MonoCloth: Reconstruction and Animation of Cloth-Decoupled Human Avatars from Monocular Videos

Daisheng Jin, Ying He*

S-Lab, Nanyang Technological University, Singapore
daisheng001@e.ntu.edu.sg, YHe@ntu.edu.sg

Abstract

Reconstructing realistic 3D human avatars from monocular videos is a challenging task due to the limited geometric information and complex non-rigid motion involved. We present **MonoCloth**, a new method for reconstructing and animating clothed human avatars from monocular videos. To overcome the limitations of monocular input, we introduce a part-based decomposition strategy that separates the avatar into body, face, hands, and clothing. This design reflects the varying levels of reconstruction difficulty and deformation complexity across these components. Specifically, we focus on detailed geometry recovery for the face and hands. For clothing, we propose a dedicated cloth simulation module that captures garment deformation using temporal motion cues and geometric constraints. Experimental results demonstrate that MonoCloth improves both visual reconstruction quality and animation realism compared to existing methods. Furthermore, thanks to its part-based design, MonoCloth also supports additional tasks such as clothing transfer, underscoring its versatility and practical utility.

Project — https://kingjg.github.io/MonoCloth.github.io

Introduction

3D human avatars are used in many areas, including film production, game development, and AR/VR applications. However, creating high-quality avatars often requires manual modeling or expensive 3D capture setups, which limits their use in everyday settings.

To make avatar creation more accessible, recent work has focused on using low-cost inputs such as images and videos. The development of implicit neural representations has made it possible to reconstruct both 3D geometry and appearance from 2D images (Mildenhall et al. 2021; Jin et al. 2025). When applied to multi-view videos, these methods enable the reconstruction of high-quality deformable 3D human avatars. For example, Animatable NeRF (Peng et al. 2021) and AvatarRex (Zheng et al. 2023) build implicit fields in a canonical space and animate them using deformation functions. Although effective, these methods require synchronized multi-view capture, which is difficult to set up outside of controlled environments.

*Corresponding author

To reduce input cost, some recent methods focus on reconstructing avatars from monocular videos. HumanNeRF (Weng et al. 2022) maps frames with different poses to a unified T-pose space and learns an implicit neural field for rendering from new viewpoints. InstantAvatar (Jiang et al. 2023) improves training and rendering speed by using Instant-NGP (Müller et al. 2022), an efficient NeRF backbone. Instead of implicit fields, 3DGS-Avatar (Qian et al. 2024) represents avatars using 3D Gaussian primitives (Kerbl et al. 2023), which improves both efficiency and visual quality. ExAvatar (Moon, Shiratori, and Saito 2024) adds detailed modeling of facial expressions and hand articulations to enrich avatar expressiveness. Vid2Avatar-Pro (Guo et al. 2025) uses pretraining on a large multi-view video dataset to help the model learn useful priors for reconstruction from monocular input.

While these methods focus on detailed appearance modeling, they often overlook how the avatar moves over time. As a result, animations may show inconsistencies between frames or unnatural changes in appearance. This happens because the model does not fully capture how shape and appearance evolve across time, making it harder to maintain a stable avatar and recover fine details.

To address these challenges, we present **MonoCloth**, a method designed to better model human motion from monocular videos, leading to more consistent and accurate reconstructions. Building upon the parametric body model SMPL-X (Pavlakos et al. 2019), MonoCloth decomposes the human avatar into two major components: the body and the clothing. The body moves in mostly rigid, joint-driven ways, while clothing undergoes soft, non-rigid motion influenced by physical effects such as inertia and gravity.

We optimize these two parts separately to account for their different behaviors. For the body, where motion is more rigid, we improve reconstruction using strong priors from face and hand parametric models. For clothing, we introduce a cloth simulation module (**CloSim**), which models how garments move during motion. To train CloSim, we first use 2D clothing segmentation to decouple the garment region. Instead of using only a single pose frame as input, we feed in a sequence of poses to capture motion over time. Clothing and motion features are passed through a graph convolutional network (GCN) (Kipf 2016) to collect local spatial information, followed by a lightweight gated recurrent unit

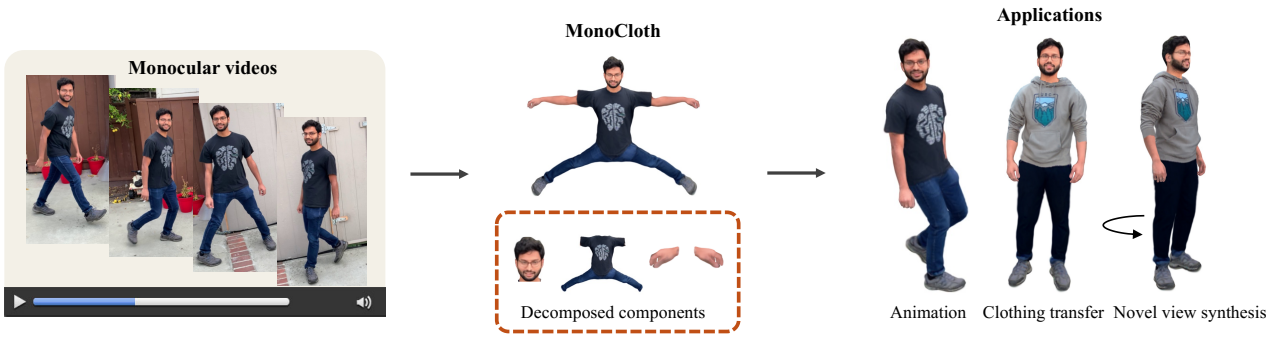

Figure 1: **MonoCloth** reconstructs a human avatar from monocular videos by dividing it into separate components. Each component is optimized using a strategy suited to its geometric and motion characteristics, which improves reconstruction quality. The reconstructed avatar supports natural animation and can be rendered from novel viewpoints. The modular design also allows for part-level editing, such as clothing transfer.

(GRU) (Chung et al. 2014) to learn how the motion evolves across frames. This setup helps the model learn both spatial and temporal effects in garment deformation.

Because modeling clothing dynamics from monocular input is difficult, we use additional information to guide the model. Specifically, we apply vision foundation models to predict depth and surface normals from the video frames. These serve as rough 3D cues during training. We also pretrain on monocular videos of different people to learn general patterns in appearance and clothing motion. With this design, MonoCloth reconstructs high-quality avatars from monocular videos and supports realistic animation under new motion sequences.

Our main contributions are summarized as follows:

- We propose a part-based avatar design built on SMPL-X and 3D Gaussians, allowing targeted optimization for each component.
- Based on this design, we introduce a cloth simulation module that integrates spatial and temporal cues, aiming to produce more physically consistent clothing motion.
- Extensive experiments demonstrate that MonoCloth improves both reconstruction quality and animation consistency. The part-based design also supports additional tasks such as clothing transfer.

Related Work

3D Human Avatar

Recent research has explored the reconstruction of 3D human avatars from a variety of input sources, including multi-view videos (Lin et al. 2024; Zhang et al. 2025), monocular videos (Song et al. 2025), and single images (Pang et al. 2025; Qiu et al. 2025).

Multi-view video-based methods typically rely on controlled laboratory environments with ideal lighting, synchronized camera setups, and known calibration parameters, providing rich and accurate 3D information. With the advancement of multi-view 3D reconstruction techniques, methods such as AvatarRex (Zheng et al. 2023) and AnimatableGaussian (Li et al. 2024) have adopted Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) and 3D Gaussian splatting (Kerbl et al. 2023), respectively, to reconstruct detailed human avatars. Building on these foundations, they further learn dynamic motion patterns from video sequences, enabling the creation of fully animatable 3D avatars. In contrast, single-image reconstruction offers a more accessible and cost-effective solution but presents significant challenges in recovering fine-grained geometry and maintaining 3D consistency (Zhang et al. 2024; Hu 2024).

By contrast, in-the-wild monocular videos are convenient to obtain and naturally contain more 3D cues and motion information. However, as the input is inherently 2D, accurately inferring 3D structures remains challenging. GaussianAvatar (Hu et al. 2024) and HUGS (Kocabas et al. 2024) address this by representing avatars using 3D Gaussians and introducing learnable offsets to model appearance variations associated with human motion. ExAvatar (Moon, Shiratori, and Saito 2024) tackles the common limitation of prior works that often neglect facial expressions and hand movements by explicitly modeling these regions, resulting in significantly enhanced avatar realism. Vid2Avatar-Pro (Guo et al. 2025) and PGHM (Peng et al. 2025) enhance 3D perception and robustness by pretraining on multi-view video data to compensate for the limited 3D cues in monocular inputs. Nevertheless, most monocular video-based methods treat the outfit and human body as a unified whole during optimization, overlooking their distinct motion patterns. In contrast, our method explicitly models garment dynamics through a dedicated CloSim module that captures clothing motion across both spatial and temporal dimensions, enabling more accurate and detailed avatar reconstructions.

Clothing Reconstruction and Motion Simulation

Clothing reconstruction is essential for clothed human avatar modeling, as garments often cover most of the body. However, the complex, non-rigid deformations of clothing during motion lead to significant challenges in modeling its geometry and appearance. To address this, many studies focus on garment reconstruction. Garment4D (Hong et al. 2021) recovers fine cloth geometry directly from 3D point

Method	Inputs		Design		Outputs	
	Vid.	Pre.	Temp.	Decomp.	Anim.	Trans.
AvatarRex	Multi	None	✗	✗	✓	✗
PhysAvatar	Multi	None	✓	✓	✓	✓
D^3-Human	Mono	None	✓	✓	✗	✓
ExAvatar	Mono	None	✗	✗	✓	✗
Vid2Avatar-Pro	Mono	Multi	✗	✗	✓	✗
Ours	Mono	Mono	✓	✓	✓	✓

Table 1: **Comparison of video-based human avatar reconstruction methods.** We compare each method based on input type (Vid.), use of pretraining (Pre.), temporal modeling (Temp.), clothing decomposition (Decomp.), animation support (Anim.), and ability of clothing transfer (Trans.).

clouds, leveraging their rich 3D information. GaussianGarments (Rong et al. 2024) uses multi-view videos and 3D Gaussians to jointly reconstruct garment geometry and appearance. PhysAvatar (Zheng et al. 2024) extends beyond garments to reconstruct dressed humans with high-fidelity geometry and appearance, incorporating physical fabric parameters to produce more realistic cloth dynamics.

In contrast, monocular videos provide limited 3D information, making accurate reconstruction significantly more challenging. Typically, recent works aim to infer human geometry directly from 2D input. Reloo (Guo et al. 2024) models loose clothing with virtual bone deformation for high-quality monocular normal prediction. D^3-Human (Chen et al. 2025) reconstructs full-body garment geometry from monocular input and supports dynamic cloth simulation during animation. While these methods mainly focus on geometric reconstruction, achieving photorealistic rendering still requires deeper integration of appearance modeling.

In this work, we propose a method for reconstructing high-quality, animatable avatars from low-cost monocular videos. By explicitly decoupling body components and introducing a cloth-specific motion model, our approach optimizes Gaussian avatars to directly render dynamic human animations with fine geometric and appearance details.

Method

Decomposed Clothed Human Avatar

Different parts of the human avatar present distinct challenges for reconstruction and animation, such as modeling subtle facial expressions, fine-grained hand articulations, and the complex deformations of flexible garments. To effectively address these challenges, we decompose the human avatar into functionally distinct components and apply dedicated optimization strategies to each. Inspired by (Zheng et al. 2023) and (Moon, Shiratori, and Saito 2024), we use FLAME (Li et al. 2017) to reconstruct facial geometry and adopt joint-based optimization to capture detailed hand articulations, enabling more accurate and robust reconstruction of these critical regions.

On the other hand, clothing poses a significant challenge for modeling due to its diverse styles, soft materials, and non-rigid deformation patterns, which lead to substantial variations in geometry and appearance. These effects are difficult to capture using parametric models and rigid articulations. To enable more realistic reconstruction, we decouple clothing from the body and model it independently. Given the limited 3D cues in monocular videos, we adopt SMPL-X (Pavlakos et al. 2019) as a coarse 3D prior. We place one 3D Gaussian at each upsampled SMPL-X mesh vertex and assign each to a semantic component. Formally, the complete set of Gaussians $\mathcal{G}$ is decomposed as:

$$\mathcal{G} = \mathcal{G}^{\text{face}} \cup \mathcal{G}^{\text{hands}} \cup \mathcal{G}^{\text{cloth}} \cup \mathcal{G}^{\text{body}}, \tag{1}$$

where $\mathcal{G}^{\text{face}}$, $\mathcal{G}^{\text{hands}}$, and $\mathcal{G}^{\text{cloth}}$ denote the Gaussians associated with the face, hands, and clothing, respectively. The remaining Gaussians are assigned to $\mathcal{G}^{\text{body}}$.

Specifically, $\mathcal{G}^{\text{face}}$ and $\mathcal{G}^{\text{hands}}$ are obtained through vertex correspondences between SMPL-X and the FLAME/-MANO (Romero, Tzionas, and Black 2022) models, while $\mathcal{G}^{\text{cloth}}$ is derived by 2D clothing segmentations. These 2D segmentations serve as pseudo-ground-truth labels to train a lightweight classifier that assigns semantic labels to Gaussian points. To address errors caused by occlusions and limited viewpoints in monocular videos, we introduce a connectivity-based refinement step that corrects misclassifications and enforces spatial consistency. Please see the supplementary material for more details.

MonoCloth Pipeline

Two-stage strategy. To address the challenge of learning complex non-rigid motion dynamics from short monocular video clips, we adopt a two-stage strategy that leverages a broader collection of monocular videos to capture shared patterns of human appearance and cloth motion. To enable identity-specific representation, each subject is assigned a compact latent code $\mathbf{z}_i \in \mathbb{R}^{64}$.

In Stage 1, we train on monocular videos of multiple subjects to jointly reconstruct avatars of each individual. The only identity-specific component is the latent code, enabling the shared network to learn generalized priors for human geometry, appearance, and motion. Leveraging these learned priors, we then fine-tune the pretrained model on a single subject in Stage 2 to further enhance reconstruction fidelity.

Static reconstruction. In both training stages, avatar reconstruction begins by decoding the latent code into a triplane feature representation. Specifically, each identity i is associated with a latent code, which is decoded by a shared decoder $\mathcal{D}_\phi$ parameterized by ϕ, as follows:

$$\mathcal{T}_i = \mathcal{D}_\phi(\mathbf{z}_i) = \{\mathcal{T}_i^x,\ \mathcal{T}_i^y,\ \mathcal{T}_i^z\}, \tag{2}$$

where each $\mathcal{T}_i^* \in \mathbb{R}^{C\times H\times W}$ is a 2D feature map ($C = 32$, $H = 128$, $W = 128$), and the set $\mathcal{T}_i$ forms a triplane encoding both geometry and appearance information.

For feature extraction, inspired by (Kocabas et al. 2024) and (Moon, Shiratori, and Saito 2024), we use an upsampled SMPL-X mesh as the geometric base. To eliminate shape variation across subjects, we set all identity-dependent SMPL-X shape parameters to zero, resulting in a standard reference mesh in the canonical space, denoted as $\mathbf{v}^{\text{cano}}$. Given the vertices of $\mathbf{v}^{\text{cano}}$, we extract per-vertex features $\text{feat}^{\text{cano}}$ by bilinearly sampling across the three orthogonal

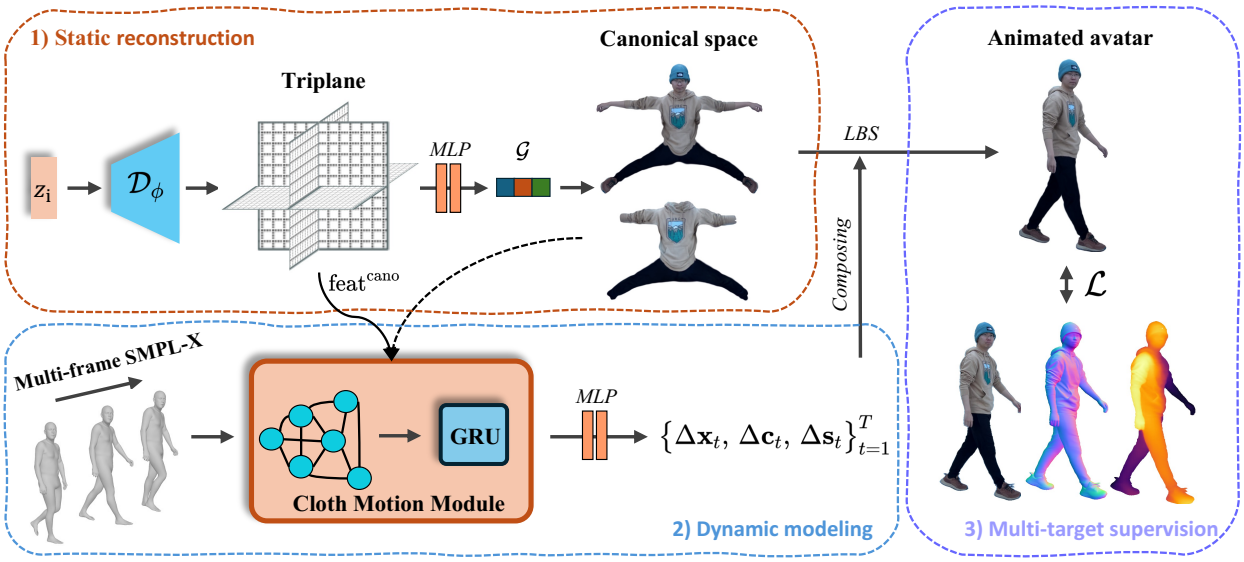

Figure 2: **MonoCloth pipeline. 1)** We first reconstruct the static geometry and appearance in the canonical pose, where Gaussian attributes are computed and decomposed into different components. **2)** Combining static avatar features with multi-frame SMPL-X parameters, we incorporate both spatial and temporal information to predict motion-dependent offsets that enrich avatar details. **3)** The reconstructed avatar is supervised using ground-truth RGB images, normal maps, depth maps, and more auxiliary targets to jointly optimize appearance and geometry.

2D planes. For region-specific features such as $\text{feat}^{\text{cloth}}$, we select the corresponding vertices based on the Gaussian set $\mathcal{G}^*$ of each component.

The extracted features are passed through geometry and appearance decoder MLPs to predict per-vertex Gaussian attributes, including the spatial displacement $\Delta\mathbf{x}$ from the subject-specific SMPL-X mesh $\mathbf{v}^{\text{id}}$, base color $\mathbf{c}$, and scale $\mathbf{s}$. The identity-dependent mesh in the canonical pose is then computed as $\mathbf{v}^{\text{id}} = \mathbf{v}^{\text{cano}} + \Delta\mathbf{x}$. Following prior works (Hu et al. 2024), we assume isotropic Gaussians with fixed opacity of 1 to mitigate the limited information inherent in monocular videos. The final Gaussian set $\mathcal{G} = \{\mathbf{x}, \mathbf{c}, \mathbf{s}\}$ is rendered via Gaussian Splatting to reconstruct the avatar's static geometry and appearance in canonical space.

Dynamic modeling. To simulate motion-dependent variations, we model the dynamics of different body regions independently. For exposed body parts such as arms and lower legs, which undergo relatively minor non-rigid deformation, we use a pose-dependent module to predict small displacement offsets. In contrast, clothing exhibits complex, non-rigid deformations driven by flexible material properties and motion history. To better capture these dynamics, we introduce the cloth motion simulation module (**CloSim**) that integrates both spatial and temporal information.

For the input to Closim, we combine the clothing feature $\text{feat}^{\text{cano}}$ from the canonical space with a sequence of SMPL-X poses from multiple frames, enabling the input to capture both geometric context and temporal dynamics. To ensure sufficient motion variation, frames are sampled at a low frequency (i.e., 5 FPS), with one frame before and one after the current time step included to provide bidirectional temporal context. MonoCloth adopts a temporal window as:

$$\mathcal{P}_T = \{\boldsymbol{\theta}_{T-\Delta t},\ \boldsymbol{\theta}_T,\ \boldsymbol{\theta}_{T+\Delta t}\}, \tag{3}$$

where $\boldsymbol{\theta}$ denotes the SMPL-X pose parameters, and Δt represents the temporal offset, set to 0.2 seconds.

Through CloSim, we obtain per-time-step offsets that encode motion dynamics relative to the canonical avatar.

Multi-target supervision. After obtaining the motion offsets, we adopt a deformation strategy that first applies the offsets to the canonical pose and then performs Linear Blend Skinning (LBS). This design ensures that the predicted offsets are defined independently of articulation, allowing the model to learn generalizable motion patterns across individuals without being affected by the transformations introduced by LBS. The deformation process is defined as:

$$\mathbf{v}_t = \text{LBS}(\mathbf{v}^{\text{id}} + \Delta\mathbf{x}_t,\ \boldsymbol{\theta}_t), \tag{4}$$

where $\mathbf{v}^{\text{id}}$ is the mesh in the canonical pose and $\Delta\mathbf{x}_t$ denotes the predicted per-vertex offset at time t.

To enhance the model's 3D perception, we compute per-vertex normals $\mathbf{n}_t$ and depth values $\mathbf{d}_t$ from the deformed mesh $\mathbf{v}_t$. These quantities are then rendered into 2D normal and depth maps, denoted as $\hat{\mathbf{N}}_t$ and $\hat{\mathbf{D}}_t$, using Gaussian Splatting. The rendered maps are supervised by pseudo ground-truth $\mathbf{N}_t^{\text{gt}}$ and $\mathbf{D}_t^{\text{gt}}$, generated by the vision foundation model Sapiens (Khirodkar et al. 2024). This additional supervision improves the geometric accuracy of the reconstructed avatar and enhances the fidelity of motion-dependent surface details.

Cloth Simulation Module

Clothing is particularly challenging to model in human motion due to its flexible and loose nature, resulting in com-

plex, non-rigid variations in geometry and appearance across frames. To address this, we propose CloSim to capture detailed garment dynamics. Leveraging avatar decomposition, we first decouple the clothing regions from the full-body avatar. Then, to capture the rich geometric and appearance details of garments, we employ a dedicated feature extraction branch, which enables fine-grained representations for subsequent reconstruction and refinement.

Spatio-temporal modeling. To capture cloth dynamics during avatar motion, we model overall deformation as the collective behavior of clothing-related Gaussian points $\mathcal{G}^{\text{cloth}}$. Due to the spatial continuity of cloth geometry, the motion of each point is influenced by its neighbors, making it essential to model their interactions. To this end, we employ a Graph Convolutional Network (GCN) (Kipf 2016) that enables information propagation across connected points, allowing each point to incorporate local deformation context and better capture collective motion patterns:

$$Z_t = \mathrm{GCN}\big(\mathrm{Concat}(\,\mathrm{feat}^{\text{cano}}, \mathrm{feat}_t^{\text{pose}}),\, \mathcal{E}\big), \tag{5}$$

where $Z^t \in \mathbb{R}^{N\times d}$ denotes the encoded features for all Gaussians, each with dimension $d = 128$. The GCN processes the concatenation of the canonical-space features $\mathrm{feat}^{\text{cano}} \in \mathbb{R}^{N\times 96}$ and body pose feature $\mathrm{feat}_t^{\text{pose}} \in \mathbb{R}^{N\times 126}$, using the mesh connectivity $\mathcal{E}$ to guide message passing.

Monocular avatar reconstruction methods typically rely solely on body pose to drive deformation. However, this is insufficient for modeling the complex dynamics of garments, which are influenced not only by pose but also by inertia, gravity, and other physical factors. Capturing temporal continuity is therefore essential. To model these sequential dependencies, we integrate a Gated Recurrent Unit (GRU) (Chung et al. 2014), enabling the network to track motion trajectories and evolving garment states over time:

$$\big\{\Delta\mathbf{x}_t,\, \Delta\mathbf{c}_t,\, \Delta\mathbf{s}_t\big\}_{t=1}^{T} = \Psi\big(\mathrm{GRU}(\{Z_t\}_{t=1}^{T},\, h_0)\big)\,, \tag{6}$$

where $\Delta\mathbf{x}_t \in \mathbb{R}^3$, $\Delta\mathbf{c}_t \in \mathbb{R}^3$, and $\Delta\mathbf{s}_t \in \mathbb{R}$ represent the predicted residuals for position, color, and scale of each Gaussian point at time t, respectively. Given the sequence of encoded features $\{Z_t\}_{t=1}^{T}$ and an initial hidden state h_0, the GRU generates temporally-aware embeddings, which are subsequently decoded by a lightweight MLP Ψ. The output lies in $\mathbb{R}^{T\times N\times 7}$, representing per-frame offsets for all N points across T frames.

Sampling strategy. To improve the model's robustness to temporal variation, we adopt a data augmentation strategy based on random supervision during training. Specifically, while CloSim consistently receives inputs at $\mathcal{P}_T$ and outputs motion offsets $\{\Delta\mathbf{x}_t,\, \Delta\mathbf{c}_t,\, \Delta\mathbf{s}_t\}$ for the frames $\{T - \Delta t,\ T,\ T + \Delta t\}$, supervision is applied only at randomly sampled time points $t \in [T - \Delta t,\ T + \Delta t]$, with the corresponding offset computed via linear interpolation as:

$$\Delta\hat{\mathbf{x}}_t = (1-\alpha)\Delta\mathbf{x}_{T-\Delta t} + \alpha\Delta\mathbf{x}_{T+\Delta t}, \tag{7}$$

where α represents the normalized position of t within the interval. The offsets $\Delta\hat{\mathbf{c}}_t$ and $\Delta\hat{\mathbf{s}}_t$ are computed similarly.

Loss Functions

Rendering loss. To supervise the rendered images I_{pred} against the ground truth I_{gt}, we employ a combination of L1 loss $\mathcal{L}_{\text{rgb}}$, structural similarity loss $\mathcal{L}_{\text{ssim}}$, and perceptual loss $\mathcal{L}_{\text{lpips}}$, encouraging both pixel-level accuracy and perceptual fidelity. To further enhance appearance reconstruction in clothing regions, we introduce an additional loss term $\mathcal{L}_{\text{cloth}}$, computed between the rendered 3D clothing and the masked 2D clothing image, to better capture fine-grained garment details.

Geometry loss. To capture the fine-grained dynamics of cloth motion, we introduce explicit supervision on 3D geometry. Given the deformed mesh, we compute vertex normals and encode them as RGB values for the corresponding Gaussian points. The rendered normal map $\mathbf{N}_{\text{pred}}$ is compared with the reference map $\mathbf{N}_{\text{gt}}$ from Sapiens (Khirodkar et al. 2024) using a cosine similarity loss: $\mathcal{L}_{\mathbf{N}} = 1 - \langle \mathbf{N}_{\text{pred}}, \mathbf{N}_{\text{gt}} \rangle$. Similarly, the depth map $\mathbf{D}_{\text{pred}}$ is supervised using an L1 loss against the ground truth $\mathbf{D}_{\text{gt}}$: $\mathcal{L}_{\mathbf{D}} = \|\mathbf{D}_{\text{pred}} - \mathbf{D}_{\text{gt}}\|_1$. To improve boundary alignment, we also apply a silhouette loss between predicted and ground-truth masks: $\mathcal{L}_{\mathbf{S}} = \|\mathbf{S}_{\text{pred}} - \mathbf{S}_{\text{gt}}\|_2^2$. The final geometry loss is defined as:

$$\mathcal{L}_{\text{geo}} = \lambda_{\mathbf{N}}\mathcal{L}_{\mathbf{N}} + \lambda_{\mathbf{D}}\mathcal{L}_{\mathbf{D}} + \lambda_{\mathbf{S}}\mathcal{L}_{\mathbf{S}}, \tag{8}$$

where the weights are set to $\lambda_{\mathbf{N}} = 5$, $\lambda_{\mathbf{D}} = 1$, and $\lambda_{\mathbf{S}} = 2$.

Temporal loss. To enforce temporal coherence across frames, we introduce a temporal consistency loss that penalizes temporal discontinuities in geometry, color, and scale offsets between consecutive frames. Specifically, let $\Delta\mathbf{x}_t$, $\Delta\mathbf{c}_t$, and $\Delta\mathbf{s}_t$ denote the predicted offsets at frame t. The temporal loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{temp}} = \lambda_{\text{temp}} \sum_{t=1}^{T-1} \big(&\|\Delta\mathbf{x}_{t+1} - \Delta\mathbf{x}_t\|_2^2 \\ &+ \|\Delta\mathbf{c}_{t+1} - \Delta\mathbf{c}_t\|_2^2 + \|\Delta\mathbf{s}_{t+1} - \Delta\mathbf{s}_t\|_2^2\big),\end{aligned} \tag{9}$$

where $\lambda_{\text{temp}} = 0.1$ controls the strength of temporal regularization. This encourages smooth transitions over time and reduces artifacts in cloth animation.

Final loss. In addition to the core loss terms described above, we incorporate several auxiliary losses to improve training stability and reconstruction accuracy. Inspired by (Moon, Shiratori, and Saito 2024), we introduce regularization terms such as L2-norm penalties on the predicted offsets to stabilize geometry optimization. For facial and hand regions, $\mathcal{G}^{\text{face}}$ and $\mathcal{G}^{\text{hands}}$, we further apply vertex-level supervision based on parametric models, directly constraining the 3D positions of the associated Gaussian points. This enhances the robustness and precision of geometry reconstruction in these high-detail areas. All loss components are combined to form the final training objective $\mathcal{L}$. Please refer to the supplementary material for more details.

Experiments

Experimental Setups

Datasets and metrics. We conduct training and evaluation on two monocular human video datasets: NeuMan (Jiang

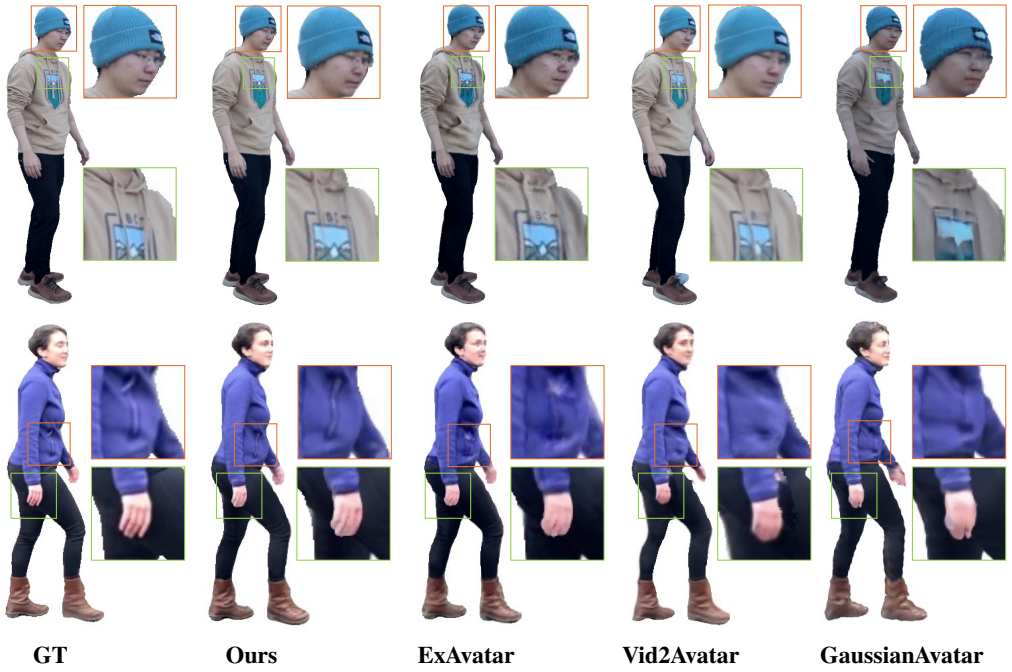

Figure 3: **Qualitative results on NeuMan.** Our method achieves the highest overall visual quality, particularly in reconstructing fine clothing textures as well as facial and hand details.

et al. 2022) and X-Humans (Shen et al. 2023). The NeuMan dataset contains in-the-wild motion videos of six individuals, with frames down-sampled to increase pose variation between adjacent frames. Given its low frame rate and non-consecutive testing frames, NeuMan serves as a benchmark for evaluating avatar reconstruction and generalization to novel poses. In contrast, X-Humans provides a larger collection of high-quality monocular videos captured in controlled laboratory settings. It contains motion data from 20 subjects, along with corresponding RGB-D videos, 3D meshes, and SMPL parameters. In our experiments, we only use the RGB videos of 10 subjects as input to evaluate the model's performance on temporally continuous video sequences.

We evaluate the rendering quality of all methods using PSNR, SSIM, and LPIPS (Zhang et al. 2018) metrics. Since the NeuMan dataset contains outdoor backgrounds, we apply a mask to replace the background with a white canvas before the evaluation. This ensures that the evaluation focuses solely on the quality of the avatar reconstruction.

Implementation details. We use only five subjects from the X-Humans dataset with large appearance and motion variations for Stage 1 training, aiming to assess the potential of pretraining to enhance model performance. The model is optimized using the Adam optimizer with an initial learning rate of 10^{-3}. On a single NVIDIA RTX 4090 GPU, our method takes approximately 5 minutes to complete 3D clothing segmentation and around 3.5 hours to reconstruct a single subject from the NeuMan dataset on average. We provide more details in the supplementary material.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
HumanNeRF	27.06	0.967	1.92
InstantAvatar	28.47	0.972	2.77
Vid2Avatar	29.48	0.976	1.85
3DGS-Avatar	29.75	0.975	1.75
GaussianAvatar	28.90	0.974	1.81
ExAvatar	31.70	0.982	1.47
*Vid2Avatar-Pro	32.71	0.983	**1.19**
Ours (w/o pretrain)	<u>33.18</u>	<u>0.985</u>	1.28
Ours (full)	**33.53**	**0.986**	<u>1.20</u>

Table 2: **Quantitative results on NeuMan.** LPIPS is measured on the scale of 10^{-2}. The best and second-best results are highlighted in boldface and underlined, respectively.

Evaluation and Comparison

We first evaluate the quality of avatar reconstruction on the NeuMan dataset. Following previous works (Hu et al. 2024), we use four high-quality sequences (e.g., Seattle, Bike, Citron, and Jogging) for quantitative evaluation. Tab. 2 presents comparisons between our method and several state-of-the-art approaches. Some results are partially sourced from (Hu et al. 2024) and (Guo et al. 2025). While all methods reconstruct avatars from monocular videos, Vid2Avatar-Pro (Guo et al. 2025) leverages pretraining on a large-scale multi-view dataset, providing stronger 3D priors at the cost of increased data collection complexity. To highlight this distinction, we mark it with a star in Tab. 2. To ensure a fair comparison, we report both the results of our full method and an ablated

version without any pretrained model. Benefiting from our part-based optimization design, our approach achieves state-of-the-art performance even without monocular video pretraining. Fig. 3 visualizes the reconstruction and animation results, showing that our method produces the highest overall quality, especially in capturing detailed clothing textures as well as facial and hand details.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
ExAvatar	<u>29.41</u>	<u>0.973</u>	<u>2.24</u>
Ours	**30.68**	**0.976**	**2.21**

Table 3: **Quantitative results on X-Humans.** Our method achieves superior performance in video animation quality.

Tab. 3 presents the evaluation results on the X-Humans dataset, where we compare our method against ExAvatar (Moon, Shiratori, and Saito 2024), which is the strongest baseline on the NeuMan dataset. Our method consistently outperforms ExAvatar across all metrics. Since both the training and testing sequences in X-Humans are temporally continuous, modeling clothing dynamics over time becomes especially critical. As shown in Fig. 4, ExAvatar suffers from noticeable artifacts when handling drastic motions, while our method demonstrates superior robustness and temporal stability.

Figure 4: **X-Humans comparison.** Temporal modeling improves clothing stability.

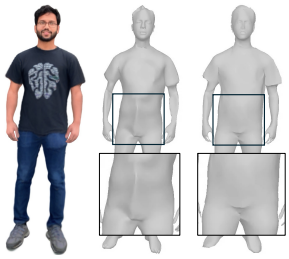

Figure 5: **Geometry loss ablation.** Geometry supervision reduces 3D artifacts.

Ablation Studies

We conduct ablation studies on several key components of our method to assess their individual contributions. The results are summarized in Tab. 4, where we evaluate the impact of CloSim, temporal modeling, and geometry supervision. Removing motion offsets Δ_* from CloSim and relying solely on LBS in canonical space leads to the loss of motion-dependent details and a significant drop in reconstruction quality. Without temporal sampling (i.e., multi-frame sampling strategy), the model cannot effectively capture temporal context, resulting in degraded rendering quality and animation smoothness. Fig. 5 illustrates the effect of the geometry loss. Without geometry supervision, the inherent lack of 3D information in monocular video can lead to inaccurate estimation of clothing geometry. While the model may overfit the training views to produce visually plausible results, it often results in distorted or implausible geometry when rendered from novel viewpoints.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
w/o Δ_*	31.93	0.983	1.40
w/o temporal	<u>33.00</u>	<u>0.985</u>	<u>1.29</u>
w/o $\mathcal{L}_{\text{geo}}$	32.55	0.985	1.30
Ours	**33.53**	**0.986**	**1.20**

Table 4: **Ablation studies**. Evaluation on the impact of motion offsets Δ_*, temporal sampling, and geometry loss $\mathcal{L}_{\text{geo}}$.

Applications

Our part-based decomposition and targeted optimization not only enhance reconstruction quality but also offer greater flexibility for downstream applications such as avatar editing. As illustrated in Fig. 6, we showcase a virtual try-on example by transferring garments from the subject "Bike" to "Citron" in the NeuMan dataset. By adjusting the SMPL-X parameters and refining the garment-body attachment, our method enables effective clothing transfer across different subjects. Moreover, since both the body and clothing avatars in our method are animatable, the virtual try-on results can be rendered as videos, providing a compelling visual experience for dynamic clothing transfer.

Figure 6: **Clothing transfer.** Avatars with new garments support animation and novel-view synthesis as the original.

Conclusion

We propose a novel framework for reconstructing clothed human avatars from monocular videos. To handle the varying reconstruction difficulty and motion characteristics across body parts, we decompose the avatar into components and apply targeted optimization strategies. For the clothing region, which involves complex non-rigid motion, we introduce the cloth simulation module that incorporates spatio-temporal information and geometric supervision to enable more physically plausible garment dynamics. The resulting avatars support high-quality animation, novel-view synthesis, and applications such as clothing transfer, demonstrating the flexibility and generality of our approach.

Acknowledgments

This work was supported in part by the Ministry of Education, Singapore, under its Academic Research Fund Grant (RT19/22), as well as cash and in-kind funding from NTU S-Lab and the industry partner(s).

References

Chen, H.; Peng, B.; Tao, Y.; and Zhang, J. 2025. Dˆ 3-Human: Dynamic Disentangled Digital Human from Monocular Video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10836–10846.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Guo, C.; Jiang, T.; Kaufmann, M.; Zheng, C.; Valentin, J.; Song, J.; and Hilliges, O. 2024. Reloo: Reconstructing humans dressed in loose garments from monocular video in the wild. In *European Conference on Computer Vision*, 21–38. Springer.

Guo, C.; Li, J.; Kant, Y.; Sheikh, Y.; Saito, S.; and Cao, C. 2025. Vid2avatar-pro: Authentic avatar from videos in the wild via universal prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5559–5570.

Hong, F.; Pan, L.; Cai, Z.; and Liu, Z. 2021. Garment4d: Garment reconstruction from point cloud sequences. *Advances in Neural Information Processing Systems*, 34: 27940–27951.

Hu, L. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8153–8163.

Hu, L.; Zhang, H.; Zhang, Y.; Zhou, B.; Liu, B.; Zhang, S.; and Nie, L. 2024. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 634–644.

Jiang, T.; Chen, X.; Song, J.; and Hilliges, O. 2023. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16922–16932.

Jiang, W.; Yi, K. M.; Samei, G.; Tuzel, O.; and Ranjan, A. 2022. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, 402–418. Springer.

Jin, D.; Hu, J.; Xu, B.; Dai, Y.; Qian, C.; and He, Y. 2025. SFDM: Robust Decomposition of Geometry and Reflectance for Realistic Face Rendering from Sparse-view Images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 26409–26419.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Khirodkar, R.; Bagautdinov, T.; Martinez, J.; Zhaoen, S.; James, A.; Selednik, P.; Anderson, S.; and Saito, S. 2024. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, 206–228. Springer.

Kipf, T. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907*.

Kocabas, M.; Chang, J.-H. R.; Gabriel, J.; Tuzel, O.; and Ranjan, A. 2024. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 505–515.

Li, T.; Bolkart, T.; Black, M. J.; Li, H.; and Romero, J. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.*, 36(6): 194–1.

Li, Z.; Zheng, Z.; Wang, L.; and Liu, Y. 2024. 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*, 19711–19722.

Lin, S.; Li, Z.; Su, Z.; Zheng, Z.; Zhang, H.; and Liu, Y. 2024. Layga: Layered gaussian avatars for animatable clothing transfer. In *ACM SIGGRAPH 2024 Conference Papers*, 1–11.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Moon, G.; Shiratori, T.; and Saito, S. 2024. Expressive whole-body 3d gaussian avatar. In *European Conference on Computer Vision*, 19–35. Springer.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4): 1–15.

Pang, H. E.; Liu, S.; Cai, Z.; Yang, L.; Zhang, T.; and Liu, Z. 2025. Disco4d: Disentangled 4d human generation and animation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 26331–26344.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10975–10985.

Peng, C.; Sun, J.; Chen, Y.; Su, Z.; Su, Z.; and Liu, Y. 2025. Parametric Gaussian Human Model: Generalizable Prior for Efficient and Realistic Human Avatar Modeling. *arXiv preprint arXiv:2506.06645*.

Peng, S.; Dong, J.; Wang, Q.; Zhang, S.; Shuai, Q.; Zhou, X.; and Bao, H. 2021. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14314–14323.

Qian, Z.; Wang, S.; Mihajlovic, M.; Geiger, A.; and Tang, S. 2024. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5020–5030.

Qiu, L.; Gu, X.; Li, P.; Zuo, Q.; Shen, W.; Zhang, J.; Qiu, K.; Yuan, W.; Chen, G.; Dong, Z.; et al. 2025. Lhm: Large animatable human reconstruction model from a single image in seconds. *arXiv preprint arXiv:2503.10625*.

Romero, J.; Tzionas, D.; and Black, M. J. 2022. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*.

Rong, B.; Grigorev, A.; Wang, W.; Black, M. J.; Thomaszewski, B.; Tsalicoglou, C.; and Hilliges, O. 2024. Gaussian garments: Reconstructing simulation-ready clothing with photorealistic appearance from multi-view video. *arXiv preprint arXiv:2409.08189*.

Shen, K.; Guo, C.; Kaufmann, M.; Zarate, J. J.; Valentin, J.; Song, J.; and Hilliges, O. 2023. X-avatar: Expressive human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16911–16921.

Song, W.; Ding, Y.; Hou, F.; Li, S.; Hao, A.; and Hou, X. 2025. CtrlAvatar: Controllable Avatars Generation via Disentangled Invertible Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6959–6967.

Weng, C.-Y.; Curless, B.; Srinivasan, P. P.; Barron, J. T.; and Kemelmacher-Shlizerman, I. 2022. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, 16210–16220.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, W.; Wu, S.; Liao, M.; and Yan, Y. 2025. Disentangled clothed avatar generation with layered representation. *arXiv preprint arXiv:2501.04631*.

Zhang, Y.; Gu, J.; Wang, L.-W.; Wang, H.; Cheng, J.; Zhu, Y.; and Zou, F. 2024. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*.

Zheng, Y.; Zhao, Q.; Yang, G.; Yifan, W.; Xiang, D.; Dubost, F.; Lagun, D.; Beeler, T.; Tombari, F.; Guibas, L.; et al. 2024. Physavatar: Learning the physics of dressed 3d avatars from visual observations. In *European Conference on Computer Vision*, 262–284. Springer.

Zheng, Z.; Zhao, X.; Zhang, H.; Liu, B.; and Liu, Y. 2023. Avatarrex: Real-time expressive full-body avatars. *ACM Transactions on Graphics (TOG)*, 42(4): 1–19.