9612	35.46
GauHuman [22]	29.47	0.9571	39.46	30.42	0.9585	38.87	28.51	0.9546	49.46	32.10	0.9546	53.67
Ours	31.81	0.9659	27.03	31.98	0.9660	27.84	31.53	0.9685	30.23	33.62	0.9651	34.02

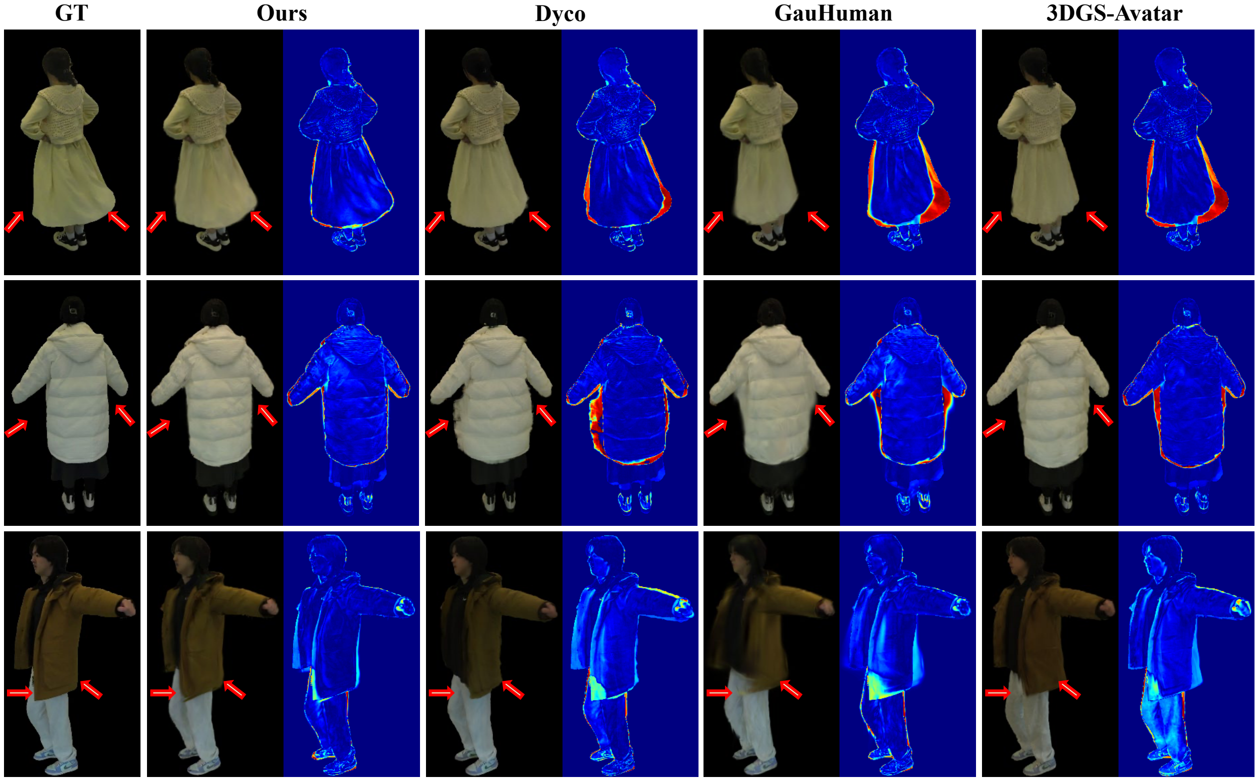

Figure 4. **Novel Pose Qualitative Results on I3D-Human.** We compare the proposed method with previous SOTA approaches. The rendered images and error maps demonstrate our robustness on novel pose rendering.

mations, allowing our method to capture detailed appearance variations caused by human motions more accurately.

Comparison on I3D-Human. The quantitative results in Tab. 2 show that our method surpasses the previous SOTA method on most metrics, demonstrating our competitive performance on the I3D-Human dataset. Previous NeRF-based Dyco [8] utilizes pose variation as a condition to model human motions. However, it is limited to capturing only overall body motions and lacks the capacity for finer-grained modeling in local regions. For example, in ID1_1 and ID2_1 shown in Fig. 8 (in the Supp. Mat.), our method renders results that accurately reflect the true motion, while Dyco fails to model the information in regions far from the human body skeleton. Tab. 3 reports the quantitative results of novel pose rendering on the I3D-Human dataset. Our method achieves best performance among 3DGS-based human modeling methods and also enables real-time rendering (∼ 45 FPS on I3D-Human) compared to NeRF-based Dyco (∼ 0.7 FPS). Fig. 4 shows the qualitative comparisons of novel pose rendering. Different from Dyco [8], which relies solely on joint motions as conditions, our method leverages the motion state of each Gaussian primitive to predict deformations, enabling more flexible modeling of motions in regions distant from the human body.

To evaluate the generalizability, we also conduct experiments on ZJU-MoCap [53]. Please refer to the supplementary materials for more details.

Out-of-Distribution Poses. To evaluate the generalization ability to novel poses, we conduct experiments using animations with large pose variations. Specifically, we use a model trained on one sequence to render target poses sampled from other unseen sequences, as different sequences in DNA-Rendering and I3D-Human contain distinct motion patterns. As illustrated in Fig. 6, the *Target Pose* is

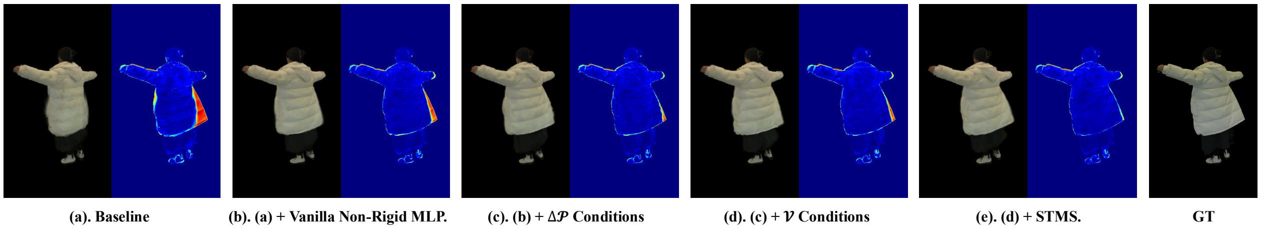

Figure 5. **Qualitative Ablation Results.** We compare the visual influence of different components on ID2_1 scene of I3D-Human dataset.

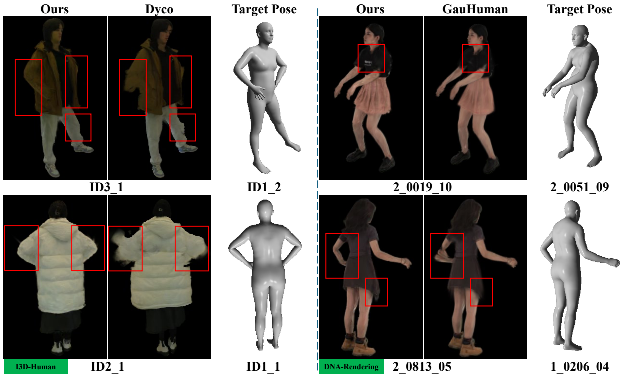

Figure 6. **Out-of-distribution Pose Animation.**

drawn from an unseen sequence exhibiting different human motions. The results suggest that our method is capable of generalizing to these significant pose changes, even though such poses were not observed during training.

Table 3. **Metrics of Novel Pose Renderings on I3D-Human.**

Methods	PSNR ↑	SSIM ↑	LPIPS* ↓
Dyco [8]	30.09	0.9569	35.32
3DGS-Avatar [56]	29.54	0.9555	38.59
GauHuman [22]	29.31	0.9526	48.29
Ours	30.28	0.9583	36.07

5.4. Ablation Study

In this section, we ablate the influence of various components in our methods on I3D-Human dataset. Tab. 4 shows the average metrics of 4 sequences under different settings.

Skeleton motion condition $\Delta\mathcal{P}$ denotes the overall human body movement. Compared to predicting non-rigid deformation only with the Gaussian position, incorporating this additional global temporal information allows our method to better model the overall human shape and achieve higher rendering quality, as demonstrated in Tab. 4 (c).

Fine-grained point-wise motion condition $\mathcal{V}$ describes the local region's variation in a more detailed manner. Such point-based motion information captures the relationship between motion and appearance variations in areas beyond the human skeleton. Tab. 4 (d) and Fig. 5 (d) show that our method achieves higher rendering details in local regions.

Spatio-temporal Multi-scale Sampling (STMS) is designed to enhance the robustness of human motion conditions by incorporating local region motion information and motion patterns across different temporal windows. With such a hierarchical motion context embedding design, our method is able to predict each Gaussian's deformation more robustly, leading to better rendering quality as shown in Tab. 4 (e) and Fig. 5 (e). Please refer to the supplementary materials for more detailed ablation experiments.

Table 4. **Quantitative Ablation Results on I3D-Human.** (a) denotes experiments without non-rigid deformation. (b) only utilizes Gaussian positions $\mathbf{x}$ and the current frame pose P as the conditions for non-rigid MLP.

Methods	PSNR ↑	SSIM ↑	LPIPS* ↓
(a) Baseline	29.76	0.9569	38.35
(b) (a)+Vanilla Non-Rigid MLP+$\mathcal{P}$ Conditions.	31.05	0.9617	34.35
(c) (b)+$\Delta\mathcal{P}$ Conditions.	31.89	0.9645	32.17
(d) (c)+ $\mathcal{V}$ Conditions.	32.01	0.9651	31.23
(e) (d)+STMS. (Ours)	**32.24**	**0.9664**	**29.78**

6. Limitations and Conclusion

Limitations. The Gaussian-based representation used in our method may introduce slight blur artifacts during rendering, whereas NeRF's ray-based volumetric integration tends to produce sharper results. Moreover, our local velocity cues are derived from the coarse SMPL model rather than dense surface tracking [79], which may limit the accuracy of fine-scale garment deformation. Addressing these limitations remains an open challenge for future work.

Conclusion. In this paper, we propose a 3DGS-based framework that integrates hierarchical motion context for 3D human modeling. Specifically, the non-rigid deformation for the Gaussian primitive is learned based on the global human skeleton variations and fine-grained Gaussian's point-wise motions. To capture each Gaussian primitive's motion information more robustly, we introduce a spatio-temporal multi-scale sampling strategy to incorporate both local region motion features and motion patterns across different temporal intervals. Through the above design, the proposed method achieves state-of-the-art rendering of human avatars with complex garments and motions.

Acknowledgment

This research was supported by Shanghai Artificial Intelligence Laboratory.

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-NeRF: A Multiscale Representation for Anti-aliasing Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 2

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 2

[3] Ang Cao and Justin Johnson. Hexplane: A Fast Representation for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023. 2

[4] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, pages 333–350. Springer, 2022. 2

[5] Jianchuan Chen, Ying Zhang, Di Kang, Xuefei Zhe, Linchao Bao, Xu Jia, and Huchuan Lu. Animatable neural radiance fields from monocular rgb videos. *arXiv preprint arXiv:2106.13629*, 2021. 2

[6] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11594–11604, 2021.

[7] Xu Chen, Tianjian Jiang, Jie Song, Max Rietmann, Andreas Geiger, Michael J Black, and Otmar Hilliges. Fast-snarf: A fast deformer for articulated neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 11796–11809, 2023.

[8] Yutong Chen, Yifan Zhan, Zhihang Zhong, Wei Wang, Xiao Sun, Yu Qiao, and Yinqiang Zheng. Within the dynamic context: Inertia-aware 3d human modeling with pose sequence. *arXiv preprint arXiv:2403.19160*, 2024. 2, 3, 5, 7, 8, 1

[9] Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, et al. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19982–19993, 2023. 2, 5, 1, 3

[10] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 2

[11] Junting Dong, Qi Fang, Wen Jiang, Yurou Yang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation and tracking from multiple views. In *T-PAMI*, 2021. 2

[12] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, pages 5501–5510, 2022. 2

[13] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit Radiance Fields in Space, Time, and Appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 2

[14] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 2

[15] Wanshui Gan, Hongbin Xu, Yi Huang, Shifeng Chen, and Naoto Yokoya. V4D: Voxel for 4D Novel View Synthesis. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 2

[16] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic View Synthesis from Dynamic Monocular Video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021. 2

[17] Qingzhe Gao, Yiming Wang, Libin Liu, Lingjie Liu, Christian Theobalt, and Baoquan Chen. Neural novel actor: Learning a generalized animatable neural representation for human actors. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 2

[18] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *ICCV*, pages 14346–14355, 2021. 2

[19] Chen Geng, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Learning neural volumetric representations of dynamic humans in minutes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8759–8770, 2023. 2

[20] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023. 2

[21] 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*, pages 634–644, 2024. 2

[22] 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*, pages 20418–20431, 2024. 1, 2, 3, 5, 6, 7, 8

[23] Tao Hu, Fangzhou Hong, and Ziwei Liu. Surmo: Surface-based 4d motion modeling for dynamic human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6550–6560, 2024. 2

[24] Hankyu Jang and Daeyoung Kim. D-tensorf: Tensorial Radiance Fields for Dynamic Scenes. *arXiv preprint arXiv:2212.02375*, 2022. 2

[25] Rohit Jena, Ganesh Subramanian Iyer, Siddharth Choudhary, Brandon Smith, Pratik Chaudhari, and James Gee. Splatarmor: Articulated gaussian splatting for animatable humans from monocular rgb videos. *arXiv preprint arXiv:2311.10812*, 2023. 2

[26] HyunJun Jung, Nikolas Brasch, Jifei Song, Eduardo Perez-Pellitero, Yiren Zhou, Zhihao Li, Nassir Navab, and Benjamin Busam. Deformable 3d gaussian splatting for animatable human avatars. *arXiv preprint arXiv:2312.15059*, 2023. 2

[27] 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. 1, 2, 3, 5

[28] Martin Kilian, Niloy J Mitra, and Helmut Pottmann. Geometric modeling in shape space. In *ACM SIGGRAPH 2007 papers*, pages 64–es. 2007. 1

[29] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 505–515, 2024. 2

[30] 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*, 34:24741–24752, 2021. 2

[31] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19876–19887, 2024. 2, 5, 6

[32] Mingwei Li, Jiachen Tao, Zongxin Yang, and Yi Yang. Human101: Training 100+ fps human gaussians in 100s from 1 view. *arXiv preprint arXiv:2312.15258*, 2023.

[33] Mengtian Li, Shengxiang Yao, Zhifeng Xie, Keyu Chen, and Yu-Gang Jiang. Gaussianbody: Clothed human reconstruction via 3d gaussian splatting. *arXiv preprint arXiv:2401.09720*, 2024. 2

[34] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3D Video Synthesis from Multi-view Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5521–5531, 2022. 2

[35] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural Scene Flow Fields for Space-time View Synthesis of Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021. 2

[36] 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. 2

[37] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *ICCV*, pages 5741–5751, 2021. 2

[38] Haotong Lin, Sida Peng, Zhen Xu, Tao Xie, Xingyi He, Hujun Bao, and Xiaowei Zhou. High-fidelity and real-time novel view synthesis for dynamic scenes. In *SIGGRAPH Asia Conference Proceedings*, 2023. 2

[39] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. pages 15651–15663, 2020. 2

[40] Xinqi Liu, Chenming Wu, Xing Liu, Jialun Liu, Jinbo Wu, Chen Zhao, Haocheng Feng, Errui Ding, and Jingdong Wang. Gea: Reconstructing expressive 3d gaussian avatar from monocular video. *arXiv preprint arXiv:2402.16607*, 2024. 2

[41] Yang Liu, Xiang Huang, Minghan Qin, Qinwei Lin, and Haoqian Wang. Animatable 3d gaussian: Fast and high-quality reconstruction of multiple human avatars. *arXiv preprint arXiv:2311.16482*, 2023. 2

[42] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust Dynamic Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13–23, 2023. 2

[43] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6), 2015. 1, 2, 3

[44] 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 *European Conference on Computer Vision*, pages 405–421, 2020. 2

[45] 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*, pages 788–798, 2024. 2

[46] 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

[47] Muyao Niu, Yifan Zhan, Qingtian Zhu, Zhuoxiao Li, Wei Wang, Zhihang Zhong, Xiao Sun, and Yinqiang Zheng. Bundle adjusted gaussian avatars deblurring. *arXiv preprint arXiv:2411.16758*, 2024. 2

[48] Muyao Niu, Mingdeng Cao, Yifan Zhan, Qingtian Zhu, Mingze Ma, Jiancheng Zhao, Yanhong Zeng, Zhihang Zhong, Xiao Sun, and Yinqiang Zheng. Anicrafter: Customizing realistic human-centric animation via avatar-background conditioning in video diffusion models. *arXiv preprint arXiv:2505.20255*, 2025. 2

[49] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2

[50] Sungheon Park, Minjung Son, Seokhwan Jang, Young Chun Ahn, Ji-Yeon Kim, and Nahyup Kang. Temporal Interpolation Is All You Need for Dynamic Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4212–4221, 2023. 2

[51] 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)*, pages 10975–10985, 2019. 1, 2, 3

[52] Rui Peng, Wangze Xu, Luyang Tang, Jianbo Jiao, Ronggang Wang, et al. Structure consistent gaussian splatting with matching prior for few-shot novel view synthesis. *Advances in Neural Information Processing Systems*, 37:97328–97352, 2024. 2

[53] 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, 7, 1

[54] Sergey Prokudin, Qianli Ma, Maxime Raafat, Julien Valentin, and Siyu Tang. Dynamic point fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7964–7976, 2023. 1

[55] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-neRF: Neural Radiance Fields for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2

[56] 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. 1, 2, 3, 5, 6, 7, 8

[57] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient Neural 4D Decomposition for High-fidelity Dynamic Reconstruction and Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16632–16642, 2023. 2

[58] Shihe Shen, Huachen Gao, Wangze Xu, Rui Peng, Luyang Tang, Kaiqiang Xiong, Jianbo Jiao, and Ronggang Wang. Disentangled generation and aggregation for robust radiance fields. In *European Conference on Computer Vision*, pages 218–236. Springer, 2024. 2

[59] Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. Novel view synthesis of human interactions from sparse multi-view videos. In *SIGGRAPH Conference Proceedings*, 2022. 2

[60] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, pages 5459–5469, 2022. 2

[61] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4190–4200, 2023. 2

[62] Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. Stableanimator: High-quality identity-preserving human image animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21096–21106, 2025. 2

[63] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9065–9076, 2023. 2

[64] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *European conference on computer vision*, pages 1–19. Springer, 2022. 2

[65] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *arXiv preprint arXiv:2406.01188*, 2024. 2

[66] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image Quality Assessment: from Error Visibility to Structural Similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 3, 5

[67] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2

[68] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 16210–16220, 2022. 2, 5

[69] Jiahao Wu, Rui Peng, Zhiyan Wang, Lu Xiao, Luyang Tang, Jinbo Yan, Kaiqiang Xiong, and Ronggang Wang. Swift4d: Adaptive divide-and-conquer gaussian splatting for compact and efficient reconstruction of dynamic scene. *arXiv preprint arXiv:2503.12307*, 2025. 2

[70] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time Neural Irradiance Fields for Free-viewpoint Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9421–9431, 2021. 2

[71] Wangze Xu, Huachen Gao, Shihe Shen, Rui Peng, Jianbo Jiao, and Ronggang Wang. Mvpgs: Excavating multi-view priors for gaussian splatting from sparse input views. In *European Conference on Computer Vision*, pages 203–220. Springer, 2024. 2

[72] Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 4k4d: Real-time 4d view synthesis at 4k resolution. In *CVPR*, 2024. 2

[73] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 2

[74] Jinbo Yan, Rui Peng, Zhiyan Wang, Luyang Tang, Jiayu Yang, Jie Liang, Jiahao Wu, and Ronggang Wang. Instant gaussian stream: Fast and generalizable streaming of dynamic scene reconstruction via gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16520–16531, 2025. 2

[75] Yifan Zhan, Qingtian Zhu, Muyao Niu, Mingze Ma, Jiancheng Zhao, Zhihang Zhong, Xiao Sun, Yu Qiao, and Yinqiang Zheng. Tomie: Towards modular growth in enhanced smpl skeleton for 3d human with animatable garments. *arXiv preprint arXiv:2410.08082*, 2024. 2

[76] Yifan Zhan, Wangze Xu, Qingtian Zhu, Muyao Niu, Mingze Ma, Yifei Liu, Zhihang Zhong, Xiao Sun, and Yinqiang Zheng. R3-avatar: Record and retrieve temporal codebook for reconstructing photorealistic human avatars. *arXiv preprint arXiv:2503.12751*, 2025. 2

[77] 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. 5

[78] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gpsgaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19680–19690, 2024. 2

[79] Yang Zheng, Qingqing Zhao, Guandao Yang, Wang Yifan, Donglai Xiang, Florian Dubost, Dmitry Lagun, Thabo Beeler, Federico Tombari, Leonidas Guibas, et al. Physavatar: Learning the physics of dressed 3d avatars from visual observations. *arXiv preprint arXiv:2404.04421*, 2024. 2, 8

[80] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *European conference on computer vision*, pages 145–163. Springer, 2024. 2

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