Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PoseRoute: Selective Pose Conditioning for Anchor-Structured Gaussian Avatars from Monocular Video

Xianguo Liu

xgliu@fosu.edu.cn

Foshan University

Weina Cao

Foshan University

Fen Liu

Foshan University

Wenyin Yang

Foshan University

Li Ma

Foshan University

Haoliang Qi

Foshan University

Research Article

Keywords:

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11499241/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

PoseRoute: Selective Pose Conditioning for Anchor-Structured Gaussian Avatars from Monocular Video

Xianguo Liu[1,2*] · Weina Cao[1] · Fen Liu[1,2] · Wenyin Yang[1,2] · Li Ma[1,2] · Haoliang Qi[1]

Abstract

Reconstructing animatable humans from monocular video becomes particularly challenging when avatars are driven by poses beyond those observed during training, as locally generated Gaussian attributes may become inconsistent with articulated body support. To address this challenge, we present PoseRoute, an anchor-structured Gaussian avatar framework with selective pose conditioning. Within articulated anchor–child groups, body-aligned anchors follow skeletal motion, while shared decoders generate local child Gaussians. Pose information is routed only to dynamic child offsets through a root-canonicalized local pose representation, while opacity, appearance, and canonical Gaussian shape are decoded from anchor features without direct pose conditioning. Anchor-relative offset and scale constraints further bound local Gaussian geometry during novel-pose driving. Experiments on People-Snapshot, UBC-Fashion, and ZJU-MoCap demonstrate competitive reconstruction quality with real-time rendering at up to 119 FPS. In paired evaluations against our reproduced GART baseline, PoseRoute improves articulated-boundary fidelity. Controlled root-yaw tests further yield 4.6×, 4.7×, and 10.5× reductions in alpha inconsistency, boundary Chamfer distance, and Gaussian drift, respectively. These results show that structured support with controlled pose-dependent Gaussian generation improves avatar consistency under novel-pose driving.

Keywords 3D Gaussian splatting · Animatable human avatars · Monocular video reconstruction · Anchor-structured representation · Selective pose conditioning

1 Introduction

Reconstructing animatable human avatars from monocular videos is an important problem for telepresence, AR/VR, gaming, and digital content creation. This task is challenging because monocular observations provide limited viewpoint and pose coverage, while the human body exhibits complex articulation and non-rigid deformation. Parametric body models such as SMPL [1] provide strong priors for body shape and articulation, offering a useful foundation for human reconstruction. However, these models alone are insufficient for capturing personalized geometry and appearance, motivating richer and more flexible representations. Neural human rendering methods address this requirement through canonical radiance fields [2] or structured neural representations [3], enabling more detailed modeling of individual subjects. Meanwhile, interactive avatar applications impose strict computational constraints and require efficient rendering and animation strategies for real-time use.

3D Gaussian Splatting [4] provides an appealing explicit representation by combining differentiable Gaussian primitives with high-quality real-time rendering. Recent Gaussian avatar approaches integrate this representation with body templates or learned deformation fields to

* Xianguo Liu (Corresponding author)
xgliu@fosu.edu.cn

[1] School of Computer Science and Artificial Intelligence, Foshan University, Foshan, China

[2] Guangdong Engineering Research Center for Manufacturing Data Security Governance, Foshan, China

reconstruct animatable humans from monocular videos [5–8]. However, novel-pose animation imposes stronger requirements on the representation: local support and Gaussian attributes must remain coherent beyond the observed pose distribution. The key challenge is therefore to preserve consistent appearance, local geometric support, and articulated boundaries when animating under unseen poses.

Structured Gaussian representations provide a natural way to enhance local spatial organization by coupling multiple primitives to shared geometric support. Recent scene- and avatar-level methods demonstrate that anchor- or surface-supported Gaussians can improve local coherence and provide structured deformation priors [9–12]. For articulated human avatars, skeletal transformations provide a basis for global body motion, while local non-rigid variations require additional deformation modeling beyond body-scale articulation. Pose information offers an effective control signal for these local changes, but novel-pose animation requires careful control over how pose-dependent information is incorporated into Gaussian representations, so that deformation adapts to articulation while canonical appearance and local geometry remain consistent. This motivates a structured Gaussian representation that explicitly controls pose-responsive local deformation under an articulated anchor framework.

We propose PoseRoute, an anchor-structured Gaussian avatar framework that uses selective pose conditioning for monocular human animation under novel poses. Within articulated anchor–child groups, body-aligned anchors follow skeletal motion through voxel-field skinning, while shared decoders generate local child Gaussians. Pose information is routed only to dynamic child offsets through a local pose representation, while appearance, opacity, and canonical Gaussian shape are decoded from anchor features without direct pose conditioning. Anchor-relative geometry constraints further regulate local deformation, and the framework retains explicit Gaussian rasterization for real-time animation.

We evaluate PoseRoute on People-Snapshot [13], UBC-Fashion [14], and ZJU-MoCap [3]. Across these benchmarks, PoseRoute achieves competitive reconstruction quality while maintaining real-time rendering at up to 119 FPS. Paired evaluations against our reproduced GART baseline and controlled robustness analyses demonstrate improved consistency under novel-pose driving. These results show that PoseRoute provides structured and efficient Gaussian generation for monocular human animation under novel poses.

In summary, our main contributions are as follows:

1. We design an articulated anchor–child Gaussian avatar framework that combines body-aligned local support with voxel-field skinning, providing a structured basis for local Gaussian generation under monocular supervision.
2. We introduce selective pose conditioning and anchor-relative geometry control: pose modulates only dynamic child offsets, while appearance, opacity, and canonical Gaussian shape remain anchor-conditioned, with offset and scale constraints preserving local geometry under novel poses.
3. We validate PoseRoute on People-Snapshot, UBC-Fashion, and ZJU-MoCap, demonstrating strong reconstruction quality and real-time rendering, together with targeted improvements in articulated-boundary consistency and reduced sensitivity to global orientation changes under novel-pose driving.

2 Related Work

Neural Human Avatars. Reconstructing animatable humans from monocular observations requires jointly modeling articulation, non-rigid clothing motion, self-occlusion, and personalized appearance under limited viewpoint and pose coverage. Parametric body models such as SMPL [1] provide strong geometric and kinematic priors, and performance-capture and neural rendering methods [2, 3, 13, 15, 16] build on these priors with learned geometry, appearance, or canonical representations. Many neural avatar approaches adopt radiance-field representations derived from NeRF [17]. For deformation between canonical and posed spaces, SNARF and Fast-SNARF introduce forward-skinning formulations [18, 19], whereas HumanNeRF learns backward deformation mappings [2]. Articulated implicit surfaces [20] and template-free volumetric actors [21]

provide additional representation choices for animatable humans. More recent monocular techniques further refine deformation and appearance modeling by using bidirectional deformation constraints and keyframe feature correspondences in MonoHuman [22], as well as cross-frame temporal information in TMIHuman [23]. While these neural representations support high-quality novel-view synthesis and re-animation, explicit rasterization provides a more direct path to real-time avatar rendering.

3D Gaussian Human Avatars. 3D Gaussian Splatting [4] replaces volumetric sampling with explicit anisotropic primitives and differentiable rasterization, enabling efficient dynamic and human representations. Dynamic Gaussian methods model time-varying scenes through persistent Gaussian tracking [24] and learned deformation fields [25], including deformations predicted from 4D neural voxel features [26]. For scene reconstruction, Adaptive Control [27] improves 3D Half-Gaussian Splatting through adaptive opacity consistency, selective normal smoothness, and normal anchoring. This motivates examining attribute and geometry control in articulated avatars, where local Gaussian structure must remain coherent under body motion.

Existing Gaussian-avatar methods can be broadly grouped into three complementary directions. The first direction focuses on articulated deformation: methods such as *GART* [5], GauHuman [6], 3DGS-Avatar [7], and Drivable 3D Gaussian Avatars [28] combine canonical Gaussians with skinning, learned non-rigid deformation, or parametric control. A second direction emphasizes geometric support: methods including SplattingAvatar [29], GoMAvatar [30], RMAvatar [31], and GenericAvatar [32] organize or constrain Gaussians with respect to tracked human geometry, with GenericAvatar further applying pose-conditioned Gaussian adjustment for non-rigid detail. The third direction addresses limited monocular observation and pose coverage: VGA [8] improves pose alignment and Gaussian distribution, whereas KungFuCap [33] augments monocular observations with pose-conditioned complementary views and temporal guidance for rapid motion. Adjacent work further explores structure priors, pose guidance, and spatially distributed predictors through HumanSplat [34], PoseGaussian [35], and related real-time Gaussian-avatar models [36]. These studies show that novel-pose behavior depends on articulated deformation, geometric support, and the information used to condition Gaussian attributes.

Anchor-Structured Gaussian Representations. A key question for structured Gaussian avatars is how to preserve explicit rendering efficiency while organizing primitives into coherent local groups. Scaffold-GS [9] introduces anchors with learnable features from which shared decoders generate child Gaussians, establishing anchor-level parameter sharing. Scene-level extensions such as Octree-GS [37] and SOGS [38] further develop hierarchical and feature-rich anchor representations. For avatars, ScaffoldAvatar [10] adapts anchor-conditioned generation to expression-driven head modeling, while related point- and rigged-Gaussian approaches [39, 40] model local head dynamics. Full-body methods bring structured support into articulated motion: StruGauAvatar [12] constrains Gaussians with a learned surface and dual-space optimization, whereas LoGAvatar [11] predicts local Gaussian groups around SMPL-surface anchors and animates them through learnable forward skinning. SubGAva [41] explores Gaussian primitive subdivision, predicting child attributes from fused geometry, appearance, and pose features while constraining child offsets and anisotropy. *PoseRoute* combines body-aligned local support with selective pose conditioning: pose information is routed only to dynamic child offsets, while appearance, opacity, and canonical Gaussian shape are decoded from anchor features without direct pose conditioning. Anchor-relative offset and scale controls further bound local geometry under novel poses.

3 Method

PoseRoute separates skeletal motion from pose-dependent local displacement within an articulated anchor-child representation. Shared decoders generate local Gaussian attributes from anchor features, while a root-canonicalized pose latent modulates child offsets. Offset and scale controls couple this selective pose dependence with bounded local geometry. Fig. 1 illustrates

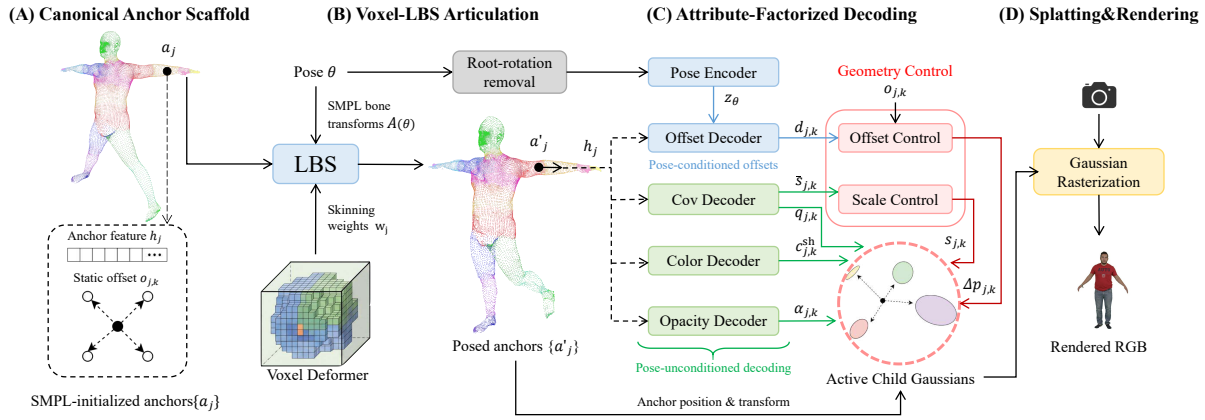


Fig. 1 Overview of *PoseRoute*. *PoseRoute* represents an animatable avatar through a structured anchor-child Gaussian scaffold that separates skeletal motion from pose-dependent local deformation. Left: A subdivided SMPL surface initializes body-aligned anchors with shared features and candidate child Gaussians. Middle: Voxel-LBS articulates the scaffold, while attribute-factorized decoders preserve canonical appearance and Gaussian shape by routing pose information only to dynamic offsets. Geometry controls constrain anchor-relative displacement and decoded scale anisotropy. Right: The resulting posed Gaussians are rendered through differentiable Gaussian splatting

the complete pipeline for monocular human reconstruction and animation.

3.1 Preliminaries: Template-Based Gaussian Avatars

Template-based Gaussian avatars represent a human in canonical space and deform its geometry to a target pose using a parametric body template. The canonical representation consists of a set of Gaussians:

$$\mathcal{G} = \{(\mu^{(i)}, R^{(i)}, s^{(i)}, \alpha^{(i)}, f^{(i)})\}_{i=1}^{N}, \tag{1}$$

where each Gaussian is parameterized by its center μ, rotation R, scale s, opacity α, and color feature f.

Given a target pose θ, the body template provides bone transforms $\{A_k(\theta)\}_{k=1}^{n_b}$ from canonical to posed space. Linear blend skinning combines these transforms according to the skinning weights at a canonical point x:

$$T(x,\theta) = \sum_{k=1}^{n_b} W_k(x) A_k(\theta), \qquad \bar{x}' = T(x,\theta)\bar{x}. \tag{2}$$

Here, $W_k(x)$ is the skinning weight associated with bone k, and $\bar{x}$ and $\bar{x}'$ denote the homogeneous coordinates of the canonical and posed points. The blended transform determines how the point follows the articulated body.

GART [5] implements this template-based formulation with voxel-based skinning and neighborhood regularization of canonical Gaussian attributes. These representation and deformation components provide the foundation for the anchor-child formulation described below.

3.2 Canonical Anchor Scaffold

Anchor Initialization. As shown in Fig. 1(A), we initialize a canonical anchor scaffold

$$\mathcal{A} = \left\{\left(\mathbf{a}_j,\, \mathbf{h}_j,\, \{\mathbf{o}_{j,k}\}_{k=1}^{K}\right)\right\}_{j=1}^{M}, \tag{3}$$

where $\mathbf{a}_j \in \mathbb{R}^3$ denotes the canonical position of anchor j, $\mathbf{h}_j \in \mathbb{R}^{d_f}$ is its learnable feature shared by its child Gaussians, and $\mathbf{o}_{j,k} \in \mathbb{R}^3$ is the learnable static offset of child k relative to that anchor. We subdivide the SMPL mesh once and use all resulting vertices as canonical anchor positions. At each position, we instantiate an anchor with a learnable feature vector and $K = 20$ learnable static offsets, one for each candidate child Gaussian.

Anchor-Based Local Generation. Each anchor supports a local group of K candidate

child Gaussians that represent the surrounding appearance and geometry. The attributes of children attached to the same anchor are generated by shared Gaussian decoders using their common parent feature, linking attribute predictions within the local group. This organization provides a body-aligned spatial prior for local Gaussian generation under limited monocular observations.

Adaptive Anchor Refinement. We adapt *GART*'s gradient-driven density-control strategy to refine complete anchor-child groups [5]. Image-space center-gradient magnitudes are accumulated from active child Gaussians and averaged at their parent anchors to guide anchor cloning and splitting. An anchor is pruned when the activated opacity of every candidate child falls below τ_{prune}. This process adjusts the anchor count M while preserving K candidate children per anchor. The refinement schedule, cloning and splitting operations, and dataset-specific thresholds are described in Appendix B.3.

3.3 Voxel-LBS Articulation

Given a target pose θ, we combine each anchor's skinning coefficients with the corresponding bone transforms to obtain its posed position and affine transform, as shown in Fig. 1(B).

A Voxel Deformer represents a canonical-space skinning-coefficient field with a base voxel grid initialized from normalized SMPL skinning weights and a learnable correction voxel grid. Interpolating the sum of these grids at canonical anchor position $\mathbf{a}_j$ gives $\mathbf{w}_j = \mathcal{W}_{\text{vox}}(\mathbf{a}_j)$, whose entries are the blending coefficients for the $n_b = 24$ SMPL bones.

Applying the forward-skinning formulation in Eq. 2 at each anchor and incorporating the per-frame global translation gives the blended transform and posed anchor position:

$$\begin{aligned} \widehat{T}_j(\theta) &= \sum_{k=1}^{n_b} w_{j,k} A_k(\theta), \\ \mathbf{a}'_j &= \mathbf{B}_j \mathbf{a}_j + \mathbf{t}_j + \mathbf{g}. \end{aligned} \tag{4}$$

Here, $\mathbf{B}_j = \widehat{T}_{j,1:3,1:3}$ and $\mathbf{t}_j = \widehat{T}_{j,1:3,4}$ are the linear and translation blocks, and $\mathbf{g}$ is the per-frame global translation shared by all anchors. Blending multiple bone transforms allows $\mathbf{B}_j$ to include local stretching or shear.

For the optional loose-garment configuration, we adopt *GART*'s residual-bone formulation [5]. At each anchor, we weight the pose-dependent rigid transforms of the residual bones using additional coefficients queried from the voxel field and add their weighted sum to the SMPL bone blend in Eq. 4. We extract $\mathbf{B}_j$ and $\mathbf{t}_j$ from the resulting augmented transform and compute $\mathbf{a}'_j$ using the position expression in Eq. 4.

3.4 Attribute-Factorized Decoding

Our decoding design aims to capture pose-dependent local deformation while maintaining a consistent canonical description of appearance and Gaussian shape. We achieve this through attribute-factorized decoding, conditioning dynamic child offsets on pose while decoding opacity, SH color coefficients, and canonical scale and rotation from anchor features.

Root-Rotation-Removed Pose Encoding. The full pose θ supplies global orientation to LBS, while the Offset Decoder receives a latent describing local articulation. For the flattened axis-angle SMPL pose vector θ, we zero the first three root-rotation entries while retaining the remaining local joint rotations to obtain $\widetilde{\theta}$, which we then encode using the Pose Encoder E_{pose}:

$$\begin{aligned} \widetilde{\theta}_{1:3} &= \mathbf{0}, \qquad \widetilde{\theta}_{4:} = \theta_{4:}, \\ \mathbf{z}_\theta &= E_{\text{pose}}(\widetilde{\theta}). \end{aligned} \tag{5}$$

Selective Pose Conditioning. For each anchor, the Offset Decoder conditions dynamic-offset residuals on both the anchor feature and the local-articulation latent:

$$\{\mathbf{d}_{j,k}\}_{k=1}^{K} = F_{\text{off}}(\mathbf{h}_j, \mathbf{z}_\theta). \tag{6}$$

Canonical shape, SH color coefficients, and opacity are decoded from the anchor feature through separate pose-unconditioned branches:

$$\begin{aligned} \{(\bar{\mathbf{s}}_{j,k}, \mathbf{q}_{j,k})\}_{k=1}^{K} &= F_{\text{cov}}(\mathbf{h}_j), \\ \{\mathbf{c}^{\text{sh}}_{j,k}\}_{k=1}^{K} &= F_{\text{color}}(\mathbf{h}_j), \\ \{\widetilde{\alpha}_{j,k}\}_{k=1}^{K} &= F_{\alpha}(\mathbf{h}_j). \end{aligned} \tag{7}$$

The Cov Decoder includes scale activation and quaternion normalization, yielding positive canonical scales $\bar{\mathbf{s}}_{j,k}$ and normalized rotation quaternions $\mathbf{q}_{j,k}$. The logits $\widetilde{\alpha}_{j,k}$ output by the Opacity Decoder are converted by a sigmoid to opacities $\alpha_{j,k} = \sigma(\widetilde{\alpha}_{j,k})$. We select active child Gaussians based on their decoded opacities, with index set $\mathcal{I}_{\text{act}} = \{(j,k) \mid \alpha_{j,k} > \tau_{\text{act}}\}$ and $\tau_{\text{act}} = 0.1$. If the set is empty, we retain all candidate child Gaussians.

Geometry Control. Offset Control combines the learnable static offset and pose-conditioned dynamic residual as $\mathbf{r}_{j,k} = \mathbf{o}_{j,k} + \gamma \mathbf{d}_{j,k}$ in canonical anchor-relative coordinates, with $\gamma = 0.05$. A numerically stabilized radial projection bounds the resulting child-center displacement $\Delta\mathbf{p}_{j,k}$ relative to its parent anchor:

$$\Delta\mathbf{p}_{j,k} = \mathbf{r}_{j,k} \frac{\min(\|\mathbf{r}_{j,k}\|_2, \delta_{\max})}{\|\mathbf{r}_{j,k}\|_2 + \epsilon}. \tag{8}$$

Here, $\delta_{\max} = 0.01\,\text{m}$ bounds the offset within a canonical neighborhood centered on the parent anchor with a radius of 1 cm, and $\epsilon = 10^{-9}$ ensures a nonzero denominator.

Scale Control constrains the ratio between the largest and smallest canonical scales. The decoded scales $\bar{\mathbf{s}}_{j,k}$ lie within $[s_{\min}, s_{\max}] = [0.0005, 0.05]\,\text{m}$. We obtain the controlled scales $\mathbf{s}_{j,k}$ by raising each axis scale to at least $1/\rho_{\max}$ of the largest decoded scale:

$$s_{j,k,\ell} = \max\left(\bar{s}_{j,k,\ell}, \frac{\max_{u\in\{1,2,3\}} \bar{s}_{j,k,u}}{\rho_{\max}}\right), \quad \ell \in \{1,2,3\}. \tag{9}$$

Here, $\rho_{\max} = 20$ bounds the largest-to-smallest scale ratio of each canonical child Gaussian.

Posed Child Gaussian Construction. We construct each child Gaussian in posed space by combining its controlled local geometry with the parent anchor's posed position $\mathbf{a}'_j$ and affine linear block $\mathbf{B}_j$ from Sec. 3.3. The linear block transforms the child offset and canonical rotation, yielding the child center and covariance:

$$\begin{aligned} \mu'_{j,k} &= \mathbf{a}'_j + \mathbf{B}_j \Delta\mathbf{p}_{j,k}, \\ \mathbf{\Sigma}'_{j,k} &= \mathbf{F}'_{j,k}\,\text{diag}(\mathbf{s}^2_{j,k})(\mathbf{F}'_{j,k})^{\mathsf{T}}. \end{aligned} \tag{10}$$

Here, $\mathbf{F}'_{j,k} = \mathbf{B}_j \mathcal{R}(\mathbf{q}_{j,k})$ is the posed affine Gaussian frame, and $\mathcal{R}(\mathbf{q}_{j,k})$ converts the normalized quaternion into a canonical local rotation matrix.

3.5 Rendering and Optimization

We transform the active child Gaussians into the target camera coordinates and evaluate their decoded SH coefficients $\mathbf{c}^{\text{sh}}_{j,k}$ using view-dependent SH evaluation as in *GART* [5]. The resulting colors, together with the Gaussian centers, covariances, and opacities $\alpha_{j,k}$, are passed to the differentiable Gaussian rasterizer to form the target image [4], as shown in Fig. 1(D).

Training Objective. Given a ground-truth monocular RGB frame I^* and rendered image $\hat{I}$, we optimize

$$L = L_1(\hat{I}, I^*) + \lambda_{\text{SSIM}} L_{\text{SSIM}}(\hat{I}, I^*) + L_{\text{reg}}, \tag{11}$$

where $L_{\text{SSIM}}(\hat{I}, I^*) = 1 - \text{SSIM}(\hat{I}, I^*)$ uses the structural similarity index [42]. The regularization term is

$$\begin{aligned} L_{\text{reg}} = \lambda_{\text{off}} L_{\text{off}} + \lambda_{\text{ratio}} L_{\text{scale}} + \lambda_{\max} L_{\max} \\ + \lambda_{\text{feat}} L^{\text{KNN}}_{\text{feat}} + \lambda_{\text{aoff}} L^{\text{KNN}}_{\text{aoff}} + L_{\text{vox}}. \end{aligned} \tag{12}$$

We adapt *GART*'s neighborhood regularization to anchor features and static child offsets through $L^{\text{KNN}}_{\text{feat}}$ and $L^{\text{KNN}}_{\text{aoff}}$, respectively. The two terms average neighborhood-wise sample standard deviations over attribute components and anchors. We retain *GART*'s voxel-correction regularization L_{vox}, which combines weighted TV and magnitude penalties. Loss weights and optimization schedules are given in Appendix B.3.

Geometry Regularization. The hard controls bound child offsets and decoded scale ratios, while the regularizers penalize static-offset magnitude, scale anisotropy above a lower threshold, and large canonical scales during optimization. We apply a magnitude penalty to all static child offsets:

$$L_{\text{off}} = \frac{1}{MK} \sum_{j=1}^{M} \sum_{k=1}^{K} \|\mathbf{o}_{j,k}\|_2^2. \tag{13}$$

For the controlled canonical scales of active child Gaussians, we penalize anisotropy ratios

above a threshold:

$$L_{\text{scale}} = \frac{1}{|\mathcal{I}_{\text{act}}|} \sum_{(j,k)\in\mathcal{I}_{\text{act}}} \max(0, \rho_{j,k} - \rho_0)^2. \quad (14)$$

Here, $\rho_{j,k} = \max(\mathbf{s}_{j,k})/(\min(\mathbf{s}_{j,k}) + \epsilon_{\text{ratio}})$ is the stabilized scale ratio, $\rho_0 = 10$ is the penalty threshold, and $\epsilon_{\text{ratio}} = 10^{-8}$ provides numerical stability.

Following *GART*, $L_{\max}$ is the mean squared largest controlled canonical scale over active child Gaussians, discouraging excessively large local support.

4 Experiments

We assess reconstruction quality on three human-avatar benchmarks and boundary quality under novel poses on the four People-Snapshot subjects. Component ablations examine the effects of anchor initialization density, selective pose conditioning, and geometry control. Controlled root-yaw tests evaluate global-orientation robustness, while joint-bending tests isolate the geometric effects of scale control. Finally, we report representation size and runtime to characterize computational cost.

4.1 Experimental Setup

Datasets and Protocols. We evaluate four People-Snapshot subjects [13] using the InstantAvatar-preprocessed splits and poses [16], and six loose-clothing sequences from UBC-Fashion [14] following *GART*'s protocol [5]. These datasets are evaluated on held-out frames with white backgrounds. On ZJU-MoCap [3], we evaluate six subjects using the *GART* protocol, training each avatar from one camera and testing on the remaining cameras with black backgrounds.

Implementation and Evaluation. All *PoseRoute* models are optimized for 30,000 iterations using Adam [43] on a single NVIDIA A800 GPU. For quantitative evaluation, we refine the SMPL pose and global translation for the test images while keeping the learned avatar parameters fixed. We report PSNR, SSIM, and LPIPS [44] averaged equally across sequences, with LPIPS on its original scale. Implementation details are provided in Appendix B.

For paired evaluation on People-Snapshot, *PoseRoute* and our *GART* reproduction use the same 190 test frames, evaluator, and test-time pose-refinement protocol. We report 95% hierarchical-bootstrap confidence intervals over the four subjects and their paired samples. Complete results and evaluation details are provided in Appendix A.

4.2 Reconstruction Quality

People-Snapshot. *PoseRoute* improves PSNR, SSIM, and LPIPS over the published *GART* results on all four People-Snapshot subjects, achieving the highest PSNR and SSIM among the compared methods on male-3-casual (Table 1). Averaged equally across the four subjects, LPIPS decreases by approximately 5.2% relative to *GART*. Fig. 2 compares test-frame reconstructions with canonical- and novel-pose renderings, with enlarged regions highlighting differences in reconstructed appearance and garment boundaries.

UBC-Fashion. UBC-Fashion evaluates reconstruction of dynamic loose garments whose surfaces extend beyond the SMPL body template. In this setting, *PoseRoute* achieves higher PSNR and SSIM than both *GART* variants, even in its default configuration (Table 2). For this loose-garment setting, we further evaluate an extended *PoseRoute* configuration with learned residual bones to provide additional deformation flexibility. With $n_r = 32$ learned residual bones, *PoseRoute* reaches 26.60 PSNR and 0.943 SSIM, exceeding the reported *GART* T-Table configuration by 0.64 dB and 0.008, respectively, while achieving nearly identical LPIPS. Fig. 3 highlights reduced reconstruction errors around the arms and garment boundaries in the displayed examples. These results indicate that the anchor-child representation captures loose-garment appearance and geometry, while the extended configuration benefits from the additional deformation flexibility provided by residual bones.

ZJU-MoCap. On held-out ZJU-MoCap views, *PoseRoute* achieves 31.79 PSNR, 0.975 SSIM, and 0.03018 LPIPS, outperforming GauHuman on all three metrics and yielding higher PSNR

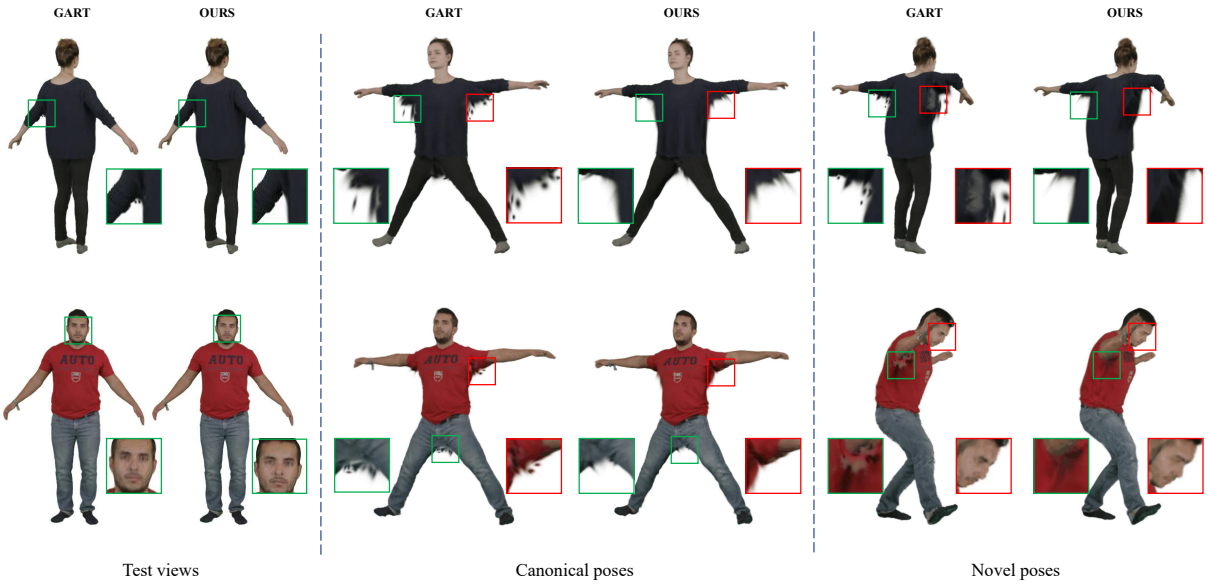


Fig. 2 Qualitative comparison of *GART* and *PoseRoute* on People-Snapshot [13]. Close-ups highlight appearance details and garment boundaries, with matching green and red boxes indicating corresponding regions

Table 1 Quantitative comparison on People-Snapshot [13]. Baseline results are taken from published literature. Best and second-best results are highlighted in **bold** and underlined, respectively.

	male-3-casual			male-4-casual			female-3-casual			female-4-casual		
Method	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Neural-field and hybrid baselines												
Neural Body [3]	24.94	0.9428	0.0326	24.71	0.9469	0.0423	23.87	0.9504	0.0346	24.37	0.9451	0.0382
InstantAvatar [16]	29.65	0.9730	0.0192	27.97	0.9649	0.0346	**27.90**	**0.9722**	0.0249	28.92	0.9692	0.0180
Gaussian-avatar methods												
GART [5]	30.40	0.9769	0.0377	27.57	0.9657	0.0607	26.26	0.9656	0.0498	29.23	0.9720	0.0378
LoGAvatar [11]	30.56	0.9791	**0.0163**	**28.31**	**0.9701**	**0.0294**	27.60	0.9700	**0.0223**	**29.99**	**0.9763**	**0.0164**
PoseRoute	**30.73**	**0.9799**	0.0376	27.66	0.9697	0.0564	26.34	0.9677	0.0466	29.44	0.9749	0.0357

Table 2 Quantitative comparison of view synthesis on UBC-Fashion [14]. Metrics are averaged across six sequences. n_r denotes the number of residual bones; the unqualified *PoseRoute* row uses the default $n_r = 0$ configuration. Best results are in **bold**, and second-best results are underlined.

Methods	PSNR↑	SSIM↑	LPIPS↓
InstantAvatar [16] Test Pose	19.87	0.880	0.15770
InstantAvatar [16] Training Pose	19.97	0.882	0.15705
GART MLP [5]	25.65	0.934	0.08188
GART T-Table [5]	25.96	0.935	0.08057
PoseRoute	26.31	0.942	0.08760
PoseRoute ($n_r = 32$)	**26.60**	**0.943**	**0.08055**

and lower LPIPS than the listed neural-rendering baselines (Table 3). Fig. 4 complements these results with novel-view reconstructions and pose-driven renderings. The animation examples preserve subject-specific clothing appearance under external motion driving and when rendered at varying global orientations.

4.3 Generalization to Novel Poses

We first evaluate animation under external driving motions that are not paired with target images. Following *GART*, we use four external motion sequences: eating, fencing, walking, and the `aist_demo` sequence (Appendix A.2). Fig. 5 compares reproduced *GART* and *PoseRoute* under these shared external motions, using representative frames from the fencing and `aist_demo` sequences. As the arms move away from the torso,

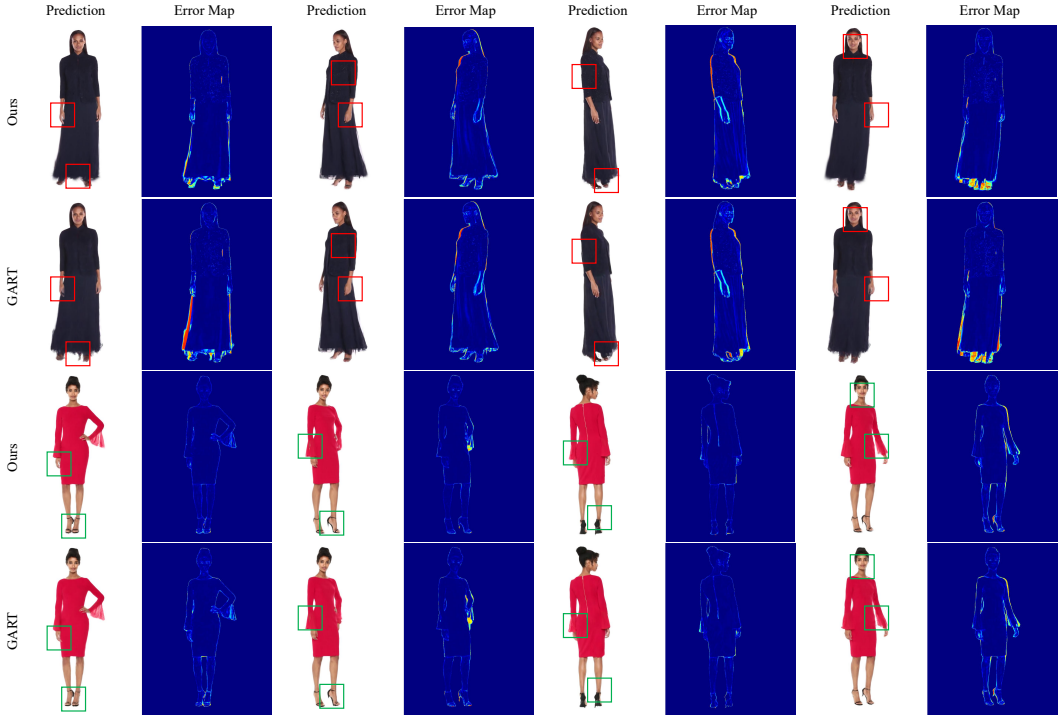


Fig. 3 Qualitative comparison on UBC-Fashion [14]. Rendered images and RGB reconstruction error maps are shown with a shared error color scale, from blue (low error) to red (high error). *PoseRoute* uses $n_r = 32$ in both this figure and the accompanying UBC-Fashion supplementary video

Table 3 Quantitative comparison on ZJU-MoCap [3], averaged over six subjects. Baseline results are taken from published literature. Best and second-best results are highlighted in bold and underlined, respectively.

Methods	PSNR↑	SSIM↑	LPIPS↓
Neural-rendering baselines			
HumanNeRF [2]	30.66	0.969	0.03338
NB [3]	29.03	0.964	0.04247
AN [45]	29.77	0.965	0.04689
NHP [46]	28.25	0.955	0.06477
PixelNeRF [47]	24.71	0.892	0.12186
Instant-NVR [48]	31.01	0.971	0.03845
Gaussian-avatar methods			
GauHuman [6]	31.34	0.965	0.03051
GART (30 s) [5]	31.76	<u>0.976</u>	0.03401
GART (2.5 min) [5]	**32.22**	**0.977**	**0.02921**
PoseRoute	<u>31.79</u>	0.975	<u>0.03018</u>

reproduced *GART* exhibits elongated dark structures below the raised arms, whereas *PoseRoute* maintains more compact underarm and sleeve boundaries in the displayed examples. Because both methods are driven by identical poses and cameras, these comparisons isolate differences in how their learned representations deform under the same novel articulations. We quantify structural artifacts under external motions using disconnected foreground components, a measure that does not require paired ground-truth images. Across all four subjects and four actions, with 480 evaluated frames per method, *PoseRoute* produces no disconnected foreground components of at least 25 pixels. In contrast, reproduced *GART* yields a small number of extra components on the eating, fencing, and `aist_demo` motions, while both methods produce none on walking.

We further evaluate boundary fidelity under novel poses on held-out People-Snapshot frames. Across the 190 held-out frames, *PoseRoute* reduces boundary-band RGB MAE by 0.004761 and foreground spill ratio by 0.000527 relative to reproduced *GART*; both confidence intervals lie entirely above zero, and all four subjects show improvement. On the high pose novelty subset, defined by the upper tertile of pose novelty among

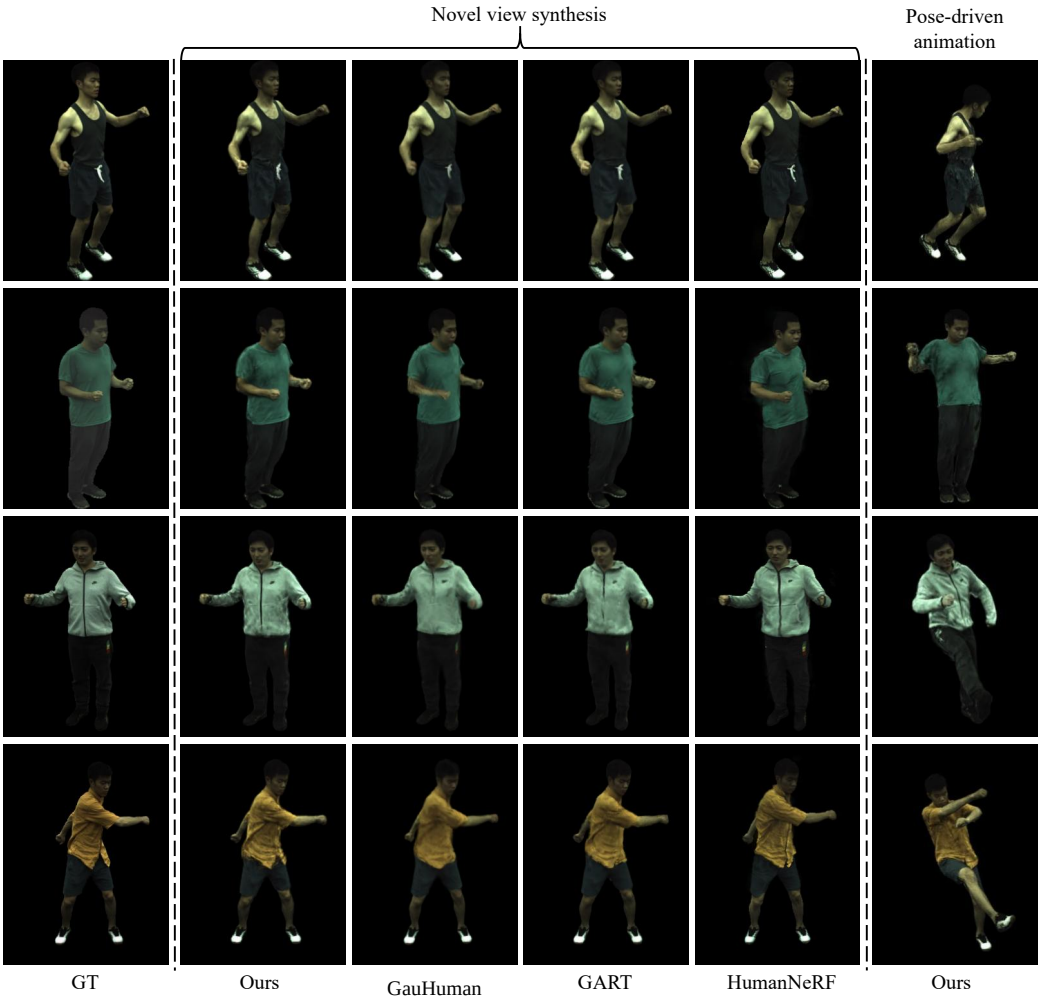


Fig. 4 Qualitative results on ZJU-MoCap [3]. Left: novel-view reconstruction comparisons. Right: *PoseRoute* animations driven by the external `aist_demo` motion (top two rows) and training motions rendered under varying global orientations (bottom two rows)

all test frames, boundary-band MAE decreases from 0.07068 to 0.06797, corresponding to a paired reduction of 0.002709 with a 95% confidence interval of $[0.000784, 0.004746]$. Pose novelty is defined by the nearest-training-pose distance over the 23 non-root joints, and boundary-band MAE is measured within a 5-pixel band around the ground-truth foreground contour. The boundary-aware evaluation confirms reduced errors near the silhouette under novel articulation. The external-motion comparisons, structural diagnostic, and GT-based evaluation indicate that *PoseRoute* maintains more stable body and garment boundaries under novel articulation. Appendix Tables 10 and 11 report the complete paired and external-motion results.

4.4 Ablations and Robustness Analysis

We analyze key design choices underlying anchor organization, pose conditioning, and local geometry control through component ablations and targeted robustness evaluations. Tables 4 and 5 report four-subject reconstruction means and paired effects on the high-pose-novelty subset, while Fig. 6 compares all six variants at an identical external eating pose and camera. Table 6



Fig. 5 Animation under shared external motions. Six uniformly spaced frames of female-3-casual are shown for the fencing motion (top panel) and the `aist_demo` motion (bottom panel), with *GART* above *PoseRoute* in each panel. Underarm and sleeve artifacts visible in *GART* are reduced in the displayed *PoseRoute* renderings. Both methods use the same driving inputs and camera. The corresponding motion sequences are shown in the supplementary video.

Table 4 Quantitative ablation study on the four People-Snapshot sequences. Values are unweighted averages of the four sequence-level means, with reconstruction metrics computed on held-out test frames. Full denotes the complete PoseRoute model. The Full row is shown in **bold**.

Variant	PSNR↑	SSIM↑	LPIPS↓	FPS↑	Train (min)↓
Full	**28.54**	**0.9731**	**0.0441**	**115.97**	**23.26**
w/o selective pose conditioning	28.63	0.9734	0.0449	120.58	22.02
w/o root-rotation removal	28.59	0.9731	0.0445	117.67	22.69
w/o anchor subdivision	28.52	0.9725	0.0482	129.11	17.19
w/o offset constraints	23.89	0.9456	0.0889	142.81	13.89
w/o scale constraint	28.57	0.9732	0.0432	117.46	22.68

Table 5 Component degradation under high pose novelty. Each entry reports Δ = w/o − Full; positive values indicate degradation after removing a component. Bold entries denote statistically significant degradation.

Removed component	Δ LPIPS	Δ B-band MAE
selective pose conditioning	**+0.00148**	+0.00032
root-rotation removal	+0.00044	-0.00022
anchor subdivision	**+0.00394**	**+0.00108**
offset constraints	**+0.04357**	**+0.09284**
scale constraint	-0.00063	-0.00066

summarizes the root-yaw robustness results and the scale-control results.

4.4.1 Component Ablations

Anchor Organization and Locality. Dense anchor coverage and constrained anchor-relative displacement jointly support local Gaussian generation. Initializing anchors from the original SMPL vertices instead of the subdivided surface raises mean LPIPS to 0.0482. In this setting, the high-novelty LPIPS and boundary-band penalties both have confidence intervals above zero, supporting denser anchor initialization for

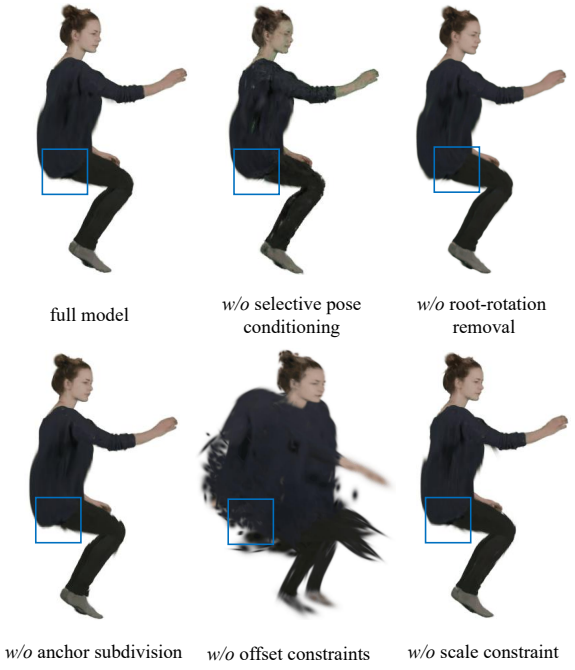


Fig. 6 Qualitative ablation study. People-Snapshot female-3-casual is rendered from the same frame of the external eating motion using the same camera for all variants. Blue boxes mark a shared rear garment-boundary region for direct comparison

improved local coverage. Offset constraints further preserve locality after anchor initialization. Removing dynamic-offset attenuation, the 1 cm projection, and static-offset regularization lowers PSNR from 28.54 to 23.89 and increases LPIPS from 0.0441 to 0.0889. Under these conditions, the high-novelty boundary-band penalty reaches 0.09284, and Fig. 6 shows pronounced Gaussian spillover. These observations support the role of dense anchors in establishing local support and of offset constraints in keeping child Gaussians spatially tied to that support.

Selective Pose Conditioning. Routing pose only to the Offset Decoder improves perceptual reconstruction under novel articulation. In contrast, supplying pose additionally to the opacity, color, and covariance decoders increases high-novelty LPIPS by 0.00148, with a 95% interval of $[0.00059, 0.00247]$. The corresponding rendering under the external eating motion shows appearance corruption around the garment boundary (Fig. 6). The paired error and visible artifact support separating pose-dependent displacement from the canonical description of appearance and Gaussian shape.

Table 6 Root-orientation robustness and scale control. Equal-subject means for the controlled root-yaw and joint-bending evaluations. The root-orientation panel compares PoseRoute with its w/o root-rotation removal ablation, whereas the scale panel compares PoseRoute with its w/o scale constraint ablation. Complete diagnostics and confidence intervals appear in Appendices A.4 and A.5.

Root-orientation robustness	Full	w/o root-rotation removal
Alpha inconsistency	0.001933	0.008913
Boundary Chamfer (px)	0.04611	0.21474
Mean Gaussian drift (mm)	0.33422	3.52212
Scale and contour statistics	**Full**	**w/o scale constraint**
Scale ratio > 20 (%)	0.00	43.77
Min. scale < 0.5 mm (%)	0.00	46.31
Thin-protrusion ratio	0.0010797	0.0012904
Local boundary roughness	4.453854	4.481102

4.4.2 Root-Orientation Robustness

Removing root rotation from the pose-encoder input reduces sensitivity to global orientation: alpha inconsistency and boundary Chamfer decrease by factors of 4.6 and 4.7, respectively, and mean Gaussian drift decreases by a factor of 10.5 (Table 6). The test rotates each of 480 external base poses through 11 nonzero yaw angles, applies the inverse yaw to the generated Gaussians, and compares the aligned output with the same model's zero-yaw output. Local articulation and the camera remain fixed, isolating the effect of global orientation on local Gaussian generation. Alpha inconsistency measures foreground-mask disagreement ($1 - \text{IoU}$), boundary Chamfer measures contour alignment, and drift measures the displacement of corresponding child centers after alignment. *PoseRoute* improves all three reported orientation measures in every subject, with paired confidence intervals above zero; the additional image and drift diagnostics show the same trend (Appendix Table 13). These controlled results show that excluding root rotation makes local Gaussian generation less sensitive to global orientation already handled by LBS.

Table 7 Representation budget and reconstruction quality on People-Snapshot. Values are averaged over four final checkpoints. Parameter counts are in millions; active #Gs are reported in thousands. *GART* is our reproduction.

Variant	Rep. total	Optim. total	Active #Gs	PSNR↑	SSIM↑	LPIPS↓
GART (reprod.)	0.479	2.536	20.8	28.37	0.9711	0.0466
PoseRoute-Compact	0.539	2.112	78.6	28.40	0.9714	0.0477
PoseRoute	1.951	3.524	113.0	**28.54**	**0.9731**	**0.0441**

4.4.3 Scale Control under Joint Bending

Scale control bounds decoded Gaussian anisotropy and preserves local geometry under strong articulation. Across the four subject checkpoints, the fraction of active Gaussians with scale ratios above 20:1 decreases from 43.77% to zero; the fraction with a minimum axis scale below 0.5 mm decreases from 46.31% to zero (Table 6). These statistics indicate that the scale constraints prevent highly anisotropic Gaussian configurations. To assess performance under extreme articulation, we evaluate 96 high-flexion frames selected from knee and elbow rotations. On these frames, joint-local boundary roughness decreases from 4.4811 to 4.4539, with a paired reduction of 0.0272 and a 95% interval of $[0.0062, 0.0479]$. The scale ablation removes the minimum-scale floor, hard ratio cap, and soft ratio penalty, directly isolating the contribution of these constraints. The Gaussian-scale statistics and joint-bending evaluation indicate that scale control bounds canonical Gaussian anisotropy while improving contour regularity under strong articulation.

4.5 Efficiency and Model Capacity

PoseRoute supports real-time rendering across all three evaluation settings, achieving approximately 116 FPS on People-Snapshot, 119 FPS on ZJU-MoCap, and 88 FPS for the residual-bone UBC-Fashion configuration. These A800 measurements include Gaussian generation, SH evaluation, covariance construction, and rasterization. Optimization takes 23.26 minutes on People-Snapshot and approximately 20.5 minutes on ZJU-MoCap.

Table 7 evaluates how representation capacity affects reconstruction quality using *PoseRoute*-Compact, which reduces the child count from 20 to 4 and the feature dimension from 32 to 8 while keeping the other model components unchanged. *PoseRoute*-Compact has a representation budget close to reproduced *GART* (0.539M versus 0.479M) and uses fewer optimized parameters overall (2.112M versus 2.536M), while maintaining comparable reconstruction quality. With 1.951M representation parameters, the full *PoseRoute* model improves all three reconstruction metrics over both compact configurations. The corresponding active primitive counts are 20.8K, 78.6K, and 113.0K for *GART*, *PoseRoute*-Compact, and full *PoseRoute*, respectively. These results suggest that the proposed architecture remains competitive under a representation budget comparable to *GART*, while increasing representation capacity further improves reconstruction quality. Parameter decomposition and timing details are provided in Appendix B.4.

5 Conclusion

In this paper, we presented *PoseRoute*, an anchor-structured Gaussian avatar framework for monocular human animation under novel poses. By organizing Gaussian primitives with articulated anchor-child structures and introducing selective pose conditioning, *PoseRoute* separates body-scale articulation from local appearance and geometry modeling, enabling coherent avatar deformation while preserving high-frequency details. Extensive experiments on People-Snapshot, UBC-Fashion, and ZJU-MoCap demonstrate that *PoseRoute* achieves high-quality reconstruction and real-time rendering. Targeted robustness evaluations show improved boundary consistency and reduced sensitivity to global orientation variations. These results highlight the effectiveness of structured local Gaussian representations with controlled pose dependence for building reliable monocular human avatars. Future work may explore more expressive Gaussian deformation strategies to better model diverse clothing dynamics and complex motion scenarios.

Statements and Declarations

Author contributions All authors participated in discussions of the research design. Xianguo Liu proposed the overall research direction, supervised the whole work, revised the manuscript, and was responsible for project administration and funding acquisition. Weina Cao designed

and developed the core algorithm, implemented the code, conducted all experiments, and drafted the manuscript. Fen Liu, Wenyin Yang, Li Ma, and Haoliang Qi provided guidance on research methodology, offered critical comments and revisions on the manuscript, and participated in in-depth academic discussions of the experimental results. All authors read and approved the final manuscript.

Funding This work was supported by the National Natural Science Foundation of China (Grant No. 662276064), the General Program of the National Social Science Fund of China (Grant No. 22BTQ101), the Regional Joint Fund Project of Basic and Applied Basic Research Foundation of Guangdong Province (Grant No. 2024A1515110038), and the Guangdong Provincial University Natural Science Platform and Project (Key Area Special Project) (Grant No. 2025ZDZX1033).

Data availability The datasets analyzed in this study are publicly available from their providers: People-Snapshot [13], UBC-Fashion [14], and ZJU-MoCap [3]. Access and reuse are subject to the providers' terms.

Competing interests The authors declare no competing interests.

References

[1] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Transactions on Graphics **34**(6), 1–16 (2015) https://doi.org/10.1145/2816795.2818013 . Article 248

[2] Weng, C.-Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16189–16199 (2022). https://doi.org/10.1109/CVPR52688.2022.01573

[3] Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9050–9059 (2021). https://doi.org/10.1109/CVPR46437.2021.00894

[4] Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4), 1–14 (2023) https://doi.org/10.1145/3592433 . Article 139

[5] Lei, J., Wang, Y., Pavlakos, G., Liu, L., Daniilidis, K.: GART: Gaussian articulated template models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19876–19887 (2024). https://doi.org/10.1109/CVPR52733.2024.01879

[6] Hu, S., Hu, T., Liu, Z.: GauHuman: Articulated gaussian splatting from monocular human videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20418–20431 (2024). https://doi.org/10.1109/CVPR52733.2024.01930

[7] Qian, Z., Wang, S., Mihajlovic, M., Geiger, A., Tang, S.: 3DGS-avatar: Animatable avatars via deformable 3D gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5020–5030 (2024). https://doi.org/10.1109/CVPR52733.2024.00480

[8] Liu, X., Wu, C.: VGA: Reconstructing vivid 3D gaussian avatars from monocular videos. In: Didyk, P., Hou, J. (eds.) Computational Visual Media. Lecture Notes in Computer Science, vol. 15664, pp. 172–193. Springer, Singapore (2025). https://doi.org/10.1007/978-981-96-5812-1_10

[9] Lu, T., Yu, M., Xu, L., Xiangli, Y., Wang, L., Lin, D., Dai, B.: Scaffold-GS: Structured 3D gaussians for view-adaptive rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20654–20664 (2024). https://doi.org/10.1109/CVPR52733.2024.01952

[10] Aneja, S., Weiss, S., Baeza, I., Chandran, P.,

Zoss, G., Nießner, M., Bradley, D.: ScaffoldAvatar: High-fidelity gaussian avatars with patch expressions. In: ACM SIGGRAPH Conference Papers, pp. 1–11 (2025). https://doi.org/10.1145/3721238.3730729 . Article 90

[11] Zhang, J., Li, X., Jia, H., Li, J., Su, Z., Wang, G., Li, K.: LoGAvatar: Local gaussian splatting for human avatar modeling from monocular video. Computer-Aided Design **190**, 103973 (2026) https://doi.org/10.1016/j.cad.2025.103973

[12] Zhi, Y., Sun, W., Chang, J., Ye, C., Feng, W., Han, X.: StruGauAvatar: Learning structured 3D gaussians for animatable avatars from monocular videos. IEEE Transactions on Visualization and Computer Graphics **31**(10), 7820–7833 (2025) https://doi.org/10.1109/TVCG.2025.3557457

[13] Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3D people models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8387–8397 (2018). https://doi.org/10.1109/CVPR.2018.00875

[14] Zablotskaia, P., Siarohin, A., Zhao, B., Sigal, L.: DwNet: Dense warp-based network for pose-guided human video generation. In: 30th British Machine Vision Conference (BMVC) (2019). Paper 51. BMVA Press

[15] Habermann, M., Xu, W., Zollhofer, M., Pons-Moll, G., Theobalt, C.: DeepCap: Monocular human performance capture using weak supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5051–5062 (2020). https://doi.org/10.1109/CVPR42600.2020.00510

[16] Jiang, T., Chen, X., Song, J., Hilliges, O.: InstantAvatar: Learning avatars from monocular video in 60 seconds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16922–16932 (2023). https://doi.org/10.1109/CVPR52729.2023.01623

[17] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2022) https://doi.org/10.1145/3503250

[18] Chen, X., Zheng, Y., Black, M.J., Hilliges, O., Geiger, A.: SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11574–11584 (2021). https://doi.org/10.1109/ICCV48922.2021.01139

[19] Chen, X., Jiang, T., Song, J., Rietmann, M., Geiger, A., Black, M.J., Hilliges, O.: Fast-snarf: A fast deformer for articulated neural fields. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 11796–11809 (2023) https://doi.org/10.1109/TPAMI.2023.3271569

[20] Wang, S., Schwarz, K., Geiger, A., Tang, S.: ARAH: Animatable volume rendering of articulated human SDFs. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022, Part XXXII. Lecture Notes in Computer Science, vol. 13692, pp. 1–19. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19824-3_1

[21] Li, R., Tanke, J., Vo, M., Zollhöfer, M., Gall, J., Kanazawa, A., Lassner, C.: TAVA: Template-free animatable volumetric actors. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022, Part XXXII. Lecture Notes in Computer Science, vol. 13692, pp. 419–436. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19824-3_25

[22] Yu, Z., Cheng, W., Liu, X., Wu, W., Lin, K.-Y.: MonoHuman: Animatable human neural field from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16943–16953 (2023). https://doi.org/10.1109/CVPR52729.2023.01625

[23] Zhang, R., Wang, X., Baciu, G., Li, P.: Distilling complementary information from temporal context for enhancing human appearance in human-specific NeRF. The Visual Computer **41**(9), 6479–6491 (2025) https://doi.org/10.1007/s00371-025-03948-z

[24] Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3D gaussians: Tracking by persistent dynamic view synthesis. In: 2024 International Conference on 3D Vision (3DV), pp. 800–809 (2024). https://doi.org/10.1109/3DV62453.2024.00044

[25] Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3D gaussians for high-fidelity monocular dynamic scene reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20331–20341 (2024). https://doi.org/10.1109/CVPR52733.2024.01922

[26] Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4D gaussian splatting for real-time dynamic scene rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20310–20320 (2024). https://doi.org/10.1109/CVPR52733.2024.01920

[27] Xiong, W., Wang, F., Liu, W., Li, X., Zhu, Z., Xiong, B., Rao, Z.: Adaptive control for 3D gaussian splatting: A systematic regularization framework. The Visual Computer **42**(8), 328 (2026) https://doi.org/10.1007/s00371-026-04532-9

[28] Zielonka, W., Bagautdinov, T., Saito, S., Zollhöfer, M., Thies, J., Romero, J.: Drivable 3D gaussian avatars. In: International Conference on 3D Vision, pp. 979–990 (2025). https://doi.org/10.1109/3DV66043.2025.00095

[29] Shao, Z., Wang, Z., Li, Z., Wang, D., Lin, X., Zhang, Y., Fan, M., Wang, Z.: SplattingAvatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1606–1616 (2024). https://doi.org/10.1109/CVPR52733.2024.00159

[30] Wen, J., Zhao, X., Ren, Z., Schwing, A.G., Wang, S.: GoMAvatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2059–2069 (2024). https://doi.org/10.1109/CVPR52733.2024.00201

[31] Peng, S., Xie, W., Wang, Z., Guo, X., Chen, Z., Yang, B., Dong, X.: RMAvatar: Photorealistic human avatar reconstruction from monocular video based on rectified mesh-embedded gaussians. Graphical Models **139**, 101266 (2025) https://doi.org/10.1016/j.gmod.2025.101266

[32] Peng, S., Fu, Y., Miu, R., Lv, T., Yang, B., Dong, X.: GenericAvatar: Generic human modeling from monocular video based on mesh-guided gaussians. The Visual Computer **41**(9), 6657–6670 (2025) https://doi.org/10.1007/s00371-025-03966-x

[33] Yang, Z., Liu, X., Xu, Y., Lin, S., Lin, Z., Luo, Z., Chen, H., Zhang, Q.: KungFuCap: Monocular martial 3D reconstruction via gaussian splatting and diffusion-guided view synthesis. In: International Conference on Computer Vision and Image Computing (CVIC 2025). Proceedings of SPIE, vol. 14070 (2026). https://doi.org/10.1117/12.3107140 . Article 1407015

[34] Pan, P., Su, Z., Lin, C., Fan, Z., Zhang, Y., Li, Z., Shen, T., Mu, Y., Liu, Y.: HumanSplat: Generalizable single-image human gaussian splatting with structure priors. In: Advances in Neural Information Processing Systems, vol. 37, pp. 74383–74410 (2024). https://doi.org/10.52202/079017-2367

[35] Shen, J., Chen, C., Nguyen, T.V., Asari, V.K.: PoseGaussian: Pose-driven novel view synthesis for robust 3D human reconstruction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5004–5015 (2026). https://doi.org/10.1109/WACV61042.2026.00486

[36] Zhan, Y., Shao, T., Yang, Y., Zhou, K.: Real-time high-fidelity gaussian human avatars with position-based interpolation of spatially distributed MLPs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26297–26307 (2025). https://doi.org/10.1109/CVPR52734.2025.02449

[37] Ren, K., Jiang, L., Lu, T., Yu, M., Xu, L., Ni, Z., Dai, B.: Octree-GS: Towards consistent real-time rendering with LOD-structured 3D gaussians. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–15 (2025) https://doi.org/10.1109/TPAMI.2025.3568201

[38] Zhang, J., Zhan, F., Shao, L., Lu, S.: SOGS: Second-order anchor for advanced 3D gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11167–11176 (2025). https://doi.org/10.1109/CVPR52734.2025.01043

[39] Zheng, Y., Yifan, W., Wetzstein, G., Black, M.J., Hilliges, O.: PointAvatar: Deformable point-based head avatars from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21057–21067 (2023). https://doi.org/10.1109/CVPR52729.2023.02017

[40] Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: GaussianAvatars: Photorealistic head avatars with rigged 3D gaussians. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20299–20309 (2024). https://doi.org/10.1109/CVPR52733.2024.01919

[41] Wang, P., Kong, D., Li, J., Yin, B.: SubGAva: 3D gaussian primitive subdivision for photo-realistic and animatable human avatars from monocular video. In: Proceedings of the 2026 International Conference on Multimedia Retrieval, pp. 544–552. Association for Computing Machinery, New York, NY, USA (2026). https://doi.org/10.1145/3805622.3810803

[42] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004) https://doi.org/10.1109/TIP.2003.819861

[43] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations (ICLR) (2015). https://arxiv.org/abs/1412.6980

[44] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068

[45] Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 14294–14303 (2021). https://doi.org/10.1109/ICCV48922.2021.01405

[46] Kwon, Y., Kim, D., Ceylan, D., Fuchs, H.: Neural human performer: Learning generalizable radiance fields for human performance rendering. In: Advances in Neural Information Processing Systems, vol. 34, pp. 24741–24752 (2021). https://proceedings.neurips.cc/paper/2021/hash/cf866614b6b18cda13fe699a3a65661b-Abstract.html

[47] Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelNeRF: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4576–4585 (2021). https://doi.org/10.1109/CVPR46437.2021.00455

[48] Geng, C., Peng, S., Xu, Z., Bao, H., Zhou, X.: Learning neural volumetric representations of dynamic humans in minutes. In: Proceedings of the IEEE/CVF Conference

on Computer Vision and Pattern Recognition, pp. 8759–8770 (2023). https://doi.org/10.1109/CVPR52729.2023.00846



Xianguo Liu received his B.S. degree in Computer Science and Technology from Qingdao University of Technology, Shandong, China, in 2005, and his M.S. degree in Computer Software and Theory and Ph.D. degree in Computer Application Technology from Harbin University of Science and Technology, Heilongjiang, China, in 2008 and 2011, respectively. He is currently a Lecturer at the School of Computer Science and Artificial Intelligence, Foshan University, Guangdong, China. His research focuses on 3D vision and Spatial AI.



Weina Cao is currently pursuing her master's degree at the School of Computer Science and Artificial Intelligence, Foshan University, Guangdong, China. She received her bachelor's degree in Software Engineering from Hebei Normal University in 2025. Her research focuses on 3D vision, with particular interests in 3D Gaussian reconstruction and animatable digital human modeling.



Fen Liu received her B.S. and M.S. degrees from Xidian University in 2014 and 2017, respectively, and earned her Ph.D. degree from the Institute of Information Engineering, Chinese Academy of Sciences, in 2024. She is currently a Lecturer at the School of Computer Science and Artificial Intelligence, Foshan University, Guangdong, China. Her research interests lie in Cryptography-enabled Data Security, Privacy Preservation, and Embodied Intelligence.



Wenyin Yang received her B.Sc. degree in Computer Science in 2005 and her M.Sc. degree in Computer Science in 2007, both from Jinan University, Guangzhou, China, and her Ph.D. degree in Computer Application Technology in 2017 from Central South University, Changsha, China. She has spearheaded several provincial and ministerial-level scientific research initiatives. Her contributions were recognized with the 2021 Science and Technology Award of Guangdong Computer Society (Second Prize). She is currently an Associate Professor at the School of Computer Science and Artificial Intelligence, Foshan University, Guangdong, China. Her research interests include Federated Learning, Spatial AI and Intelligent Storage.



Li Ma received the B.E. degree from Central China Normal University, China, in 2001 and the M.S. degree from South China University of Technology, China, in 2008. She is currently a Professor at the School of Computer Science and Artificial Intelligence, Foshan University, Guangdong, China. She has published more than 40 peer-reviewed journal and conference papers. Her research interests include Large Model Security and World Models.



Haoliang Qi received the B.S. degree in Computer Application from Harbin Institute of Electrical Engineering, Heilongjiang, China, in 1995, and the M.S. degree in Computer Application from Harbin University of Science and Technology, Heilongjiang, China, in 1999, and the Ph.D. degree in Computer Application Technology from Harbin Institute of Technology, Heilongjiang, China, in 2007. He is currently a Professor at the School of Computer Science and Artificial Intelligence, Foshan University, Guangdong, China. His research interests include Fundamental Large Models and World Models.

A Evidence for Novel-Pose Robustness

This appendix presents pose-novelty analysis, external-motion evaluation, boundary-aware comparisons, and root-orientation and scale diagnostics on People-Snapshot. External-motion inputs are paired across methods by subject, action, sample ID, source-frame ID, pose, translation, camera, background, and resolution.

A.1 Pose Novelty Analysis

This section evaluates robustness under increasing pose deviation from the training distribution. Pose novelty is defined as the minimum mean SO(3) geodesic distance of the 23 non-root SMPL joints to all training poses. Global tertiles over the *PoseRoute* input poses define shared strata for all methods, with the high-novelty stratum containing 63 frames above 5.7476°. Paired effects are reported as equal-subject macro means with percentile 95% confidence intervals obtained from 10,000 hierarchical-bootstrap resamples over subjects and paired frames.

Table 8 Pose-novelty-stratified evaluation on People-Snapshot. Results are reported for shared low, medium, and high novelty strata, with equal-subject macro means.

Subset	Method	LPIPS↓	B-F1@3px↑	Spill↓	B-band MAE↓
Low	Full	0.04363	0.99094	0.01607	0.04926
	w/o selective pose conditioning	0.04317	0.99181	0.01541	0.04801
	w/o root-rotation removal	0.04422	0.98983	0.01625	0.04957
	w/o anchor subdivision	0.04845	0.99108	0.01558	0.05040
	w/o offset constraints	0.08896	0.91510	0.03820	0.10292
	w/o scale constraint	0.04255	0.99131	0.01570	0.04840
	GART	0.04901	0.98801	0.01693	0.05562
Medium	Full	0.04509	0.98424	0.01922	0.05749
	w/o selective pose conditioning	0.04609	0.98497	0.01895	0.05781
	w/o root-rotation removal	0.04542	0.98480	0.01897	0.05749
	w/o anchor subdivision	0.04912	0.98544	0.01888	0.05895
	w/o offset constraints	0.09181	0.84249	0.05686	0.12862
	w/o scale constraint	0.04410	0.98497	0.01886	0.05671
	GART	0.04812	0.98322	0.01987	0.06243
High	Full	0.04351	0.97327	0.02269	0.06797
	w/o selective pose conditioning	0.04499	0.97520	0.02229	0.06829
	w/o root-rotation removal	0.04396	0.97488	0.02221	0.06776
	w/o anchor subdivision	0.04745	0.97315	0.02217	0.06905
	w/o offset constraints	0.08709	0.74167	0.08350	0.16081
	w/o scale constraint	0.04288	0.97347	0.02226	0.06732
	GART	0.04379	0.97458	0.02277	0.07068

Across increasing pose-novelty levels, *PoseRoute* maintains stable perceptual and boundary quality, while the ablation patterns reveal the role of anchor locality and offset control under challenging poses. We stratify test poses to examine how performance changes as poses move away from the training distribution.

Table 9 Confidence intervals for component effects under high pose novelty. Each entry reports Δ = w/o − Full and its hierarchical-bootstrap 95% confidence interval on the high pose novelty subset.

Removed component	Δ LPIPS	95% CI	Δ B-band MAE	95% CI
selective pose conditioning	+0.00148	[+0.00059,+0.00247]	+0.00032	[-0.00124,+0.00162]
root-rotation removal	+0.00044	[-0.00039,+0.00147]	-0.00022	[-0.00252,+0.00246]
anchor subdivision	+0.00394	[+0.00314,+0.00480]	+0.00108	[+0.00001,+0.00257]
offset constraints	+0.04357	[+0.03903,+0.04830]	+0.09284	[+0.07555,+0.10999]
scale constraint	-0.00063	[-0.00127,-0.00003]	-0.00066	[-0.00112,-0.00019]

Table 10 Paired comparison with *GART*. The advantage is oriented so that a positive value favors *PoseRoute*. Confidence intervals are obtained by hierarchical bootstrap over the four subjects and their paired frames. The high-novelty test frames follow the shared stratification used in the component analysis.

Subset	Metric	*PoseRoute*	*GART*	*PoseRoute* advantage [95% CI]	Better subjects
Held-out	LPIPS↓	0.04409	0.04662	0.002531 [0.000923, 0.003977]	4/4
	B-F1@3px↑	0.98209	0.98123	0.000858 [-0.001259, 0.003220]	3/4
	Spill↓	0.01934	0.01987	0.000527 [0.000118, 0.001000]	4/4
	B-band MAE↓	0.05883	0.06359	0.004761 [0.004070, 0.005485]	4/4
High novelty	LPIPS↓	0.04351	0.04379	0.000272 [-0.001522, 0.002125]	2/4
	B-F1@3px↑	0.97327	0.97458	-0.001309 [-0.006152, 0.001961]	3/4
	Spill↓	0.02269	0.02277	0.000075 [-0.000728, 0.000836]	2/4
	B-band MAE↓	0.06797	0.07068	0.002709 [0.000784, 0.004746]	4/4

A.2 External Motion Evaluation

The external driving sequences are taken from the `novel_poses` directory of the *GART* codebase [5][1]: `eating.npy`, `fencing.npy`, `walking.npy`, and `aist_demo.npy`. These filenames identify the motion clips used in our experiments, each evaluated on 30 uniformly spaced frames per subject (480 frames per method). All methods share pose, global translation, camera, background, and resolution. The six-frame visual summaries in Fig. 5 are uniformly spaced over the rendered clips; their displayed frame indices identify the visual samples. Extra-component statistics count connected regions of at least 25 pixels. Extra components count significant connected regions beyond the largest foreground component; disconnected area is their foreground-area fraction; boundary roughness normalizes contour length by the perimeter of an equal-area circle; soft-edge ratio measures pixels with $0.05 < \alpha < 0.95$; and alpha change is the foreground-normalized difference between adjacent sampled frames. These unpaired diagnostics are summarized in Table 11 and interpreted together with the fixed-ID comparisons in Fig. 5.

[1] https://github.com/JiahuiLei/GART

Table 11 Structural diagnostics under external driving motions. Each action uses 30 uniformly spaced frames per subject, and the table reports equal-subject macro means. Extra components and disconnected area are lower-is-better artifact indicators. Boundary roughness, soft-edge ratio, and alpha change are descriptive statistics interpreted with the fixed-ID comparisons in Fig. 5.

Action	Method	Extra comp.↓	Disconnected area↓	Boundary rough.	Soft edge	Alpha change
eating	**Full**	0.0000	0.000000	2.1248	0.08193	0.01571
	w/o selective pose conditioning	0.1000	0.001107	3.0367	0.21590	0.01859
	w/o root-rotation removal	0.0000	0.000000	2.1248	0.08079	0.01568
	w/o anchor subdivision	0.0000	0.000004	2.1073	0.09131	0.01563
	w/o offset constraints	7.0417	0.010583	3.1991	0.28747	0.02483
	w/o scale constraint	0.0000	0.000000	2.1336	0.07988	0.01573
	GART	0.0083	0.000026	2.1203	0.08184	0.01563
fencing	**Full**	0.0000	0.000000	2.6493	0.09473	0.04460
	w/o selective pose conditioning	0.0000	0.000000	2.6450	0.11781	0.04446
	w/o root-rotation removal	0.0000	0.000000	2.6471	0.09213	0.04450
	w/o anchor subdivision	0.0000	0.000000	2.6367	0.10459	0.04434
	w/o offset constraints	3.9500	0.006389	3.1871	0.27576	0.04583
	w/o scale constraint	0.0000	0.000001	2.6576	0.09400	0.04451
	GART	0.1333	0.000197	2.6563	0.09491	0.04441
walking	**Full**	0.0000	0.000000	2.4642	0.06745	0.03431
	w/o selective pose conditioning	0.0000	0.000000	2.4436	0.07881	0.03419
	w/o root-rotation removal	0.0000	0.000000	2.4499	0.06576	0.03406
	w/o anchor subdivision	0.0000	0.000000	2.4573	0.07433	0.03431
	w/o offset constraints	2.4000	0.003722	2.8702	0.22570	0.03500
	w/o scale constraint	0.0000	0.000000	2.4673	0.06498	0.03421
	GART	0.0000	0.000000	2.4468	0.07401	0.03414
aist_demo	**Full**	0.0000	0.000000	2.4488	0.07864	0.03293
	w/o selective pose conditioning	0.0000	0.000000	2.4608	0.10836	0.03311
	w/o root-rotation removal	0.0000	0.000000	2.4520	0.07779	0.03302
	w/o anchor subdivision	0.0000	0.000000	2.4442	0.08817	0.03291
	w/o offset constraints	4.1000	0.006518	2.9779	0.26033	0.03538
	w/o scale constraint	0.0000	0.000001	2.4571	0.07664	0.03292
	GART	0.0250	0.000035	2.4517	0.08167	0.03288

Under external driving motions, *PoseRoute* maintains structurally coherent foregrounds across all evaluated actions, with no disconnected foreground components. The additional diagnostics characterize contour and temporal stability under the same driving conditions.

A.3 Boundary-Aware Evaluation

Table 12 extends the standard reconstruction metrics with foreground coverage and boundary accuracy on the held-out People-Snapshot frames. *GART* appears as a separate controlled baseline alongside the component ablations. Masks use an alpha threshold of 0.5. Boundary-band RGB MAE is measured within 5 pixels of the GT contour; spill ratio divides predicted foreground outside the GT mask by total predicted foreground, and Boundary-F1 uses a 3-pixel matching tolerance.

Table 12 Boundary-aware evaluation on held-out People-Snapshot test frames. Values are equal-subject macro means over the four sequences. B-F1 uses a 3-pixel tolerance and boundary-band MAE uses a 5-pixel band. The full configuration is listed first, and *GART* is separated from the component ablations.

Method	PSNR↑	SSIM↑	LPIPS↓	Mask IoU↑	B-F1@3px↑	Chamfer↓	Spill↓	Miss↓	B-band MAE↓
Full	28.543	0.97307	0.04409	0.97266	0.98209	0.6264	0.01934	0.00834	0.05883
w/o selective pose conditioning	28.632	0.97337	0.04486	0.97305	0.98308	0.6192	0.01899	0.00829	0.05873
w/o root-rotation removal	28.593	0.97310	0.04449	0.97275	0.98250	0.6269	0.01910	0.00849	0.05890
w/o anchor subdivision	28.524	0.97251	0.04824	0.97270	0.98243	0.6273	0.01891	0.00873	0.06004
w/o offset constraints	23.887	0.94555	0.08889	0.92153	0.83004	1.8880	0.06032	0.02056	0.13135
w/o scale constraint	28.575	0.97316	0.04318	0.97291	0.98235	0.6222	0.01896	0.00846	0.05809
GART	28.369	0.97114	0.04662	0.97151	0.98123	0.6553	0.01987	0.00898	0.06359

A.4 Root-Orientation Robustness

This section evaluates root-orientation robustness enabled by removing root orientation from pose conditioning. We compare PoseRoute with its w/o root-rotation removal ablation on 480 base poses from four subjects and four external motions. Each pose is evaluated under synthetic yaw perturbations while local articulation, translation, camera, and rendering conditions remain fixed. The evaluation measures foreground consistency, contour alignment, and Gaussian stability after inverse-yaw alignment. PoseRoute consistently reduces orientation sensitivity across all reported measures: alpha inconsistency, Boundary Chamfer, and Gaussian drift improve for all subjects, with confidence intervals above zero. The largest improvement appears in Gaussian drift, showing that root-canonicalized conditioning improves both Gaussian stability and image-space consistency under global orientation changes.

Table 13 Root-orientation robustness under synthetic yaw perturbations. All metrics are lower-is-better. Measurements are averaged over nonzero yaw perturbations for each base pose, and confidence intervals are obtained by resampling base-pose units. Positive advantage favors Full.

Metric	Full	w/o root-rotation removal	Full advantage [95% CI]	Base-pose win rate	Better subjects
Alpha inconsistency	0.001933	0.008913	0.006979 [0.004801, 0.009131]	1.0000	4/4
Boundary Chamfer (px)	0.04611	0.21474	0.16863 [0.11759, 0.20888]	1.0000	4/4
Mean Gaussian drift (mm)	0.33422	3.52212	3.18790 [2.80137, 3.65578]	1.0000	4/4
RGB MAE	0.0002786	0.0017561	0.0014775 [0.0010132, 0.0018678]	1.0000	4/4
Boundary-band MAE	0.003501	0.018298	0.014797 [0.010715, 0.018805]	1.0000	4/4
p95 Gaussian drift (mm)	1.19438	11.30743	10.11305 [9.56165, 10.66959]	1.0000	4/4

A.5 Scale-Control Robustness

This section evaluates how scale control preserves local Gaussian geometry under strong joint articulation. The high-flexion evaluation compares Full with w/o scale constraint on pose-selected paired frames emphasizing strong knee and elbow bending. Joint-local metrics are computed in shared regions around these joints, while decoded-scale statistics are computed over the four final subject checkpoints. Figure 7 provides a qualitative comparison under increasing joint flexion, where removing scale constraints produces more irregular local geometry around highly articulated joints.

Scale control prevents extreme canonical anisotropy: across all four subject checkpoints, *PoseRoute* has no active children with a decoded scale ratio above 20 or a minimum axis scale below 0.5 mm, whereas removing the constraint produces frequent extreme-scale configurations (Table 15). Under strong joint bending, joint-local boundary roughness improves from 4.481102 with w/o scale constraint to 4.453854 with Full, with a paired reduction of 0.027248 and a 95% confidence interval of [0.006192, 0.047944] (Table 14). These results connect bounded decoded geometry with improved joint-local contour regularity.

Table 14 Joint-bending evidence for scale-ratio control. The 96 evaluation frames are selected by joint flexion. Advantage is computed as the ablation value minus the Full value, so positive values indicate that Full improves the corresponding metric over w/o scale constraint.

Metric	Full	w/o scale constraint	Full advantage [95% CI]	Frame win rate	Better subjects
Thin protrusion ratio	0.0010797	0.0012904	0.0002107 [-0.0000037, 0.0004912]	0.5104	4/4
Boundary roughness	4.453854	4.481102	0.027248 [0.006192, 0.047944]	0.6667	4/4

Table 15 Decoded scale distributions across four subject checkpoints. Values are equal-subject means, and confidence intervals resample subjects.

Metric	Full	w/o scale constraint	Full advantage [95% CI]	Better subjects
p95 scale ratio	9.31	6,658.88	6,649.56 [1,714.19, 13,533.17]	4/4
p99 scale ratio	9.93	4.3612×10^5	4.3611×10^5 [1.1952×10^5, 7.5642×10^5]	4/4
Maximum scale ratio	10.9	4.8523×10^6	4.8522×10^6 [4.7874×10^6, 4.8999×10^6]	4/4
Fraction ratio > 20	0.0000	0.4377	0.4377 [0.3605, 0.4994]	4/4
Fraction min scale < 0.5 mm	0.0000	0.4631	0.4631 [0.4001, 0.5258]	4/4

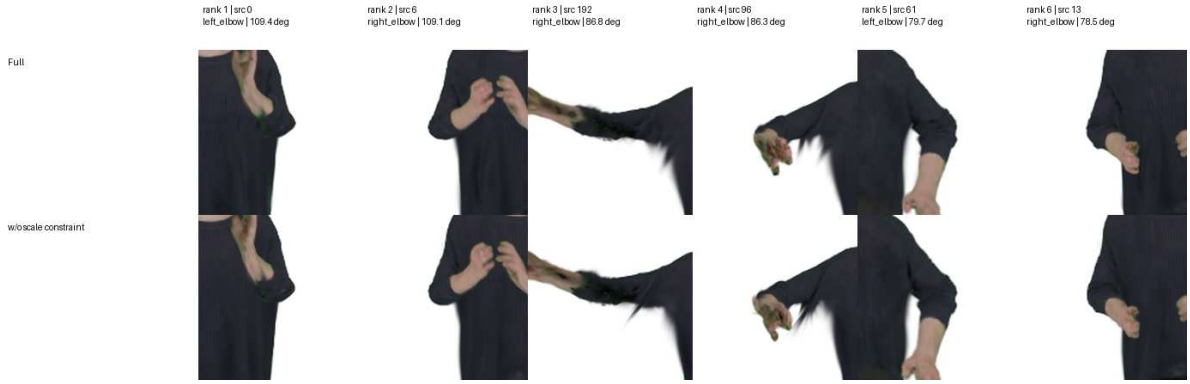


Fig. 7 High-flexion scale-control comparison. Fixed pose-selected crops from female-3-casual under the fencing motion are ordered by decreasing joint flexion. Full is shown above w/o scale constraint to visualize the effect of scale control on joint-local geometry

B Implementation and Computational Details

The settings below apply to the full PoseRoute model; compact and ablation variants are specified in Secs. 4.5 and 4.4, respectively.

B.1 Frame Sampling and Pose Refinement

People-Snapshot uses male-3-casual, male-4-casual, female-3-casual, and female-4-casual, with 55, 36, 51, and 48 test frames, respectively; Table 16 gives the sampling indices. UBC-Fashion takes every fourth frame from index 0 for training and from index 2 for testing; test poses are initialized from the corresponding optimized training-pose sequence before refinement. ZJU-MoCap uses subjects 377, 386, 387, 392, 393, and 394, with 100 training time steps at stride 5 from index 0 in camera 4. Evaluation uses the remaining cameras and every sixth time step of this sampled sequence. Evaluation resolutions are 540×540 for People-Snapshot, 720×940 for UBC-Fashion, and 512×512 for ZJU-MoCap.

Table 16 People-Snapshot frame-index bounds and strides. Bounds are inclusive; indices refer to the preprocessed image order.

	Training			Testing		
Subject	Start	End	Stride	Start	End	Stride
female-3-casual	0	445	4	446	647	4
female-4-casual	0	335	4	335	523	4
male-3-casual	0	455	4	456	675	4
male-4-casual	0	659	6	660	872	6

Test-time pose refinement uses 300 steps at learning rate 3×10^{-3}, halved every 70 steps, for People-Snapshot; UBC-Fashion and ZJU-MoCap use 50 steps at 4×10^{-3}, halved every 20 steps. Following *GART*'s evaluation paths, People-Snapshot and UBC-Fashion use the InstantAvatar evaluator with AlexNet LPIPS, and ZJU-MoCap uses the Instant-NVR evaluator with VGG LPIPS.

B.2 Network and Deformation Configuration

One-level SMPL mesh subdivision initializes 27,554 anchors, each with a 32-dimensional feature and $K = 20$ candidate children. Table 17 gives the network dimensions and output conventions.

Table 17 Full-model decoder dimensions and activation conventions. L denotes the maximum SH order and $K = 20$.

Module	Layer widths	Output convention
Pose Encoder	$72 \to 32 \to 16$	Linear
Offset Decoder	$48 \to 32 \to 60$	tanh
Opacity Decoder	$32 \to 32 \to 20$	Sigmoid after logit selection
Cov Decoder	$32 \to 32 \to 140$	Scale sigmoid and quaternion normalization
Color Decoder	$32 \to 32 \to K\,3(L+1)^2$	SH coefficients

Hidden layers use ReLU; training adds Gaussian noise with standard deviation 0.02 to the pose latent. Quaternions use a nonnegative scalar component. SH coefficients are clamped to $[-2, 2]$; evaluated colors are shifted by 0.5 and clamped to $[0, 1]$. People-Snapshot and ZJU-MoCap use maximum SH order 1, enabled at iteration 500; UBC-Fashion enables orders 1–4 at iterations 3000, 5000, 6000, and 7000.

Voxel and Residual-Bone Configuration. The default configuration uses 24 SMPL bones without residual bones ($n_r = 0$). The additional UBC-Fashion configuration uses $n_r = 32$ learned residual bones. The human Voxel Deformer uses a $16\times64\times64$ grid with 24 base channels. Initialization uses inverse-distance interpolation from the 30 nearest SMPL vertices followed by 30 smoothing iterations. Queries use trilinear interpolation with border padding. The optional residual-bone MLP receives the 69-dimensional root-rotation-removed body pose, uses four 256-wide layers with LeakyReLU (slope 0.01), and outputs $32\times7 = 224$ quaternion/translation values. Its output layer is initialized with standard deviation 10^{-3}.

B.3 Optimization and Refinement

Training uses one frame per iteration and foreground-mask compositing with random backgrounds. The photometric weights in Eq. 11 are 1.0 for L_1 and 0.1 for SSIM. Adam uses the PyTorch default momentum parameters and zero weight decay. Table 18 gives module learning rates; arrows denote exponential decay over training. Pose and translation refinement begins at iteration 500.

Table 18 Learning rates during avatar optimization.

Parameter group	Learning rate
Anchor features	7.5×10^{-3}
Offset Decoder	4×10^{-3}
Static child offsets	$10^{-2} \rightarrow 10^{-4}$
Opacity Decoder	$2\times10^{-3} \rightarrow 2\times10^{-5}$
Cov Decoder and Pose Encoder	$4\times10^{-3} \rightarrow 4\times10^{-5}$
Color Decoder	$8\times10^{-3} \rightarrow 8\times10^{-5}$
Root/body pose and translation, People-Snapshot/ZJU-MoCap	10^{-4}
Root pose and translation, UBC-Fashion	$10^{-4} \rightarrow 10^{-5}$
Body pose, UBC-Fashion	$3\times10^{-4} \rightarrow 10^{-5}$
Base voxel correction (from iteration 1500)	$5\times10^{-5} \rightarrow 5\times10^{-6}$
Residual-bone network	3×10^{-5}
Additional voxel correction	$3\times10^{-5} \rightarrow 3\times10^{-6}$

For People-Snapshot and ZJU-MoCap, anchor refinement starts at iteration 100, with center-gradient threshold 1.5×10^{-4}, density coefficient 0.01, and pruning threshold $\tau_{\text{prune}} = 0.1$. UBC-Fashion starts at iteration 500 and uses 3×10^{-4}, 0.005, and 0.01, respectively. Among anchors satisfying the gradient criterion, the mean static-offset length determines cloning or splitting. Compact offset groups are cloned; wider groups are replaced by two anchors at the parent's canonical position, with perturbed features and rescaled, perturbed static offsets. Cloning and splitting run every 200 iterations until iteration 10,000, while opacity-based pruning continues at the same interval throughout optimization.

The geometry regularization weights are $\lambda_{\text{off}} = 5.0$, $\lambda_{\text{ratio}} = 0.01$, and $\lambda_{\text{max}} = 0.05$. People-Snapshot and ZJU-MoCap use $\lambda_{\text{feat}} = 0.005$, $\lambda_{\text{aoff}} = 0.01$, and $K_{\text{KNN}} = 3$; UBC-Fashion uses 0.01, 0.01, and 6, respectively. Euclidean KNN neighborhoods include each query anchor. Base voxel TV and magnitude weights are both 0.8; additional residual-bone channels use 0.5 and 0.3.

Scale-Control Measurement Protocol. For each subject and action, the 30 fixed frames are ranked by the maximum SO(3) rotation angle among the left and right knees and elbows (SMPL joints 4, 5, 18, and 19), and the top six are selected using pose alone. This yields 96 paired frames for a 10,000-sample hierarchical bootstrap over subjects and frames. Decoded-scale statistics are computed once per subject checkpoint.

Joint-local regions are defined as the union of four circles centered on the projected knees and elbows and are shared by both methods. The circle radius is $\max(12, \text{round}(0.08H_{\text{fg}}))$ pixels, where H_{fg} is the clipped projected SMPL vertex bounding-box height from the *PoseRoute* reference. Binary opening uses a disk with radius $\max(1, \text{round}(0.01H_{\text{fg}}))$ pixels, with half-up rounding in both expressions and an alpha threshold of 0.5. The thin-protrusion ratio is the fraction of foreground removed within these joint-local regions. Local roughness is computed by dividing the inner-boundary pixel count by the square root of the local foreground area. Scale ratios are obtained by dividing the largest decoded axis scale by the smallest, with the denominator lower-bounded at 10^{-8}, and are summarized over active children. Checkpoint confidence intervals are estimated by resampling the four subjects.

B.4 Model Capacity and Timing

Table 19 decomposes the parameter budget reported in Table 7.

Table 19 Detailed representation-capacity decomposition on People-Snapshot. Values are averages over four final checkpoints and parameter counts are in millions. The reproduced *GART* configuration uses $n_r = 4$, while both *PoseRoute* variants use $n_r = 0$. Rep. total counts direct Gaussian attributes for *GART*, and anchor features, static offsets, the Pose Encoder, and four attribute MLPs for *PoseRoute*. Optim. total adds trainable voxel-deformation correction fields and residual-bone networks, when present, to Rep. total; per-frame SMPL pose and translation refinements are omitted. Active #Gs denotes the final pruned set for *GART* and the opacity-filtered child set for *PoseRoute*.

Variant	Final anchors	Anchor feat.	Static offsets	Shared networks	Rep. total	Optim. total	Active #Gs
GART (reprod.)	–	–	–	–	0.479	2.536	20.8K
PoseRoute-Compact	26.8K	0.214	0.321	0.004	0.539	2.112	78.6K
PoseRoute	21.0K	0.671	1.258	0.023	1.951	3.524	113.0K

Timing Protocol. FPS uses batch size one and maximum SH order, over ten passes through each sequence after one warm-up forward-and-render frame. The timed interval includes Gaussian generation, SH evaluation, covariance construction, and rasterization. Data and checkpoint loading, test-time pose fitting, disk output, and metric computation are excluded. Our timings use the NVIDIA A800, whereas literature entries retain the hardware and optimization budgets reported by their sources.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- peoplefemale3gartvsposeroute.mp4
- ubcselectedgartvsposeroute.mp4