DSAR: Dual-Stream Autoregressive Modeling of Temporal Cloth Dynamics for Photorealistic Animatable Avatars

Haozhong Xiong[1], Yao Yu[1]*, Yu Zhou[1]*, and Sidan Du[1]

Nanjing University, Nanjing, China
502024230011@smail.nju.edu.cn, {allanyu,nackzhou,coff128}@nju.edu.cn

Fig. 1: We propose DSAR, which explicitly models temporal causality in cloth dynamics through dual-stream autoregressive architecture, achieving realistic deformations on loose clothing and robust generalization to out-of-distribution poses.

Abstract. Creating photorealistic and temporally coherent animatable human avatars from RGB videos remains challenging. Current methods struggle to capture realistic cloth dynamics, producing over-smoothed appearance or severe artifacts on out-of-distribution poses. This limitation stems from a fundamental oversight: existing approaches neglect the temporal causality inherent in cloth physics, where current states emerge from previous states through temporal evolution rather than instantaneous skeletal configurations alone. Without explicit modeling of this causal structure, networks learn pose-appearance correlations instead of motion evolution, leading to poor generalization. We introduce a dual-stream autoregressive framework that explicitly models both observable geometric information and implicit internal state. The geometric stream propagates surface displacement from the previous frame, while the state stream fuses current features with historical states retrieved from a memory bank. Motion-adaptive aggregation handles spatially-varying dynamics, and adaptive regularization balances smoothness with flexibility.Experiments on challenging datasets demonstrate significant improvements in rendering quality, temporal consistency, and generalization to motion patterns beyond training distributions, validating that dual-stream temporal modeling enables realistic cloth dynamics.

Keywords: Animatable avatars · Neural rendering · Dual-Stream

* Corresponding authors.

1 Introduction

The creation of photorealistic, animatable human avatars has emerged as a cornerstone technology for virtual reality, telepresence, film production, and interactive media. The ultimate goal is to capture the dynamic interplay between human motion and garment appearance—subtle wrinkles forming during arm bending, fabric momentum during rapid spinning, and gradual cloth settling after motion ceases. These temporal phenomena are fundamental to visual realism yet remain challenging for current neural rendering approaches.

Recent advances in neural rendering have enabled impressive progress in avatar creation from multi-view RGB videos. Building on foundational techniques in neural radiance fields [3,40] and 3D Gaussian Splatting [23], numerous methods [17,24,27,31,64] have achieved high-fidelity avatar reconstruction. Despite differences in underlying representations, these methods share a common modeling paradigm: they model garment appearance as functions of current or recent skeletal poses, where pose information is derived from either single frames or concatenated across temporal window. While this design enables efficient per-frame reconstruction, it exhibits fundamental limitations in practice—oversmoothed appearance on training poses, temporal flickering during motion, and severe artifacts including unrealistic wrinkles on novel pose sequences.

We identify that this limitation stems from ignoring the temporal causal structure inherent in cloth motion. In reality, garment appearance at time t is not uniquely determined by skeletal pose alone. Consider a standing pose reached after rapid spinning versus from rest: while skeletal configurations are nearly identical, cloth states differ drastically due to motion history—the post-spinning garment exhibits residual momentum and dynamic wrinkles absent in the static case. This reveals a fundamental one-to-many mapping where identical poses correspond to different cloth states depending on temporal context.

Analysis of real-world cloth motion reveals that resolving this ambiguity requires two complementary types of information. The first is **explicit kinematic information**—the observable geometric configuration (where cloth surfaces are positioned) and motion patterns (how surfaces are moving, characterized by velocity and directional momentum), which together describe the current observable state of cloth surfaces. The second is **implicit internal state**, encompassing latent properties such as fabric tension, deformation history, and material memory that influence cloth appearance but cannot be directly observed from surface geometry alone. Only when both are accounted for can the cloth state at time t be uniquely determined.

Skeletal poses provide only body joint positions, lacking both cloth kinematic information (geometric configuration and motion patterns) and internal state. This dual deficiency creates fundamental prediction ambiguity—without explicit cloth state feedback, networks must infer both "where and how cloth surfaces are moving" and "what internal factors govern their evolution" from skeletal motion alone, leading to the observed one-to-many mapping problem.

Various approaches have attempted to address this issue. RealityAvatar [28] and MonoHuman [70] optimize per-frame latent codes to disambiguate iden-

tical poses during training, but rely on pose-based interpolation at test time that cannot capture motion-dependent variations. InstantAvatar [21] and HumanRF [20] incorporate historical skeletal poses through concatenation or transformer aggregation to provide temporal context, but skeletal representations capture only body-centric motion—where the skeleton was but not how cloth has evolved—creating ambiguity when identical poses correspond to different cloth states. Test-time alignment strategies [9, 31] improve generalization by mapping novel poses back to training distributions, but remain effective primarily when test poses exhibit small distributional shifts from training data.

Despite their diversity in architectural design, these learning-based avatar rendering methods do not establish explicit temporal causality between successive cloth states, instead learning correlations between pose patterns and appearance rather than temporal evolution rules.

Physics-based cloth simulation [4, 5, 13, 29, 53] naturally incorporates temporal causality through explicit state evolution, tracking cloth configurations across timesteps to model momentum and dynamic deformations. However, these methods are fundamentally limited by human modeling capabilities—the precision of manually-crafted material models, the completeness of physical constraints, and the approximations necessary for computational tractability may fail to capture the full complexity of real-world cloth behavior. Our method addresses cloth dynamics from a complementary data-driven perspective, learning temporal evolution rules directly from multi-view photometric observations. This image-based supervision enables the network to discover dynamics from visual evidence of real-world phenomena, potentially capturing effects that extend beyond manually-modeled physics.

We address this through explicit dual-stream modeling. The geometric stream establishes cloth geometric information propagation via autoregressive conditioning on the previous frame's predicted deformation $\Delta\mu_{t-1}$, providing explicit feedback on cloth surface configuration and movement rather than inferring them from body pose alone. The state stream maintains a memory bank of historical temporal states that encode implicit internal state, capturing latent properties that are difficult to model explicitly but govern temporal evolution. These two streams address the dual requirements: the geometric stream tracks observable kinematics deformation, while the state stream captures implicit internal state learned from visual data.

Contributions. Our work makes the following contributions:

- We identify that temporally coherent cloth rendering requires modeling both observable kinematic information and implicit internal state, which cannot be inferred from skeletal poses alone due to the one-to-many mapping problem.
- We propose a dual-stream autoregressive architecture that explicitly addresses these dual requirements from photometric supervision: a geometric stream propagates kinematic information through autoregressive conditioning on previous frame's cloth deformation and motion-adaptive temporal

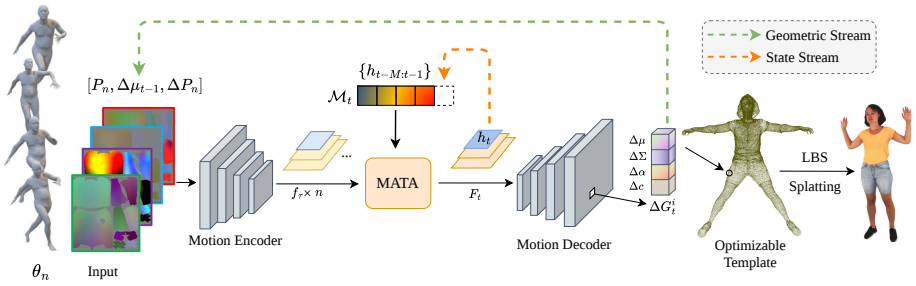

Fig. 2: Overview. Our dual-stream framework processes a temporal window and previous deformation $\Delta\mu_{t-1}$ through two complementary pathways. The geometric stream aggregates multi-scale features via MATA. The state stream fuses aggregated features with historical states from memory bank $\mathcal{M}_t$ to obtain temporal state h_t, which is decoded into Gaussian deformations ΔG_t and stored for next-frame inference.

aggregation, while a state stream encodes implicit internal state through memory-based retrieval of historical temporal features.
– Our framework achieves significant improvements in rendering quality, temporal consistency, and generalization to motion patterns beyond training distributions.

2 Related Work

Animatable Human Avatar Representation. Early reconstruction methods employ implicit functions, such as occupancy fields [16, 19, 37–39, 50, 51] and signed distance functions [10, 44, 52, 60, 63, 74], or utilize explicit mesh representations [15, 25, 68, 69] to reconstruct clothed humans from scans or depth sequences. However, these representations often struggle to model view-dependent appearance effects and complex materials, limiting photorealistic rendering quality and temporal consistency.

Building on neural rendering techniques [34, 48, 55, 56, 59], neural radiance field-based methods have revolutionized avatar creation through differentiable volumetric rendering with view-dependent appearance. Recent advances [6, 11, 43] have enabled numerous human modeling approaches [14, 27, 30, 32, 47, 57, 64, 75]. HumanNeRF [67] learns pose-dependent deformations using skeletal motion priors, while Neural Body [47] associates latent codes with SMPL vertices for pose-conditioned rendering. AnimatableNeRF [46] establishes a canonical NeRF with pose refinement for pose-driven animation. TAVA [27] adopts explicit warping fields for fine-grained control. However, these implicit representations are computationally expensive, and their MLP-based architectures struggle to capture high-frequency geometric details due to spectral bias [58].

3D Gaussian Splatting [23] enables real-time rendering through explicit point-based primitives, inspiring various avatar modeling approaches [2, 17, 18, 41, 54,

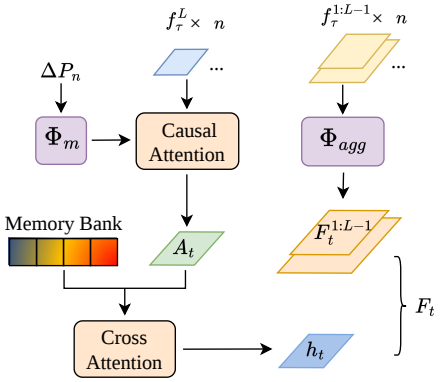

Fig. 3: Motion-Adaptive Temporal Aggregation (MATA). Features are aggregated hierarchically: motion-adaptive causal attention at the deepest level (f_τ^L) and lightweight convolutions (Φ_{agg}) at finer levels ($f_\tau^{1:L-1}$) .

76]. GaussianAvatar [17] binds Gaussians to UV maps with learnable features for pose-driven animation. Animatable-GS [31] employs StyleUNet [22, 62] for multi-scale deformation prediction with learnable skinning weights. To enhance geometric quality, several methods [33, 72] introduce geometric priors including as-rigid-as-possible constraints and normal consistency regularization. Despite achieving impressive rendering performance, these methods fundamentally model clothing appearance as functions of current or recent skeletal poses, formulated as $\text{Appearance}_t = F(\text{pose}_{t-n:t})$, neglecting the temporal causality where current cloth states emerge from previous states through evolution.

Temporal Modeling in Neural Avatar Rendering. Various neural rendering approaches have explored temporal modeling to capture cloth dynamics beyond instantaneous pose-to-appearance mappings.

Per-frame latent encoding approaches [7, 28, 70, 73] optimize frame-specific codes to disambiguate identical poses during training, capturing variability that skeletal pose alone cannot explain. RealityAvatar [28] learns per-frame latent codes jointly with neural rendering, while MonoHuman [70] employs 4D human representation with spatio-temporal features. However, these codes are optimized independently without temporal constraints between frames. At test time, codes for novel poses must be inferred through pose-based interpolation—finding training frames with similar skeletal configurations and blending their optimized codes—which cannot capture motion-dependent variations where identical poses correspond to different cloth states depending on motion history.

Skeletal pose concatenation methods [20, 21, 67] incorporate historical skeletal poses as network input to provide temporal context through body motion sequences. InstantAvatar [21] concatenates multiple consecutive pose parameters through temporal encoders, while HumanRF [20] employs transformers to aggregate pose sequences for dynamic human modeling. Neural Cloth Simulation [5] processes body motion through recurrent encoders with disentangled

static and dynamic branches, enabling fast cloth animation through learned temporal dynamics in latent space. However, these skeletal representations provide only body-centric motion information—they capture how the skeleton moved but not the actual cloth geometric state from previous frames. This creates ambiguity when identical poses correspond to different cloth states due to motion history.

Test-time alignment strategies [9,31] address generalization to novel poses by mapping out-of-distribution poses back to the training distribution. Animatable-GS [31] employs PCA-based dimensionality reduction to project test poses into the training manifold, while RANA [9] introduces explicit pose space alignment through learned transformations. These methods remain effective primarily when test poses exhibit small distributional shifts from training data, as pose similarity-based alignment cannot capture motion-dependent cloth state variations.

Despite their diversity in architectural design, these learning-based methods do not establish explicit temporal causality between successive cloth states, instead learning correlations between pose patterns and appearance rather than temporal evolution rules, limiting their ability to generalize beyond training sequences.

Physics-based cloth simulation [4, 5, 8, 13, 29, 42, 53] models garment dynamics through explicit state evolution, naturally incorporating temporal causality but bounded by manually-modeled physical constraints. Our work focuses on photorealistic avatar rendering from multi-view images, learning cloth dynamics directly from photometric observations while establishing explicit temporal causality through autoregressive geometric state propagation and distributed memory-based state retrieval.

3 Method

3.1 Preliminary

3D Gaussian Splatting. We adopt 3D Gaussian Splatting [23] as our rendering representation. Each Gaussian is parameterized by its 3D center $\mu \in \mathbb{R}^3$, rotation quaternion $q \in \mathbb{R}^4$, anisotropic scale $s \in \mathbb{R}^3$, opacity $\alpha \in \mathbb{R}$, and spherical harmonics coefficients c for view-dependent color. The covariance matrix is factorized as $\Sigma = RSS^TR^T$, where R is the rotation matrix derived from quaternion q, and S is a diagonal scaling matrix. During rendering, 3D Gaussians are projected onto the 2D image plane and blended via differentiable alpha compositing.

SMPL Model and Linear Blend Skinning. We utilize SMPL [35] and SMPL-X [45] as the underlying body model, parameterized by shape $\beta \in \mathbb{R}^{10}$ and pose $\theta \in \mathbb{R}^{J\times 3}$, where J denotes the number of body joints. SMPL provides a template mesh $\mathcal{T}(\beta)$ in canonical space and joint locations $\mathcal{J}(\beta)$ that enable articulated motion. To transform points from canonical space to posed space, we employ Linear Blend Skinning (LBS). Given a canonical point x_c with skinning weights $w = \{w_1, \ldots, w_K\}$ and bone transformation matrices $\{B_1, \ldots, B_K\}$

derived from pose θ with K joints, the posed position is:

$$x_p = \sum_{k=1}^{K} w_k B_k x_c \tag{1}$$

3.2 Overview

We represent the avatar using a learnable Gaussian template G_{tmp} in canonical space. At each frame, the network predicts deformations ΔG_t that modify the template's Gaussian attributes. These deformations are added to the template to obtain deformed canonical Gaussians, which are then transformed to posed space via Linear Blend Skinning and rendered through differentiable splatting.

Our deformation network Ψ adopts a StyleUNet [22, 49, 62] architecture and learns cloth dynamics through a dual-stream autoregressive process. As shown in Fig. 2, the geometric stream extracts multi-scale features through an encoder and aggregates them temporally using Motion-Adaptive Temporal Aggregation (MATA) to produce geometric features A_t. The state stream fuses A_t with historical states from a memory bank to obtain temporal state h_t, which is decoded to predict canonical-space deformations. During training, multi-view photometric supervision and adaptive temporal regularization guide the network to discover temporal evolution patterns.

3.3 Learnable Gaussian Template

We model avatar dynamics by predicting deformations of a learnable Gaussian template in canonical space, which are then transformed to posed space via Linear Blend Skinning.

Template Initialization. We initialize an optimizable Gaussian template G_{tmp} in canonical space by binding Gaussians to the SMPL-X surface using its UV parameterization. Each Gaussian is placed at a UV coordinate location with skinning weights inherited from nearby vertices. The number of Gaussians depends on the UV map resolution and remains fixed throughout training, ensuring spatial correspondence across frames.

Template and Deformation Factorization. We factorize the canonical Gaussians into a learnable template and frame-specific deformations:

$$G_t^{\text{cano}} = G_{\text{tmp}} + \Delta G_t \tag{2}$$

where $\Delta G_t = \{\Delta\mu_t, \Delta q_t, \Delta s_t, \Delta\alpha_t, \Delta c_t\}$ represents deformations in Gaussian attributes. The template captures average canonical geometry and appearance, while deformations encode pose-dependent and temporally-driven variations.

Posed Space Transformation. To render under driving pose θ_t, we transform the deformed canonical Gaussians to posed space via LBS. For a canonical Gaussian with center μ_c and covariance Σ_c:

$$\mu_p = \sum_{k=1}^{K} w_k B_k \mu_c, \quad \Sigma_p = J \Sigma_c J^T \tag{3}$$

where J is the rotation component of the bone transformation matrices. The posed Gaussians are rendered to the target viewpoint through splatting-based rasterization.

3.4 Dual-Stream Autoregressive Architecture

At each time step t, we predict deformations through two complementary autoregressive streams. The geometric stream explicitly conditions on the previous frame's predicted geometric deformation $\Delta\mu_{t-1}$ and processes the current temporal window through motion-adaptive aggregation. The state stream maintains a memory bank of historical temporal states and fuses them with current geometric features to capture long-term temporal evolution.

Our deformation prediction conditions on the immediate previous $\Delta\mu_{t-1}$ rather than a full historical sequence. This single-frame design is sufficient because $\Delta\mu_{t-1}$ was itself predicted from earlier frames through the autoregressive chain, thereby carrying cloth geometric history.

Geometric Stream. The network input consists of UV-unwrapped position maps derived from skeletal pose and gaussian template. For each frame τ in the temporal window $\{t-n+1,\ldots,t\}$, the input comprises the position map p_τ, velocity map Δp_τ, and the geometric information $\Delta\mu_{t-1}$:

$$I_\tau = \text{Concat}[p_\tau, \Delta p_\tau, \Delta\mu_{t-1}] \tag{4}$$

where p_τ captures surface configuration, $\Delta p_\tau = p_\tau - p_{\tau-1}$ captures motion dynamics, and $\Delta\mu_{t-1}$ is shared across all frames in the window, providing cloth geometric deformation from the previous frame.

We extract multi-scale features through StyleUNet encoder Ψ_{enc}, producing an L-level feature pyramid for each frame, where $l=1$ denotes the finest resolution and $l=L$ the coarsest:

$$\{f_\tau^l\}_{\tau=t-n+1}^{t}, \quad l \in \{1,\ldots,L\} \tag{5}$$

This yields n feature pyramids, one for each frame in the temporal window.

Motion-Adaptive Temporal Aggregation. Cloth exhibits spatially heterogeneous motion characteristics, requiring adaptive processing rather than uniform aggregation. At the deepest semantic level (Level L), we employ motion-adaptive temporal aggregation to process the temporal window.

We first quantify motion magnitude at each spatial location from the velocity maps to capture motion heterogeneity across different body regions:

$$V_\tau[u,v] = \|\Delta p_\tau[u,v]\|_2 \tag{6}$$

where $V_\tau[u,v]$ represents the motion magnitude at spatial location (u,v) in frame τ. A lightweight network Φ_{m} transforms motion magnitude maps into feature-space embeddings:

$$m_\tau = \Phi_{\text{m}}(V_\tau), \quad \tau \in \{t-n+1,\ldots,t\} \tag{7}$$

where Φ_{m} consists of lightweight convolutions that project spatial motion patterns into the feature dimension. Following [26], we incorporate motion embeddings and temporal position encodings into the deepest features:

$$\tilde{f}_\tau^L = f_\tau^L + m_\tau + e_\tau, \quad \tau \in \{t-n+1, \ldots, t\} \tag{8}$$

where e_τ are position embeddings. This motion-enhanced representation enables the network to attend adaptively based on local motion characteristics—high-motion regions emphasize recent dynamic information while low-motion regions maintain broader temporal context.

We then compute causal self-attention [1, 61] over the n motion-enhanced frames:

$$Q = \Phi_Q(\{\tilde{f}_\tau^L\}), \quad K = \Phi_K(\{\tilde{f}_\tau^L\}), \quad V = \Phi_V(\{\tilde{f}_\tau^L\}) \tag{9}$$

where Φ_Q, Φ_K, Φ_V are projection networks. We apply causal masking to enforce temporal causality:

$$M_{\text{causal}}[\tau, \tau'] = \begin{cases} -\infty, & \text{if } \tau < \tau' \\ 0, & \text{if } \tau \geq \tau' \end{cases} \tag{10}$$

where $\tau, \tau' \in \{t-n+1, \ldots, t\}$ denote frame indices in the temporal window. The final attention output aggregates temporal information within the window while respecting causal constraints:

$$A_t = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M_{\text{causal}}\right)V \tag{11}$$

This motion guidance enables spatially-adaptive attention patterns, where actively moving regions focus on recent frames to capture transient dynamics, while stable regions distribute attention more uniformly to maintain temporal coherence.

State Stream. While A_t captures surface configuration and motion patterns observable from the temporal window, cloth dynamics also depend on implicit internal state that is not directly observable from geometry alone. The state stream addresses this by maintaining a memory bank of historical temporal states that encode this implicit internal state and fusing them with current geometric features.

Specifically, we maintain a memory bank $\mathcal{M}_t = \{h_{t-m}, \ldots, h_{t-1}\}$ storing the most recent M temporal states, where each h_i encodes the implicit internal state from previous timesteps. At time t, we retrieve relevant temporal information from the memory bank and fuse it with the current geometric feature A_t through cross-attention:

$$h_t^* = \text{CrossAttn}(Q = A_t,\, KV = \mathcal{M}_t) \tag{12}$$

$$h_t = (1-\lambda)A_t + \lambda \cdot h_t^* \tag{13}$$

where λ is a learnable parameter that balances current geometric information and historical temporal patterns.

After obtaining h_t, we update the memory bank for the next timestep:

$$\mathcal{M}_{t+1} = \{h_{t-M+1}, \ldots, h_{t-1}, h_t\} \tag{14}$$

Hierarchical Aggregation and Decoding. The fused temporal state h_t serves as the aggregated feature at the deepest level (Level L). At shallower levels (Levels 1 to $L-1$), we use lightweight convolutional networks Φ^l_{agg} for efficient temporal aggregation:

$$F^l_t = \begin{cases} h_t, & l = L \\ \Phi^l_{\text{agg}}(\{f^l_\tau\}^t_{\tau=t-n+1}), & l \in \{1, \ldots, L-1\} \end{cases} \tag{15}$$

The hierarchical aggregation produces the final feature pyramid $\{F^l_t\}^L_{l=1}$, combining dual-stream fusion at the deepest semantic level with efficient convolutions at finer detail levels. The aggregated multi-scale features are fed into StyleUNet decoder Ψ_{dec} with skip connections from the encoder, which predicts the canonical-space deformations:

$$\Delta G_t = \Psi_{\text{dec}}(\{F^l_t\}^L_{l=1}) \tag{16}$$

3.5 Adaptive Temporal Regularization

Uniform temporal smoothness constraints create fundamental tension: they stabilize static poses but over-smooth dynamic deformations, suppressing natural momentum effects. We resolve this through motion-magnitude-aware weighting that dynamically adapts regularization strength based on spatially-varying motion characteristics.

Motion-Magnitude-Aware Weighting. For each Gaussian i at frame t, we compute an adaptive weight based on its local motion magnitude. Let (u_i, v_i) denote the UV coordinates of Gaussian i in the canonical template. We compute per-Gaussian adaptive weights:

$$w_t[i] = \exp\left(-\frac{V_t[u_i, v_i]}{\tau_{\text{reg}}}\right) \tag{17}$$

where $V_t[u_i, v_i]$ is the motion magnitude at the Gaussian's spatial location and τ_{reg} is a hyperparameter controlling sensitivity. This formulation achieves spatially-varying adaptive regularization: high weights for static regions enforce strong temporal consistency, while low weights for high-motion regions permit flexibility.

Multi-Level Regularization. We apply motion-weighted regularization at three hierarchical levels:

Feature-Level Smoothness regularizes intermediate features at each pyramid level:

$$\mathcal{L}_{\text{feat}} = \sum_{l=1}^{L} \sum_{t} \gamma_l \cdot \|F^l_t - F^l_{t-1}\|^2_2 \tag{18}$$

where $\gamma_l = 0.1$ controls regularization strength at each level. This prevents high-frequency temporal variations in learned representations that could lead to flickering artifacts.

Prediction-Level Consistency regularizes predicted deformations with motion-adaptive weighting:

$$\mathcal{L}_{\text{pred}} = \sum_t \sum_i w_t[i] \cdot \Big(\|\Delta\mu_t[i] - \Delta\mu_{t-1}[i]\|_2^2 + \|\Delta q_t[i] - \Delta q_{t-1}[i]\|_2^2 + \|\Delta s_t[i] - \Delta s_{t-1}[i]\|_2^2 \Big) \tag{19}$$

The spatially-varying weights $w_t[i]$ enforce strong consistency in static regions (e.g., torso during standing) while permitting flexibility in dynamic regions (e.g., sleeves during arm swinging).

Cross-Frame Position Smoothness penalizes second-order temporal differences in posed positions:

$$\text{accel}_t[i] = \mu_{t+1}^{\text{p}}[i] - 2\mu_t^{\text{p}}[i] + \mu_{t-1}^{\text{p}}[i] \tag{20}$$

$$\mathcal{L}_{\text{sm}} = \sum_t \sum_i w_t[i] \cdot \|\text{accel}_t[i]\|_2^2 \tag{21}$$

where $\mu_t^{\text{p}}[i]$ denotes the posed position of Gaussian i at frame t. This acceleration penalty minimizes jerk for smooth trajectories while motion-adaptive weighting permits natural momentum-driven motion in active regions.

Training Objective. Our complete training loss combines multi-view photometric supervision with adaptive temporal regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{render}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}} + \lambda_{\text{pred}}\mathcal{L}_{\text{pred}} + \lambda_{\text{sm}}\mathcal{L}_{\text{sm}} \tag{22}$$

The rendering loss comprises:

$$\mathcal{L}_{\text{render}} = \lambda_{\text{rgb}}\mathcal{L}_{\text{rgb}} + \lambda_{\text{ssim}}\mathcal{L}_{\text{ssim}} + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}} \tag{23}$$

where $\mathcal{L}_{\text{rgb}}$ is the L1 loss between rendered and ground-truth images, $\mathcal{L}_{\text{ssim}}$ [66] measures structural similarity, and $\mathcal{L}_{\text{lpips}}$ [71] captures perceptual quality. These photometric losses provide supervision from multiple camera views, enabling the network to learn accurate cloth deformations from visual observations alone without requiring explicit geometric supervision.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on 4D-DRESS [65] and AvatarREX [75]. Each subject contains 5-6 motion sequences (approximately 1000 frames total, captured by 8 calibrated cameras). We train on 3-4 sequences and test on a

separate held-out sequence. To further validate generalization to challenging out-of-distribution poses, we additionally evaluate on motion sequences from AMASS [36], which provides diverse motion patterns including sports, dancing, and acrobatic movements substantially different from training data.

Distribution Divergence Quantification. To systematically evaluate generalization under varying distribution shifts, we quantify pose distribution divergence between training and test sets using Maximum Mean Discrepancy (MMD) [12]. Given pose feature sets $\mathcal{X}_{\text{train}}$ and $\mathcal{X}_{\text{test}}$ extracted from SMPL-X parameters, MMD measures distributional distance via kernel embeddings:

$$\begin{aligned} \text{MMD}^2 =& \|\mu_{\mathcal{X}} - \mu_{\mathcal{Y}}\|_{\mathcal{H}}^2 \\ =& \mathbb{E}[k(x, x')] - 2\mathbb{E}[k(x, y)] + \mathbb{E}[k(y, y')] \end{aligned} \tag{24}$$

where k is an RBF kernel and $\mathcal{H}$ is the reproducing kernel Hilbert space. We categorize test sequences as Near-Distribution (ND) with MMD < 0.3 (similar motion patterns) or Far-Distribution (FD) with MMD > 0.6 (substantially different dynamics), enabling rigorous evaluation of out-of-distribution generalization.

Baselines. We compare against HumanNeRF [67], GaussianAvatar [17], and Animatable-GS [31], representing NeRF and Gaussian Splatting paradigms. All baselines use official implementations trained on identical data with recommended settings.

Metrics. We evaluate rendering quality using PSNR, SSIM [66] and LPIPS [71].

Implementation Details. We employ a StyleUNet architecture with temporal window size $n = 10$, feature pyramid levels $L = 4$, UV map resolution $H \times W = 512 \times 512$, $\sim$200k Gaussians per subject, memory bank size $m = 10$ frames. Training uses Adam optimizer (learning rate $lr = 6 \times 10^{-4}$). Loss weights: $\lambda_{\text{render}} = 1$, $\lambda_{\text{ssim}} = 0.2$, $\lambda_{\text{lpips}} = 0.5$, $\lambda_{\text{feat}} = 0.01$, $\lambda_{\text{pred}} = 0.1$, $\lambda_{\text{sm}} = 0.05$.

4.2 Comparisons

Table 1 presents quantitative evaluation. Our method consistently outperforms baselines, confirming that explicit modeling of kinematic information and implicit internal state enables robust generalization.

Figure 4 demonstrates our method's ability to capture realistic cloth dynamics. In post-rotation stopping (row 1), our method correctly models momentum-driven evolution: the jacket continues swinging after the body stops, then gradually settles. Baselines produce artifacts due to instantaneous modeling—without explicit kinematic feedback and internal state encoding, they infer appearance solely from current pose. Similar improvements appear in arm movements (row 2) and jumping (row 3), where our dual-stream architecture captures natural dynamics while baselines exhibit flickering and over-smoothing.

4.3 Ablation Studies

We systematically validate each component's contribution through ablation studies. Table 2 and Figure 5 compare five variants: (a) Ground Truth, (b) Full model,

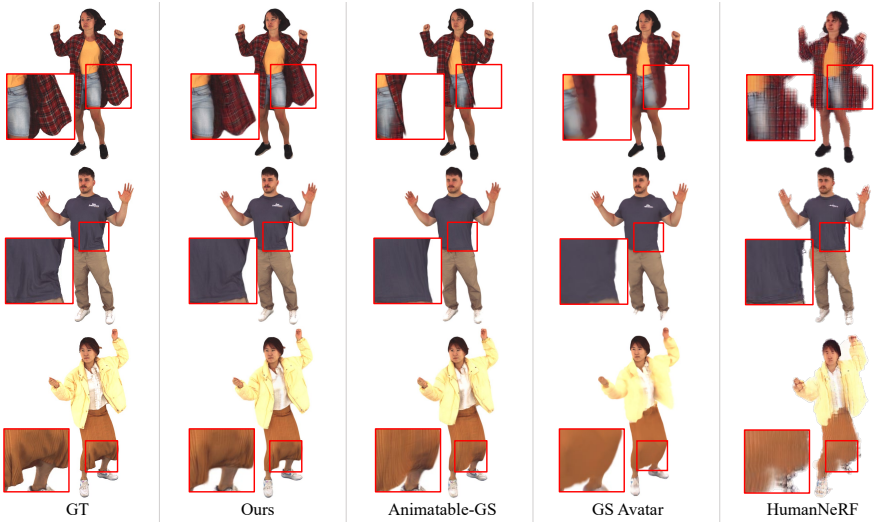

Fig. 4: Qualitative comparison with other methods on challenging test poses. Our method better captures momentum-driven cloth dynamics and generalizes to novel poses.

Table 1: Quantitative comparison. All methods are evaluated at 940×1280 resolution.

Method	PSNR↑	SSIM↑	LPIPS↓
HumanNeRF	20.0567	0.9121	0.1292
GaussianAvatar	26.2311	0.9575	0.0715
Animatable-GS	<u>29.5138</u>	<u>0.9724</u>	<u>0.0413</u>
Ours	**31.0765**	**0.9839**	**0.0305**

(c) w/o Adaptive Regularization, (d) w/o Geometric Stream (removing $\Delta\mu_{t-1}$ input), (e) w/o State Stream (removing memory bank and cross-attention).

Results reveal the complementary necessity of both streams, particularly on far-distribution (FD) poses. The w/o Geometric Stream variant (d) exhibits the most severe degradation, confirming that explicit geometric deformation feedback is essential—without $\Delta\mu_{t-1}$ and motion-adaptive aggregation, the network must infer cloth configuration and motion from body pose alone, leading to accumulated errors in position and velocity estimation. The w/o State Stream variant (e) also degrades significantly, demonstrating that implicit internal state cannot be reliably recovered from body pose alone.

Notably, the substantially larger performance degradation on FD versus ND sets validates that explicit temporal causality modeling is essential for achieving robust generalization beyond the training distribution.

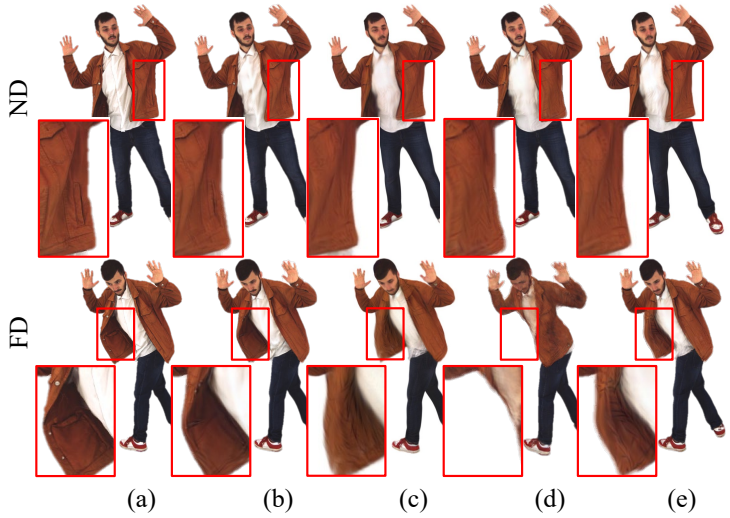

Fig. 5: Ablation study. (a) Ground Truth, (b) Full model with both streams, (c) w/o adaptive regularization, (d) w/o Geometric Stream, (e) w/o State Stream.

Table 2: Ablation study on ND (MMD=0.23) and FD (MMD=0.87) test sequences.

Split	Variant	PSNR↑	SSIM↑	LPIPS↓
ND	Full Model	**31.7153**	**0.9802**	**0.0283**
	w/o Adaptive Reg.	29.3130	<u>0.9680</u>	0.0453
	w/o GeoStream.	28.9622	0.9651	0.0476
	w/o StateStream.	<u>30.2593</u>	0.9676	<u>0.0386</u>
FD	Full Model	**30.6720**	**0.9739**	**0.0302**
	w/o Adaptive Reg.	<u>28.8778</u>	<u>0.9658</u>	<u>0.0492</u>
	w/o GeoStream.	24.2694	0.9557	0.0623
	w/o StateStream.	28.2286	0.9632	0.0511

5 Conclusion

We present a dual-stream autoregressive framework for temporally coherent animatable avatars. We identify that cloth rendering requires modeling both observable kinematic information and implicit internal state—dual requirements that skeletal poses alone cannot satisfy. By explicitly modeling kinematic information through autoregressive conditioning and implicit internal state through memory-based fusion, our method learns temporally coherent, photorealistic cloth dynamics directly from multi-view photometric observations.

Experiments demonstrate robust generalization to motion patterns beyond training distributions, validating that dual-stream temporal modeling is essential for realistic cloth dynamics in neural rendering.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Bao, Y., Ding, T., Huo, J., Liu, Y., Li, Y., Li, W., Gao, Y., Luo, J.: 3d gaussian splatting: Survey, technologies, challenges, and opportunities. IEEE Transactions on Circuits and Systems for Video Technology (2025)
3. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5855–5864 (2021)
4. Bertiche, H., Madadi, M., Escalera, S.: Cloth3d: clothed 3d humans. In: European Conference on Computer Vision. pp. 344–359. Springer (2020)
5. Bertiche, H., Madadi, M., Escalera, S.: Neural cloth simulation. ACM Transactions on Graphics (TOG) **41**(6), 1–14 (2022)
6. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: European conference on computer vision. pp. 333–350. Springer (2022)
7. Chen, Y., Wang, X., Chen, X., Zhang, Q., Li, X., Guo, Y., Wang, J., Wang, F.: Uv volumes for real-time rendering of editable free-view human performance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16621–16631 (2023)
8. Choi, K.J., Ko, H.S.: Stable but responsive cloth. In: ACM SIGGRAPH 2005 Courses. ACM (2005)
9. Deng, X., Zheng, Z., Zhang, Y., Sun, J., Xu, C., Yang, X., Wang, L., Liu, Y.: Ram-avatar: Real-time photo-realistic avatar from monocular videos with full-body control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1996–2007 (2024)
10. Dong, Z., Guo, C., Song, J., Chen, X., Geiger, A., Hilliges, O.: Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20470–20480 (2022)
11. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5501–5510 (2022)
12. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. Journal of Machine Learning Research **13**, 723–773 (2012)
13. Grigorev, A., Black, M.J., Hilliges, O.: Hood: Hierarchical graphs for generalized modelling of clothing dynamics. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16965–16974 (2023)
14. Guo, C., Jiang, T., Chen, X., Song, J., Hilliges, O.: Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12858–12868 (2023)
15. Habermann, M., Xu, W., Zollhoefer, M., Pons-Moll, G., Theobalt, C.: A deeper look into deepcap. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4009–4022 (2021)
16. He, T., Xu, Y., Saito, S., Soatto, S., Tung, T.: Arch++: Animation-ready clothed human reconstruction revisited. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11046–11056 (2021)

17. Hu, L., Zhang, H., Zhang, Y., Zhou, B., Liu, B., Zhang, S., Nie, L.: 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. pp. 634–644 (2024)
18. 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)
19. Huang, Z., Xu, Y., Lassner, C., Li, H., Tung, T.: Arch: Animatable reconstruction of clothed humans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3093–3102 (2020)
20. Işık, M., Rünz, M., Georgopoulos, M., Khakhulin, T., Starck, J., Agapito, L., Nießner, M.: Humanrf: High-fidelity neural radiance fields for humans in motion. ACM Transactions on Graphics (TOG) **42**(4), 1–12 (2023)
21. 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)
22. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
23. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
24. Kocabas, M., Chang, J.H.R., Gabriel, J., Tuzel, O., Ranjan, A.: Hugs: Human gaussian splats. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 505–515 (2024)
25. Kwon, Y., Liu, L., Fuchs, H., Habermann, M., Theobalt, C.: Deliffas: Deformable light fields for fast avatar synthesis. Advances in neural information processing systems **36**, 40944–40962 (2023)
26. Li, D., Zhong, W., Yu, W., Pan, Y., Zhang, D., Yao, T., Han, J., Mei, T.: Pursuing temporal-consistent video virtual try-on via dynamic pose interaction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22648–22657 (2025)
27. Li, R., Tanke, J., Vo, M., Zollhöfer, M., Gall, J., Kanazawa, A., Lassner, C.: Tava: Template-free animatable volumetric actors. In: European Conference on Computer Vision. pp. 419–436. Springer (2022)
28. Li, Y., Zeng, Z., Pang, L., Zhang, G., Zhang, S.: Realityavatar: Towards realistic loose clothing modeling in animatable 3d gaussian avatars. arXiv preprint arXiv:2504.01559 (2025)
29. Li, Y., Du, T., Wu, K., Xu, J., Matusik, W.: Diffcloth: Differentiable cloth simulation with dry frictional contact. ACM Transactions on Graphics (TOG) **42**(1), 1–20 (2022)
30. Li, Z., Zheng, Z., Liu, Y., Zhou, B., Liu, Y.: Posevocab: Learning joint-structured pose embeddings for human avatar modeling. In: ACM SIGGRAPH 2023 conference proceedings. pp. 1–11 (2023)
31. Li, Z., Zheng, Z., Wang, L., Liu, Y.: 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. pp. 19711–19722 (2024)
32. Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: Neural free-view synthesis of human actors with pose control. ACM transactions on graphics (TOG) **40**(6), 1–16 (2021)

33. Liu, X., Wu, C.: Vga: Reconstructing vivid 3d gaussian avatars from monocular videos. In: International Conference on Computational Visual Media. pp. 172–193. Springer (2025)
34. Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y.: Neural volumes: learning dynamic renderable volumes from images. ACM Transactions on Graphics (TOG) **38**(4), 1–14 (2019)
35. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM TOG (2015)
36. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5442–5451 (2019)
37. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4460–4470 (2019)
38. Mihajlovic, M., Saito, S., Bansal, A., Zollhoefer, M., Tang, S.: Coap: Compositional articulated occupancy of people. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13201–13210 (2022)
39. Mihajlovic, M., Zhang, Y., Black, M.J., Tang, S.: Leap: Learning articulated occupancy of people. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10461–10471 (2021)
40. 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 (2021)
41. Moreau, A., Song, J., Dhamo, H., Shaw, R., Zhou, Y., Pérez-Pellitero, E.: Human gaussian splatting: Real-time rendering of animatable avatars. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 788–798 (2024)
42. Müller, M., Heidelberger, B., Hennix, M., Ratcliff, J.: Position based dynamics. Journal of Visual Communication and Image Representation **18**(2), 109–118 (2007)
43. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM transactions on graphics (TOG) **41**(4), 1–15 (2022)
44. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 165–174 (2019)
45. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: 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. pp. 10975–10985 (2019)
46. 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. pp. 14314–14323 (2021)
47. 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. 9054–9063 (2021)
48. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10318–10327 (2021)

49. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
50. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2304–2314 (2019)
51. Saito, S., Simon, T., Saragih, J., Joo, H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 84–93 (2020)
52. Saito, S., Yang, J., Ma, Q., Black, M.J.: Scanimate: Weakly supervised learning of skinned clothed avatar networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2886–2897 (2021)
53. Santesteban, I., Otaduy, M.A., Casas, D.: Snug: Self-supervised neural dynamic garments. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8140–8150 (2022)
54. 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)
55. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. Advances in neural information processing systems **33**, 7462–7473 (2020)
56. Sitzmann, V., Zollhöfer, M., Wetzstein, G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. Advances in neural information processing systems **32** (2019)
57. Su, S.Y., Yu, F., Zollhöfer, M., Rhodin, H.: A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. Advances in neural information processing systems **34**, 12278–12291 (2021)
58. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. Advances in neural information processing systems **33**, 7537–7547 (2020)
59. Tewari, A., Fried, O., Thies, J., Sitzmann, V., Lombardi, S., Sunkavalli, K., Martin-Brualla, R., Simon, T., Saragih, J., Nießner, M., et al.: State of the art on neural rendering. In: Computer graphics forum. vol. 39, pp. 701–727. Wiley Online Library (2020)
60. Tiwari, G., Sarafianos, N., Tung, T., Pons-Moll, G.: Neural-gif: Neural generalized implicit functions for animating people in clothing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11708–11718 (2021)
61. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
62. Wang, L., Zhao, X., Sun, J., Zhang, Y., Zhang, H., Yu, T., Liu, Y.: Styleavatar: Real-time photo-realistic portrait avatar from a single video. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–10 (2023)
63. Wang, S., Mihajlovic, M., Ma, Q., Geiger, A., Tang, S.: Metaavatar: Learning animatable clothed human models from few depth images. Advances in Neural Information Processing Systems **34**, 2810–2822 (2021)

64. Wang, S., Schwarz, K., Geiger, A., Tang, S.: Arah: Animatable volume rendering of articulated human sdfs. In: European conference on computer vision. pp. 1–19. Springer (2022)
65. Wang, W., Ho, H.I., Guo, C., Rong, B., Grigorev, A., Song, J., Zarate, J.J., Hilliges, O.: 4d-dress: A 4d dataset of real-world human clothing with semantic annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 550–560 (2024)
66. 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)
67. 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. 16210–16220 (2022)
68. Xiu, Y., Yang, J., Cao, X., Tzionas, D., Black, M.J.: Econ: Explicit clothed humans optimized via normal integration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 512–523 (2023)
69. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: Icon: Implicit clothed humans obtained from normals. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13286–13296. IEEE (2022)
70. 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)
71. 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. pp. 586–595 (2018)
72. Zheng, S., Zhou, B., Shao, R., Liu, B., Zhang, S., Nie, L., Liu, Y.: Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19680–19690 (2024)
73. Zheng, Z., Huang, H., Yu, T., Zhang, H., Guo, Y., Liu, Y.: Structured local radiance fields for human avatar modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15893–15903 (2022)
74. Zheng, Z., Yu, T., Liu, Y., Dai, Q.: Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. IEEE transactions on pattern analysis and machine intelligence **44**(6), 3170–3184 (2021)
75. Zheng, Z., Zhao, X., Zhang, H., Liu, B., Liu, Y.: Avatarrex: Real-time expressive full-body avatars. ACM Transactions on Graphics (TOG) **42**(4), 1–19 (2023)
76. Zielonka, W., Bagautdinov, T., Saito, S., Zollhöfer, M., Thies, J., Romero, J.: Drivable 3d gaussian avatars. In: 2025 International Conference on 3D Vision (3DV). pp. 979–990. IEEE (2025)