RRA-AVATAR: REPRESENTATION RESPONSIBILITY ALLOCATION FOR GAUSSIAN AVATARS WITH LOOSE GARMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Animatable Gaussian avatars typically use a single body-guided representation for both body-supported surfaces and off-body structures such as loose garments, hair, and accessories. This creates a representation-responsibility ambiguity: regions with substantially different levels of body-prior support are forced to share geometry, deformation, and optimization capacity. We introduce RRA-Avatar, a support-aware dual-layer Gaussian avatar framework that assigns representation responsibility according to whether foreground geometry is adequately explained by the fitted body prior, rather than according to garment semantics. The body-aligned layer preserves stable SMPL-guided modeling, while the residual-surface layer provides dedicated capacity for unexplained off-body structures. To instantiate this decomposition without garment templates or semantic labels, we recover residual-surface seeds from temporally fused multi-view foreground evidence. We further introduce a responsibility-allocation mechanism that coordinates layer-specific supervision, gradient flow, and topology evolution, reducing inter-layer competition while retaining joint refinement near uncertain boundaries. Experiments on DNA-Rendering, I3D-Human, and ZJU-MoCap demonstrate substantially improved reconstruction of loose and off-body structures while preserving competitive body-region and novel-pose performance. On six DNA-Rendering subjects, RRA-Avatar reduces LPIPS* from 30.91 to 22.98 relative to the strongest compared baseline under the same protocol.

1 INTRODUCTION

Recent advances in neural rendering (Mildenhall et al., 2022; Müller et al., 2022) have enabled high-quality and animatable human avatars from multi-view videos. In particular, 3D Gaussian Splatting (3DGS) provides an explicit representation with efficient rasterization and strong rendering quality (Kerbl et al., 2023). Most Gaussian avatar methods build canonical primitives around SMPL or SMPL-X (Loper et al., 2015; Pavlakos et al., 2019) and deform them with body-guided motion (Hu et al., 2024b; Qian et al., 2024; Lei et al., 2024; Moon et al., 2024). Such a design is highly effective for surfaces whose geometry and motion are well supported by the fitted body prior. However, it becomes substantially less reliable for off-body structures, such as loose garments, hair, and accessories, whose geometry often extends beyond the body and whose motion is only weakly coupled to the skeleton. As shown in Figure 1, body-supported regions are reconstructed faithfully, whereas loose sleeves, skirt hems, and other distant structures frequently become blurred, contracted, or incorrectly attached to the underlying body. These failures reveal a representational mismatch: a unified body-guided representation is required to explain foreground structures with fundamentally different levels of body-prior support.

Recent methods target avatar deformation or decomposition. SeqAvatar (Xu et al., 2025) combines coarse skeletal and point-wise motion but retains a single body-guided Gaussian set. It improves primitive deformation but does not provide independent support for foreground geometry outside the fitted body. Semantic-layer avatars instead separate body and clothing (Feng et al., 2022; Lin et al., 2024; Peng et al., 2025; Jin & He, 2026), but asset identity differs from body-prior support: one garment may be supported at the waist but off-body at the hem. FreeCloth (Ye et al., 2025), a related point-based method, partitions surface points for replication, LBS-based deforma-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

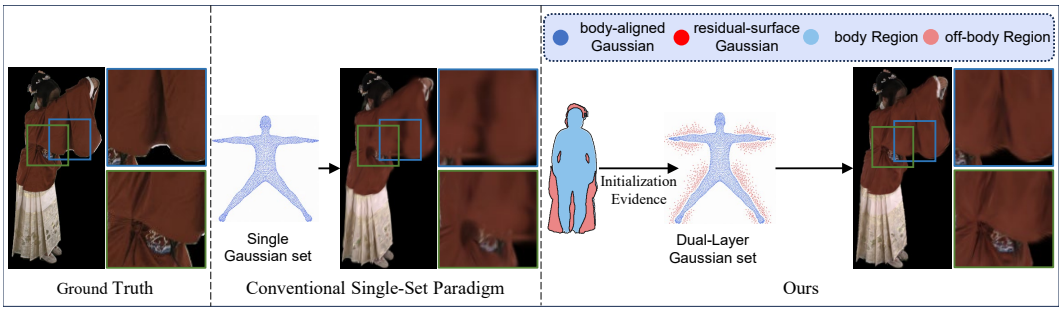


Figure 1: RRA-Avatar separates body-aligned and residual-surface Gaussians according to body-prior support, improving off-body regions that are blurred by a conventional single-set representation.

tion, or free-form generation, but does not allocate Gaussian representations according to body-prior support. Consequently, existing methods still lack a principled mechanism for determining which representation should explain each foreground structure and receive the corresponding deformation capacity, optimization signals, and topology growth.

Accordingly, we formulate Gaussian avatar reconstruction as a representation-responsibility allocation problem. Representation responsibility determines not only which representation should explain a foreground structure, but also where deformation capacity, optimization signals, and topology growth should be allocated during optimization. Rather than separating body and clothing by semantic identity, we allocate representations according to body-prior support. Structures adequately explained by the fitted body remain in a body-aligned representation, whereas unexplained foreground structures receive dedicated representational capacity in a residual-surface representation.

Realizing this formulation requires solving three coupled problems. First, residual representations must be initialized from noisy, occluded multi-view observations without direct 3D supervision. Second, the two representations require different local deformation models. Third, representation assignment alone cannot prevent responsibility leakage, where the same foreground residual updates both representations and fixed assignments hinder joint refinement near uncertain body–garment boundaries.

To address these challenges, we propose **RRA-Avatar**, a Gaussian avatar framework based on **R**epresentation **R**esponsibility **A**llocation. As shown in Figure 2, RRA-Avatar derives representation ownership from body-prior support and maintains this ownership throughout initialization, deformation, and optimization. A body-prior-guided residual-surface discovery module constructs residual Gaussian seeds from temporally fused multi-view foreground evidence without garment templates or semantic labels. Layer-specific deformation heads model the distinct local motion of the two representations. Finally, a responsibility-allocation mechanism coordinates supervision, gradient flow, and topology evolution, allowing the two representations to specialize before jointly refining uncertain boundaries.

We evaluate RRA-Avatar on DNA-Rendering (Cheng et al., 2023), I3D-Human (Chen et al., 2024), and ZJU-MoCap (Peng et al., 2021) for novel-view rendering, and additionally on I3D-Human for novel-pose animation. It improves loose-garment and off-body reconstruction over single-set body-guided avatars while remaining competitive in body-centric and novel-pose settings.

Our main contributions are summarized as follows:

- We formulate animatable Gaussian avatar reconstruction as a representation-responsibility allocation problem and propose RRA-Avatar, a support-aware dual-layer Gaussian framework that allocates representations according to body-prior support rather than semantic identity. This enables dedicated off-body representation while preserving reliable body-guided modeling.
- We introduce a body-prior-guided residual-surface discovery strategy that samples SMPL-guided shells, evaluates candidates using multi-view unexplained-foreground evidence, and

fuses temporally consistent observations into residual Gaussian seeds, without garment partitions or direct 3D supervision.

- We introduce a responsibility-allocation mechanism that jointly coordinates supervision, gradient flow, and topology evolution, preventing responsibility leakage while enabling cooperative refinement between body-aligned and residual-surface representations.

2 Related Work

Animatable human avatars. Neural human avatar methods often use parametric body models such as SMPL (Loper et al., 2015) to build correspondences across poses. Neural Body (Peng et al., 2021), HumanNeRF (Weng et al., 2022), and ARAH (Wang et al., 2022) represent dynamic humans with neural fields and body-guided deformation. More recently, 3D Gaussian Splatting (Kerbl et al., 2023) has enabled explicit avatars with efficient rasterization. Existing Gaussian avatar methods use body priors or learnable articulation (Hu et al., 2024b; Qian et al., 2024; Lei et al., 2024; Kocabas et al., 2024), embed Gaussians in meshes or templates (Hu et al., 2024a; Shao et al., 2024), or model pose-dependent appearance and deformation (Li et al., 2024; Moreau et al., 2024; Xu et al., 2025). These methods improve avatar geometry, appearance, or motion modeling, but most of them still optimize a unified canonical representation for both body-supported and off-body regions.

Loose-garment and layered modeling. Loose garments, hair, and accessories may slide, lag, or deform based on motion history. Prior work studied clothed-avatar reconstruction (Saito et al., 2021; Luo et al., 2024; Guo et al., 2024) and modeled loose dynamics with garment-adaptive Gaussian maps (Li et al., 2024), adaptive skeletons (Zhan et al., 2025), physics (Zheng et al., 2024; Lee et al., 2025), or temporal models (Lee et al., 2026; Li et al., 2025). D3GA (Zielonka et al., 2025) uses garment- and part-specific layers; LayGA (Lin et al., 2024), Gaussian Wardrobe (Chen et al., 2026), and MonoCloth (Jin & He, 2026) separate body and clothing, while FreeCloth (Ye et al., 2025) partitions points by deformation strategy. DynaAvatar (Kwon et al., 2026) predicts motion-dependent cloth deformation from one image without subject-specific avatar optimization; PERSONA (Sim & Moon, 2025) fits a pose-driven avatar to diffusion-generated videos. Both infer unobserved subject-specific geometry and appearance from learned or generated priors, which may not match the target. We instead use observed multi-view residual evidence to assign Gaussians by body-prior support and coordinate deformation, supervision, gradients, and topology without garment templates or 3D surface supervision.

Geometry initialization. SMPL-based Gaussian avatars commonly initialize primitives from body templates, articulated point sets, or template-aligned maps (Hu et al., 2024b; Lei et al., 2024; Li et al., 2024). Such template-based initialization provides stable support for body-supported regions, but does not directly cover foreground structures that extend beyond it. Classical visual-hull methods recover 3D occupancy from multi-view silhouettes (Laurentini, 1994), but silhouette constraints alone do not determine accurate surface depth. RRA-Avatar uses a lighter strategy: it keeps the fitted body as a prior and uses multi-view silhouette residuals only to initialize off-body Gaussian candidates.

3 Method

3.1 Overall Architecture

RRA-Avatar follows a simple principle: preserve body guidance where it is reliable and allocate separate representational capacity where it is not. As shown in Figure 2, RRA-Avatar takes as input synchronized multi-view images $\{\mathbf{I}_{t,v}\}$ captured by calibrated cameras, foreground masks $\{M^{\mathrm{fg}}_{t,v}\}$, and per-frame fitted SMPL meshes $\{\mathcal{S}_t\}$, where t and v index video frames and camera views. Our goal is to learn a canonical Gaussian avatar that can be driven by novel poses and rendered from novel views.

To this end, RRA-Avatar assigns Gaussians to body-aligned and residual-surface layers according to body-prior support. This split uses body-prior support as a proxy for skeletal-motion coupling rather than garment semantics. The body-aligned layer is initialized from canonical SMPL vertices,

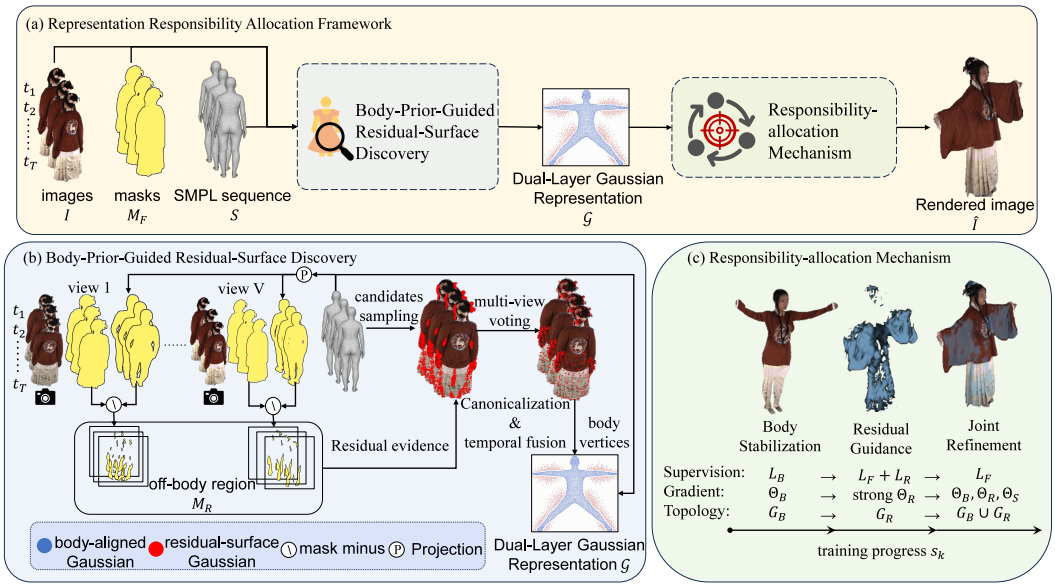


Figure 2: Representation Responsibility Allocation Framework. (a) The framework builds and optimizes a dual-layer Gaussian avatar. (b) Body-prior-guided discovery initializes residual Gaussians from multi-view foreground evidence. (c) The responsibility-allocation mechanism coordinates supervision, gradient flow, and topology evolution from body stabilization through residual guidance to joint refinement.

and the residual-surface layer is built from multi-view residual evidence. A responsibility-allocation mechanism then controls layer-specific supervision, gradient flow, and topology evolution during optimization.

3.2 Dual-Layer Gaussian Representation

Canonical dual-layer representation. Representation responsibility defines which representation should explain a foreground structure, while representation ownership denotes its concrete assignment to one of the two Gaussian layers. We instantiate this ownership using a body-aligned representation for structures adequately explained by the fitted body prior and a residual-surface representation for foreground structures that remain unexplained by the body.

We represent the avatar as a set of canonical 3D Gaussian primitives. Each primitive g_i stores its position $\boldsymbol{\mu}_i$, rotation quaternion $\mathbf{q}_i$, scale $\mathbf{s}_i$, opacity α_i, color coefficients $\mathbf{c}_i$, and a layer label ℓ_i:

$$g_i = (\boldsymbol{\mu}_i, \mathbf{q}_i, \mathbf{s}_i, \alpha_i, \mathbf{c}_i, \ell_i), \quad \ell_i \in \{\mathrm{B}, \mathrm{R}\}. \tag{1}$$

Here B and R denote the body-aligned and residual-surface layers, respectively. For each layer label $\ell \in \{\mathrm{B}, \mathrm{R}\}$, we define $\mathcal{G}_\ell = \{g_i \mid \ell_i = \ell\}$, and the full representation is $\mathcal{G} = \mathcal{G}_\mathrm{B} \cup \mathcal{G}_\mathrm{R}$.

Layer-specific deformation heads. For frame t, each primitive receives a shared motion feature

$$\mathbf{h}_i^t = \mathcal{E}(\boldsymbol{\mu}_i, \boldsymbol{\theta}_t, \mathbf{z}_t), \tag{2}$$

where $\boldsymbol{\theta}_t$ is the current SMPL pose and $\mathbf{z}_t$ denotes multi-scale skeletal and point-wise motion context. The encoder $\mathcal{E}$ is shared by both layers, whereas the non-rigid deformation heads are layer-specific:

$$(\Delta\boldsymbol{\mu}_i^t, \Delta\mathbf{r}_i^t, \Delta\mathbf{s}_i^t) = H_{\ell_i}(\mathbf{h}_i^t), \quad \ell_i \in \{\mathrm{B}, \mathrm{R}\}. \tag{3}$$

The two heads are introduced to model different deformation assumptions rather than simply increasing network capacity. Body-aligned Gaussians remain strongly coupled to skeletal motion, whereas residual-surface Gaussians accommodate larger local deviations beyond the fitted body. The shared encoder preserves common pose-dependent motion context, while the layer-specific heads specialize the residual deformation associated with each representation. This separation allows the

two representations to share global motion context while specializing local residual deformation. Here $\Delta\boldsymbol{\mu}_i^t$ and $\Delta\mathbf{s}_i^t$ denote additive center and scale updates, respectively. The predicted residual quaternion $\Delta\mathbf{r}_i^t$ is left-composed with the canonical Gaussian rotation quaternion $\mathbf{q}_i$ using the Hamilton product, and the result is normalized to unit length. The heads H_{B} and H_{R} are used for body-aligned and residual-surface primitives, respectively. We apply the predicted offsets to each Gaussian before warping it to observation space with the coarse SMPL-guided deformation.

Joint and layer-wise rendering. At frame t, let $\mathcal{G}_{\mathrm{B}}^t$ and $\mathcal{G}_{\mathrm{R}}^t$ be the deformed body-aligned and residual-surface layers. For image reconstruction, both layers are rendered together by a single depth-ordered Gaussian rasterizer:

$$(\hat{\mathbf{I}}_{t,v}, \hat{\mathbf{A}}_{t,v}) = \mathcal{R}_v(\mathcal{G}_{\mathrm{B}}^t \cup \mathcal{G}_{\mathrm{R}}^t), \tag{4}$$

where $\mathcal{R}_v$ denotes the differentiable Gaussian rasterizer for view v. The final image is produced by joint compositing rather than by summing separately rendered layers. We also perform layer-wise rendering to obtain the residual-layer opacity map $\hat{\mathbf{A}}_{t,v}^{\mathrm{R}}$ for the responsibility-allocation mechanism (Sec. 3.4).

3.3 Body-Prior-Guided Residual-Surface Discovery

The residual-surface layer should cover foreground geometry that is not well supported by the fitted SMPL body. Since these regions have no direct 3D supervision, we use the body prior to reduce the search space and multi-view masks to select reliable off-body candidates. The objective is not to reconstruct a complete off-body surface, but to provide reliable geometric support from which the residual representation can be optimized.

SMPL-guided shell candidates. For frame t, let $\mathbf{s}_{t,a} \in \mathbb{R}^3$ be the position of the a-th vertex on the posed SMPL mesh $\mathcal{S}_t$, and let $\mathbf{n}_{t,a} \in \mathbb{R}^3$ be its unit outward normal, with $\|\mathbf{n}_{t,a}\|_2 = 1$. For each shell index j, we translate every SMPL vertex along its normal:

$$\mathbf{x}_{t,a,j} = \mathbf{s}_{t,a} + \rho_j d_t \mathbf{n}_{t,a}, \quad \rho_j \in \mathcal{D}, \tag{5}$$

where d_t scales the predefined dimensionless offsets $\mathcal{D}$ to the body size. For each j, candidates from all anchors form one shell; varying j yields a multi-shell search around the fitted body instead of the full volume.

Image-space residual evidence. We evaluate each candidate by projecting it into each calibrated camera view using the projection operator Π_v. For view v, let $\phi_{t,v}^{\mathrm{F}}$ and $\phi_{t,v}^{\mathrm{B}}$ be the signed distance maps of the foreground mask and the projected SMPL silhouette, where negative values denote interior pixels. The residual evidence region is

$$\Omega_{t,v}^{\mathrm{R}} = \{\mathbf{u} \mid \phi_{t,v}^{\mathrm{F}}(\mathbf{u}) < 0, \phi_{t,v}^{\mathrm{B}}(\mathbf{u}) > \delta\}. \tag{6}$$

Here δ specifies the margin around the projected body silhouette, so $\Omega_{t,v}^{\mathrm{R}}$ corresponds to the foreground region outside the dilated SMPL mask. Pixels inside $\Omega_{t,v}^{\mathrm{R}}$ provide positive evidence, while background pixels and body-supported pixels provide negative evidence. We also use a narrow soft band around $\Omega_{t,v}^{\mathrm{R}}$ to reduce sensitivity to mask noise and uncertain boundaries. Positive evidence alone is insufficient because a candidate may agree with a small subset of views while contradicting body-supported or background observations elsewhere. Negative evidence therefore suppresses geometrically inconsistent hypotheses.

Multi-view candidate voting. A view is valid if the candidate is in front of the camera, its projection lies inside the image, and its normal faces the camera. Let $\mathcal{V}_t(\mathbf{x})$ be the valid views of candidate $\mathbf{x}$. We aggregate positive and negative evidence as

$$E_t^{\pm}(\mathbf{x}) = \sum_{v \in \mathcal{V}_t(\mathbf{x})} w_v^t(\mathbf{x}) e_{t,v}^{\pm}(\Pi_v(\mathbf{x})). \tag{7}$$

Here $e_{t,v}^{+}$ rewards residual projections, $e_{t,v}^{-}$ penalizes background or body-supported projections, and w_v^t is a cosine weight favoring front-facing views. We then compute the support and contradiction ratios:

$$S_t(\mathbf{x}) = \frac{E_t^{+}(\mathbf{x})}{Z_t(\mathbf{x})}, \quad C_t(\mathbf{x}) = \frac{E_t^{-}(\mathbf{x})}{Z_t(\mathbf{x})}, \tag{8}$$

Here $Z_t(\mathbf{x})$ is the total valid-view weight plus a small stabilizer. High S_t and low C_t identify cross-view-consistent off-body candidates; boundary filtering removes spurious hypotheses, while outer-shell preference improves off-body coverage.

Canonicalization and temporal fusion. Temporal fusion suppresses frame-specific mask and body-fitting noise while retaining residual structures that persist across motion. The retained candidates are first obtained in the observation space of each frame. We map them to the canonical SMPL space using local frames attached to their anchor vertices. We then cluster canonical candidates with the same anchor vertex across time and discard clusters with weak temporal support. The remaining cluster centers are used as residual-surface Gaussian seeds with layer label R. This initialization gives the residual-surface layer consistent geometric support for off-body regions, while the responsibility-allocation mechanism further refines the geometry.

3.4 Responsibility-allocation Mechanism

Responsibility allocation. A dual-layer representation alone does not guarantee a dual-layer explanation. Under a shared image reconstruction objective, the same foreground residual may simultaneously update both representations, causing responsibility leakage, where body-supported and off-body structures compete for the same representational capacity. Consequently, representation ownership established during initialization must also be enforced throughout optimization. To this end, supervision allocates which representation should explain each region, gradient scaling controls which representation receives the dominant parameter updates, and topology gating determines where new Gaussian capacity is created.

For clarity, we omit frame and view indices. Let $\mathbf{I}$ denote the ground-truth RGB image. Let M^{F}, M^{B}, and M^{R} denote the full foreground mask, the projected SMPL body mask, and the residual mask outside the dilated SMPL silhouette, respectively. Let $(\hat{\mathbf{I}}^{\mathrm{B}}, \hat{\mathbf{A}}^{\mathrm{B}})$ be the body-only render, $(\hat{\mathbf{I}}, \hat{\mathbf{A}})$ be the joint render of both layers, and $\hat{\mathbf{A}}^{\mathrm{R}}$ be the residual-layer opacity map. We use

$$\begin{aligned} \mathcal{L}_{\mathrm{B}} &= \mathcal{L}_{\mathrm{img}}(\hat{\mathbf{I}}^{\mathrm{B}}, \hat{\mathbf{A}}^{\mathrm{B}}; \mathbf{I}, M^{\mathrm{B}}), \\ \mathcal{L}_{\mathrm{F}} &= \mathcal{L}_{\mathrm{img}}(\hat{\mathbf{I}}, \hat{\mathbf{A}}; \mathbf{I}, M^{\mathrm{F}}). \end{aligned} \tag{9}$$

Here $\mathcal{L}_{\mathrm{img}}$ is a weighted sum of ℓ_1 RGB, ℓ_2 opacity, SSIM, and LPIPS losses. The pixel-wise losses are evaluated over the mask-defined supervision region, while SSIM and LPIPS are computed on its bounding-box crop. We supervise the residual-layer opacity by minimizing the following loss:

$$\mathcal{L}_{\mathrm{R}} = \frac{\sum_{\mathbf{u}} M^{\mathrm{F}}(\mathbf{u}) \left(\hat{\mathbf{A}}^{\mathrm{R}}(\mathbf{u}) - M^{\mathrm{R}}(\mathbf{u}) \right)^2}{\sum_{\mathbf{u}} M^{\mathrm{F}}(\mathbf{u}) + \epsilon_{\mathrm{m}}}. \tag{10}$$

For binary masks, this term drives residual opacity toward one on M^{R} and zero on $M^{\mathrm{F}} \setminus M^{\mathrm{R}}$; ϵ_{m} prevents division by zero. The full objective is

$$\mathcal{L} = w_{\mathrm{B}}(s_k)\mathcal{L}_{\mathrm{B}} + w_{\mathrm{F}}(s_k)\mathcal{L}_{\mathrm{F}} + w_{\mathrm{R}}(s_k)\mathcal{L}_{\mathrm{R}} + \lambda_{\mathrm{A}}(s_k)\mathcal{L}_{\mathrm{AIAP}}. \tag{11}$$

Here $s_k = k/K$, where k is the current iteration and K is the total number of training iterations. $\mathcal{L}_{\mathrm{AIAP}}$ is the as-isometric-as-possible regularizer for preserving local distances during deformation. The weights w_{B}, w_{F}, w_{R}, and λ_{A} change with s_k. For small s_k, only $\mathcal{L}_{\mathrm{B}}$ is active, so the body-aligned layer first learns stable body-supported geometry and appearance. As training progresses, the full foreground loss $\mathcal{L}_{\mathrm{F}}$ and residual opacity loss $\mathcal{L}_{\mathrm{R}}$ are enabled. The former trains the full avatar; the latter directs residual opacity toward off-body regions. After the residual-guidance interval, the residual opacity loss $\mathcal{L}_{\mathrm{R}}$ is disabled, so both layers are refined under the full foreground loss without a fixed residual constraint.

Gradient control. We also scale gradients according to the same progress variable. Let $\Theta = (\Theta_{\mathrm{B}}, \Theta_{\mathrm{R}}, \Theta_{\mathrm{S}})$ collect the body-aligned, residual-surface, and shared motion parameters. The update can be written as

$$\Theta \leftarrow \Theta - \eta \mathbf{P}(s_k) \nabla_{\Theta} \mathcal{L}, \tag{12}$$

where η is the learning rate and $\mathbf{P}(s_k)$ is a block-diagonal gradient scaling matrix for the three parameter groups. When $w_{\mathrm{R}}(s_k) > 0$, we reduce the update strength of the body-aligned layer

Table 1: Average results on six DNA-Rendering subjects. RRA-Avatar performs best on all metrics; best and second-best values are bold and underlined. Off-LPIPS uses the same method-independent off-body proxy for all methods, and LPIPS* = LPIPS $\times 10^3$.

	Full-image quality			Off-body quality
Method	PSNR↑	SSIM↑	LPIPS* ↓	Off-LPIPS↓
3DGS-Avatar (Qian et al., 2024)	28.63	0.9565	41.43	0.5132
GART (Lei et al., 2024)	29.00	0.9597	44.55	0.5621
GauHuman (Hu et al., 2024b)	29.55	0.9616	40.96	0.4662
ToMiE (Zhan et al., 2025)	30.81	0.9664	35.23	0.4438
SeqAvatar (Xu et al., 2025)	<u>32.05</u>	<u>0.9711</u>	<u>30.91</u>	<u>0.3962</u>
Ours	**32.52**	**0.9760**	**22.98**	**0.3452**

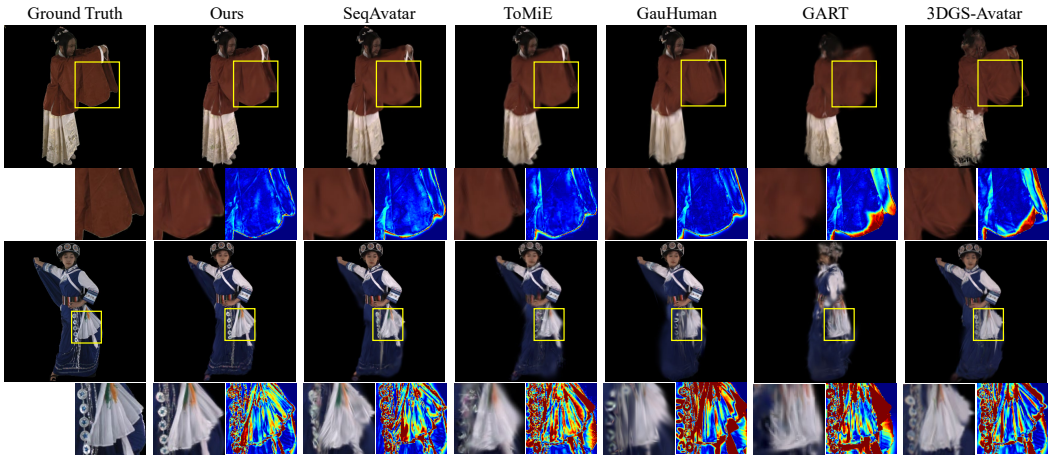


Figure 3: RRA-Avatar preserves loose-garment boundaries and details while reducing RGB error. Each subject shows the rendering, zoomed region, and MAE map on a shared color scale.

and shared motion modules while keeping the residual layer at full strength. This keeps the body-aligned layer stable and lets the residual-surface layer take the main responsibility for off-body regions. Gradient scaling therefore enforces representation responsibility during optimization rather than merely stabilizing training.

Topology control. The same mechanism controls Gaussian densification, pruning, and opacity reset. Topology updates are applied to body Gaussians before residual-opacity guidance, to residual Gaussians while this guidance is active, and to all Gaussians after the guidance is disabled. This allocates capacity to the currently responsible layer and reduces responsibility leakage during subsequent joint refinement. Unlike gradients, topology updates permanently change representational capacity. Aligning densification with representation responsibility prevents irreversible capacity growth in the wrong representation.

4 Experiments

We evaluate RRA-Avatar on novel-view rendering, unseen-pose animation, off-body regional quality, layer allocation, and component ablations.

4.1 Experimental Setup

Datasets. We evaluate on DNA-Rendering (Cheng et al., 2023), I3D-Human (Chen et al., 2024), and ZJU-MoCap (Peng et al., 2021). DNA-Rendering has substantial off-body regions; I3D-Human

Table 2: Average I3D-Human results over four subjects. RRA-Avatar achieves the best PSNR and SSIM in both settings, together with the best novel-view LPIPS and second-best novel-pose LPIPS.

	Novel-view rendering			Novel-pose animation		
Method	PSNR↑	SSIM↑	LPIPS* ↓	PSNR↑	SSIM↑	LPIPS* ↓
3DGS-Avatar (Qian et al., 2024)	30.86	0.9608	34.07	29.54	0.9555	38.59
Dyco (Chen et al., 2024)	31.04	0.9619	30.69	30.09	0.9569	**35.32**
GauHuman (Hu et al., 2024b)	30.13	0.9562	45.37	29.31	0.9526	48.29
SeqAvatar (Xu et al., 2025)	<u>32.24</u>	<u>0.9664</u>	<u>29.78</u>	<u>30.28</u>	<u>0.9583</u>	36.07
Ours	**32.35**	**0.9671**	**29.28**	**30.36**	**0.9589**	<u>35.47</u>

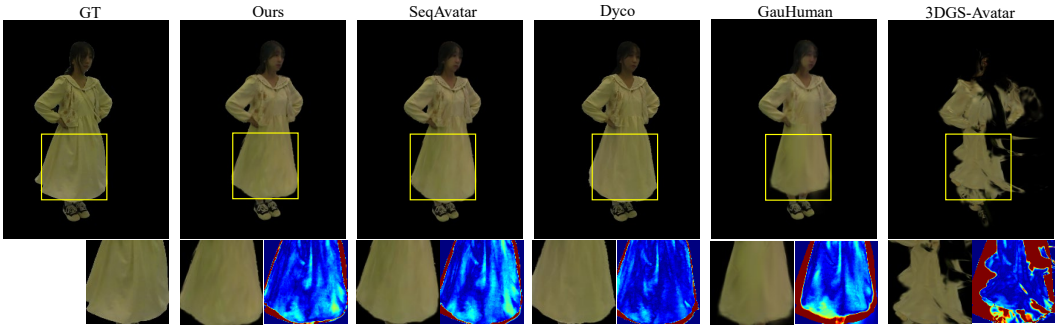


Figure 4: RRA-Avatar preserves loose-garment details and reduces artifacts under novel I3D-Human poses.

covers moderate garment deformation and novel poses; ZJU-MoCap represents body-centric settings. We use SeqAvatar's established DNA/I3D splits and the standard ZJU-MoCap protocol.

Implementation details. RRA-Avatar follows SeqAvatar's hierarchical motion encoder and is trained from scratch using the same SMPL/SMPL-X fits and masks. With matched densification windows and criteria, residual discovery raises the Gaussian count by 6.6% on average; the final model has $1.18\times$ SeqAvatar's Gaussian count and runs at 237.7 FPS.

Baselines and metrics. We retrained and evaluated all compared baselines using the same data splits, camera views, and evaluation protocol. Quantitative comparisons are restricted to methods evaluated under the same data and evaluation protocol. Off-LPIPS averages spatial LPIPS over foreground pixels lying more than 1% of the projected-body bounding-box height outside its mask; all methods use identical proxy regions. We report PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and Off-LPIPS. We exclude methods without matched-protocol results.

4.2 Comparison with State-of-the-Art Methods

DNA-Rendering. Table 1 shows that RRA-Avatar ranks first on all four metrics. Relative to SeqAvatar, it reduces LPIPS* from 30.91 to 22.98 and Off-LPIPS from 0.3962 to 0.3452, while achieving the lowest Off-LPIPS on every subject (Appendix C.2). Figure 3 further shows sharper loose-garment boundaries and lower off-body errors; the explicit separation also reduces interference in body-supported regions.

I3D-Human. On I3D-Human, RRA-Avatar ranks first on all novel-view metrics (Table 2), with lower LPIPS* than SeqAvatar on all four subjects. For novel poses, it ranks first in PSNR and SSIM and second in LPIPS (35.47 versus the best 35.32); Figure 4 shows sharper details and fewer artifacts.

Table 3: Ablation on DNA-Rendering. The full allocation mechanism achieves the best PSNR and LPIPS* and ties for the best SSIM.

Variant	PSNR↑	SSIM↑	LPIPS* ↓
(a) Base model	32.30	0.9742	28.40
(b) + Residual discovery	32.32	0.9743	28.22
(c) + Layer-specific deformation heads	32.36	0.9746	27.62
(d_1) + Allocation w/o supervision	32.49	0.9757	23.27
(d_2) + Allocation w/o topology gating	32.51	0.9759	23.16
(d_3) + Allocation w/o gradient scaling	32.50	**0.9760**	23.02
(d) + Full allocation mechanism	**32.52**	**0.9760**	**22.98**



Figure 5: Responsibility allocation reduces residual-opacity leakage inside the SMPL mask from 51.56% to 32.86% (left) while improving off-body support (right).

On ZJU-MoCap, RRA-Avatar improves all metrics over SeqAvatar and remains competitive with the best methods (31.06 PSNR, 0.9623 SSIM, and 27.47 LPIPS*; see Appendix C.9). Its larger gains on DNA-Rendering suggest particular benefits for loose garments while retaining competitive body-centric performance.

4.3 Ablation Study

Our base adopts SeqAvatar's core architecture. Table 3 shows discovery and separate heads enable responsibility allocation, reducing LPIPS* from 27.62 to 22.98. Removing residual supervision, topology gating, or gradient scaling gives 23.27, 23.16, and 23.02, respectively. The first two provide most gains, while gradient scaling gives a smaller improvement by biasing dominant updates toward the responsible representation. Figure 5 visualizes the resulting allocation.

5 Limitations

RRA-Avatar targets synchronized, calibrated multi-view capture; sparse-view capture remains untested. Mask or body-fit errors can impair residual discovery. Body-prior support only approximates deformation responsibility, so fixed labels may preserve misassignments; occluded or thin structures may remain incomplete.

6 Conclusion

We have presented RRA-Avatar, a dual-layer Gaussian avatar framework for representation-responsibility allocation. It assigns representation ownership by body-prior support and coordinates deformation, supervision, gradients, and topology. Experiments show it reduces LPIPS* from 30.91 to 22.98 on DNA-Rendering and remains competitive in novel-pose and body-centric settings. It improves off-body reconstruction without garment templates or semantic labels. Future work will reduce body-fit and mask dependence through adaptive ownership.

AI USE STATEMENT

In this work, we did not use generative AI tools for any tasks requiring mandatory disclosure under the ICLR 2027 AI Policy. We used generative AI tools to improve the readability of the manuscript, review LaTeX formatting, refine the presentation of figures and tables, and assist with information retrieval and discovery, including identifying potentially relevant related work. All AI-assisted revisions were reviewed by the authors, and all suggested literature was independently checked against the original publications for relevance, accuracy, and bibliographic correctness. We take responsibility for the final content of this work, including all text, claims, citations, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Implementation and training details are provided in the appendix. An anonymized supplementary code package containing training, rendering, evaluation, and dataset-processing scripts is included with the submission.

REFERENCES

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. In *European Conference on Computer Vision (ECCV)*, pp. 491–508. Springer, 2024.

Zhiyi Chen, Hsuan-I Ho, Tianjian Jiang, Jie Song, Manuel Kaufmann, and Chen Guo. Gaussian Wardrobe: Compositional 3D Gaussian avatars for free-form virtual try-on. In *2026 International Conference on 3D Vision (3DV)*, pp. 739–749, 2026. doi: 10.1109/3DV69130.2026.00076. URL https://doi.org/10.1109/3DV69130.2026.00076.

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 (ICCV)*, pp. 19982–19993, 2023.

Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, SA '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394703. doi: 10.1145/3550469.3555423. URL https://doi.org/10.1145/3550469.3555423.

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

Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. GaussianAvatar: Towards realistic human avatar modeling from a single video via animatable 3D Gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 634–644, 2024a.

Shoukang Hu, Tao Hu, and Ziwei Liu. GauHuman: Articulated Gaussian splatting from monocular human videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20418–20431, 2024b.

Daisheng Jin and Ying He. MonoCloth: Reconstruction and animation of cloth-decoupled human avatars from monocular videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5503–5511, 2026. doi: 10.1609/aaai.v40i7.37468.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. doi: 10.1145/3592433.

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 (CVPR)*, pp. 505–515, 2024. doi: 10.1109/CVPR52733.2024.00055.

Joohyun Kwon, Geonhee Sim, and Gyeongsik Moon. Zero-shot reconstruction of animatable 3D avatars with cloth dynamics from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Aldo Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on pattern analysis and machine intelligence*, 16(2):150–162, 1994.

Changmin Lee, Jihyun Lee, and Tae-Kyun Kim. MPMAvatar: Learning 3D Gaussian avatars with accurate and robust physics-based dynamics. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Seungeun Lee, SeungJun Moon, Hah Min Lew, Ji-Su Kang, and Gyeong-Moon Park. SMAGA: Secondary motion-aware 3D clothed Gaussian avatars from monocular videos. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2A3Q2EtGTF.

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 (CVPR)*, pp. 19876–19887, 2024.

Yahui Li, Zhi Zeng, Liming Pang, Guixuan Zhang, and Shuwu Zhang. RealityAvatar: Towards realistic loose clothing modeling in animatable 3D Gaussian avatars, 2025. arXiv preprint arXiv:2504.01559.

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 (CVPR)*, pp. 19711–19722, 2024.

Siyou Lin, Zhe Li, Zhaoqi Su, Zerong Zheng, Hongwen Zhang, and Yebin Liu. LayGA: Layered Gaussian avatars for animatable clothing transfer. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024. doi: 10.1145/3641519.3657501.

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):1–16, 2015. doi: 10.1145/2816795.2818013.

Chunjie Luo, Fei Luo, Yusen Wang, Enxu Zhao, and Chunxia Xiao. DLCA-Recon: Dynamic loose clothing avatar reconstruction from monocular videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3963–3971, 2024. doi: 10.1609/aaai.v38i4.28189.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2022.

Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3D Gaussian avatar. In *European Conference on Computer Vision (ECCV)*, pp. 19–35. Springer, 2024.

Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human Gaussian splatting: Real-time rendering of animatable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 788–798, 2024.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4):1–15, 2022. doi: 10.1145/3528223.3530127.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10975–10985, 2019.

Bo Peng, Yunfan Tao, Haoyu Zhan, Yudong Guo, and Juyong Zhang. Pica: Physics-integrated clothed avatar. *IEEE Transactions on Visualization and Computer Graphics*, 31(12):10762–10776, 2025. doi: 10.1109/TVCG.2025.3614642.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9054–9063, 2021.

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*, pp. 5020–5030, 2024.

Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2885–2896, 2021. doi: 10.1109/CVPR46437.2021.00291.

Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. SplattingAvatar: Realistic real-time human avatars with mesh-embedded Gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1606–1616, 2024.

Geonhee Sim and Gyeongsik Moon. PERSONA: Personalized whole-body 3D avatar with pose-driven deformations from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12670–12680, 2025.

Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. ARAH: Animatable volume rendering of articulated human SDFs. In *European Conference on Computer Vision (ECCV)*, pp. 1–19. Springer, 2022.

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. doi: 10.1109/TIP.2003.819861.

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 (CVPR)*, pp. 16210–16220, 2022.

Wangze Xu, Yifan Zhan, Zhihang Zhong, and Xiao Sun. Sequential gaussian avatars with hierarchical motion context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13592–13603, 2025.

Hang Ye, Xiaoxuan Ma, Hai Ci, Wentao Zhu, and Yizhou Wang. FreeCloth: Free-form generation enhances challenging clothed human modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15987–15997, 2025.

Yifan Zhan, Qingtian Zhu, Muyao Niu, Mingze Ma, Jiancheng Zhao, Zhihang Zhong, Xiao Sun, Yu Qiao, and Yinqiang Zheng. Towards explicit exoskeleton for the reconstruction of complicated 3d human avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14259–14269, 2025. doi: 10.1109/ICCV51701.2025.01323.

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, June 2018.

Yang Zheng, Qingqing Zhao, Guandao Yang, Yifan Wang, Donglai Xiang, Florian Dubost, Dmitry Lagun, Thabo Beeler, Federico Tombari, Leonidas J. Guibas, and Gordon Wetzstein. Physavatar: Learning the physics of dressed 3d avatars from visual observations. In *Computer Vision – ECCV 2024*, pp. 262–284. Springer, 2024. doi: 10.1007/978-3-031-72913-3_15.

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

A Additional Method Details

A.1 Dual-Layer Gaussian Representation

Recall that each canonical Gaussian primitive is defined as $g_i = (\boldsymbol{\mu}_i, \mathbf{q}_i, \mathbf{s}_i, \alpha_i, \mathbf{c}_i, \ell_i)$, where $\ell_i \in \{\mathrm{B}, \mathrm{R}\}$ is its layer label. Here we provide additional details for this representation.

Representation ownership and layer sets. Let $\mathcal{G}_\ell = \{g_i \mid \ell_i = \ell\}$ for $\ell \in \{\mathrm{B}, \mathrm{R}\}$. The full avatar is the union of the body-aligned and residual-surface layers:

$$\mathcal{G} = \mathcal{G}_\mathrm{B} \cup \mathcal{G}_\mathrm{R}. \tag{13}$$

Since each primitive has one layer label, the two sets are disjoint. Following the main paper, representation responsibility specifies which representation should explain a foreground structure, whereas ℓ_i instantiates its concrete representation ownership. The label does not encode a semantic clothing class. Instead, it assigns each primitive to either the body-aligned representation supported by the fitted body prior or the residual-surface representation initialized from unexplained foreground evidence. This ownership selects the primitive's deformation head and determines the layer-specific supervision, gradient scaling, and topology updates applied during optimization.

Applying non-rigid offsets. As described in the main paper, the shared motion encoder produces a motion feature $\mathbf{h}_i^t$, and the layer-specific head H_{ℓ_i} predicts the offsets $(\Delta\boldsymbol{\mu}_i^t, \Delta\mathbf{r}_i^t, \Delta\mathbf{s}_i^t)$. We apply these offsets to the canonical Gaussian before the coarse SMPL-guided warp:

$$\begin{aligned} \bar{\boldsymbol{\mu}}_i^t &= \boldsymbol{\mu}_i + \Delta\boldsymbol{\mu}_i^t, \\ \bar{\mathbf{q}}_i^t &= \mathcal{N}\left(\Delta\mathbf{r}_i^t \otimes \mathbf{q}_i\right), \\ \bar{\mathbf{s}}_i^t &= \mathbf{s}_i + \Delta\mathbf{s}_i^t, \end{aligned} \tag{14}$$

where $\otimes$ denotes the Hamilton product for scalar-first quaternions (w, x, y, z), and $\mathcal{N}(\mathbf{a}) = \mathbf{a}/\|\mathbf{a}\|_2$ enforces a unit quaternion. The remaining Gaussian attributes are unchanged. The updated primitive is then warped to the observation space by the fitted SMPL mesh:

$$g_i^t = \mathcal{W}_\mathrm{SMPL}\left(\bar{g}_i^t; \mathcal{S}_t\right), \tag{15}$$

where $\mathcal{S}_t$ is the posed SMPL mesh at frame t. This design shares global pose-dependent motion context across both representations while allowing the two heads to specialize their local residual deformation. The separate heads implement different deformation assumptions rather than merely serving as additional network capacity.

Layer-wise opacity rendering. Let $\mathcal{G}_\ell^t$ be the deformed Gaussian set of layer ℓ at frame t. The final RGB image is rendered by joint depth-ordered compositing of both layers, as in the main paper. In addition, we render each layer alone to obtain its opacity support:

$$\hat{\mathbf{A}}_{t,v}^\ell = \mathcal{R}_v^\alpha\left(\mathcal{G}_\ell^t\right), \quad \ell \in \{\mathrm{B}, \mathrm{R}\}. \tag{16}$$

Here $\mathcal{R}_v^\alpha$ denotes the opacity output of the same Gaussian rasterizer used for joint rendering. These opacity maps are not composed separately to produce the final image. Instead, the residual-layer opacity provides layer-specific supervision for the responsibility-allocation mechanism, directing unexplained off-body evidence toward the residual-surface representation.

A.2 Body-Prior-Guided Residual-Surface Discovery

The main paper defines the SMPL-guided shell candidates, the residual evidence region $\Omega_{t,v}^\mathrm{R}$, and the multi-view support and contradiction ratios. Here we provide the full definitions and implementation details for pixel evidence, view weighting, filtering, and temporal fusion.

Pixel-level residual evidence. Let $\Omega_{t,v}^\mathrm{F}$ denote the foreground region and $\Omega_{t,v}^{\mathrm{B},\delta}$ denote the foreground pixels inside the dilated SMPL body silhouette. The residual region $\Omega_{t,v}^\mathrm{R}$ is the foreground area outside this body margin, as defined in the main paper. We use a narrow soft band around $\Omega_{t,v}^\mathrm{R}$ to reduce sensitivity to mask noise. Let

$$d_{t,v}^\mathrm{R}(\mathbf{u}) = \inf_{\mathbf{q} \in \Omega_{t,v}^\mathrm{R}} \|\mathbf{u} - \mathbf{q}\|_2 \tag{17}$$

be the distance from pixel $\mathbf{u}$ to the residual region. With $\mathbf{1}[\cdot]$ denoting an indicator function, the positive pixel evidence is

$$e^{+}_{t,v}(\mathbf{u}) = \mathbf{1}[\mathbf{u} \in \Omega^{\mathrm{R}}_{t,v}] + \alpha \mathbf{1}[\mathbf{u} \notin \Omega^{\mathrm{R}}_{t,v}]\mathbf{1}[d^{\mathrm{R}}_{t,v}(\mathbf{u}) \le \epsilon], \tag{18}$$

where α gives weaker support to pixels near the residual boundary. The negative pixel evidence is

$$e^{-}_{t,v}(\mathbf{u}) = \mathbf{1}[d^{\mathrm{R}}_{t,v}(\mathbf{u}) > \epsilon] \cdot \left(\mathbf{1}[\mathbf{u} \notin \Omega^{\mathrm{F}}_{t,v}] + \beta \mathbf{1}[\mathbf{u} \in \Omega^{\mathrm{B},\delta}_{t,v}]\right). \tag{19}$$

Here ϵ is the soft-band width, and β controls the penalty for candidates projected into the body-supported region.

View weighting and filtering. For the valid views defined in the main paper, we use a normal-aware view weight:

$$w^t_v(\mathbf{x}) = m^t_v(\mathbf{x}) \max\left(0, \mathbf{n}^T_{t,a}\mathbf{d}_{t,v}(\mathbf{x})\right). \tag{20}$$

Here $\mathbf{x} = \mathbf{x}_{t,a,j}$ is a shell candidate anchored at SMPL vertex a, $m^t_v(\mathbf{x}) \in \{0,1\}$ is the valid-view indicator, $\mathbf{n}_{t,a}$ is the anchor normal, and $\mathbf{d}_{t,v}(\mathbf{x})$ is the unit direction from $\mathbf{x}$ to the camera center. The normalizer used in the support and contradiction ratios is

$$Z_t(\mathbf{x}) = \sum_v w^t_v(\mathbf{x}) + \epsilon_0, \tag{21}$$

where ϵ_0 avoids division by zero.

After the support and contradiction tests in the main paper, we apply two extra filters. First, we remove candidates whose supporting projections are far from the foreground boundary or too close to the dilated SMPL silhouette. Second, when several shell candidates at the same SMPL vertex have similar evidence, we keep the outer one because it better covers off-body structures. The numerical thresholds are listed in Section B.

Canonicalization and temporal fusion. For the local-frame mapping mentioned in the main paper, let a retained candidate $\mathbf{x}$ be anchored at vertex a. Let (η_1, η_2, η_3) be its coordinates in the observed local frame formed by two tangent directions and the outward normal. Its canonical position is

$$\mathbf{x}^c = \mathbf{s}^c_a + \eta_1 \mathbf{t}^c_{1,a} + \eta_2 \mathbf{t}^c_{2,a} + \max(\eta_3, 0)\mathbf{n}^c_a, \tag{22}$$

where $\mathbf{s}^c_a$, $\mathbf{t}^c_{1,a}$, $\mathbf{t}^c_{2,a}$, and $\mathbf{n}^c_a$ are the canonical anchor position, tangent directions, and outward normal. The outward coordinate is clamped to avoid moving residual candidates inside the body.

We then group canonical candidates by their anchor vertices and cluster them in canonical space. For each cluster, we average candidate positions using their evidence scores and record how many different frames support the cluster. Clusters with weak temporal support are discarded. The remaining cluster centers form the initial residual-surface Gaussian seeds:

$$\mathcal{G}^0_{\mathrm{R}} = \{g_i \mid \boldsymbol{\mu}_i \in \mathcal{X}^c_{\mathrm{R}}, \ell_i = \mathrm{R}\}, \tag{23}$$

where $\mathcal{X}^c_{\mathrm{R}}$ denotes the fused canonical residual points. These seeds provide reliable geometric support for unexplained off-body structures before the responsibility-allocation mechanism further refines the residual-surface representation.

A.3 Responsibility-allocation Mechanism

The main paper describes how layer-specific supervision, gradient scaling, and topology gating enforce representation ownership throughout optimization. Here we provide the detailed schedules and formulations for the gradient and topology controls.

The duration of residual-opacity guidance is adapted to the amount of reliable residual evidence in each dataset. For DNA-Rendering, it is activated after 11,000 iterations and applied for 5,000 iterations. For I3D-Human and ZJU-MoCap, it is activated after 500 and 100 iterations, respectively,

and applied for 100 iterations. DNA-Rendering contains larger and more reliable off-body regions, for which a longer guidance interval helps establish representation responsibility for unexplained foreground structures. The residual regions in I3D-Human and ZJU-MoCap are smaller, so the residual mask serves as a brief ownership-allocation cue before both representations continue joint refinement under the full-foreground loss.

Gradient scaling matrix. Following the update rule in the main paper, we use a block-diagonal gradient scaling matrix for $\Theta = (\Theta_{\mathrm{B}}, \Theta_{\mathrm{R}}, \Theta_{\mathrm{S}})$, which collects body-layer parameters, residual-layer parameters, and shared motion parameters:

$$\mathbf{P}(s_k) = \begin{bmatrix} \kappa_{\mathrm{B}}(s_k) I_{\mathrm{B}} & 0 & 0 \\ 0 & \kappa_{\mathrm{R}}(s_k) I_{\mathrm{R}} & 0 \\ 0 & 0 & \kappa_{\mathrm{S}}(s_k) I_{\mathrm{S}} \end{bmatrix}. \tag{24}$$

Here I_{B}, I_{R}, and I_{S} are identity matrices for the three parameter blocks. The scalars κ_{B}, κ_{R}, and κ_{S} control the update strength of the body layer, residual layer, and shared motion modules.

When residual supervision is active, we reduce $\kappa_{\mathrm{B}}(s_k)$ and $\kappa_{\mathrm{S}}(s_k)$ while keeping $\kappa_{\mathrm{R}}(s_k)$ at full strength. This prevents the body-aligned representation and shared motion modules from absorbing off-body errors, allowing the residual-surface representation to receive the dominant updates for unexplained foreground structures. Gradient scaling therefore enforces representation responsibility during optimization rather than serving only as numerical stabilization. After residual-opacity guidance is disabled, the scaling factors return to their joint-refinement values so that both representations can improve the final foreground rendering.

Topology control. We use the same progress variable s_k to decide which layer can densify, prune, and reset opacity. Early in training, topology updates are applied to body-aligned Gaussians. This gives the body layer enough capacity to form stable SMPL-supported geometry. When residual supervision is active, topology updates are applied to residual-surface Gaussians. This lets the residual layer grow or remove primitives where multi-view residual evidence supports off-body structures. After residual-opacity guidance is disabled, topology updates are applied to all Gaussians, allowing the two layers to jointly refine uncertain boundaries.

Newly cloned or split Gaussians inherit the layer label of their parent, and existing Gaussians do not switch labels. Thus, topology evolution changes the number and placement of primitives without changing representation ownership. This preserves layer specialization during responsibility allocation while allowing later joint refinement.

B Implementation Details

B.1 Training Details

Each model is trained for 30,000 iterations on DNA-Rendering, 13,800 iterations on I3D-Human, and 3,000 iterations on ZJU-MoCap. DNA-Rendering uses SMPL-X fits, while I3D-Human and ZJU-MoCap use SMPL fits. All experiments use the provided or preprocessed foreground masks and fitted body models under the same data protocol as the compared methods. We implement our method using PyTorch 2.7.0 on an RTX 5090.

Residual-surface discovery settings. For body-prior-guided residual-surface discovery, we sample four exterior shells with normalized offsets $\mathcal{D} = \{0.01, 0.03, 0.06, 0.10\}$, measured by the diagonal length of the fitted body bounding box. We use every second vertex of the fitted body mesh as an anchor. The residual mask is obtained by dilating the projected body silhouette with a 5×5 elliptical kernel for one iteration and taking the foreground area outside the dilated silhouette. We further use 3×3 morphological opening and closing operations, and remove connected components smaller than 40 pixels.

For pixel-level evidence, the soft positive band width is $\epsilon = 2$ pixels, the soft positive weight is $\alpha = 0.3$, and the body-region negative weight is $\beta = 0.5$ relative to the unit background penalty. For multi-view voting, we use a minimum support ratio of 0.30 and a maximum contradiction ratio of 0.50. The same hit-count threshold is applied to residual, foreground-boundary, and body-gap hits.

Table 4: Initial Gaussian counts. Body and Residual denote the body-aligned and residual-surface layers, respectively.

Subject	Total	Body	Residual
1_0206_04	11,268	10,475	793
2_0007_04	12,038	10,475	1,563
2_0019_10	10,684	10,475	209
2_0044_11	10,934	10,475	459
2_0051_09	11,142	10,475	667
2_0813_05	10,927	10,475	452
Average	11,166	10,475	691

Table 5: Final Gaussian counts after training on DNA-Rendering. RRA-Avatar averages $1.18\times$ the number of Gaussians used by SeqAvatar under matched densification iteration settings.

Subject	SeqAvatar	RRA-Avatar
1_0206_04	121,801	132,141
2_0007_04	95,456	127,872
2_0019_10	94,207	125,604
2_0044_11	122,070	138,349
2_0051_09	122,449	135,059
2_0813_05	107,235	126,685
Average	110,536	130,952

Foreground-boundary hits require the projection to lie near the foreground boundary, while body-gap hits require it to stay outside the dilated body silhouette. For DNA-Rendering, this threshold is two supporting views. For I3D-Human and ZJU-MoCap, it is one view. DNA-Rendering contains looser garments and larger off-body regions, so stricter multi-view agreement helps reduce false residual seeds. The other two datasets have smaller residual regions, where a lower threshold keeps useful candidates. The foreground-boundary threshold is 4 pixels, and the body-gap threshold is 5 pixels. When several shell candidates at the same anchor pass the filters, we use an outer-shell selection tolerance of 0.08. Candidates whose scores are within 8% of the best score are treated as similarly supported, and we keep the outermost one.

For temporal fusion, canonical candidates are clustered with voxel size 0.01 times the canonical body bounding-box diagonal. We keep at most one cluster for each anchor vertex. A fused seed must have a positive evidence score and enough temporal support. The minimum support count is set to 2 when the number of fused frames is at most 10 or the provisional seed count is below 300, and is set to 3 otherwise.

As shown in Table 4, the body initialization is fixed across subjects, whereas the residual seed count adapts to detected off-body evidence and averages 691, adding 6.6% to the 10,475 body seeds. As reported in Table 5, after training, RRA-Avatar averages 130,952 Gaussians versus 110,536 for SeqAvatar, corresponding to a $1.18\times$ increase.

Responsibility-allocation mechanism settings. We use the Adam optimizer. For Gaussian parameters, densification, pruning, opacity reset, and spherical-harmonic color scheduling, we follow the standard 3D Gaussian Splatting setting. The maximum spherical harmonic degree is set to 3 for all experiments.

During the responsibility-allocation mechanism, the residual layer keeps full gradient strength when residual supervision is active. We use dataset-level gradient scales for four parameter groups: body-layer Gaussian parameters, the body-layer deformation head, shared motion modules, and SMPL pose/LBS modules. For DNA-Rendering, the gradient scales are $(0.2, 0.2, 0.5, 0.5)$ and the residual-opacity loss weight is 1.0. For I3D-Human and ZJU-MoCap, where residual regions are smaller, the gradient scales are $(0.8, 0.8, 0.8, 0.8)$ and the residual-opacity loss weight is 0.001. The first two entries separately instantiate the body-parameter block for the Gaussian parameters and deformation

Table 6: Training and test camera splits on I3D-Human. Frame stride denotes the temporal sampling interval used for evaluation.

Subject	Training views	Test views	Frame stride
ID1_1	1, 3, 6, 8	2, 4, 7, 9	1
ID1_2	0, 2, 5, 7	1, 6, 8, 9	1
ID2_1	1, 3, 5, 7, 9	2, 4, 6, 10	3
ID3_1	1, 3, 5, 7, 9	2, 4, 6, 8, 10	3

head, while the last two entries instantiate the shared-parameter block for the motion modules and SMPL pose/LBS modules; the residual-parameter block keeps unit gradient strength. The relative AIAP multipliers $\lambda_{\mathrm{A}}(s_k)$ before, during, and after residual-opacity guidance are $(1.0, 0.25, 1.0)$ for DNA-Rendering, $(1.0, 0.02, 10.0)$ for I3D-Human, and $(1.0, 0.8, 1.0)$ for ZJU-MoCap, respectively. These dataset-level configurations are selected using only training-view residual-mask reliability and then fixed across all subjects and splits; no held-out test metric or subject-specific tuning is used. The residual loss and gradient scales follow two regimes: DNA-Rendering uses stronger guidance for its large and stable off-body evidence, whereas I3D-Human and ZJU-MoCap share weaker guidance for their smaller residual regions.

Architecture and shared settings. The shared hierarchical motion encoder follows the SeqAvatar architecture. The two layer-specific deformation heads have identical architectures but separate parameters, and all model parameters are trained from scratch. Exact per-dataset configurations and execution commands are included in the submitted code.

B.2 Dataset and Evaluation Details

Dataset splits. We follow the same splits as the prior work. On DNA-Rendering, we evaluate six subjects: 1_0206_04, 2_0007_04, 2_0019_10, 2_0044_11, 2_0051_09, and 2_0813_05. We use even-indexed training cameras from views 0 to 46, and evaluate novel-view rendering on even-indexed held-out cameras from views 48 to 58. Training uses the first 100 frames, while novel-view evaluation uses every fifth frame from the same range.

On I3D-Human, we evaluate ID1_1, ID1_2, ID2_1, and ID3_1. Following the dataset protocol, each subject has a training split, a novel-view split, and a novel-pose split. The camera splits and frame sampling steps are shown in Table 6.

On ZJU-MoCap, we evaluate CoreView_377, CoreView_386, CoreView_387, CoreView_392, CoreView_393, and CoreView_394. We train on camera 0 and test on cameras 1 to 22. The test frames are sampled every 30 frames following the common ZJU-MoCap protocol.

Evaluation protocol. All results are rendered with the original camera calibration and image resolution of each dataset. Ground-truth images are filled with the dataset background outside the foreground mask during data loading. At evaluation time, rendered images are filled with the same background outside the rectangular foreground bounding region before metric computation. We report PSNR, SSIM, and LPIPS. LPIPS is computed with the VGG backbone. Following the main paper, we report LPIPS* $=$ LPIPS $\times$ 10^3 for compact tables. All metrics are averaged over the evaluated images of each split. For qualitative comparisons, we use the same test frames and camera views as the quantitative evaluation.

C Additional Experimental Results

C.1 Detailed Quantitative Results on DNA-Rendering

The main paper reports the average novel-view rendering performance over six DNA-Rendering subjects. Table 7 provides the complete per-subject results. Among the evaluated methods, RRA-Avatar achieves the best PSNR, SSIM, and LPIPS* on all six subjects. Relative to SeqAvatar, the per-subject LPIPS* reductions range from 4.59 to 14.26 (mean 7.92), showing that the average gain is not driven by a single sequence.

Table 7: Per-subject novel-view rendering results on DNA-Rendering. Best and second-best results for each subject and metric are marked in bold and underlined, respectively. LPIPS* = LPIPS$\times 10^3$.

Subject	Metric	3DGS-Avatar	GART	GauHuman	ToMiE	SeqAvatar	**Ours**
1_0206_04	PSNR↑	26.64	27.19	27.96	29.64	<u>30.81</u>	**31.74**
	SSIM↑	0.9439	0.9454	0.9500	0.9581	<u>0.9649</u>	**0.9733**
	LPIPS* ↓	50.09	56.47	50.42	44.47	<u>39.47</u>	**25.21**
2_0007_04	PSNR↑	28.06	28.22	27.93	29.02	<u>29.63</u>	**29.72**
	SSIM↑	0.9492	0.9525	0.9496	0.9533	<u>0.9566</u>	**0.9605**
	LPIPS* ↓	56.70	56.46	56.93	49.39	<u>43.72</u>	**33.32**
2_0019_10	PSNR↑	32.06	31.13	31.63	32.56	<u>34.97</u>	**35.38**
	SSIM↑	0.9710	0.9686	0.9698	0.9719	<u>0.9777</u>	**0.9821**
	LPIPS* ↓	31.12	35.16	32.82	28.64	<u>23.53</u>	**17.16**
2_0044_11	PSNR↑	28.23	28.90	30.38	32.02	<u>32.62</u>	**33.36**
	SSIM↑	0.9512	0.9561	0.9641	0.9715	<u>0.9747</u>	**0.9801**
	LPIPS* ↓	37.15	41.27	34.19	26.52	<u>23.83</u>	**18.06**
2_0051_09	PSNR↑	25.04	26.47	25.99	27.41	<u>28.45</u>	**28.88**
	SSIM↑	0.9534	0.9628	0.9599	0.9645	<u>0.9685</u>	**0.9725**
	LPIPS* ↓	42.22	42.21	42.00	36.92	<u>34.06</u>	**27.91**
2_0813_05	PSNR↑	31.75	32.06	33.43	34.22	<u>35.81</u>	**36.03**
	SSIM↑	0.9701	0.9729	0.9764	0.9790	<u>0.9843</u>	**0.9873**
	LPIPS* ↓	31.30	35.74	29.42	25.41	<u>20.83</u>	**16.24**

C.2 Off-Body Regional Evaluation on DNA-Rendering

Because off-body surface ground truth and semantic annotations are unavailable, we define a method-independent image-space proxy using the ground-truth foreground mask and the projected fitted-body mask. Let $M^{\mathrm{F}}_{t,v}$ and $M^{\mathrm{B}}_{t,v}$ denote these two masks for frame t and view v, respectively. The off-body proxy is

$$\Omega^{\mathrm{off}}_{t,v} = \left\{ \mathbf{u} \;\middle|\; M^{\mathrm{F}}_{t,v}(\mathbf{u}) = 1,\ \mathrm{dist}\left(\mathbf{u}, M^{\mathrm{B}}_{t,v}\right) > \tau H_{t,v} \right\}, \tag{25}$$

where $H_{t,v}$ is the height of the tight bounding box of the projected body mask. We use $\tau = 0.01$ as the default margin for the per-subject evaluation and additionally examine sensitivity to this choice below. For every value of τ, the evaluation region is derived independently of the evaluated method and is shared by all methods. The Off-LPIPS proxy shares the foreground and projected-body masks used by residual-opacity supervision, but uses a $0.01H$ distance margin rather than a 5×5 body-mask dilation. The regions therefore overlap. We interpret Off-LPIPS as held-out-view appearance quality in a common off-body proxy, not as independent evidence of garment geometry.

We compute a full-image spatial LPIPS map $D_{t,v}(\mathbf{u})$ using the VGG backbone and LPIPS version 0.1. The images are not cropped or masked before computing this map. Let $m_{t,v}(\mathbf{u}) = \mathbf{1}[\mathbf{u} \in \Omega^{\mathrm{off}}_{t,v}]$. We define

$$\text{Off-LPIPS}_\tau = \frac{\sum_{(t,v)\in\mathcal{Q}} \sum_{\mathbf{u}} m_{t,v}(\mathbf{u}) D_{t,v}(\mathbf{u})}{\sum_{(t,v)\in\mathcal{Q}} \sum_{\mathbf{u}} m_{t,v}(\mathbf{u})}, \tag{26}$$

where $\mathcal{Q}$ is the set of evaluated frame–view pairs. This pixel-weighted aggregation gives more weight to views containing larger off-body regions. Views with an empty proxy region contribute to neither the numerator nor the denominator. All methods use the same images, masks, proxy regions, and evaluation resolution.

Table 8 reports the per-subject results using the default margin $\tau = 0.01$. RRA-Avatar achieves the lowest Off-LPIPS on all six subjects. Its average score is 0.3452, compared with 0.3962 for SeqAvatar, corresponding to a relative reduction of 12.9%.

Sensitivity to the proxy margin. To examine whether the regional comparison depends on the default choice of τ, we recompute Off-LPIPS using $\tau \in \{0.005, 0.01, 0.02, 0.03\}$. At each margin, all methods are evaluated using the same images, fitted-body masks, and proxy regions. Table 9 reports the average results over the six DNA-Rendering subjects.

Table 8: Per-subject Off-LPIPS on DNA-Rendering using $\text{Off-LPIPS}_{\tau=0.01}$; lower is better. The proxy regions are derived solely from the ground-truth foreground and projected fitted-body masks and are identical across methods. Best and second-best results are marked in bold and underlined, respectively.

Method	1_0206_04	2_0007_04	2_0019_10	2_0044_11	2_0051_09	2_0813_05	Average
3DGS-Avatar	0.4783	0.5535	0.5786	0.5227	0.5075	0.4386	0.5132
GART	0.5745	0.5623	0.5537	0.5799	0.5375	0.5648	0.5621
GauHuman	0.5059	0.4822	0.4273	0.4935	0.5054	0.3831	0.4662
ToMiE	0.4662	0.4482	0.4290	0.4539	0.4724	0.3931	0.4438
SeqAvatar	<u>0.4124</u>	<u>0.4336</u>	<u>0.3410</u>	<u>0.4432</u>	<u>0.4312</u>	<u>0.3159</u>	<u>0.3962</u>
Ours	**0.3296**	**0.3435**	**0.3014**	**0.4175**	**0.4011**	**0.2781**	**0.3452**

Table 9: Average Off-LPIPS on six DNA-Rendering subjects under different off-body proxy margins; lower is better. For each τ, all methods use identical proxy regions. Best and second-best results are marked in bold and underlined, respectively.

Method	τ=0.005	τ=0.01	τ=0.02	τ=0.03
3DGS-Avatar	0.4971	0.5132	0.5156	0.5279
GART	0.5599	0.5621	0.5676	0.5759
GauHuman	0.4472	0.4662	0.4855	0.5090
ToMiE	0.4312	0.4438	0.4666	0.4895
SeqAvatar	<u>0.3735</u>	<u>0.3962</u>	<u>0.4060</u>	<u>0.4256</u>
Ours	**0.3330**	**0.3452**	**0.3635**	**0.3825**

RRA-Avatar achieves the lowest Off-LPIPS at every evaluated margin. Relative to SeqAvatar, it reduces Off-LPIPS by 10.8%, 12.9%, 10.5%, and 10.1% for increasing values of τ, respectively. Although the absolute scores increase as the proxy is restricted to foreground pixels farther from the projected body, the ranking of all compared methods remains unchanged. The regional improvement is therefore stable across the evaluated range rather than being specific to the default choice $\tau = 0.01$.

Because the evaluation region remains an image-space proxy rather than a semantic off-body annotation, we use Off-LPIPS as complementary evidence alongside the full-image metrics and qualitative comparisons.

C.3 Additional Qualitative Results on DNA-Rendering

Figure 6 shows more novel-view qualitative results on DNA-Rendering. These examples cover different subjects, viewpoints, and loose-garment shapes. Across these cases, RRA-Avatar preserves wide sleeves, skirts, and garment hems more faithfully than competing methods. The rendered boundaries are cleaner, and off-body regions are less likely to be blurred, contracted, or attached to the fitted body.

The close-up views and error maps further show the benefit of the residual-surface layer. Regions that are weakly supported by the body prior have lower local errors and more stable appearance details. These results provide additional visual evidence that separating body-aligned and residual-surface Gaussians improves loose-garment rendering under novel views.

C.4 Detailed Novel-View Results on I3D-Human

The main paper reports the average novel-view rendering performance over four I3D-Human subjects. Table 10 provides the complete per-subject results. RRA-Avatar ranks first in 10 out of the 12 subject-metric comparisons and second in the remaining two, suggesting that its average improvement is consistent across subjects rather than being dominated by a single sequence.

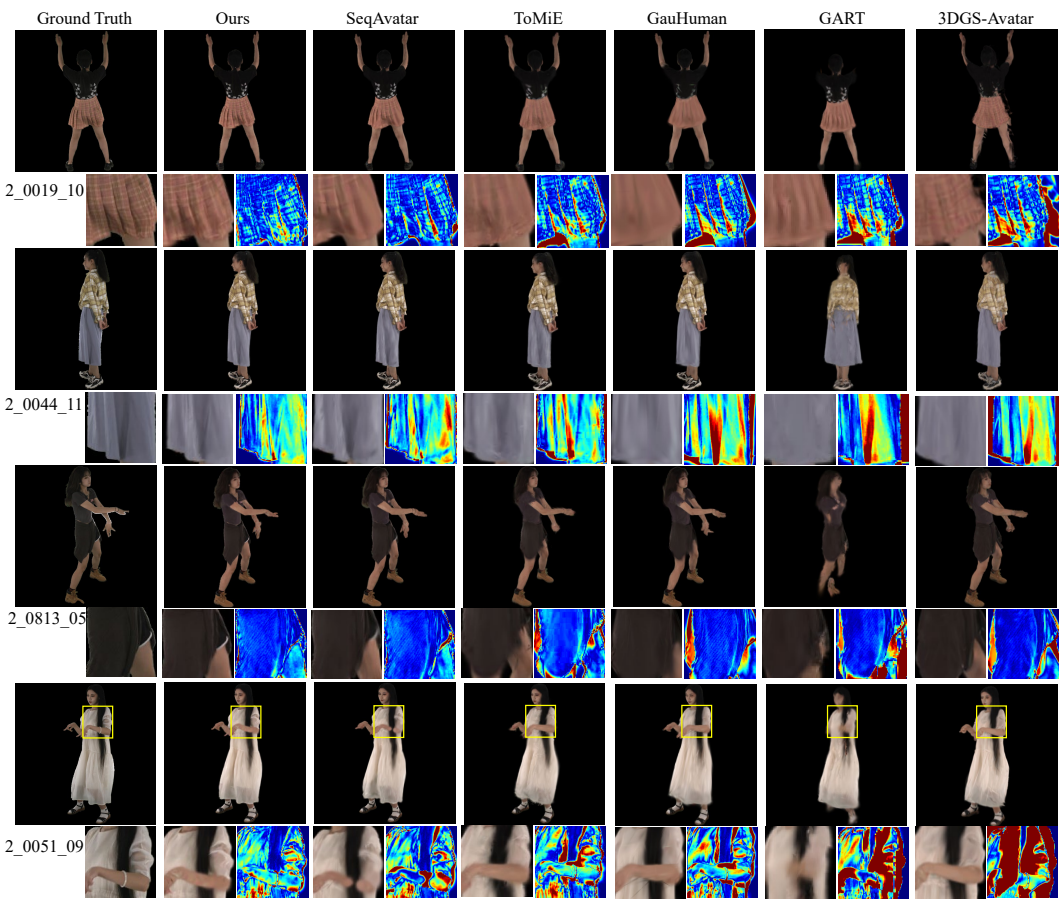


Figure 6: Additional qualitative results on DNA-Rendering.

Table 10: Per-subject novel-view rendering results on I3D-Human. Best and second-best results for each subject and metric are marked in bold and underlined, respectively. LPIPS* = LPIPS $\times 10^3$.

Subject	Metric	3DGS-Avatar	Dyco	GauHuman	SeqAvatar	**Ours**
ID1_1	PSNR↑	29.78	30.81	29.47	<u>31.81</u>	**31.96**
	SSIM↑	0.9600	0.9617	0.9571	<u>0.9659</u>	**0.9665**
	LPIPS* ↓	31.40	28.29	39.46	<u>27.03</u>	**26.41**
ID1_2	PSNR↑	31.29	31.45	30.42	<u>31.98</u>	**32.13**
	SSIM↑	0.9627	0.9628	0.9585	<u>0.9660</u>	**0.9665**
	LPIPS* ↓	28.09	**25.75**	38.87	27.84	<u>27.51</u>
ID2_1	PSNR↑	29.63	29.41	28.51	**31.53**	<u>31.51</u>
	SSIM↑	0.9604	0.9618	0.9546	<u>0.9685</u>	**0.9695**
	LPIPS* ↓	37.33	33.26	49.46	<u>30.23</u>	**29.56**
ID3_1	PSNR↑	32.73	32.50	32.10	<u>33.62</u>	**33.79**
	SSIM↑	0.9602	0.9612	0.9546	<u>0.9651</u>	**0.9660**
	LPIPS* ↓	39.47	35.46	53.67	<u>34.02</u>	**33.64**

C.5 Novel-View Qualitative Results on I3D-Human

Figure 7 shows novel-view qualitative comparisons on I3D-Human. Compared with DNA-Rendering, this dataset contains more everyday clothing and moderate garment motion, so the visual differences are often smaller. Even in this setting, RRA-Avatar produces cleaner body and garment boundaries, especially around clothing edges and hair.

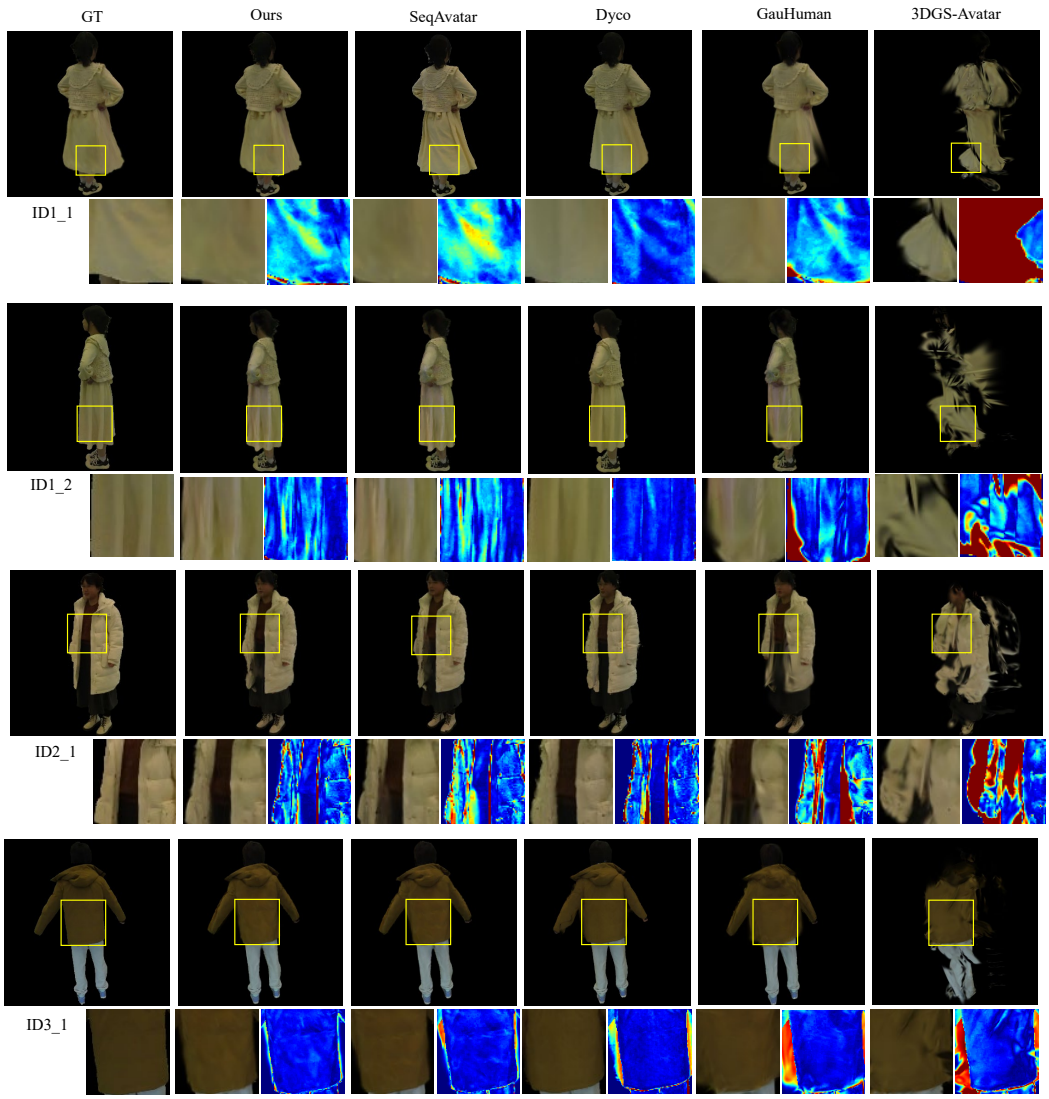


Figure 7: Qualitative comparison on I3D-Human for novel-view rendering.

The close-up regions show that our method keeps local texture and shape details more stable across held-out views. Baselines sometimes blur thin structures or introduce small artifacts near body-clothing boundaries. RRA-Avatar reduces these artifacts by separating body-supported surfaces from residual off-body regions while still rendering both layers jointly. These results complement the quantitative novel-view results in the main paper.

C.6 Additional Novel-Pose Qualitative Results on I3D-Human

Figure 8 provides more qualitative results for novel-pose generalization on I3D-Human. These examples include unseen body poses and different garment motions. RRA-Avatar keeps the body shape stable while preserving loose cloth and hair. The rendered silhouettes remain clean under pose changes, and local details are less blurred around body-clothing boundaries.

These results further show that the proposed dual-layer representation is not limited to the training poses. The body-aligned layer follows SMPL-supported motion, while the residual-surface layer helps represent regions that move differently from the fitted body. Joint rendering of the two layers produces coherent avatars under new poses.

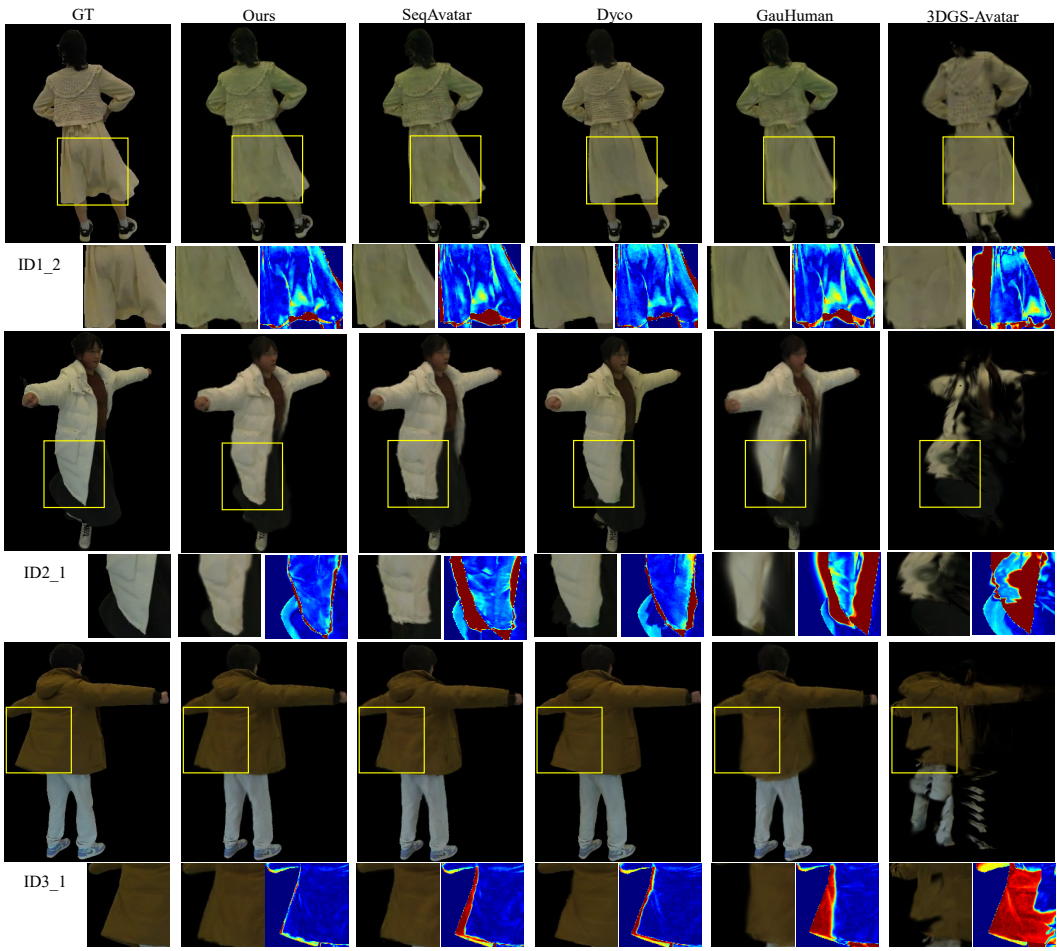


Figure 8: Additional qualitative results on I3D-Human for novel-pose generalization.

Table 11: Controlled comparisons on six DNA-Rendering subjects. The three baselines use the same fixed initial Gaussian budget, matched to the dataset-average initial budget of RRA-Avatar. Best results are marked in bold.

Variant	PSNR↑	SSIM↑	LPIPS* ↓
Single set (matched budget)	32.39	0.9748	28.11
Random two-layer split	32.39	0.9744	28.07
Distance-based split	32.37	0.9745	28.08
ours	**32.52**	**0.9760**	**22.98**

C.7 Controlled Capacity and Naive Layer-Partitioning Baselines

To separate the effect of representation-responsibility allocation from the effects of Gaussian count, naive layer assignment, and deformation capacity, we construct three controlled variants. Table 11 compares these variants with the full RRA-Avatar model under the same training and evaluation protocol. The three controlled variants use the same fixed total number of initialized Gaussians, chosen to match the dataset-average initial budget of the full RRA-Avatar model. Results are averaged over the six DNA-Rendering subjects.

Controlled variants. Variant (a) uses a single Gaussian set and one deformation head. To approximately match the initial Gaussian budget of the full dual-layer model, we sample an additional 691

Gaussians from the fitted SMPL-X mesh vertices. This number equals the average residual-surface seed count over the six DNA-Rendering subjects, as reported in Table 4. This variant tests whether the improvement can be explained by a larger representation budget alone.

Variant (b) randomly selects 691 initialized Gaussians for one layer and assigns all remaining Gaussians to the other layer. Each layer is processed by an independent deformation head with the same architecture as the layer-specific heads in the full model. The random assignment is fixed before training. Variant (b) uses the same number of Gaussians in each layer as variant (c), but assigns the Gaussians randomly.

Variant (c) replaces the random assignment with a deterministic radial partition. Let $\boldsymbol{\mu}_i = (\mu_{i,x}, \mu_{i,y}, \mu_{i,z})$ denote the canonical position of the i-th initialized Gaussian. We first compute a robust coordinate-wise center:

$$\mathbf{c}_{\mathrm{med}} = (\mathrm{median}_i\, \mu_{i,x}, \mathrm{median}_i\, \mu_{i,y}, \mathrm{median}_i\, \mu_{i,z})\,. \tag{27}$$

The radial distance of each Gaussian from this center is

$$d_i = \|\boldsymbol{\mu}_i - \mathbf{c}_{\mathrm{med}}\|_2\,. \tag{28}$$

We sort all initialized Gaussians by their radial distances in descending order. Let π be the resulting permutation such that

$$d_{\pi(1)} \geq d_{\pi(2)} \geq \cdots \geq d_{\pi(N)}, \tag{29}$$

where N is the total number of initialized Gaussians. We assign the 691 Gaussians farthest from the robust center to the outer layer and the remaining $N - 691$ Gaussians to the inner layer:

$$\ell^{\mathrm{dist}}_{\pi(k)} = \begin{cases} \mathrm{outer}, & 1 \leq k \leq 691, \\ \mathrm{inner}, & 692 \leq k \leq N. \end{cases} \tag{30}$$

The two layers use separate deformation heads, as in variant (b). However, this assignment depends only on global radial distance and does not use body-prior support or multi-view unexplained-foreground evidence.

Analysis. Matching the Gaussian budget in variant (a) produces strong PSNR and SSIM, but its LPIPS* remains at 28.11, compared with 22.98 for the full model. Therefore, increasing the number of Gaussians alone does not account for the improvement in perceptual quality.

Randomly introducing two layers and separate deformation heads in variant (b) changes the results only marginally: PSNR remains at 32.39, SSIM decreases from 0.9748 to 0.9744, and LPIPS* changes from 28.11 to 28.07. This indicates that additional layer-specific deformation capacity is insufficient when representation ownership is assigned arbitrarily.

Variant (c) evaluates a representative geometry-based assignment strategy that separates Gaussians according to their distances from a global robust center. It performs almost identically to the random assignment, with an LPIPS* of 28.08 versus 28.07. This indicates that global radial distance is not an adequate criterion for assigning representation responsibility: body-supported extremities may be far from the global center, while loose garment regions near the torso may remain relatively close despite being poorly explained by the body prior.

Together, these controlled comparisons show that the improvement cannot be explained solely by a larger Gaussian budget or additional layer-specific deformation capacity. They further show that the choice of layer partition is important. Random assignment does not encode meaningful representation ownership, while global radial distance cannot reliably distinguish body-supported and off-body structures: body extremities can be far from the global center, whereas loose garment regions near the torso can remain relatively close.

C.8 Qualitative Ablation Study

Figure 9 shows the visual effect of each component on DNA-Rendering. The baseline reconstructs the main body, but loose garment regions are often blurred or incomplete. Residual-surface discovery provides initial geometric support for foreground structures that are not adequately explained by the fitted body prior. However, without layer-specific deformation and responsibility allocation, both representations can still compete to explain the same image evidence.

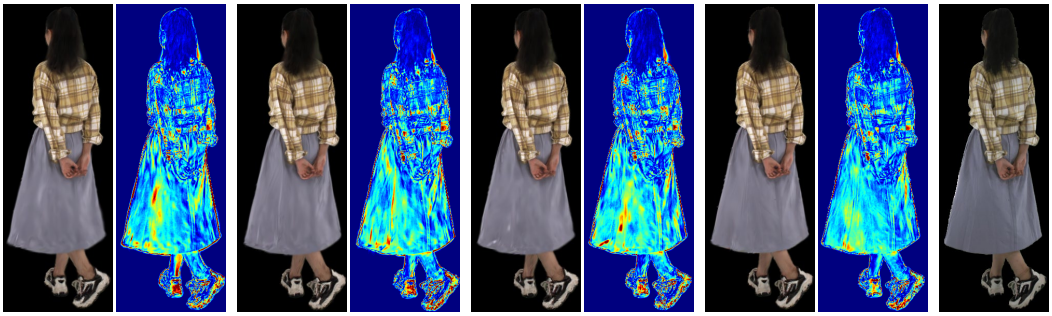


Figure 9: Qualitative ablation study on DNA-Rendering. We compare the visual effect of each component on subject 2_0044_11.

Table 12: Quantitative results on ZJU-MoCap. Best and second-best results are marked in bold and underlined, respectively. LPIPS* = LPIPS $\times$ 10^3.

Method	PSNR↑	SSIM↑	LPIPS* ↓
Dyco	30.37	0.9599	29.47
3DGS-Avatar	30.59	0.9609	**27.09**
SeqAvatar	31.02	0.9619	28.89
GauHuman	31.04	0.9620	31.81
ToMiE	**31.10**	**0.9630**	31.20
Ours	<u>31.06</u>	<u>0.9623</u>	<u>27.47</u>

Adding the layer-specific heads allows the two representations to specialize their local residual deformation, but does not yet enforce representation ownership throughout optimization. The full model coordinates supervision, dominant gradient updates, and topology evolution, producing sharper loose-garment boundaries while preserving stable body-supported regions. This visual trend is consistent with the quantitative ablation results in the main paper.

C.9 Results on ZJU-MoCap

Table 12 reports quantitative results on ZJU-MoCap. This dataset is more controlled and mostly body-supported, so it is less focused on large loose-garment deformation than DNA-Rendering. RRA-Avatar still achieves competitive results. It obtains the second-best PSNR and SSIM, and LPIPS* close to the best result. This shows that the proposed dual-layer representation remains competitive on standard body-centric human rendering benchmarks.

Figure 10 shows qualitative results on ZJU-MoCap. Since most visible structures are adequately explained by the fitted body prior, the body-aligned representation receives most of the representation responsibility, while the residual-surface representation provides limited additional support. These results complement the DNA-Rendering and I3D-Human results by showing that RRA-Avatar remains competitive when unexplained off-body structures are sparse.

C.10 Rendering Speed

We evaluate the rendering throughput of RRA-Avatar and SeqAvatar under the same hardware and evaluation protocol. Both methods are tested on an RTX 5090 GPU with batch size one, using the original image resolution, frames, and camera views of DNA-Rendering. We measure the complete forward pass used to produce each rendered image, including Gaussian rasterization, while excluding data loading, disk I/O, and metric computation. CUDA is synchronized before and after the timed forward passes, and FPS is averaged over all evaluated frames after a warm-up period.

As shown in Table 13, RRA-Avatar achieves an average rendering speed of 237.70 FPS, compared with 321.23 FPS for SeqAvatar. The dual-layer representation introduces additional computation,

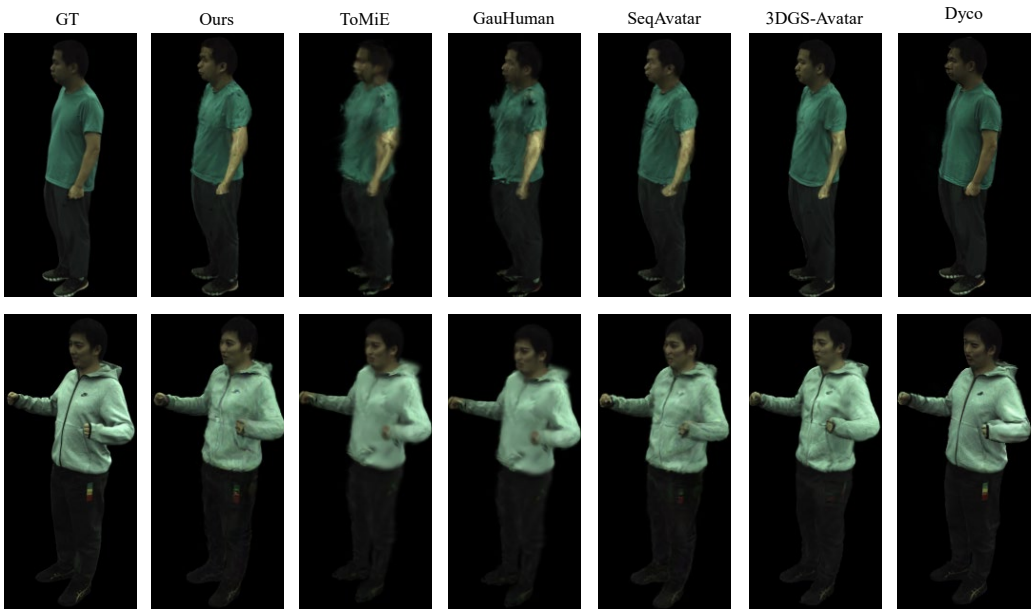


Figure 10: Qualitative results on ZJU-MoCap.

Table 13: Rendering throughput in FPS on six DNA-Rendering subjects; higher is better. Both methods are evaluated on the same RTX 5090 GPU with identical image resolutions, frames, and camera views.

Subject	SeqAvatar	RRA-Avatar
1_0206_04	330.33	236.41
2_0007_04	345.07	274.95
2_0019_10	347.58	227.69
2_0044_11	307.87	237.32
2_0051_09	274.15	227.09
2_0813_05	322.39	222.74
Average	321.23	237.70

but RRA-Avatar remains above 220 FPS on every subject and therefore retains high-throughput real-time rendering.

C.11 Video Comparisons

We provide synchronized RGB comparison videos on DNA-Rendering. Each comparison video contains pose IDs 0–99 at 15 FPS, with identical frames and held-out test viewpoints across methods. We also provide Ours-only videos rendered along smoothly interpolated virtual camera trajectories.