DYNAMIC TEXTURE MODELING OF 3D CLOTHED GAUSSIAN AVATARS FROM A SINGLE VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

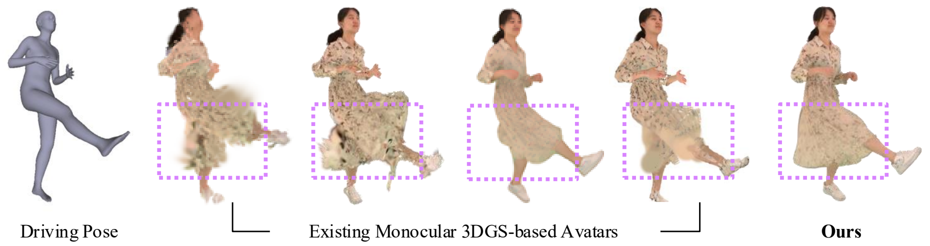
Recent advances in neural rendering, particularly 3D Gaussian Splatting (3DGS), have enabled animatable 3D human avatars from single videos with efficient rendering and high fidelity. However, current methods struggle with dynamic appearances, especially in loose garments (e.g., skirts), causing unrealistic cloth motion and needle artifacts. This paper introduces a novel approach to dynamic appearance modeling for 3DGS-based avatars, focusing on loose clothing. We identify two key challenges: (1) limited Gaussian deformation under pre-defined template articulation, and (2) a mismatch between body-template assumptions and the geometry of loose apparel. To address these issues, we propose a motion-aware autoregressive structural deformation framework for Gaussians. We structure Gaussians into an approximate graph and recursively predict structure-preserving updates, yielding realistic, template-free cloth dynamics. Our framework enables view-consistent and robust appearance modeling under the single-view constraint, producing accurate foreground silhouettes and precise alignment of Gaussian points with clothed shapes. To demonstrate the effectiveness of our method, we introduce an in-the-wild dataset featuring subjects performing dynamic movements in loose clothing, and extensive experiments validate that our approach significantly outperforms existing 3DGS-based methods in modeling dynamic appearances from single videos.

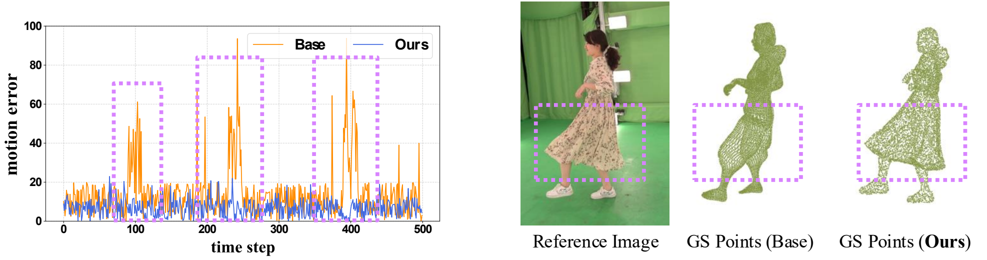
1 INTRODUCTION

Creating an animatable 3D avatar from a single video involves reconstructing a lifelike, controllable representation of a person capable of replicating both primary motions (i.e., movements of major body parts), and secondary motions (i.e., time-varying cloth dynamics). Achieving this capability is critical for immersive experiences in fields such as virtual reality, telepresence, and interactive entertainment, where realistic human representations significantly enhance user engagement (Sutherland et al., 1965; Lee et al., 2024). With the advent of 3D Gaussian Splatting (3DGS) (Kerbl & et al., 2023), high-quality neural rendering becomes feasible, substantially improving avatar realism and efficiency in synthesis from a monocular video.

However, existing 3DGS-based avatar methods (Lei et al., 2024; Hu et al., 2024a; Moon et al., 2024; Qian et al., 2024b) predominantly excel at modeling primary motion but exhibit limitations in robustly capturing secondary motion, as illustrated in Fig. 1a, where it shows the animation result of the 3D Gaussian avatar in a novel pose exhibiting a dynamic posture unseen during training. This is due to the fact that they rely on skeletal skinning of coarse meshes for animation (Loper et al., 2015), which inherently lacks subtle deformation effects such as inertia-driven soft-tissue dynamics. Consequently, it remains challenging to consistently reproduce these nuanced motions using neural networks alone.

There are two major challenges that hinder the secondary motion-aware dynamic appearance modeling in creating an animatable 3DGS avatar from a single video: (1) temporal context-unaware Gaussian deformation, and (2) cloth shape-agnostic Gaussian point initialization. First, existing methods deform the Gaussians of a clothed avatar as a function of the current body pose, which is aligned with the sampled images, to learn the appearance of a 3D Gaussian avatar. While this approach effectively captures the primary motion of main body parts, it is limited in representing secondary motion in garments such as dresses and skirts, which is strongly entangled with temporal continuity (Fig. 1b).


(a) Challenge: Appearance Modeling Including Secondary-Motion Dynamics.


(b) Motivation 1: Temporally Inconsistent Deformation.

(c) Motivation 2: Difficulty in Representing Loose-Fitting Clothes.

Figure 1: Conventional 3DGS-based avatars (Base) fail to model the dynamic appearance of subjects wearing loose garments, particularly in scenarios involving secondary motion (Lei et al., 2024; Qian et al., 2024b; Hu et al., 2024a; Moon et al., 2024). (a) Our method overcomes these limitations, enabling high-fidelity rendering of Gaussian avatars exhibiting dynamic motion from a single video. (b) In contrast, conventional methods define deformation through pre-specified articulation models such as linear blend skinning, which operate independently on each frame without accounting for temporal context—often leading to motion error spikes that indicate poor alignment with driving signals. (c) Furthermore, for initializing 3D Gaussians, they rely on a parametric template model resembling a naked body, which struggles to generate Gaussians for clothing regions deviating from the body surface, especially with loose-fitting garments.

Secondly, explicit representations based on 3D Gaussians are highly sensitive to the accuracy of their initial point placement, as widely discussed in primary 3DGS studies (Yu et al., 2024; Luiten et al., 2024). However, existing methods for creating 3DGS-based avatars from a single video (Lei et al., 2024; Hu et al., 2024a; Moon et al., 2024; Qian et al., 2024b), rely on parametric template models to initialize the shape of articulated subjects. These template models represent a naked body shape, leading to significant discrepancies between the initialized points and the actual shape when dealing with subjects wearing loose-fitting garments (Fig. 1c). Therefore, a few Gaussians should represent not only the body parts but also the appearance of the clothing, causing artifacts in novel pose animations where the model has not observed similar poses during training.

In this paper, we present a novel framework for modeling dynamic appearances of loose-fitting garments in 3D avatars, explicitly addressing the challenges posed by secondary motion. Central to our approach is a Secondary Motion-Aware Gaussian Deformation (**SMAD**) module, which constructs a velocity-encoded Gaussian graph over canonical Gaussians and autoregressively predicts second-order Gaussian dynamics. This enables realistic modeling of fine-grained cloth motion while preserving structural coherence during deformation. To achieve robust and view-consistent appearance modeling from a single-view video, we employ a confidence-aware feature fusion mechanism that aggregates multi-frame evidence and maintains silhouette fidelity throughout deformation. Finally, we contribute a new in-the-wild video dataset featuring subjects in diverse garments undergoing dynamic motion, filling the gap in benchmarks for evaluating secondary motion in animatable avatars.

Our contributions are summarized as follows:

- We propose a novel method for animatable 3D avatar generation based on 3DGS, which enables dynamic appearance modeling of dressed avatars.
- We propose a secondary motion-aware Gaussian deformation, introducing a velocity-encoded Gaussian graph representation that autoregressively estimates Gaussian dynamics.
- Extensive experiments demonstrate that our method outperforms existing methods on the subjects wearing loose-fitting clothes with dynamic movement.

2 RELATED WORK

Animatable 3D Avatars from Multi-view Videos. It has long been a major focus in vision and graphics. Early systems (Stoll et al., 2010; Alldieck et al., 2018; Joo et al., 2015; Pons-Moll et al., 2017; Habermann et al., 2019) reconstructed actors in multi-view studios and animated meshes via multi-view geometry and hand-crafted articulation designs. While these approaches empower the controllability, it required substantial expert intervention. The shift to implicit neural representations, especially neural radiance fields (Mildenhall et al., 2021), introduced photorealistic neural avatars (Peng et al., 2021b;a; Habermann & et al., 2021; Zheng et al., 2023; Shen et al., 2023b; Li et al., 2023; Zhu et al., 2024; Shen et al., 2023a; Yin et al., 2023b; Chen et al., 2024; Saito et al., 2024) and free-view synthesis (Kwon et al., 2021; Liu et al., 2021; Işık et al., 2023; Kwon et al., 2024b), though often with slow training and additional structural constraints for stable driving. The 3DGS (Kerbl & et al., 2023) further achieved efficient rendering with high fidelity (Li et al., 2024; Zielonka et al., 2025; Zheng et al., 2024; Kwon et al., 2024a; Lin et al., 2024; Zhan et al., 2025; Liao et al., 2024). Yet, their high-fidelity performance fundamentally relies on dense, calibrated multi-view supervision, geometry constraints, and explicit subject-specific ground-truth template meshes. By contrast, our method is deliberately designed for the single-video setting, aiming to create user-friendly animatable clothed avatars directly from casual monocular footage.

Animatable 3D Gaussian Avatars from Monocular Videos. Advances in neural rendering and markerless motion-capture techniques have enabled the construction of user-friendly 3D avatars from monocular videos. With these advancements, it has been to learn a neural implicit representation defined in a continuous canonical space near a template mesh, and to deform this representation into the observation space using predefined articulations driven by motion inputs (Su et al., 2021; Weng et al., 2020; Chen et al., 2021; Weng et al., 2022; Wang et al., 2022; Yu et al., 2023; Jiang et al., 2023a;b). The advent of 3DGS has further accelerated photo-realistic modeling of 3d avatars; several works attach Gaussian primitives to a skeletal model and learn pose-conditioned deformations from monocular videos (Qian et al., 2024b; Hu et al., 2024a; Moon et al., 2024; Shao et al., 2024; Lei et al., 2024; Hu et al., 2024b; Zhai et al., 2025; Guo et al., 2025). However, existing methods assume template (Loper et al., 2015)-based initialization and its pre-defined articulation, which struggle to capture subtle, temporally coherent non-rigid effects. We build upon other line, introducing two key aspects: a template-free initialization that directly aligns the Gaussian primitives, eliminating the need for naked-body templates; and a physics-inspired autoregressive deformation module that predicts velocities and accelerations with finite difference method (Xie et al., 2024), and captures second-order dynamics through a velocity-encoded Gaussian graph.

Dynamic Clothed Human Modeling. Beyond primary motion driven by the main body movement, several works have also considered secondary motion, such as cloth dynamics. One line of works (Habermann et al., 2020; Habermann & et al., 2021; Habermann et al., 2021; Liao et al., 2024; Feng et al., 2022; 2023; Guo et al., 2023; 2024) reconstructs clothed surfaces using neural implicit representations, but canonicalization with predefined articulation struggles with loose garments. Another approaches combines non-rigid deformation with LBS and neural networks, yet requires subject-specific ground-truth clothed meshes. A separate direction incorporates physics simulation by numerically solving differential equations of the dynamic systems (Terzopoulos et al., 1987; Müller et al., 2007; Macklin et al., 2016). However, they are computationally expensive and difficult to parametrize; To alleviate this, several works (Pan et al., 2022; Santesteban et al., 2022; Grigorev et al., 2023; 2024) approximate dynamic systems with neural networks—e.g., augmenting models with virtual bones or employing recurrent architectures to predict garment-deformation sequences, but depend on high-level supervision such as 4D scans, cloth–body segmentation, and explicit colliders, and operate only on polygonal meshes. Our work is different lines of this work; we aim to create avatars represented as 3D Gaussian primitives and model its dynamics and appearances, given only from monocular videos, without access to any 3D ground-truth or prior geometric knowledge.

3 METHOD

Given a monocular RGB video $\mathcal{V} = \{I_t\}_{t=1}^{T}$ capturing a human subject in motion, our goal is to reconstruct a fully *animatable 3D Gaussian avatar* that faithfully models dynamic appearances of loose-fitting clothed subjects. We adopt a dynamic set of *3D Gaussian primitives* whose spatiotempo-

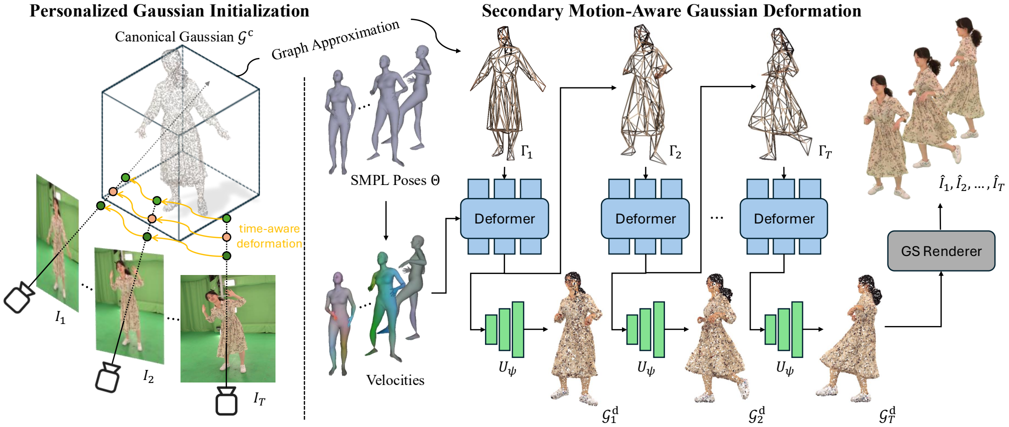


Figure 2: To model secondary motions in 3DGS-based avatars, we introduce a two-stage framework: (1) Personalized Gaussian Initialization using a deformable NeRF to estimate canonical Gaussians $\mathcal{G}^c$, and (2) Secondary Motion-Aware Deformation. $\mathcal{G}^c$ are structured as a Gaussian graph Γ, processed by a GNN-based autoregressive deformer, and decoded via U_ψ into deformed Gaussians $\mathcal{G}^d$. Motion descriptors derived from SMPL poses Θ guide temporally coherent deformation. Then GS Renderer then synthesizes the final images.

ral properties evolve over time to capture complex non-rigid deformations, such as clothing dynamics. Formally, we represent the avatar at time t as a set of N deformed Gaussians:

$$\mathcal{G}^{\mathrm{d}}_t = \{(\mu_{t,i}, \Sigma_{t,i}, c_{t,i}, \alpha_{t,i})\}_{i=1}^{N}\,, \tag{1}$$

where $\mu_{t,i} \in \mathbb{R}^3$ is the 3D mean position, $\Sigma_{t,i} \in \mathbb{R}^{3\times3}$ is the covariance matrix modeling spatial extent and orientation, $c_{t,i} \in \mathbb{R}^3$ denotes RGB color, and $\alpha_{t,i} \in \mathbb{R}$ represents opacity. The number of primitives N remains fixed across time, but their parameters are dynamically updated to reflect conditioning motion, a set of SMPL poses $\Theta = \{\theta_t\}_{t=1}^{T}$. We then obtain the animated rendering video $\hat{\mathcal{V}} = \{\hat{I}_t = \mathcal{R}(\mathcal{G}^{\mathrm{d}}_t)\}_{t=1}^{T}$ by projecting the deformed Gaussians through a differentiable splatting renderer $\mathcal{R}$. Fig.2 illustrate the overall process of the proposed method.

Baselines. We adopt a simple baseline that obtains dense canonical Gaussian primitives using a 4D NeRF Gao et al. (2021). Concretely, we train a deformable neural radiance field on the input monocular video and map each observation-space point $\mathbf{x}_t$ at time t to a canonical space (reference time). By querying color and density in the canonical space, we recover a dense canonical density field that captures both body and loose clothing without relying on a parametric template.

After training, we extract canonical Gaussians by thresholding the time-averaged canonical density $\bar{\sigma}(\mathbf{x}) = \frac{1}{T}\sum_t \sigma(\mathbf{x}, t)$ and clustering the surviving voxels to obtain Gaussian centers $\{\boldsymbol{\mu}^{\mathrm{c}}_i\}$, with isotropic variances $\{\Sigma^{\mathrm{c}}_i\}$ and colors $\{\mathbf{c}^{\mathrm{c}}_i\}$. This yields a dense set of canonical 3D Gaussian primitives $\mathcal{G}^{\mathrm{c}}$ that serves as the person-specific Gaussian initialization (PGI) for subsequent stages.

3.1 VELOCITY-ENCODED GAUSSIAN GRAPH

To overcome the limited capability of representing secondary motions caused by reliance on linear blend skinning of parametric template human models (Loper et al., 2015; Pavlakos et al., 2019), we propose an autoregressive Gaussian deformation method that moves beyond the template model. Furthermore, to ensure robust performance even when the number of Gaussians grows exponentially and to alleviate computational complexity, we propose a graph-based deformation approach that approximates Gaussian interactions.

Graph Construction. Given a set of N initial Gaussian points $\{\mu_1, \mu_2, ..., \mu_N\}$, we downsample the Gaussian points to $\mathbf{X} \in \mathbb{R}^{M\times3}$ ($M \ll N$) with voxel-grid downsampling (Rusu & Cousins, 2011); these M nodes serve as the final Gaussian primitives used for rendering. We then construct a graph $\Gamma = (\mathbf{X}, \mathbf{A})$, where $\mathbf{A} \in \mathbb{R}^{M\times M}$ is adjacency matrix. It is constructed via k-Nearest Neighbors (k-NN) by computing pairwise distances $\mathrm{d}(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \in X$. Each element of

$\mathbf{A}_{ij}$ is formulated as $\exp\left(-\frac{\mathrm{d}(\mathbf{x}_i,\mathbf{x}_j)^2}{\rho_a^2}\right)$, where ρ_a controls sensitivity to distances.

Velocity Encoding (VE). We build the node features $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_M\}$ at each node position $\mathbf{x}_i$. Let us consider $\mathbf{h}_i$ as a concatenation of the node position $\mathbf{x}_i$ and its velocity $\mathbf{v}_i(t) = \frac{\mathbf{x}_i(t)-\mathbf{x}_i(t-\Delta t)}{\Delta t}$ at the time state t. Furthermore, to capture long-range dependencies, we buffer the past τ_v memory vectors as a set of $\bar{\mathbf{v}}_i = \{\mathbf{v}_i(t), \mathbf{v}_i(t-1), \ldots, \mathbf{v}_i(t-\tau_v)\}$ To condition a set of body pose priors $\Theta_{t-\tau:t} = \{\theta_t, \theta_{t-1}, ..., \theta_{t-\tau}\}$ with time window τ, we additionally embed it to $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_M\}$ with MLP. At this end, the node feature $\mathbf{h}_i$ is defined as $\mathbf{h}_i = (\mathbf{x}_i, \bar{\mathbf{v}}_i, \mathbf{e}_i)$.

3.2 SECONDARY MOTION-AWARE DEFORMATION (SMAD)

Our goal is to move beyond linear blend skinning (LBS) with parametric body models (Loper et al., 2015) and learn an animatable 3DGS avatar that can faithfully reproduce *secondary motions*. Motivated by deformation methods that generalize to unseen motions without relying on pre-defined kinematic hierarchies (Zheng et al., 2021; Grigorev et al., 2023), we employ a graph neural network (GNN) deformer that autoregressively predicts the non-rigid dynamics of human bodies—hence, of Gaussian primitives.

Definition. We model each Gaussian node i as a point mass g_i whose motion follows a second-order mass–spring–damper system (Gilmer et al., 2020; Wang et al., 2020). Let $\mathbf{x}_i(t) \in \mathbb{R}^3$ and $\mathbf{v}_i(t) = \dot{\mathbf{x}}_i(t) \in \mathbb{R}^3$ denote the position and velocity at time t. The dynamics are

$$\mathbf{F}_i^{\text{ext}}(t) = g_i\,\ddot{\mathbf{x}}_i(t) \;+\; \gamma_i\,\dot{\mathbf{x}}_i(t) \;+\; \sum\nolimits_j k_{ij}\Big(\mathbf{x}_i(t) - \mathbf{x}_j(t) - \mathbf{L}_{ij}^{\text{rest}}\Big), \tag{2}$$

where $\ddot{\mathbf{x}}_i(t) = \mathbf{a}_i(t)$ is acceleration, γ_i is a damping coefficient, k_{ij} is the spring stiffness between nodes i and j, $\mathbf{L}_{ij}^{\text{rest}}$ is their rest offset in canonical space, and $\mathbf{F}_i^{\text{ext}}(t)$ is an external driving force. We treat k_{ij} as a learnable parameter that is adaptively updated during training. This allows the model to automatically disentangle rigid and non-rigid parts without any explicit supervision.

With a discrete step Δt, we apply explicit Euler integration:

$$\mathbf{a}_i(t) = \frac{1}{g_i}\Big(\mathbf{F}_i^{\text{ext}}(t) - \gamma_i\,\mathbf{v}_i(t) - \sum_j k_{ij}\,[\mathbf{x}_i(t) - \mathbf{x}_j(t) - \mathbf{L}_{ij}^{\text{rest}}]\Big), \tag{3}$$

$$\mathbf{v}_i(t+\Delta t) = \mathbf{v}_i(t) + \Delta t\,\mathbf{a}_i(t), \qquad \mathbf{x}_i(t+\Delta t) = \mathbf{x}_i(t) + \Delta t\,\mathbf{v}_i(t+\Delta t). \tag{4}$$

This second-order formulation naturally induces *secondary motion* (e.g., cloth flutter). In practice, we let a message-passing GNN (Gilmer et al., 2020) *learn* these updates rather than prescribing forces explicitly.

Architecture. The Gaussian graph deformer parameterizes the above updates with a message-passing GNN. Each node i carries a feature $\mathbf{h}_i(t) \in \mathbb{R}^{d_\mathrm{h}}$ obtained from Sec. 3.1. At time t, node i aggregates information from its neighbors using an adjacency $\mathbf{A}_{ij}(t)$. With an MLP M_θ, we define messages as

$$\mathbf{m}_{j\to i}(t) = M_\theta\big(\mathbf{h}_i(t),\,\mathbf{h}_j(t)\big) \in \mathbb{R}^{d_\mathrm{m}}, \qquad \mathbf{m}_i^{\text{agg}}(t) = \sum\nolimits_j \mathbf{A}_{ij}(t)\,\mathbf{m}_{j\to i}(t). \tag{5}$$

Two update functions then produce the next-step node feature and physical state:

$$\mathbf{h}_i(t+\Delta t) = F_\theta\big(\mathbf{h}_i(t),\,\mathbf{m}_i^{\text{agg}}(t)\big), \tag{6}$$

$$\big[\mathbf{x}_i(t+\Delta t),\,\mathbf{v}_i(t+\Delta t)\big] = G_\theta\big([\mathbf{x}_i(t),\,\mathbf{v}_i(t)],\,\mathbf{m}_i^{\text{agg}}(t)\big), \tag{7}$$

where G_θ serves as a neural surrogate for the mass–spring–damper updates in Eq. equation 2. After L message-passing layers, we obtain updated positions and velocities for all nodes. Each node corresponds to a deformed Gaussian $\mathcal{G}_i^d$, and we finally set

$$\boldsymbol{\mu}_i \leftarrow \mathbf{x}_i(t+\Delta t), \qquad \mathbf{c}_i,\;\alpha_i,\;\Sigma_i \leftarrow D_\psi\big(\mathbf{z}_i,\,\mathbf{h}_i(t+\Delta t)\big), \tag{8}$$

where $\mathbf{z}_i \in \mathbb{R}^{d_z}$ is a learned latent code and D_ψ predicts color, opacity, and covariance for *each* Gaussian.

Training Objectives. After computing $\mathcal{G}^{\mathrm{d}} = \{\mathcal{G}^{\mathrm{d}}_1, \mathcal{G}^{\mathrm{d}}_2, ..., \mathcal{G}^{\mathrm{d}}_N\}$, we render it via Gaussian Splatting-based rasterizer $\mathcal{R}$ to $\hat{I}_t = \mathcal{R}(\mathcal{G}^{\mathrm{d}}_t)$. We define a total SMAD loss term $\mathcal{L}_{\text{SMAD}}$ as:

$$\mathcal{L}_{\text{SMAD}} = \mathcal{L}_{\text{RGB}} + \lambda_{\text{iso}}\mathcal{L}_{\text{iso}} + \lambda_{\text{damp}}\mathcal{L}_{\text{damp}}. \tag{9}$$

We mainly use the common L1 rgb photometric loss between rendered images and ground-truth images, which is formulated as:

$$\mathcal{L}_{\text{RGB}} = \left\|\mathcal{R}(\mathcal{G}^{d}_t) - I_t\right\|_1. \tag{10}$$

It minimizes the pixel intensities of rendered Gaussians $\mathcal{R}(\mathcal{G}^{d}_t)$ to the ground-truth images I_t. In addition, we utilize two regularization terms:

$$\mathcal{L}_{\text{iso}} = \sum_{(i,j)\in\mathcal{E}} \left(\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2 - \left\|\mathbf{L}^{\text{rest}}_{ij}\right\|_2\right)^2, \quad \mathcal{L}_{\text{damp}} = \sum_{i=1}^{N}\sum_{t=1}^{T} \left\|\mathbf{v}_i(t)\right\|_2^2. \tag{11}$$

The isometry Loss $\mathcal{L}_{\text{iso}}$ penalizes deviations in geodesic distance to preserve local surface area. It prevents stretching or shrinking of garment regions; useful for preserving cloth realism during motion. We set $\lambda_{\text{iso}} = 0.1$, where it emphasize length preservation. The damping Loss $\mathcal{L}_{\text{damp}}$ regularizes velocity magnitudes to reduce high-frequency vibration and dynamic instability. It reduces visual fluttering or noise in motion, especially noticeable in fine cloth edges. We $\lambda_{\text{damp}} = 0.01$, where it avoids over-constraining dynamic details.

4 EXPERIMENTS

Dataset. **ZJU-MoCap** (Peng et al., 2021a) is a primary benchmark for animatable 3D avatars. Using HumanNeRF (Weng et al., 2022) split sequences, we report novel view synthesis results due to limited pose diversity. To supplement the lack of motion variation and loose-fitting garments, we additionally evaluate on two benchmarks. **4D-Dress** (Wang et al., 2024) firstly introduces real-world 4D human clothing dataset featuring dynamic cloth motions, designed to advance research in realistic garment modeling and animation. We carefully selected five subjects, each wearing loose-fitting clothing such as skirts or puffer jackets. We also introduce **LoCo-Human**, a new in-the-wild dataset featuring five **Lo**ose-**Clo**thed **Human**s performing 5 dynamic and 1 static motions per subject. The static sequence is used for training, and the others for evaluating generalization in-the-wild scenarios.

Baselines & Evaluation Metrics. We compare our method with existing approaches on 3DGS-based avatars from monocular videos. Given the extensive body of prior work in this domain, it is practically infeasible to compare against all existing methods. Therefore, we specifically focus on publicly available baseline methods (Lei et al., 2024; Hu et al., 2024a; Qian et al., 2024b; Moon et al., 2024) that explicitly address dynamic appearance modeling. We evaluate the visual fidelity of the rendered animatable avatars with widely used metrics: PSNR, SSIM, and LPIPS. PSNR and SSIM measure pixel-level similarity and structural consistency with the ground-truth images, while LPIPS captures perceptual quality based on deep feature distances. These metrics collectively assess both low-level accuracy and high-level perceptual realism. To quantitatively assess both temporal consistency and how faithfully the animated avatars follow the driving motion, we compute the motion error (Kanazawa et al., 2019) between driving signal and motion estimated from the rendered animations.

4.1 RESULTS

We comprehensively compare our proposed method to state-of-the-art animatable 3D Gaussian avatar methods on three datasets: the 4D-Dress, ZJU-Mocap, and our newly proposed LoCo-Human in-the-wild dataset. We adopt standard metrics including PSNR, SSIM, and LPIPS to quantitatively measure visual fidelity and perceptual quality of animated avatars in rendered images. We first evaluate the ability to synthesize novel poses of dressed avatars on 4D-Dress. As presented in Tab. 1 (a), our method outperforms baseline methods across all metrics, demonstrating superior reconstruction quality. Qualitative comparisons (Fig. 3) further confirm our method's capability to produce realistic cloth dynamics, mitigating common artifacts such as unrealistic garment splitting observed in baselines. Next, we formally compare our method on the widely-used ZJU-Mocap benchmark. Following the conventional evaluation protocol, we quantitatively and qualitatively assess novel-view synthesis quality (Tab. 1 (b) & Fig. 5). Results indicate that our method consistently achieves superior

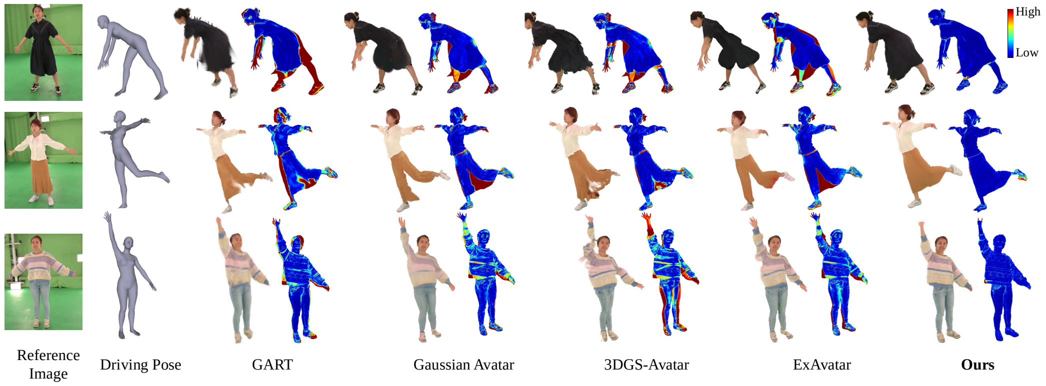


Figure 3: Qualitative comparison of novel pose synthesis on 4D-Dress dataset. We compare our method to the serveral 3D Gaussian Splatting-based Avatars (Lei et al., 2024; Hu et al., 2024a; Qian et al., 2024b; Moon et al., 2024). For each subject, we present reference image, driving pose, rendered image and error map to ground-truth image. Our method models robust dynamic appearances wearing loose-fitting clothes compared to the baselines.

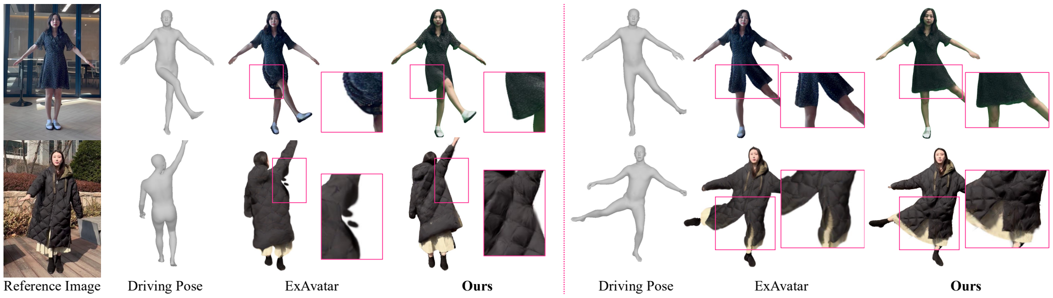


Figure 4: Qualitative comparison on the in-the-wild LoCo-Human dataset. Given the target driving poses, we animate avatars wearing loose-fitting garments. Compared to ExAvatar (Moon et al., 2024), our method better preserves cloth details and faithfully produces coherent motion under diverse poses. Insets highlight finer garment structures, showing sharper textures and more realistic deformation.

performance compared to previous single-video avatar approaches, reflecting improvements in visual sharpness and perceptual realism. In addition, we conduct extended evaluations on LoCo-Human, an in-the-wild dataset to assess the generalization ability of our method in real-world scenarios. As shown in Tab. 1 (c), our approach consistently outperforms existing baselines across diverse subjects. The qualitative results Fig. 4, further support these findings—demonstrating the robustness of our method even in scenarios involving challenging clothing, complex motions, and various confounding factors. These empirical results suggest that our deformation network, which mimics a second-order dynamic system, better captures cloth dynamics compared to conventional deformation schemes based on parametric template models. These extensive experiments validate that our approach effectively addresses critical challenges associated with dynamic appearance modeling from single monocular videos.

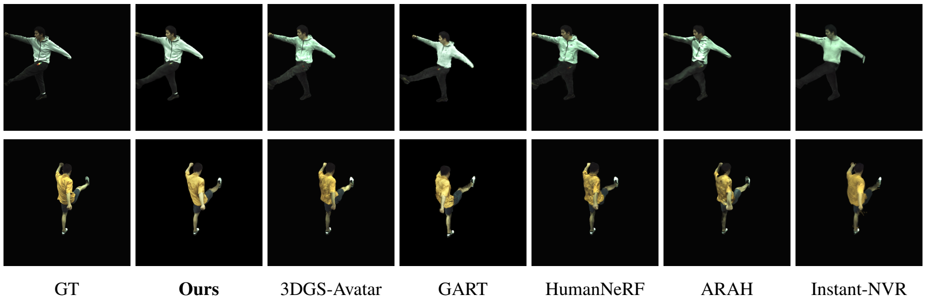


Figure 5: Qualitative comparison of novel view synthesis on ZJU-Mocap (Peng et al., 2021a). Our method yields view-consistent and artifact-free appearance modeling, even for repetitive motions in novel view synthesis scenarios.

(a) Novel Pose Synthesis on 4D-Dress Dataset

Method	*00148*			*00170*			*00185*			*00187*			*00190*		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GART (Lei et al., 2024)	20.86	0.9509	0.0661	23.52	0.9622	0.0413	26.84	0.9599	0.0488	25.81	0.9401	0.0592	29.01	0.9627	0.0375
Gaussian Avatar (Hu et al., 2024a)	20.91	0.9512	0.0657	24.12	0.9630	0.0356	26.62	0.9586	0.0500	24.96	0.9317	0.0684	26.44	0.9591	0.0512
3DGS-Avatar (Qian et al., 2024b)	22.79	0.9560	0.0471	25.49	0.9636	**0.0293**	27.54	0.9595	0.0394	25.99	0.9398	0.0457	**29.49**	0.9616	**0.0278**
ExAvatar (Moon et al., 2024)	21.93	0.9536	0.0628	26.30	0.9657	0.0367	28.35	0.9618	0.0470	25.84	0.9403	0.0620	26.12	0.9586	0.0569
Ours	**24.74**	**0.9601**	**0.0397**	**27.62**	**0.9700**	0.0301	**29.98**	**0.9673**	**0.0370**	**27.71**	**0.9548**	**0.0443**	29.44	**0.9635**	0.0347

(b) Novel View Synthesis on ZJU-MoCap

Method	*394*			*393*			*392*			*387*			*386*		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeuralBody (Peng et al., 2021a)	29.10	0.9593	54.55	28.61	0.9590	59.05	30.10	0.9642	53.27	27.00	0.9518	59.47	30.54	0.9678	46.43
HumanNeRF (Weng et al., 2022)	30.31	0.9642	32.89	28.31	0.9603	36.72	31.04	0.9705	32.12	28.18	0.9632	35.58	33.20	0.9752	28.99
MonoHuman (Yu et al., 2023)	29.15	0.9595	38.08	27.64	0.9566	43.17	29.50	0.9635	39.45	27.93	0.9601	41.76	32.94	0.9695	36.04
ARAH (Wang et al., 2022)	29.46	0.9632	40.76	28.77	0.9645	42.30	32.02	0.9742	35.28	28.49	**0.9656**	40.43	33.50	0.9781	31.40
GART (Lei et al., 2024)	29.92	0.9651	32.55	28.65	0.9620	35.55	31.36	0.9736	30.50	28.20	0.9644	34.43	33.48	**0.9850**	29.55
3DGS-Avatar (Qian et al., 2024b)	30.54	0.9661	31.21	28.88	0.9635	35.26	31.66	0.9730	30.14	28.33	0.9642	34.24	33.63	0.9773	25.77
Ours	**30.89**	**0.9677**	**31.18**	**29.48**	**0.9643**	**34.10**	**32.33**	**0.9754**	**29.47**	**28.86**	0.9650	**32.91**	**33.86**	0.9784	**25.22**

(c) LoCo-Human (In-the-Wild)

Method	*S01*			*S02*			*S03*			*S04*			*S05*		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS-Avatar (Qian et al., 2024b)	23.15	0.9374	0.0567	24.21	0.9391	0.0579	23.74	0.9349	0.0594	22.87	0.9337	0.0618	22.59	0.9312	0.0632
ExAvatar (Moon et al., 2024)	24.82	0.9478	0.0489	25.07	0.9483	0.0468	24.43	0.9465	0.0527	23.93	0.9442	0.0543	23.68	0.9426	0.0571
Ours	**26.17**	**0.9576**	**0.0423**	**26.44**	**0.9589**	**0.0409**	**25.76**	**0.9554**	**0.0441**	**25.38**	**0.9531**	**0.0467**	**24.83**	**0.9517**	**0.0484**

Table 1: Quantitative comparisons across (a) novel pose synthesis on 4D-Dress, (b) novel view synthesis on ZJU-MoCap, and (c) generalization on LoCo-Human. We highlight the best (**bold**) and second-best (underline) performance in each case.

4.2 ABLATION STUDY

Physics & Graph Design (PGD). We start from the base configuration of vanilla GNN with autoregressive deformation of predicting positions and velocities with finitie-difference method, without any physically plausible regularization $\mathcal{L}_{iso}$ and $\mathcal{L}_{damp}$. Adding physics regularization (A1) yields a clear gain of +0.84 PSNR, and -10.3% LPIPS. Introducing the spring stifness k_{ij} coefficients as a learnable parameter (A2), which adaptively distinguishes the rigid and non-rigid parts of subjects in unsupervised setting, further improves the rendering quality under dynamic motions. The advanced message-passing strategy for GNN with embedding edge features (A3) bring another boost of +0.68 PSNR, and -10.2% LPIPS. The full configuration with latent codes for time-varying dynamic appearance finally achieves the best results, which is +2.68 PSNR and a 31.0% LPIPS reduction over A0, underscoring the complementary roles of physics constraints and graph design.

Velocity Encoding (VE). Encoding an autoregressive window of past velocities markedly improves temporal fidelity. Performance rises monotonically from no VE (B0) to larger horizons, peaking at $\tau_v = 11$ (B4), with a net +5.83 PSNR and a 40.3% LPIPS drop. Very short context ($\tau_v = 1$) yields limited gains, while overly long horizons saturate; $\tau_v = 11$ strikes the best balance between temporal context and feature efficiency.

SMAD Capacity (M). Increasing the number of Gaussian graph nodes improves accuracy up to a moderate resolution. Compared to the no-SMAD baseline (C0), capacity scaling to $M = 40$k (C4) delivers +3.60 PSNR, +0.017 SSIM, and a 32.2% LPIPS reduction. Extremely small graphs ($<$ 10k) under-represent non-rigid dynamics, while very large ones (100k) underperform C4, suggesting optimization and overfitting issues at excessive capacity.

Fig. 6 (*left*) shows that VE reduces motion spikes by 35.5%, with green frames showing stable rendering and red frames showing flickering. Fig. 6 (*right*) shows PGI improves detail beyond the body, while SMAD removes skirt artifacts present in template-only results.

4.3 ANALYSES

Model Selection. To validate the effectiveness of our proposed design for SMAD module, we additionally conducted a controlled comparison against a carefully designed MLP-based autoregressive deformer, following (Zheng et al., 2021), and vanilla GNN. This baselines use the same inputs (positions, encoded velocities) as our method, ensuring a fair comparison. Table 4 shows that the MLP deformer fits the training motion but degrades significantly on unseen motion, while the GNN-

Physics & Graph Design (PGD)				Velocity Encoding (VE)				SMAD Capacity (M)			
Method	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**Method**	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**Method**	**PSNR↑**	**SSIM↑**	**LPIPS↓**
A0: Base	25.21	0.952	0.058	B0: w/o VE	22.06	0.930	0.067	C0: w/o SMAD	24.29	0.946	0.059
A1: + phys. reg	26.05	0.956	0.052	B1: $\tau_v = 1$	23.41	0.932	0.060	C1: $M = 5\text{k}$	25.36	0.950	0.053
A2: + adaptive k_{ij}	26.44	0.958	0.049	B2: $\tau_v = 7$	24.95	0.944	0.053	C2: $M = 10\text{k}$	26.47	0.958	0.048
A3: + message-passing	27.12	0.961	0.044	B3: $\tau_v = 15$	26.78	0.958	0.045	C3: $M = 100\text{k}$	27.02	0.962	0.045
A4: Full (Ours)	**27.89**	**0.963**	**0.040**	**B4:** $\tau_v = 11$ **(Ours)**	**27.89**	**0.963**	**0.040**	**C4:** $M = 40\text{k}$ **(Ours)**	**27.89**	**0.963**	**0.040**

Table 2: Ablation study on the effectiveness of our mainly proposed components. Three column blocks report (Left) physics/graph design, (Middle) velocity encoding horizon (τ_v), (Right) SMAD capacity M.

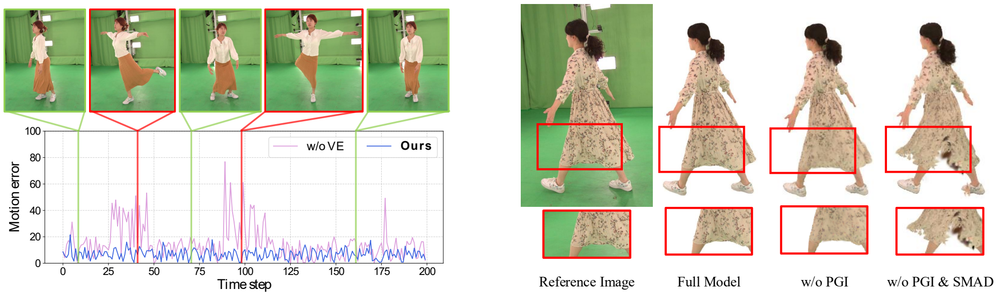


Figure 6: Ablation study on the visual effectiveness of (*left*) VE, (*right*) PGI, and SMAD. VE significantly reduces the motion error by encouraging temporal consistent deformation. PGI contributes to capturing fine-detailed clothing patterns, and SMAD sufficiently guarantees the robustness of clothing dynamics.

based deformers remains substantially more stable and accurate. This confirms that the graph-based formulation provides stronger structural priors and better generalization for clothed-human deformation. We also observe that embedding features on edges through message passing yields additional performance gains.

Generalization. Our auto-regressive deformation leverages a second-order state (x_t, v_t), where velocities are obtained via finite differences. This provides a physically meaningful motion representation that aligns with how real deformable systems evolve, enabling more stable extrapolation than pose-only models. Prior work in human and cloth dynamics similarly shows that explicit velocity states improve rollout stability. By integrating over (x_t, v_t) and regularizing with damping and local-isometry constraints, our model suppresses high-frequency drift and captures inertia-driven behavior, leading to robust generalization to unseen motions. Table 3 further supports this observation. It reports quantitative results on the 4D-Dress subjects across train, test, and out-of-distribution (OOD) motion sequences. To assess whether performance differences across these distributions are statistically significant, we conducted paired t-tests for each setting. No comparison yielded a significant difference, indicating that our model maintains consistent performance regardless of motion distribution. This empirical evidence reinforces that our approach generalizes reliably to dynamic motions unseen during training.

Data	PSNR↑	SSIM↑	LPIPS↓
OOD	26.51	0.956	0.049
Test	27.89	0.963	0.040
Train	28.64	0.984	0.037

Table 3: Quantitative results on train/test, and out-of-distribution (OOD) motion sequences to evaluate generalization capability of our method (*blue*: p-val $p > 0.05$).

Error Accumulation. It is well-known that auto-regressive models are prone to numerical error accumulation over long sequences. To analyze and reflect on this point, we captured two types of motion sequences, each lasting over 30 seconds: (a) a dynamic pose sequence, and (b) a repetitive pose sequence. We evaluated our proposed method, and also conducted a comparative analysis with and without our proposed velocity encoding scheme to evaluate its effectiveness. Our velocity encoding scheme appears to mitigate this issue by incorporating a history of multiple past states, rather than relying solely on the most recent estimate. This allows the model to remain robust even when the immediate past prediction is noisy, reducing the risk of cumulative drift.

Training Cost. Our model requires an average of 12.5 hours for personalized Gaussian initialization and 4.5 hours for training the secondary motion-aware deformation module, totaling approximately 17 hours on a single NVIDIA RTX 3090 GPU. Considering that existing state-of-the-art methods (Moon et al., 2024) typically require around 4 hours of training, our approach indeed incurs higher computa-

Model	PSNR↑	SSIM↑ Test	LPIPS↓	PSNR↑	SSIM↑ Train	LPIPS↓
MLP	25.46	0.954	0.056	27.97	0.973	0.044
vanilla GNN	28.68	0.958	0.045	28.44	0.980	0.040
Ours	**27.89**	**0.963**	**0.040**	**28.64**	**0.984**	**0.037**

Table 4: Quantitative results of difference design choices of SMAD on train/test distributions.

Model	PSNR↑	SSIM↑ (a) dynamic pose	LPIPS↓	PSNR↑	SSIM↑ (b) repetitive pose	LPIPS↓
w/o VE	24.47	0.949	0.050	24.69	0.950	0.049
w/ VE (**Ours**)	**25.65**	**0.955**	**0.044**	**26.84**	**0.960**	**0.039**

Table 5: Analyses of error accumulation ablating the velocity encoding (VE) strategy on two long sequences.

tional cost. However, we emphasize that, unlike prior methods whose limited model capacity yields only marginal gains even with extended training, our formulation continues to deliver significant performance improvements when trained longer (see Fig. 7). This suggests that our method possesses a higher effective capacity and is well suited for high-fidelity dynamic appearance modeling in personalized avatar reconstruction.

5 Discussion

On the Importance of Gaussian Initialization. Accurate initialization is fundamental for animatable 3D Gaussian avatars, especially when modeling loose-fitting clothing. Prior monocular methods rely on parametric template bodies (Loper et al., 2015; Pavlakos et al., 2019), assuming minimally clothed geometry. As seen in Fig. 1(c), this creates large mismatches between template surfaces and real garment volumes, causing undersampling, silhouette distortion, and instability under unseen poses. Because Gaussians are explicit point samples, such errors propagate into deformation and cannot be repaired by skinning alone. Our personalized Gaussian initialization avoids these issues by estimating a clothed canonical field via deformable NeRF, producing a subject-specific and geometry-faithful Gaussian distribution. This reduces the deformation network's burden, enabling it to focus on true non-rigid motion rather than fixing incorrect geometry. Fig. 6 show that PGI improves clothing detail, reduces skirt-splitting, and stabilizes secondary motion. Overall, initialization is not a preprocessing step but a critical determinant of garment fidelity and temporal stability.

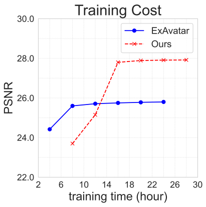


Figure 7: Analysis of the training-cost trade-off compared to an existing method.

Template-free Deformation. Even with a faithful canonical geometry, deformation remains constrained when tied to template-based articulation such as LBS. These methods define motion as a direct function of skeletal pose, which fails for loose garments that do not follow body kinematics. As visualized in Fig. 1(b), template-driven deformation creates motion-error spikes, flickering, and cloth splitting because it lacks temporal awareness and restricts non-rigid behavior. Our SMAD module departs from this paradigm by learning a template-free, autoregressive deformation field on a velocity-encoded Gaussian graph. Instead of following a fixed hierarchy, Gaussians interact through learned graph messages, enabling the model to infer how cloth regions co-move or lag independently of the body. This grants expressive, pose-agnostic deformation capability and yields coherent dynamics across diverse motions. Results in Table 2 and Fig. 6 show that removing template constraints dramatically improves robustness, generalization, and overall clothing realism.

6 Conclusion

In this paper, we introduced a novel approach for dynamic appearance modeling of 3D Gaussian Splatting-based avatars from a single video, focusing on loose-fitting clothing dynamics. We addressed two main challenges: limited Gaussian deformation from template articulation, and misalignment issues from Gaussian initialization relying on naked body templates. To resolve these, we proposed an autoregressive Gaussian deformation strategy that predicts velocities for realistic cloth dynamics, and a personalized Gaussian initialization using a deformable neural radiance field to capture clothed silhouettes. Additionally, we provided a new in-the-wild dataset featuring subjects performing dynamic movements in challenging clothing. Extensive evaluations confirmed our method improves realism and outperforms existing approaches in both controlled and unconstrained settings.

REFERENCES

Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3D people models. In *CVPR*, 2018.

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

Yushuo Chen, Zerong Zheng, Zhe Li, Chao Xu, and Yebin Liu. Meshavatar: Learning high-quality triangular human avatars from multi-view videos. In *ECCV*, 2024.

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*, pp. 1–9, 2022.

Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J Black. Learning disentangled avatars with hybrid 3d representations. *arXiv preprint arXiv:2309.06441*, 2023.

Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5712–5721, 2021.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Message passing neural networks. In *Machine learning meets quantum physics*, pp. 199–214. Springer, 2020.

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

Artur Grigorev, Giorgio Becherini, Michael Black, Otmar Hilliges, and Bernhard Thomaszewski. Contourcraft: Learning to resolve intersections in neural multi-garment simulations. In *ACM SIGGRAPH 2024 conference papers*, pp. 1–10, 2024.

Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2Avatar: 3D avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *CVPR*, 2023.

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. *arXiv:2409.15269*, 2024.

Chen Guo, Junxuan Li, Yash Kant, Yaser Sheikh, Shunsuke Saito, and Chen Cao. Vid2avatar-pro: Authentic avatar from videos in the wild via universal prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5559–5570, 2025.

Marc Habermann and et al. Real-time deep dynamic characters. In *SIGGRAPH*, 2021.

Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions On Graphics (TOG)*, 38(2):1–17, 2019.

Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5052–5063, 2020.

Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. A deeper look into deepcap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4009–4022, 2021.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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*, pp. 20418–20431, 2024b.

Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM TOG*, 42(4):1–12, 2023. doi: 10.1145/3592415. URL `https://doi.org/10.1145/3592415`.

Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. 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, 2023a.

Yuheng Jiang, Kaixin Yao, Zhuo Su, Zhehao Shen, Haimin Luo, and Lan Xu. Instant-nvr: Instant neural volumetric rendering for human-object interactions from monocular rgbd stream. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 595–605, 2023b.

Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE international conference on computer vision*, pp. 3334–3342, 2015.

Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5614–5623, 2019.

Bernhard Kerbl and et al. 3d gaussian splatting for real-time radiance field rendering. In *SIGGRAPH*, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural Human Performer: Learning generalizable radiance fields for human performance rendering. *NeurIPS*, 2021.

Youngjoong Kwon, Baole Fang, Yixing Lu, Haoye Dong, Cheng Zhang, Francisco Vicente Carrasco, Albert Mosella-Montoro, Jianjin Xu, Shingo Takagi, Daeil Kim, et al. Generalizable human gaussians for sparse view synthesis. In *European Conference on Computer Vision*, pp. 451–468. Springer, 2024a.

Youngjoong Kwon, Lingjie Liu, Henry Fuchs, Marc Habermann, and Christian Theobalt. DELIFFAS: Deformable light fields for fast avatar synthesis. *NeurIPS*, 2024b.

Lik-Hang Lee, Tristan Braud, Peng Yuan Zhou, Lin Wang, Dianlei Xu, Zijun Lin, Abhishek Kumar, Carlos Bermejo, Pan Hui, et al. All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda. *Foundations and trends® in human-computer interaction*, 18(2–3):100–337, 2024.

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*, pp. 19876–19887, 2024.

Zhe Li, Zerong Zheng, Yuxiao Liu, Boyao Zhou, and Yebin Liu. Posevocab: Learning joint-structured pose embeddings for human avatar modeling. *ACM SIGGRAPH Conference Proceedings*, 2023.

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*, pp. 19711–19722, 2024.

Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pp. 590–606. Springer, 2022.

Zhouyingcheng Liao, Vladislav Golyanik, Marc Habermann, and Christian Theobalt. Vinecs: video-based neural character skinning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1377–1387, 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.

Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, pp. 800–809. IEEE, 2024.

Miles Macklin, Matthias Müller, and Nuttapong Chentanez. Xpbd: position-based simulation of compliant constrained dynamics. In *Proceedings of the 9th International Conference on Motion in Games*, pp. 49–54, 2016.

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, 2021.

Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. *arXiv preprint arXiv:2407.21686*, 2024.

Matthias Müller, Bruno Heidelberger, Marcus Hennix, and John Ratcliff. Position based dynamics. *Journal of Visual Communication and Image Representation*, 18(2):109–118, 2007.

Xiaoyu Pan, Jiaming Mai, Xinwei Jiang, Dongxue Tang, Jingxiang Li, Tianjia Shao, Kun Zhou, Xiaogang Jin, and Dinesh Manocha. Predicting loose-fitting garment deformations using bone-driven motion networks. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

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

Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021a.

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural Body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021b.

Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (ToG)*, 36(4):1–15, 2017.

Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. *CVPR*, pp. 20299–20309, 2024a.

Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. *CVPR*, pp. 5020–5030, 2024b.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL `https://arxiv.org/abs/2408.00714`.

Radu Bogdan Rusu and Steve Cousins. 3d is here: Point cloud library (pcl). In *2011 IEEE international conference on robotics and automation*, pp. 1–4. IEEE, 2011.

Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *CVPR*, 2024.

Igor Santesteban, Miguel A Otaduy, and Dan Casas. Snug: Self-supervised neural dynamic garments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8140–8150, 2022.

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)*, 2024.

Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-avatar: Expressive human avatars. In *Computer Vision and Pattern Recognition (CVPR)*, 2023a.

Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-avatar: Expressive human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16911–16921, 2023b.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Carsten Stoll, Juergen Gall, Edilson De Aguiar, Sebastian Thrun, and Christian Theobalt. Video-based reconstruction of animatable human characters. *ACM Transactions on Graphics (TOG)*, 29(6): 1–10, 2010.

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

Ivan E Sutherland et al. The ultimate display. In *Proceedings of the IFIP Congress*, volume 2, pp. 506–508. New York, 1965.

Demetri Terzopoulos, John Platt, Alan Barr, and Kurt Fleischer. Elastically deformable models. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, pp. 205–214, 1987.

Bohan Wang, Mianlun Zheng, and Jernej Barbič. Adjustable constrained soft-tissue dynamics. In *Computer Graphics Forum*, volume 39, pp. 69–79. Wiley Online Library, 2020.

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

Wenbo Wang, Hsuan-I Ho, Chen Guo, Boxiang Rong, Artur Grigorev, Jie Song, Juan Jose Zarate, and Otmar Hilliges. 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.

Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Vid2actor: Free-viewpoint animatable person synthesis from video in the wild. *arXiv preprint arXiv:2012.12884*, 2020.

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*, pp. 16210–16220, 2022.

Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4389–4398, 2024.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9043–9053, 2023a.

Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Computer Vision and Pattern Recognition (CVPR)*, 2023b.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19447–19456, 2024.

Zhenzhen Yu, Yikai Zhang, and et al. Monohuman: Animatable human neural field from monocular video. In *CVPR*, 2023.

Zhichao Zhai, Guikun Chen, Wenguan Wang, Dong Zheng, and Jun Xiao. Taga: Self-supervised learning for template-free animatable gaussian articulated model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21159–21169, 2025.

Youyi Zhan, Tianjia Shao, Yin Yang, and Kun Zhou. Real-time high-fidelity gaussian human avatars with position-based interpolation of spatially distributed mlps. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26297–26307, 2025.

Mianlun Zheng, Yi Zhou, Duygu Ceylan, and Jernej Barbic. A deep emulator for secondary motion of 3d characters. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5932–5940, 2021.

Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. 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.

Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. AvatarRex: Real-time expressive full-body avatars. *ACM TOG*, 2023.

Heming Zhu, Fangneng Zhan, Christian Theobalt, and Marc Habermann. Trihuman: A real-time and controllable tri-plane representation for detailed human geometry and appearance synthesis. *ACM Trans. Graph.*, September 2024. ISSN 0730-0301. doi: 10.1145/3697140. URL `https://doi.org/10.1145/3697140`.

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.

Appendix

Table of Contents

A NOTATION

We summarize the key notations used in main paper in Table F.

Table F: Summary of Mathematical Notations in the Paper

#	Notation	Dimension / Set	Definition
1	$V = \{I_t\}_{t=1}^{T}$	$I_t \in \mathbb{R}^{H \times W \times 3}$	Monocular RGB video with T frames
2	T	$\mathbb{N}$	Number of video frames
3	$\mathcal{G}_t^d = \{(\mu_{t,i}, \Sigma_{t,i}, c_{t,i}, \alpha_{t,i})\}_{i=1}^{N}$	–	Deformed Gaussians at time t
4	N	$\mathbb{N}$	Number of Gaussian primitives
5	$\mu_{t,i}$	$\mathbb{R}^3$	Mean position of the i-th Gaussian
6	$\Sigma_{t,i}$	$\mathbb{R}^{3\times 3}$	Covariance (size & orientation)
7	$c_{t,i}$	$\mathbb{R}^3$	RGB color
8	$\alpha_{t,i}$	$\mathbb{R}$	Opacity
9	$\Theta = \{\theta_t\}_{t=1}^{T}$	$\mathrm{SO}(3)^K$	SMPL joint pose sequence
10	$\mathcal{R}(\cdot)$	$\left(\mathbb{R}^3, \Sigma, \alpha\right)^N \to \mathbb{R}^{H \times W \times 3}$	Differentiable 3D Gaussian renderer
11	$\mathcal{G}^c$	–	Canonical (undeformed) Gaussian set
12	$\Gamma = (X, A)$	$X \in \mathbb{R}^{M \times 3}, A \in \mathbb{R}^{M \times M}$	Gaussian graph (nodes & adjacency)
13	$X = \{x_i\}_{i=1}^{M}$	$\mathbb{R}^3$	Node positions after voxel downsampling
14	$A_{ij} = \exp\left[-\frac{\lVert x_i - x_j \rVert^2}{\rho_a^2}\right]$	$[0, 1]$	Edge weight (k-NN Gaussian kernel)
15	ρ_a	$\mathbb{R}^+$	Distance-sensitivity bandwidth
16	$h_i = (x_i, \bar{v}_i, e_i)$	$\mathbb{R}^{d_h}$	Node feature: position, velocity buffer, pose embed
17	$v_i(t) = \frac{x_i(t) - x_i(t - \Delta t)}{\Delta t}$	$\mathbb{R}^3$	Instantaneous velocity
18	τ_v	$\mathbb{N}$	Past-velocity buffer length
19	g_i	$\mathbb{R}^+$	Point mass for physics model
20	$a_i(t) = \ddot{x}_i(t)$	$\mathbb{R}^3$	Acceleration of node i
21	γ_i	$\mathbb{R}^+$	Damping coefficient
22	k_{ij}	$\mathbb{R}^+$	Spring stiffness between nodes i, j
23	L_{ij}^{rest}	$\mathbb{R}^3$	Rest offset of the spring
24	$F_i^{\text{ext}}(t)$	$\mathbb{R}^3$	External force applied to node i
25	Δt	$\mathbb{R}^+$	Simulation time-step
26	z_i	$\mathbb{R}^{d_z}$	Learnable latent code for Gaussian i
27	$\omega_{t,k} = \exp\left[-\frac{1 - d_{t,0}^{\top} d_{t,k}}{2\rho}\right]$	$(0, 1]$	Confidence of k-th synthesized view
28	$d_{t,0}, d_{t,k}$	$\mathbb{S}^2$	Unit view-direction vectors
29	ρ	$\mathbb{R}^+$	Temperature for confidence fall-off
30	$f_{t,k}$	$\mathbb{R}^{d_f}$	Pixel-aligned feature from view k
31	$(q_{t,0}, k_{t,k}, v_{t,k})$	$\mathbb{R}^d$	Query, key, value embeddings for attention
32	$\alpha_{t,k}$	$[0, 1]$	Attention weight for view k
33	$\bar{\sigma}(x) = \frac{1}{T} \sum_{t=1}^{T} \sigma(x, t)$	$\mathbb{R}$	Time-averaged density for Gaussian extraction
34	κ	$\mathbb{R}^+$	Density threshold for voxel selection
35	$L_{\text{smooth}}, L_{\text{bend}}, L_{\text{iso}}, L_{\text{damp}}$	-	Regularization losses

B IMPLEMENTATION DETAILS

B.1 VELOCITY-ENCODED GAUSSIAN GRAPH

Voxel-grid downsampling ($N \to M$). Given a set of N initial Gaussian points, we introduce autoregressively graph-based Gaussian deformation to transform the Gaussians without pre-defined articulation to template parametric model for enhancing secondary motion dynamics. To avoid an $\mathcal{O}(N^2)$ neighbourhood search and to limit graph size for the GNN, we down-sample the Gaussian cloud on an isotropic voxel grid:

1. *Grid resolution.* Let $d_{\min}$ denote the minimum distance below which two Gaussians would overlap in the 3DGS renderer (*e.g.* the renderer's splat radius at canonical scale). We choose the voxel edge length as $s = 2d_{\min}$, which empirically yields $\approx$10 Gaussians per occupied voxel.
2. *Hash insertion.* Every Gaussian is hashed into a voxel key. It retains the index with the *smallest* per-voxel rendering error, measured on a 4× subsampled depth map; all other Gaussians in that voxel are discarded.
3. *Representative pooling.* For the surviving indices we conduct average pooling, giving a single *graph node*. The total number of nodes is $M = 40k$, an order of magnitude smaller than N without noticeable quality loss.

k-Nearest-Neighbour edge set. With the down-sampled node coordinates $X = \{\mathbf{x}_j\}_{j=1}^{M}$ (which is different from the one defined in PGI), we build an undirected, symmetric k-NN graph: $A = \big\{\,(i,j) \;\big|\; \mathbf{x}_i \in \mathrm{Top}k\big(\,\|\mathbf{x}_i - \mathbf{x}_j\|_2\big)\big\}$. We set $k = 16$, which is sufficiently dense to preserve local manifold connectivity yet keeps the message-passing cost low.

B.2 SECONDARY MOTION-AWARE GAUSSIAN DEFORMATION

Architecture. Given the *velocity-encoded Gaussian graph* $- = (X, A)$ (Table G), SMAD converts the current node state at animation step t into frame-specific *Gaussian attribute deltas* $\{\Delta\boldsymbol{\mu}_i, \Delta\boldsymbol{\Sigma}_i, \Delta\mathbf{c}_i, \Delta\alpha_i\}_{i=1}^{M}$, through three conceptually simple stages. We present the architectural details of our SMAD, as illustrated in Table H.

Node projection (layer 0). Each node i consists of concatenated features $\mathbf{h}_i = \{\mathbf{x}_i, \bar{\mathbf{v}}_i, \mathbf{e}_i\}$, where embedding vector through MLP is obtained by driving pose sequences $\Theta_{t-\tau:t} = \{\theta_t, ..., \theta_{t-\tau}\}$. The $\mathbf{h}_i$ seeds the message-passing stage.

Three shared message-passing iterations (layers 1–9). At each iteration $\ell \in \{1, 2, 3\}$ we construct an edge feature vector for every directed pair $(i, j) \in A$

$$\mathbf{e}_{ij}^{\ell} = \big[h_i^{\ell-1},\, h_j^{\ell-1},\, \bar{\mathbf{x}}_j - \bar{\mathbf{x}}_i,\, \bar{\mathbf{v}}_j - \bar{\mathbf{v}}_i,\, \|\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_i\|_2,\, 1\big] \in \mathbb{R}^{519}.$$

The shared *edge-MLP* M_θ (layer 1+3($\ell - 1$)) compresses $\mathbf{e}_{ij}^{\ell}$ to a 128-D message $m_{ij}^{\ell} = M_\theta(\mathbf{e}_{ij}^{\ell})$. For every receiver node we perform mean aggregation over its k nearest neighbours: $\bar{m}_i^{\ell} = \frac{1}{k}\sum_{j\in\mathcal{N}(i)} m_{ij}^{\ell}$. The aggregated vector is fed, together with the previous hidden state $h_i^{\ell-1}$, into a GRU cell $h_i^{\ell} = H_\theta(\bar{m}_i^{\ell}, h_i^{\ell-1})$. Because M_θ and H_θ share weights across iterations, the network learns a recurrent, physics-inspired propagation of inertia without increasing parameter count.

Decoder U_ψ (layers 10–15). After three iterations we obtain the refined latent representation $h_i^3 \in \mathbb{R}^{256}$ for each node. A two-stage MLP (256→256→128 with ReLU) acts as a shared decoder whose final 128-D activations feed four independent linear heads:

$$\Delta\boldsymbol{\mu}_i,\ \Delta\mathbf{v}_i, \Delta\boldsymbol{\Sigma}_i,\ \Delta\mathbf{c}_i = U_\psi(h_i^3).$$

Here, $\Delta\boldsymbol{\mu}_i$ is a 3-D position offset, $\Delta\mathbf{v}_i$ a 3-D velocity refinement that is re-queued into the velocity ring buffer, $\Delta\boldsymbol{\Sigma}_i$ a log-diagonal covariance update, and $\Delta\mathbf{c}_i \in [0, 1]^3$ a colour residual (sigmoid-bounded). These deltas are added to the canonical attributes before the Gaussian splatting renderer is invoked for the current frame.

SMAD therefore (i) embeds pose and recent motion into a compact latent space, (ii) injects neighbourhood cues through three message-passing steps that emulate mass–spring–damper interactions, and (iii) decodes temporally coherent, view-aware adjustments to every Gaussian's geometry and appearance.

Layer	Operation	Input	Output
1-2	FC + LN + ReLU	$\{\theta_t, ..., \theta_{t-\tau}\}$	256
3	FC + LN + ReLU + skip	256	256
4	Head (Linear)	256	128

Table G: Embedding MLP of driving motions into Gaussian graph for SMAD.

Layer	Operation	Input (dim)	Output (dim)
0	Node projection (Linear → ReLU)	$(\mathbf{x}_i, \bar{\mathbf{v}}_i, \mathbf{e}_i)$	256
Message-Passing Iteration 1–3 (shared weights)			
1,4,7	M_θ edge-MLP	edge features (519)	128
2,5,8	Message aggregation	$k \times 128$	128
3,6,9	H_θ	$(\bar{\mathbf{m}}_i^{\text{agg}}, \mathbf{h}_i)$	256
Decoder U_ψ: Gaussian attribute heads			
10	MLP-1 (Linear 256→256→ReLU)	256	256
11	MLP-2 (Linear 256→128→ReLU)	256	128
12	$\Delta\boldsymbol{\mu}$ head	128	3
13	$\Delta\mathbf{v}$ head	128	3
14	$\Delta\boldsymbol{\Sigma}$ head	128	3
15	$\Delta\mathbf{c}$ head + Sigmoid	128	3

Table H: Secondary Motion-Aware Deformation (SMAD): main network & Gaussian decoder.

B.3 TRAINING

We adopted 2-stage training. Fristly, we train personalized Gaussian initialization as pre-stage, where it locates initial Gaussians densely aligned onto the person-specific silhouette. Thereafter, we train secondary motion-aware Gaussian deformation to auto-regressively transform the canonical 3D Gaussians that are aware of clothing dynamics. We used the Adam optimizer Kingma & Ba (2015) with an initial learning rate of 0.001, decaying by a factor of 0.5 if no improvement is made in four consecutive epochs. We used PyTorch Paszke et al. (2019) for the backend processing. All experiments were conducted on AMD Ryzen Threadripper PRO 5965WX CPU and an NVIDIA GeForce RTX 3090 GPU.

C EXPERIMENTAL SETTING DETAILS

C.1 DATASET DESCRIPTION

Motivation. Existing datasets for evaluating animatable 3D avatars predominantly focus on subjects wearing tight-fitting clothing and performing repetitive, often monotonous motions. Although the recently proposed 4D-Dress dataset addresses some of these limitations by including diverse garment types, it is still collected in a controlled laboratory setting and primarily designed for multi-view capture evaluations. However, our ultimate goal is to democratize avatar generation—making it robust and accessible to everyday users using only monocular inputs. To this end, it is essential to evaluate performance under in-the-wild scenarios, where diverse factors such as occlusion, motion blur, and uncontrolled lighting can affect avatar quality. We introduce a novel dataset **LoCo-Human**, featuring (1) subjects wearing loose-fitting garments, (2) realistic clothing dynamics exhibiting secondary motion, and (3) videos captured in the wild. This setting enables evaluation of avatar reconstruction robustness under real-world conditions.

Dataset Statistics. Our dataset comprises five unique subjects, each recorded in a total of six sequences: one static-motion sequence and five dynamic-motion sequences. For each subject, one sequence captures a 360-degree rotation, while the remaining four sequences feature free-form dynamic motion, performed without scripted guidelines. All subjects wear challenging garments, such as long skirts and padded coats, designed to emphasize loose-fitting clothing dynamics. Fig. H shows qualitative results on several samples from our dataset.

Capture Setup. All sequences were captured using standard smartphone devices. We provide 3 to 10-minute-long RGB video sequences for each subject, along with corresponding segmentation masks, depth maps, and SMPL pose parameters. Segmentation masks were obtained using SAM v2 Ravi et al. (2024), with optional manual refinement to ensure silhouette accuracy. For depth information, we used smartphone depth cameras to obtain coarse estimates, which are further refined using Metric3D Yin et al. (2023a). To extract driving pose parameters, we fit the SMPL model Loper et al. (2015) to each input monocular video, following the protocol described in Moon et al. (2024).

C.2 BASELINE METHODS

Our objective is to construct animatable 3D human avatars from a single video, faithfully reflecting secondary motion, based on the 3D Gaussian Splatting representation. To validate the effectiveness of our method, we conduct a comparative analysis with existing 3DGS-based avatar approaches. Given the rapid expansion of research in this domain, an exhaustive comparison with all prior works is impractical. Therefore, we select a subset of publicly available methods, with particular emphasis on those that explicitly address dynamic appearance modeling.

GART Lei et al. (2024). GART introduces the Gaussian Articulated Template (GART) model, designed to reconstruct non-rigid articulated human subjects from monocular videos. To facilitate challenging deformation, it employs a learnable forward skinning strategy via latent bones. However, due to the lack of supervision on where to place novel bones and how to assign skinning weights for each Gaussian, the method struggles to maintain structural consistency during deformation. This often leads to a breakdown in kinematic coherence and introduces excessive degrees of freedom, making stable animation difficult.

GaussianAvatar Hu et al. (2024a). GaussianAvatar proposes an efficient method for creating realistic human avatars with dynamic 3D appearances from a single video. It utilizes UV positional maps to encode pose-dependent features and integrates them with canonical surface geometry. However, its pose representation is heavily reliant on parametric template priors. While effective for minimally clothed humans, this reliance limits its generalization to clothed avatars whose geometry deviates significantly from the template, especially in the case of skirts, which often exhibit unnatural splitting between the legs under dynamic motion.

3DGS-Avatar Qian et al. (2024a). 3DGS-Avatar also presents a framework for creating animatable human avatars from monocular video using 3D Gaussian primitives. It introduces a non-rigid deformation network that learns per-Gaussian offsets to represent dynamic clothed avatars. However, by assigning independent degrees of freedom to each Gaussian, the method neglects the underlying structural coherence of the avatar. This leads to undesired needle-like artifacts, particularly under dynamic motions.

ExAvatar Moon et al. (2024). ExAvatar proposes a hybrid representation that combines a whole-body parametric mesh with 3D Gaussian Splatting. By binding each Gaussian to corresponding mesh vertices, the model ensures stable deformation under novel motions. Nevertheless, it exhibits weak appearance modeling for clothed subjects wearing loose-fitting garments, such as coats or skirts, which significantly deviate from the shape of the minimally clothed parametric template.

In summary, existing methods largely depend on shape and articulation priors from parametric template models to synthesize and animate avatars. This reliance limits their ability to model (1) loose-fitting clothed humans with geometry far from minimally clothed templates, and (2) realistic deformation that preserves geometric structure while capturing clothing dynamics. Our approach aims to overcome these limitations by introducing a template-free formulation tailored for secondary motion-aware avatar modeling.

C.3 EVALUATION METRICS

We used PSNR, SSIM, LPIPS, and motion error as the primary evaluation metrics.

Peak Signal-to-Noise Ratio (PSNR). PSNR is a widely used metric for evaluating the reconstruction quality of compressed or reconstructed images by comparing them to the original. It quantifies the ratio between the maximum possible pixel value and the power of the distortion (error) introduced. Given an original image I and a rendered image $\hat{I}$ of animatable 3D Gaussian avatars, we first compute the Mean Squared Error (MSE). Then the PSNR is defined as:

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{255^2}{\text{MSE}}\right).$$

Structural Similarity Index Measure (SSIM). SSIM is a perceptual metric that quantifies image quality degradation based on changes in structural information, taking into account human visual perception. Unlike PSNR, it considers luminance, contrast, and structural similarity. Given two local image patches x and y, SSIM is defined as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

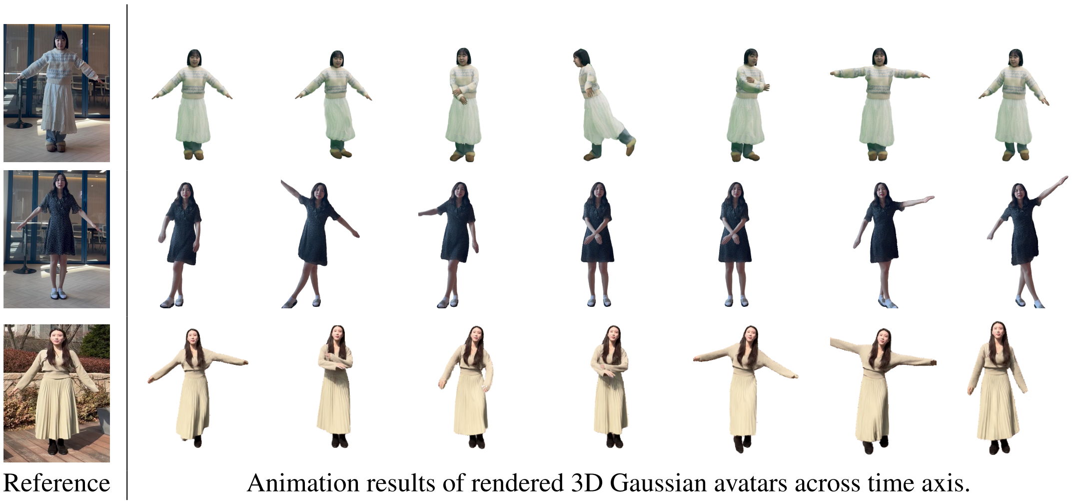


Figure H: Qualitative results on LoCo-Human consisting of the subjects wearing loose-fitting clothes with dynamic motions.

where (μ_x, μ_y) are mean intensities, (σ_x^2, σ_y^2) are variances, σ_{xy} are covariance between x and y. The final SSIM value is obtained by averaging local SSIM scores across the entire image.

Learned Perceptual Image Patch Similarity (LPIPS). LPIPS is a perceptual metric that compares images using deep features extracted from pretrained neural networks. It is designed to align closely with human perceptual judgments by evaluating similarity in a learned feature space. Given two images I and $\hat{I}$, let $\hat{f}^l(x)$ and $\hat{f}^l(y)$ denote the normalized feature maps from layer l of a pretrained network ϕ, with spatial dimensions $H_l \times W_l$ and channel dimension C_l. Then LPIPS is defined as:

$$\text{LPIPS}(I, \hat{I}) = \sum_l w_l \cdot \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \left\| \hat{f}^l_{h,w}(I) - \hat{f}^l_{h,w}(\hat{I}) \right\|_2^2$$

where w_l are learned weights that reweight the contribution of each layer to better match human perceptual similarity. We use deep features from Simonyan & Zisserman (2014). For ZJU-Mocap, following the convention of previous studies Qian et al. (2024a), we reported the LPIPS values scaled by 10^3 in the main draft to make the performance differences with the baselines more clearly distinguishable.

Motion Error. We additionally measured motion error to evaluate the temporally consistent animation and fidelity to the driving motion of the generated avatars. Specifically, this is computed as the acceleration error between the driving pose (used as the condition) and the corresponding pose of the rendered Gaussian avatar, which is estimated in reverse using a pre-trained pose estimator Li et al. (2022). To assess this, we measured acceleration error, presented in Kanazawa et al. (2019) the acceleration error metric is used. It measures the average deviation between the estimated and ground-truth joint accelerations across a temporal sequence. Given a sequence of 3D joint positions $\{\mathbf{x}^t_{\text{joints}} \in \mathbb{R}^{3J}\}_{t=1}^T$, the acceleration at time t is approximated using the second-order finite difference: $\mathbf{a}^t_{\text{joints}} = \mathbf{x}^{t+1}_{\text{joints}} - 2\mathbf{x}^t_{\text{joints}} + \mathbf{x}^{t-1}_{\text{joints}}$. The acceleration error is then computed as:

$$\text{Motion Error} = \frac{1}{T-2} \sum_{t=2}^{T-1} \left\| \hat{\mathbf{a}}^t_{\text{joints}} - \mathbf{a}^t_{\text{joints}} \right\|_2$$

where $\hat{\mathbf{a}}^t_{\text{joints}}$ and $\mathbf{a}^t_{\text{joints}}$ denote the predicted and ground-truth joint accelerations, respectively.

Method	Training Time	Inference Time	Method	Training Time	Inference Time
ExAvatar	3.0h	27 fps	**Ours**	12.5h (1) + 4.5h (2)	26 fps

Table I: Training (time: hours) and testing (fps: frames per second) cost comparison to template Pavlakos et al. (2019)-based 3D Gaussian avatar Moon et al. (2024). Our training time consists of (1) personalized Gaussian initialization module and (2) secondary motion-aware Gaussian deformation.

D ADDITIONAL ANALYSES AND DISCUSSIONS

D.1 HYPER-PARAMETER SEARCH

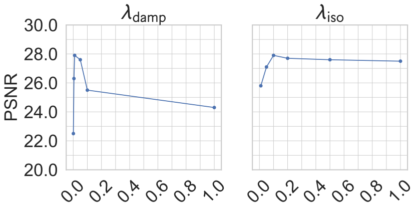


Figure I: Hyper-parameter search on weighting factor of loss functions λ_{damp} and λ_{iso}.

We ablated key hyperparameters: scaling factors of loss functions $\lambda_{\text{damp}} \in \{0.001\text{–}1.0\}$, and $\lambda_{\text{iso}} \in \{0.01\text{–}1.0\}$ in Fig. I. The results show that $\lambda_{\text{damp}} = 0.01$ and $\lambda_{\text{iso}} = 0.1$ performed best; $\lambda_{\text{damp}} = 1.0$ sharply degraded results, implying excessive temporal smoothing, while isotropy gains largely saturated past 0.1. Based on the results shown in the figure, both $\mathcal{L}_{\text{iso}}$ and $\mathcal{L}_{\text{damp}}$ help suppress excessive deformation and promote stable optimization, though they operate differently. $\mathcal{L}_{\text{iso}}$ preserves local isotropy, preventing geometric distortions, and its increase leads to a gradual improvement in PSNR. In contrast, $\mathcal{L}_{\text{damp}}$ mitigates excessive dynamic oscillations, yielding a more pronounced PSNR gain within an appropriate range. Combined, the two terms jointly enforce structural fidelity and dynamic stability, achieving a balanced improvement in both visual consistency and numerical robustness.

D.2 TIME EFFICIENCY

To evaluate the efficiency of our model, we measured the runtime of the proposed method from both the training and testing (inference) perspectives. Tab. I presents the average time costs, computed by measuring the runtimes for all subjects used in the experiments and reporting the mean values. Compared to existing 3DGS-based avatar methods that rely on parametric template model priors, our method consists of two stages. In the first stage, we train a network for personalized Gaussian initialization, and in the second stage, we train a secondary motion-aware Gaussian deformation network. On average, the first and second stages take 12.5 and 4.5 hours to train, respectively, which is relatively longer than existing methods. Importantly, the network trained in the first stage is not required during the inference phase. As a result, the inference-time cost is nearly equivalent to prior methods and operates at a near real-time speed.

D.3 STATISTICAL SIGNIFICANCE

We conducted a two-sided paired t-test, conservatively setting $p = 0.05$ to relieve a multi-comparison issue. The test was based on SSIM metric scores evaluated across the test sequences of all subjects used for evaluation on 4D-Dress. We performed (1) statistical significance analysis against comparison methods Lei et al. (2024); Hu et al. (2024a); Qian et al. (2024b); Moon et al. (2024), and (2) significance testing against ablated versions of our proposed main components. When compared with 3DGS-based baseline methods, all resulting p-values were lower than 0.005, demonstrating that our method achieves significantly improved performance despite the conservative threshold (see Table J). Furthermore, to assess the effectiveness of each major component proposed in this paper, we performed two-sided paired t-tests between the full model and its ablated variants. As shown in Table K, all components were found to be statistically significant, highlighting in particular the effectiveness of our template-free Gaussian deformation module.

Method	vs GART Lei et al. (2024)	vs GaussianAvatar Hu et al. (2024a)	vs 3DGS-Avatar Qian et al. (2024a)	vs ExAvatar Moon et al. (2024)
p-value	5.6×10^{-6}	4.2×10^{-6}	6.9×10^{-5}	5.7×10^{-5}

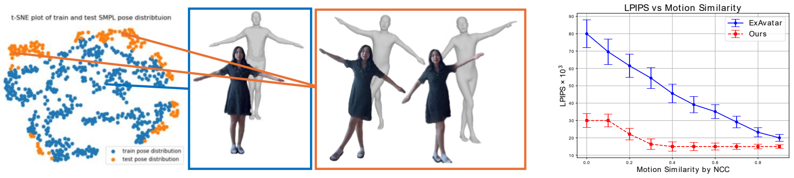
Table J: Statistical significance (p ¡ 0.05). We performed a two-sided paired t-test against each baseline method conservatively at $p = 0.05$ to relieve a multi-comparison issue. Our method exhibit statistical significance compared to the baselines suggesting that our method have significant performance improvement.

Method	vs w/o VE	vs w/o SMAD
p-value	3.3×10^{-4}	1.7×10^{-4}

Table K: Statistical significance (p ¡ 0.05) to validate the mainly proposed components. We performed a two-sided paired t-test against each baseline method conservatively at $p = 0.05$ to relieve a multi-comparison issue. The proposed components exhibit statistical significance by suggesting that each components are effective to create animatable 3D Gaussian avatars.

D.4 GENERALIZATION

We further evaluate the generalization performance. Fig. J shows the distribution of training poses and test poses on a t-SNE plot, as well as the performance on the in-the-wild dataset. The blue box indicates the distribution of the training dataset, and the orange box indicates that of the test dataset. Even though the test dataset was unseen during training, our method demonstrates excellent generalization performance. In Fig. J, we plot how the perceptual error changes relative to the motion similarity between the training and testing data, measured via normalized cross-correlation (NCC) between the time-varying 3D conditional poses. We observe a more pronounced increase in the error for Moon et al. (2024) as the testing frames deviate further from the training data. Furthermore, while ExAvatar shows a large variation in standard error that increases as motion similarity decreases, our method consistently maintains a low level of standard error. Since the standard error is computed between the train and test motions, a lower value indicates less overfitting and suggests better generalization performance. This suggests that our method exhibits robust generalization performance on par with the linear skinning model of the conventional template parametric model.



(a) (b)

Figure J: (a) Visual check of in-distribution (*blue*) and out-of-distribution (*orange*) driving poses with t-sne plot. (b) Average perceptual metric (LPIPS; Lower is better) with standard error plot of 4D-Dress over motion similarity between train and test set. Our method (*red*) maintains consistent rendering performance even for test motions with low similarity to the training motion—showing relatively less performance degradation compared to high-similarity cases—whereas the baseline Moon et al. (2024) (*blue*) exhibits a significant drop in perceptual quality when handling test motions with low motion similarity.

E MORE RESULTS

We present additional qualitative comparison results in the remainder. Please refer to the Fig. K, L, M, N, O, P.

F LIMITATIONS

First, our method struggles under dynamics that involve sudden and large motion changes. Although we designed the system to account for temporal context using velocity encoding and auto-regressive modeling, it still has difficulty predicting the emergence of node accelerations that lie outside the training distribution. Second, our method does not model multi-garment interactions. The current Gaussian graph is a single-layer structure that captures the overall clothed shape and ensures deformations that preserve this global structure. However, it does not model interactions between garments or predict their independent motions. In future work, we aim to address these limitations. To tackle the first challenge, we could introduce a deformation model that considers bidirectional temporal context. Alternatively, incorporating a generative flow matching technique that predicts a bundle of vectors (a vector field) may offer a promising way to learn and represent the distribution of complex motions. To address the second issue, we could consider constructing a hierarchical Gaussian graph and introducing a graph neural network to model interactions between different garments. However, achieving this would require highly accurate semantic segmentation between garments. Currently, such segmentation remains difficult in the presence of diverse self-occlusions and depth ambiguities from a single-view video. Therefore, enabling high-quality multi-garment segmentation from a single video alone would itself be a highly challenging yet exciting direction for future research in hierarchical Gaussian deformation modeling.

G BROADER IMPACTS

Potential Negative Societal Impacts. Our technology could be misused, leading to negative societal consequences. One major risk is Deepfake-style impersonation: a realistic avatar of a person could be created without consent and used to impersonate them, enabling misinformation or fraud. The ability to replicate someone's likeness from a single video also raises privacy concerns, as individuals could have their image replicated and misused in unwanted ways, which can erode trust in digital media. It could also impact creative industries: unauthorized digital replicas of actors might violate intellectual property rights and undermine the entertainment industry's economy, and a proliferation of lifelike fake characters could confuse audiences and devalue genuine performances. These risks underscore the need for ethical guidelines and safeguards to prevent malicious use of AI-driven avatar technology.

Broader Impact. Our work offers positive implications for research, industry, and consumers. *Research Community:* Our method introduces a new approach to animatable avatars using 3D Gaussian Splatting, advancing neural rendering, and provides an in-the-wild dynamic clothing dataset to spur further research on neural avatars and secondary motion modeling. *Industry:* The improved realism and efficiency of our approach can benefit digital human applications in entertainment, gaming, and virtual reality by enabling creators to produce lifelike characters with realistic cloth dynamics from minimal input, allowing immersive real-time experiences. *Consumers:* More realistic and animatable avatars mean more immersive virtual experiences for end-users. Users in VR and gaming will be able to interact through avatars that mirror their appearance and clothing motion, enhancing their sense of presence. By bridging real and virtual representations, our work enriches virtual experiences.

Ethics Statement. This work makes use of both publicly available datasets (e.g., ZJU-MoCap, 4D-Dress) and a newly collected dataset, LoCo-Human, which contains in-the-wild monocular video sequences of clothed human subjects. For all publicly available datasets, we adhere to their respective license terms and usage conditions.

For LoCo-Human, all participants were recruited with explicit informed consent, covering video recording, research use, and potential public release of the anonymized dataset. No minors or vulnerable populations were included. Personally identifying metadata beyond facial and body appearance was not collected, and access to raw recordings will be restricted. The dataset will be released after peer review with a research-only license prohibiting redistribution and commercial use, and with clear take-down procedures if requested by participants.

We acknowledge that technologies enabling high-fidelity 3D avatar reconstruction from monocular videos may be misused for malicious purposes (e.g., impersonation, non-consensual content generation). To mitigate such risks, we emphasize responsible use of the dataset and models, encourage watermarking or detection mechanisms for synthetic outputs, and restrict the release of model weights to verified research purposes only.

We also recognize the possibility of bias due to the limited diversity of clothing types, subjects, and motions in LoCo-Human. We report dataset composition and limitations transparently and encourage future work to expand demographic and cultural coverage for fairness and inclusivity.

No sensitive medical or financial information is used in this work. Institutional review board (IRB) approval was not required, but ethical considerations regarding informed consent, privacy, and responsible release were carefully followed.

Figure K: Qualitative Results of *00148* subjects on 4D-Dress dataset, compared to Qian et al. (2024b) with multiple motions across time axis.

Figure L: Qualitative Results of *00169* subjects on 4D-Dress dataset, compared to Moon et al. (2024) with multiple motions across time axis.

Figure M: Qualitative Results of *00170* subjects on 4D-Dress dataset, compared to Lei et al. (2024) with multiple motions across time axis.

Figure N: Qualitative Results of *00185* subjects on 4D-Dress dataset, compared to Moon et al. (2024) with multiple motions across time axis.

Figure O: Qualitative Results of *00187* subjects on 4D-Dress dataset, compared to Hu et al. (2024a) with multiple motions across time axis.

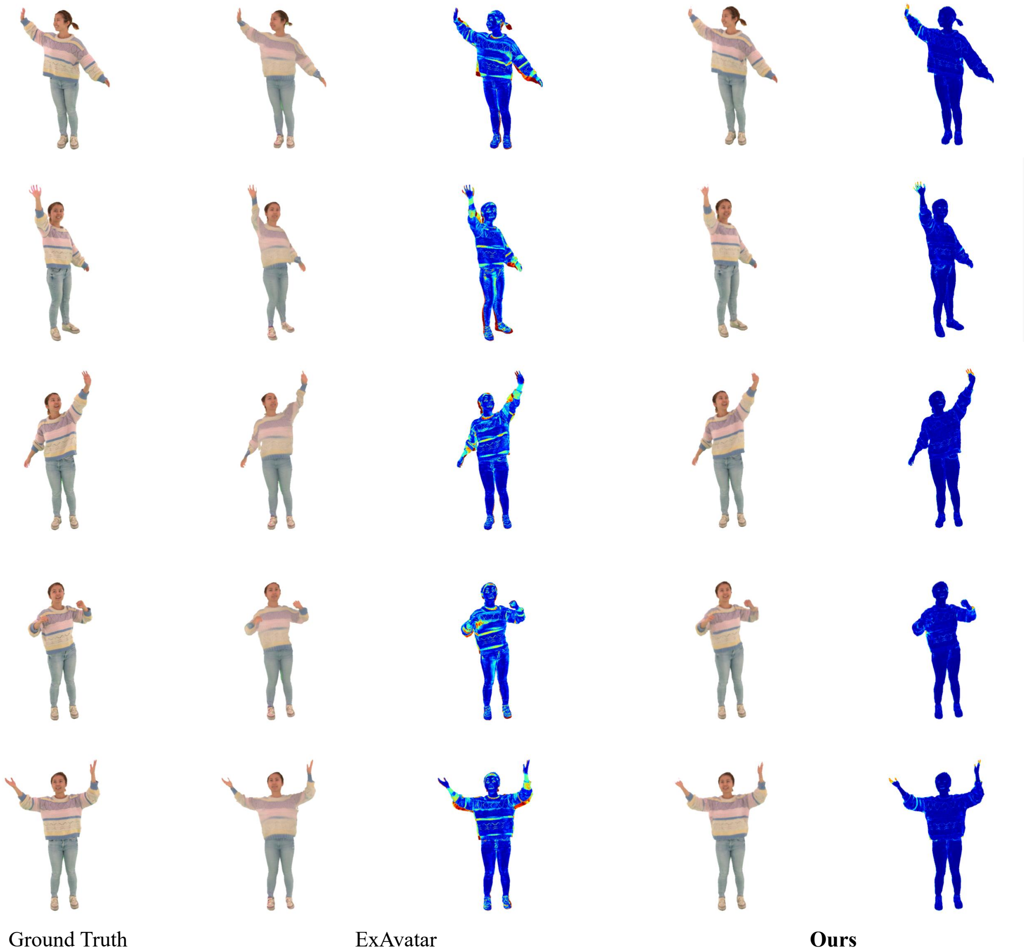


Figure P: Qualitative Results of *00190* subjects on 4D-Dress dataset, compared to Moon et al. (2024) with multiple motions across time axis.