FastAnimate: Towards Learnable Template Construction and Pose Deformation for Fast 3D Human Avatar Animation

Jian Shu*, Nanjie Yao*, Gangjian Zhang, Junlong Ren, Yu Feng, Hao Wang†

The Hong Kong University of Science and Technology (Guangzhou)
jshu704@connect.hkust-gz.edu.cn, nanjieyao@gmail.com, gzhang292@connect.hkust-gz.edu.cn,
jren686@connect.hkust-gz.edu.cn, yufeng9819@gmail.com, haowang@hkust-gz.edu.cn

Abstract

3D human avatar animation aims at transforming a human avatar from an arbitrary initial pose to a specified target pose using deformation algorithms. Existing approaches typically divide this task into two stages: canonical template construction and target pose deformation. However, current template construction methods demand extensive skeletal rigging and often produce artifacts for specific poses. Moreover, target pose deformation suffers from structural distortions caused by Linear Blend Skinning (LBS), which significantly undermines animation realism. To address these problems, we propose a unified learning-based framework to address both challenges in two phases. For the former phase, to overcome the inefficiencies and artifacts during template construction, we leverage a U-Net architecture that decouples texture and pose information in a feed-forward process, enabling fast generation of a human template. For the latter phase, we propose a data-driven refinement technique that enhances structural integrity. Extensive experiments show that our model delivers consistent performance across diverse poses with an optimal balance between efficiency and quality, surpassing state-of-the-art (SOTA) methods.

Introduction

3D human avatar animation aims to deform an individual's 3D model into a specified pose. The main challenge in 3D human animation is balancing high realism with real-time performance, personalization and natural motion. Prevailing approaches generally split the animation process into two phases: canonical template construction and target pose deformation (McManus et al. 2011; Ichim, Bouaziz, and Pauly 2015; Li et al. 2019). The first stage focuses on constructing high-quality, customizable human templates based on the provided body data; while the second stage seeks to minimize structural errors resulting from pose changes and reduce texture artifacts.

Prior studies on canonical template construction can be categorized into two approaches: pretrained general template and data-driven personalized template construction. Pretrained models such as SCAPE (Anguelov et al. 2005) and Skinned Multi-Person Linear (SMPL) model (Loper et al. 2023) use predefined mesh structures and statistical models from 3D scan data to represent human shape and pose. In contrast to traditional methods, 3D Gaussian-based methods (Liu et al. 2024b; Kocabas et al. 2024; Moreau et al. 2024; Pang et al. 2024; Yuan et al. 2024) have emerged as a promising paradigm, representing human bodies as collections of anisotropic 3D Gaussians optimized for both rendering and animation. However, these require time-consuming data preprocessing for which is not generalizable and struggle to manage self-intersecting regions on human bodies (Kwon et al. 2024; Moon, Shiratori, and Saito 2024; Wen, Schwing, and Wang 2025).

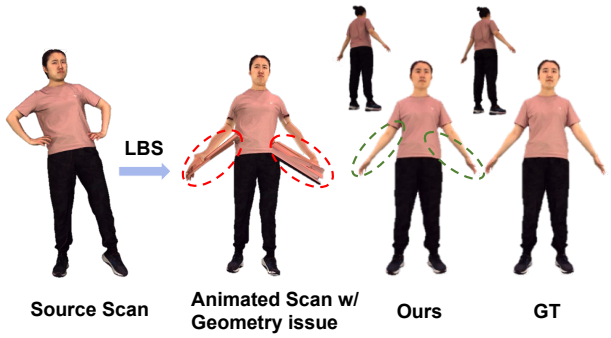

Figure 1: Visualization of the stretch-geometric problem under deformation (*Left*) and our result (*Right*).

In terms of the target pose deformation phase, it greatly determines the quality of pose transformation. Linear Blend Skinning (LBS) allows vertices to be influenced by multiple skeletons through weighted averaging, offering a balance between efficiency and flexibility. However, it suffers from volume loss and limited capacity for non-linear deformations (James and Twigg 2005; Yang, Somasekharan, and Zhang 2006). To address these limitations, Dual Quaternion Skinning (Kavan et al. 2007) replaced linear matrix blending with dual quaternions, helping to preserve volume and improve joint realism. Despite this, a key issue in this phase is that significant deformation discrepancies in joint regions can lead to geometric problems, for example, stretching or collapsing of the human mesh, illustrated in Figure 1 (*Left*). This occurs because the commonly used LBS algorithm calculates vertex positions through a linear blend of joint trans-

*Equal contribution
†Corresponding author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

formations, without any inherent constraints on human body topology (Kavan et al. 2008; Kavan, Sloan, and O'Sullivan 2010; Nuvoli et al. 2022).

In this paper, we propose a learning-based Gaussian animation method, which contains the decoupled template representation and the learnable Gaussian animation modules. In the canonical template construction stage, we extract UV and pose features from human scan and SMPL-X body mesh respectively. Then these features are fed into a U-Net to initialize the parameters of human Gaussians. As a feed-forward process that directly learned from the raw source data, it is free from data preprocessing and only cost less than 50 ms for inference, which is much more efficient than the existing methods (Ho et al. 2023; Shen et al. 2023; Moon, Shiratori, and Saito 2024; Wen, Schwing, and Wang 2025).

In the target pose deformation stage, we introduce a pre-trained model as a strong human prior for realistic human animation. This model eliminates unexpected geometry issues (such as incorrect stretching) using geometry supervision losses and, based on the personalized template constructed earlier, corrects texture issues (like edge texture fusion) via color loss, with example results shown in Figure 1 (*Right*). Extensive experiments show that our proposed two-stage framework effectively prevents quality degradation from multi-stage error accumulation, outperforming current state-of-the-art (SOTA) approaches and demonstrating impressive efficiency advantage. The key contributions can be summarized as follows:

- We propose an unified human animation framework, FastAnimate to realize pose-conditioned textured human avatar animation with balanced efficiency and fidelity. The entire inference process costs only about 0.1s.
- We propose a decoupled method to construct canonical human Gaussian templates. This approach separates UV feature and pose feature, eliminating the need for heuristic texture estimation.
- We design a data-driven, learnable Gaussian animation module capable of addressing geometry challenges arising from human deformation. This approach mitigates binding inaccuracies while preserving fine-grained geometry details.

Related Work

3D Avatar Animation. Animatable templates and animation algorithms are essential for 3D Human Avatar Animation. Early skeleton-based methods rig precise skeletons and weight surface points (Lewis, Cordner, and Fong 2023; Pons-Moll et al. 2015). Parametric human models expanded animatable template research (Anguelov et al. 2005; Hasler et al. 2009), focusing on templates easily deformed via joint transformations (Pantuwong and Sugimoto 2012; Fechteler et al. 2016; Feng, Casas, and Shapiro 2015). While SMPL enables deformation, it lacks photorealism (Varol et al. 2018). Subsequent studies use SMPL models as 3D priors, reconstructing clothed avatars from monocular/sparse inputs (Aliakbarian et al. 2023; Xu et al. 2024; Lu et al. 2024; Karthikeyan et al. 2024), but require specific data formats. Recently, mesh-based animation avoids complex preprocessing (Saito et al. 2021; Ho et al. 2023), deforming meshes while preserving details. Based on the high-quality template, human avatar could be easily animated using LBS algorithm. Our proposed method is inspired by these mesh-based approaches and goes further in both of the two main areas of avatar animation.

Gaussian Splatting for Humans. Compared to previous representations, 3D Gaussian Splatting (3DGS) introduced a novel paradigm for human representation and rendering, utilizes explicit Gaussian ellipsoids to achieve high-quality novel view synthesis with real-time performance (Kerbl et al. 2023; Zheng et al. 2024; Zhang et al. 2024; Shen et al. 2025; Zhang et al. 2025). 3DGS offers a differentiable and efficient pipeline, making it a promising tool for digital human modeling. In human reconstruction, 3DGS enables the creation of detailed avatars from multi-view images or sparse point clouds with remarkable efficiency. For instance, (Abdal et al. 2024) demonstrated that 3DGS can reconstruct photorealistic human models in minutes. Specifically, (Liu et al. 2024a) introduces structure-aware score distillation sampling to optimize the appearance and geometry of human gaussians. These studies widely explore the potential of 3DGS on human avatars. For human avatar animation, 3DGS has been adapted to support animation by integrating parametric models like SMPL. GauHuman (Hu, Hu, and Liu 2024) leverages LBS to deform Gaussians from a canonical pose to animated states, achieving high training and rendering speeds. Furthermore, through (Kocabas et al. 2024) real-time interaction is made feasible, which drive digital humans from monocular video inputs, capturing intricate details such as clothing wrinkles and hair motion with minimal latency. Impressed by the outstanding performance of 3D Human Gaussians, our research utlizes this representation as the base of our human model.

Methodology

Preliminaries

SMPL-X Model. The SMPL (Loper et al. 2023) model is a parametric model that represents 3D human body shapes and poses. Our method is build upon one of its variants, SMPL-X, which can be denote as $\mathcal{X}(\theta, \beta, \alpha, \sigma) \in \mathbb{R}^{10475\times3}$, where the $\theta \in \mathbb{R}^{63}$ controls the global orientation and relative rotations of the body joints, the $\beta \in \mathbb{R}^{10}$ represents the body shape, and the $\alpha \in \mathbb{R}^{20}$ and $\sigma \in \mathbb{R}^{20}$ control the facial expression and finger movements, respectively.

3D Gaussian Splatting. We utilize 3D Gaussians (Kerbl et al. 2023), an explicit and expressive differentiable 3D representation to model the textured and animated humans. Each human is represented as a set 3D Gaussians $\mathcal{G} = \{\mathcal{G}_i\}$, where each 3D Gaussian is parameterized as: $\mathcal{G}_i = \{\mu_i, \alpha_i, s_i, r_i, \mathbf{c}_i\} \in \mathbb{R}^{14}$. The $\mu_i \in \mathbb{R}^3$ is the geometric center, $\alpha_i \in \mathbb{R}^1$ is the opacity value, $s_i \in \mathbb{R}^3$ and $r_i \in \mathbb{R}^4$ denotes the scale and rotation parameters, respectively, and the $\mathbf{c}_i \in \mathbb{R}^3$ is the RGB color. This parametric formulation compactly encodes both spatial properties and visual attributes while maintaining differentiability.

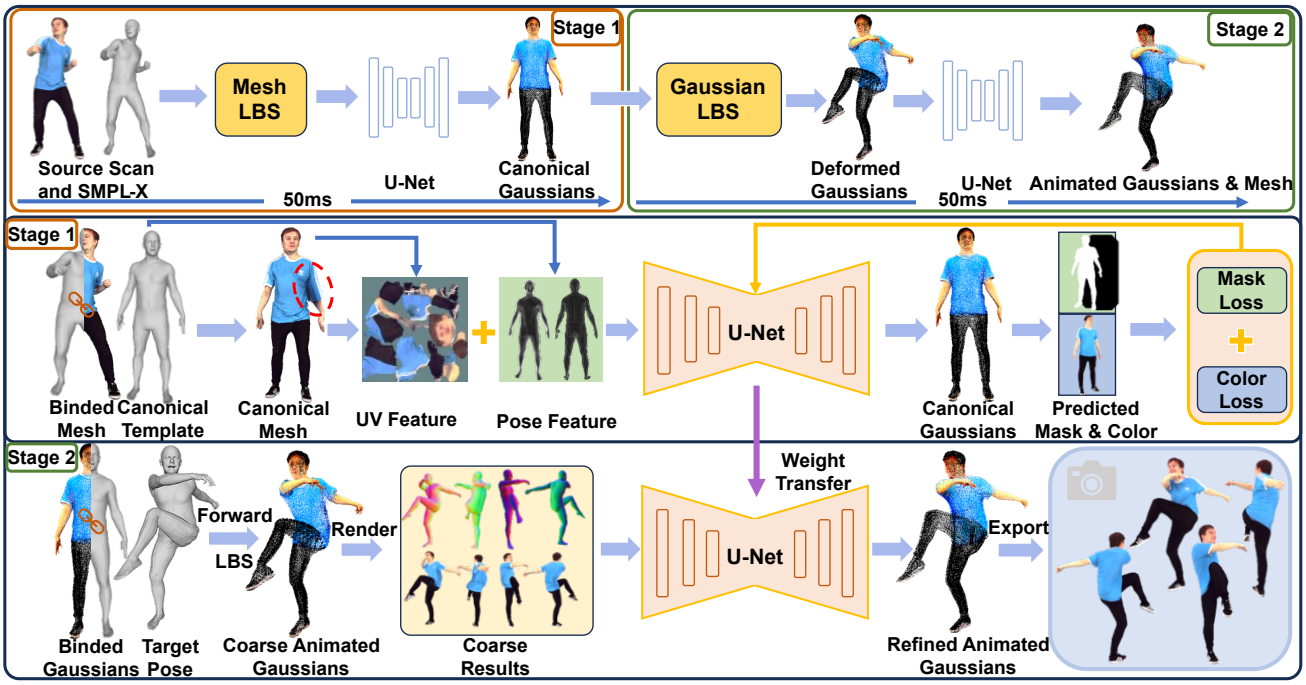

Figure 2: Overview of the proposed FastAnimate. This framework consists of two stages: In the first stage, we decouple the UV feature and pose feature from canonical mesh to build canonical Gaussians with the given human scans and SMPL-X canonical template. In the second stage, we utilize the LBS to drive canonical Gaussians to form coarse animated Gaussians. To further improve the animation quality, we utilize a coarse geometry refinement module to obtain high-fidelity refined animated human Gaussians. By leveraging FastAnimate, we can achieve robust 3D human animation results with enhanced texture quality and pose correctness.

Linear Blend Skinning. LBS is a widely adopted technique in character animation and 3D modeling for deforming a body with skeletal structure. Given its computational efficiency and compatibility with parametric models, LBS serves as a foundational component in our proposed framework. Here, we briefly outline its formulation and key properties to provide context for subsequent sections.

In this study, we consider a 3D mesh $\mathcal{M} = \{\mathbf{v}_i\}_{i=1}^{V}$ with V vertices, where $\mathbf{v}_i \in \mathbb{R}^3$ represents the position of the i-th vertex in a rest pose, alongside a 3D Gaussian representation. Both are rigged to a skeleton from the SMPL-X model, which includes J joints, each defined by a transformation matrix $\mathbf{T}_j \in \mathbb{R}^{4\times4}$ capturing rotation and translation relative to the rest state. The LBS algorithm deforms the mesh by calculating transformed vertex positions $\mathbf{v}_i'$ as:

$$\mathbf{v}_i' = \sum_{j=1}^{J} w_{i,j}\mathbf{T}_j\mathbf{v}_i, \tag{1}$$

where $w_{i,j} \in [0,1]$ denotes the blend weight of the j-th joint on the i-th vertex, with $\sum_{j=1}^{J} w_{i,j} = 1$ for normalization. Here, $\mathbf{T}_j$ is derived from SMPL-X models fitted to two distinct poses, and the weights $\mathbf{W} = \{w_{i,j}\}$ are precomputed during rigging, based on vertex-joint proximity or data-driven optimization. In the following sections, we build upon this mechanism to enhance the geometric and visual fidelity of animated human avatar representations.

Method Overview

To achieve robust and generalizable human avatar animation, we propose a learning-based framework with symmetric forward-backward architecture, an approach that leverages a two-stage transformation process with transferred weights. This method ensures pose alignment while preserving texture integrity and enables efficient animation to a target pose. The methodology is detailed as follows:

Given an input human avatar representation $\mathcal{A}$, characterized by its 3D geometry $\mathcal{M}$ for clothed human pose/shape and 2D texture $\mathcal{T}$ for clothed human appearance, we first transform it into a pre-defined intermediate canonical pose using the forward component of the symmetric structure. This step aims to standardize the pose information while retaining accurate texture details. Formally, let $\mathcal{P}_u$ denote the canonical pose, and $\mathcal{F}_{\text{fwd}}(\cdot)$ represents the forward transformation function parameterized by a learned module with weights $\mathbf{W}$. The intermediate representation $\mathcal{A}_u$ is computed as:

$$\mathcal{A}_u = (\mathcal{M}_u, \mathcal{T}_u) = \mathcal{F}_{\text{fwd}}(\mathcal{A}, \mathcal{P}_u; \mathbf{W}), \tag{2}$$

where $\mathcal{M}_u$ is the geometry aligned to $\mathcal{P}_u$, and $\mathcal{T}_u$ approximates the original texture $\mathcal{T}$ with minimal distortion. The forward transformation leverages SMPL-X geometric priors and texture mapping to ensure consistency.

Subsequently, the backward component of the symmetric structure transforms $\mathcal{A}_u$ into the target pose $\mathcal{P}_t$. This process is regarded as the inverse of the forward transformation, de-

fined by the function $\mathcal{F}_{\text{bwd}}(\cdot)$, which uses the fine-tuned $\widetilde{\mathbf{W}}$ shared from $\mathbf{W}$ due to the symmetry of the framework. The final animated avatar $\mathcal{A}_t$ is obtained as:

$$\mathcal{A}_t = (\mathcal{M}_t, \mathcal{T}_t) = \mathcal{F}_{\text{bwd}}(\mathcal{A}_u, \mathcal{P}_t; \widetilde{\mathbf{W}}), \tag{3}$$

where $\mathcal{M}_t$ corresponds to the target geometry, and $\mathcal{T}_t$ preserves the texture aligned with $\mathcal{P}_t$. The symmetry implies that $\mathcal{F}_{\text{bwd}}$ mirrors $\mathcal{F}_{\text{fwd}}$ in structure but operates in the opposite direction, $\mathcal{F}_{\text{bwd}} \approx \mathcal{F}_{\text{fwd}}^{-1}$, constrained by the shared parameterization.

The transfer of weights between the two models offers two key advantages. Firstly, by sharing weights $\mathbf{W}$ between $\mathcal{F}_{\text{fwd}}$ and $\mathcal{F}_{\text{bwd}}$, the framework reduces the number of trainable parameters and enforces consistency across the transformation pipeline. Secondly, this structure enhances the generation ability of the model under limited access to related dataset. During training, the loss function comprises two components: one for geometry and one for texture. The geometry loss computes the mean squared error (MSE) between the predicted mask $\mathcal{K}^p$ and ground truth mask $\mathcal{K}^g$, summed over n viewpoints, while the texture loss calculates the MSE between the predicted color $\mathcal{C}^p$ and ground truth color $\mathcal{C}^g$, also summed over n viewpoints. The total loss is:

$$\mathcal{L} = \sum_{v=1}^{n} \|\mathcal{K}^{p,v} - \mathcal{K}^{g,v}\|_2^2 + \sum_{v=1}^{n} \|\mathcal{C}^{p,v} - \mathcal{C}^{g,v}\|_2^2, \tag{4}$$

where v denotes the viewpoint index. This formulation ensures effective optimization of both geometric accuracy and texture fidelity across multiple perspectives.

Decoupled Template Representation

In this section, we introduce the method of generating Decoupled Template Representation. This method utilizes a human mesh in arbitrary pose at the beginning and transforms it to the canonical space with pose prior extracted from corresponding SMPL-X model. To enable efficient and pose-driven human animation, we propose a novel method termed *Decoupled Template Representation*. This method decouples human texture and geometry, extracting information from UV features and pose features to achieve a standardized, animatable 3D explicit representation based on 3D Gaussians. This animatable template could be built with the following steps:

Given a source human scan mesh $\mathcal{M}_s$, skeletal rigging is performed by binding it to the source SMPL-X template $\mathcal{X}_s$, which is parameterized by shape β_s and pose θ_s. The binding process employs LBS with pre-defined blend weights $\mathcal{W}$, associating each vertex of $\mathcal{M}_s$ with the skeletal joints of $\mathcal{T}_s$. Next, the LBS transformation matrix $\mathbf{T}_{s\to c}$ that maps the source SMPL-X template $\mathcal{X}_s$ to its canonical SMPL-X counterpart $\mathcal{X}_c$ is computed, where $\mathcal{T}_c$ is defined in a A-pose with identical shape parameters. The transformation matrix from the source to the A-pose is derived as:

$$\mathbf{T}_{s\to c} = \sum_i \mathcal{W}_i \cdot \mathbf{G}_i(\theta_c, J(\beta_s)) \cdot \mathbf{G}_i^{-1}(\theta_s, J(\beta_s)), \tag{5}$$

where $J(\beta_s)$ denotes the joint locations regressed from the shape parameters, $\mathbf{G}_i(\theta, J)$ represents the global transformation of the i-th joint, and $\mathcal{W}_i$ are the skinning weights.

Applying $\mathbf{T}_{s\to c}$ to the source mesh $\mathcal{M}_s$, we obtain a coarse standardized human mesh $\mathcal{M}_c$:

$$\mathcal{M}_c = \mathbf{T}_{s\to c} \cdot \mathcal{M}_s. \tag{6}$$

This standardized mesh aligns with the canonical SMPL-X topology, facilitating subsequent processing. Then texture features $\mathcal{F}_{\text{tex}}$ are extracted from UV maps using a pre-trained feature extractor, capturing appearance details such as color and surface patterns. These texture features are concatenated with pose-dependent features $\mathcal{F}_{\text{pose}}$ derived from $\mathcal{T}_c$, which encode geometric deformations induced by the canonical pose.

The combined feature set $\mathcal{F} = [\mathcal{F}_{\text{tex}}, \mathcal{F}_{\text{pose}}]$ is fed into a U-Net architecture with reconstruction capabilities to initialize a set of 3D Gaussians $\mathcal{G}$. Each Gaussian is parameterized by its mean μ, covariance Σ, and opacity α, representing a local volumetric distribution aligned with the canonical SMPL-X template. The U-Net outputs the Gaussian parameters as:

$$\mathcal{G} = \{\mu_i, \Sigma_i, \alpha_i\}_{i=1}^{N} = \text{U-Net}(\mathcal{F}), \tag{7}$$

where N is the number of Gaussians determined adaptively based on the mesh complexity.

Since the resulting Gaussians $\mathcal{G}$ correspond to the canonical SMPL-X template $\mathcal{X}_c$, animating the human representation becomes straightforward. Given a target SMPL-X model $\mathcal{X}_t$ with a new pose θ_t, we compute the forward LBS transformation:

$$\mathbf{T}_{c\to t} = \text{LBS}(\mathcal{T}_t, J(\beta_s), \theta_t, \mathcal{W}) \tag{8}$$

This transformation is applied to the Gaussian means:

$$\mu_i' = \mathbf{T}_{c\to t} \cdot \mu_i, \quad \forall i \in \{1, \ldots, N\}, \tag{9}$$

while the covariance Σ_i and opacity α_i remain unchanged, preserving the local structure and appearance. This process yields an decoupled animatable human template representation $\mathcal{G}'$ driven efficiently by 3D Gaussians, suitable for real-time applications.

Learnable Gaussian Animation

In this section, we introduce the proposed approach to realize learning-based gaussian animation. 3D Human Gaussians are suitable for point-level deformation. However, incorrect geometry and unreal texture still exist as the animation matrix calculated by LBS may not be prefect when applied on mesh or Gaussians.

To enhance the quality of coarse 3D Gaussian representations for human avatar animation, we introduce a method called *Learnable Gaussian Animation*. The key component of this approach is a coarse geometry refinement module, which could refine an initial coarse 3D Gaussian model by leveraging geometric alignment with a target SMPL-X model, improving both fidelity and detail in the resulting human representation. The process is outlined as follows:

Starting with a coarse 3D Gaussian set $\mathcal{G}_c = \{\mu_i, \Sigma_i, \alpha_i\}_{i=1}^{N}$, where μ_i, Σ_i, and α_i denote the mean, covariance, and opacity of each Gaussian, we render images from four distinct viewpoints $\{V_1, V_2, V_3, V_4\}$. These rendered images, denoted as $\mathcal{I}_c = \{I_c^{(k)}\}_{k=1}^{4}$, capture the

coarse geometry from multiple perspectives. Concurrently, we render images $\mathcal{I}_t = \{I_t^{(k)}\}_{k=1}^4$ from the same viewpoints using a target SMPL-X model $\mathcal{T}_t$, parameterized by shape β_t, pose θ_t, and expression ψ_t. The SMPL-X model provides a high-fidelity pose reference.

For each viewpoint k, we extract geometry features from both the coarse Gaussian renderings and the SMPL-X renderings. Let $\mathcal{F}_c^{(k)} = \text{GeoEnc}(I_c^{(k)})$ and $\mathcal{F}_t^{(k)} = \text{GeoEnc}(I_t^{(k)})$ represent the geometry features encoded by a shared geometry encoder $\text{GeoEnc}(\cdot)$, which captures surface normals and silhouette information. These features are paired to form a combined feature set:

$$\mathcal{F}_{\text{pair}}^{(k)} = [\mathcal{F}_c^{(k)}, \mathcal{F}_t^{(k)}], \quad k = 1, 2, 3, 4. \tag{10}$$

This paired feature set $\mathcal{F}_{\text{pair}} = \{\mathcal{F}_{\text{pair}}^{(k)}\}_{k=1}^4$ integrates multi-view geometric cues from both the coarse and target representations.

The paired features $\mathcal{F}_{\text{pair}}$ are input into a specialized U-Net architecture, designed to refine the coarse Gaussian parameters by leveraging the detailed geometric information from $\mathcal{T}_t$. The U-Net outputs a refined Gaussian set $\mathcal{G}_r$:

$$\mathcal{G}_r = \{\mu_i', \Sigma_i', \alpha_i'\}_{i=1}^{N'} = \text{U-Net}(\mathcal{F}_{\text{pair}}), \tag{11}$$

where N' may differ from N due to pruning or addition of Gaussians. The refinement process involves two key operations: (1) pruning artifacts via a pre-defined Gaussian opacity threshold, such as redundant Gaussians causing visual noise, by adjusting α_i' to suppress irrelevant regions; and (2) supplementing details, such as clothing wrinkles and finger geometry, by introducing new Gaussians or adjusting μ_i' and Σ_i' to align with $\mathcal{F}_t$. This alignment ensures that the refined Gaussians $\mathcal{G}_r$ closely match the target SMPL-X geometry.

Additionally, the corrected geometric features from $\mathcal{T}_t$ enable reinitialization of the Gaussians when necessary. The refined Gaussian parameters are computed as:

$$\mu_i' = \mu_i + \Delta\mu_i(\mathcal{F}_t), \quad \Sigma_i' = \Sigma_i + \Delta\Sigma_i(\mathcal{F}_t), \quad \alpha_i' = f(\alpha_i, \mathcal{F}_t), \tag{12}$$

where $\Delta\mu_i$, $\Delta\Sigma_i$, and $f(\cdot)$ are learned corrections derived from the U-Net which is pretrained as human reconstruction model, conditioned on the target SMPL-X features. This process yields a high-quality 3D Gaussian representation $\mathcal{G}_r$, capable of supporting realistic human avatar animation with improved geometric fidelity and vivid texture transformation. We also observe that this module demonstrates excellent performance in correcting point-level geometric errors. As a result, this method could provide structurally accurate, high-fidelity human deformation results.

Experiments

Experiment Setup

Dataset. Our experiments is fully conducted on the X-Humans (Shen et al. 2023) dataset. X-Humans dataset consists of 20 subjects with different garments. There are over 29K poses for training and 6.4K test poses. We follow the default setting to split the training and testing data. We sample 1000 scans from the training set as training data and follow EX-Humans (Moon, Shiratori, and Saito 2024) to select 20 scans from the test set.

Baselines. To demonstrate the superiority of our proposed method, we compare the current SOTA 3D human animation baselines including: **Editable-Humans** (Ho et al. 2023), **X-avatar** (Shen et al. 2023), **EX-avatar** (Moon, Shiratori, and Saito 2024), and **LIFE-GOM** (Wen, Schwing, and Wang 2025). These baselines represent distinct technical paradigms: Editable-Humans employs a 3D scan-based animation approach, while both X-Avatar and EX-Avatar utilize video-driven animation frameworks. The most similar approach to our method is LIFE-GOM, which establishes a 3D human Gaussian avatar from sparse-view inputs. See more details about baselines in Supplementary Material.

Metrics. We follow the metrics used in Editable-humans (Ho et al. 2023) to quantitatively evaluate the 3D pose correctness of animated human: Chamfer Distance (**CD**), Normal Consistency (**NC**), and **f-score** (Tatarchenko et al. 2019) and utilize the Peak Signal-to-Noise Ratio (**PSNR**), Structure Similarity Index Measure (**SSIM**) (Wang et al. 2004) and Learned Perceptual Image Patch Similarity score (**LPIPS**) (Zhang et al. 2018) to evaluate the 2D texture quality of animated human. Additionally, we utilize the inference **Time Cost** to assess the computational efficiency.

Evaluation

Quantitative Evaluation. We present a comprehensive quantitative comparison of our method against SOTA animation approaches in Table 1. Our framework achieves superior performance across animated pose correctness, texture quality, and computational efficiency. For pose correctness, our method attains a CD of 0.609/0.701, NC of 0.934, and F-score of 86.118, outperforming existing techniques by significant margins. These metrics validate our approach's ability to preserve topological fidelity during animation. In terms of texture quality, our method achieves a PSNR of 23.995/23.394 (F/B), SSIM of 0.9789/0.9838 (F/B), and LPIPS score of 0.0271/0.0358 (F/B) on rendered animated avatars. This demonstrates marked improvements in preserving high-frequency texture details and minimizing perceptual artifacts compared to prior work. In addition to computational efficiency, our framework animates 3D human scans in approximate 0.1s per instance—an order-of-magnitude speedup over existing methods without sacrificing output quality. These results collectively demonstrate that the superior performance of our framework on 3D human animation, simultaneously achieving unprecedented accuracy, visual quality, and practical deployability.

Qualitative Evaluation. Figure 3 showcases our framework's ability to generate animatable 3D human avatars with high-fidelity details across diverse poses. As demonstrated in Figure 3, our method generalizes robustly to novel poses while preserving intricate clothing folds, dynamic facial expressions, and articulated finger movements. Crucially, the framework remains stable even under extreme body posture changes (e.g., crouching, arm waving) where traditional

Figure 3: Qualitative comparison with state-of-the-art methods on novel pose synthesis. Please note that due to the difference in camera parameters, the results of LIFE-GOM has a marginal angle difference with others. Please zoom in for a detailed view.

methods typically fail, as evidenced by artifact-free deformations in challenging configurations. Figure 3 provides a visual comparison with SOTA animation approaches. Editable Humans fails to reconstruct fine-grained hand geometry and introduces unnatural surface jaggedness. The results of X-Avatar and EX-Avatar exhibit visible degradation in clothing wrinkles and facial expressions due to methods' limitations. While LIFE-GOM addresses some topological issues, its reliance on input images leads to blurred textures in occluded regions, particularly around armpits, palms and backs of arms. These visualization results further demonstrate the superiority of our proposed FastAnimate. More results can be seen in the Supplementary Material.

Ablation Study

In this section, we conduct ablation studies to systematically analyze the impact of the components of our proposed method. We use a simple LBS approach as our baseline in ablation study. In the setting of "w/o LGA", we remove the entire coarse geometry refinement module in Learnable Gaussian Animation module to assess the impact. In the setting of "w/o DTR", we ablate the process of decoupled template representation and generate the animated 3D human Gaussian avatars with the U-Net directly.

Effectiveness of Geometry Refinement. Table 2 highlights the effectiveness of the proposed the learnable geometry refinement module. Ablating this component results in a sig-

Methods	Publication	2D Texture Quality			3D Geometry Correctness			Efficiency
		PSNR: F/B ↑	**SSIM**: F/B ↑	**LPIPS**: F/B ↓	**CD**: P-to-S / S-to-P (cm) ↓	**NC**↑	**F-score**↑	**Time Cost**↓
Editable-Humans	CVPR 2023	19.354/16.860	0.9302/0.9383	0.1019/0.1056	1.936/2.099	0.815	37.441	≈22s
X-Avatar	CVPR 2023	20.824/19.335	0.9446/0.9443	0.0745/0.0791	0.975/0.924	0.907	65.986	≈10s
EX-Avatar†	ECCV 2024	22.201/21.984	0.9512/0.9443	0.0687/0.0712	0.732/0.717	0.921	84.231	≈10s
LIFE-GOM†	ICLR 2025	22.103/22.294	0.9540/0.9521	0.0690/0.0687	0.801/0.772	0.923	83.204	≈1s
FastAnimate †	AAAI 2026	**23.995/23.394**	**0.9789/0.9838**	**0.0271/0.0358**	**0.609/0.701**	**0.934**	**86.118**	**≈0.1s**

Table 1: We compare our proposed method with SOTA approaches in terms of 2D Texture Quality (PSNR, SSIM, LPIPS), 3D Geometry Correctness (CD, NC, F-score) and Computational Efficiency (Time Cost). The "†" denotes the method is build upon the 3D Gaussian Splatting. For GS-based methods, the mesh are exported with technique provided by LGM (Tang et al. 2024) for fair comparison. The arrow ↑/↓ represents the higher/lower is better.

Methods	2D Texture Quality		
	PSNR: F/B ↑	**SSIM**: F/B ↑	**LPIPS**: F/B ↓
w/o LGA	22.956/23.125	0.9721/0.9741	0.0472/0.0456
w/o DTR	23.183/23.284	0.9779/0.9830	0.0361/0.0366
Full Pipeline	23.995/23.394	0.9789/0.9838	0.0271/0.0358
	3D Geometry Correctness		
	CD: P-to-S / S-to-P (cm) ↓	**NC** ↑	**F-score**↑
w/o LGA	0.823/0.857	0.902	82.524
w/o DTR	0.752/0.808	0.923	84.743
Full Pipeline	0.609/0.701	0.934	86.118

Table 2: Ablation Study of Learnable Gaussian Animation module and Decoupled Template Representation.

nificant decrease in both texture quality and geometry accuracy, underscoring its critical importance. We attribute this performance gap to two key factors: First, inaccurate binding of Gaussian to canonical template causes the misalignment in fine-grained geometry. Second, canonical poses fail to account for non-rigid deformations and high-frequency details inherent in dynamic human performances. The results validate that geometry refinement is indispensable for achieving photorealistic human animation.

Effectiveness of Template Representation. Table 2 illustrates the effectiveness of our proposed template representation. Removing this component causes a decrease in SSIM and PSNR metrics and increases LPIPS score, indicating severe quality degradation in texture quality. It is likely due to the high coupling between the UV feature and the pose feature, which forces the model to estimate texture details heuristically rather than leveraging disentangled, reusable templates. Furthermore, omitting the template framework introduces redundant computations: Without persistent template storage, the model must reprocess identical inputs across varying target poses. The ablation confirms that explicit template representation is crucial for both accuracy and computational efficiency in pose-transfer scenarios.

Visual Ablation. Figure 4 demonstrates the perceptual and geometric improvements enabled by our proposed framework. The first column reveals significant degradation in animated avatar details when ablating the learnable geometry refinement module, manifesting as distortion in finger articulation and artifacts in waist clothing. The second column highlights minor yet perceptible losses in fine-grained appearance details, such as reduced fidelity in clothing wrinkles and residual finger inaccuracies. Strikingly, our full framework synthesizes avatars with high-fidelity textures and dynamically coherent wrinkles that align closely with ground truth observations, underscoring the necessity of both modules for photorealistic 3D human animation.

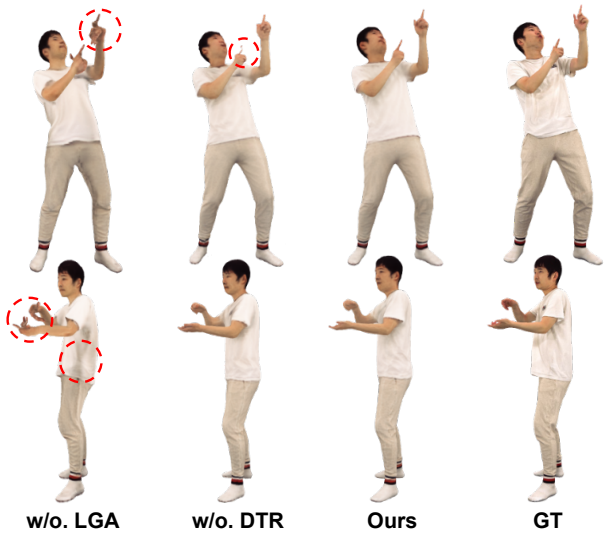

Figure 4: Visual ablation of proposed modules. Proposed modules improve fine-grained texture and geometry quality.

Conclusion

Real-time, high-quality human avatar animation under specified poses remains a persistent challenge in animation research, often yielding unrealistic results marred by significant artifacts and geometric errors. Traditional approaches prioritize improving animatable templates but overlook the animation process itself. In this study, we introduce a novel feed-forward framework that delivers easily animatable human Gaussians template while incorporating a dedicated module to refine problematic geometries. Extensive experiments demonstrate that the proposed FastAnimate generates canonical template and highly realistic reposed human avatars in approximately 0.1s, achieving an optimal balance between quality and speed.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62406267), the Guangzhou-HKUST(GZ) Joint Funding Program (No. 2025A03J3956) and the Guangzhou Municipal Education Project (No. 2024312122).

References

Abdal, R.; Yifan, W.; Shi, Z.; Xu, Y.; Po, R.; Kuang, Z.; Chen, Q.; Yeung, D.-Y.; and Wetzstein, G. 2024. Gaussian shell maps for efficient 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9441–9451.

Aliakbarian, S.; Saleh, F.; Collier, D.; Cameron, P.; and Cosker, D. 2023. Hmd-nemo: Online 3d avatar motion generation from sparse observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9622–9631.

Anguelov, D.; Srinivasan, P.; Koller, D.; Thrun, S.; Rodgers, J.; and Davis, J. 2005. SCAPE: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, SIGGRAPH '05, 408–416. ISBN 9781450378253.

Fechteler, P.; Paier, W.; Hilsmann, A.; and Eisert, P. 2016. Real-time avatar animation with dynamic face texturing. In *2016 IEEE International Conference on Image Processing (ICIP)*, 355–359. IEEE.

Feng, A.; Casas, D.; and Shapiro, A. 2015. Avatar reshaping and automatic rigging using a deformable model. In *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*, 57–64.

Hasler, N.; Stoll, C.; Sunkel, M.; Rosenhahn, B.; and Seidel, H.-P. 2009. A Statistical Model of Human Pose and Body Shape. *Computer Graphics Forum*, 28(2): 337–346.

Ho, H.-I.; Xue, L.; Song, J.; and Hilliges, O. 2023. Learning Locally Editable Virtual Humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu, S.; Hu, T.; and Liu, Z. 2024. Gauhuman: Articulated gaussian splatting from monocular human videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20418–20431.

Ichim, A. E.; Bouaziz, S.; and Pauly, M. 2015. Dynamic 3D avatar creation from hand-held video input. *ACM Transactions on Graphics (ToG)*, 34(4): 1–14.

James, D. L.; and Twigg, C. D. 2005. Skinning mesh animations. *ACM Transactions on Graphics (TOG)*, 24(3): 399–407.

Karthikeyan, A.; Ren, R.; Kant, Y.; and Gilitschenski, I. 2024. Avatarone: Monocular 3d human animation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3647–3657.

Kavan, L.; Collins, S.; Žára, J.; and O'Sullivan, C. 2007. Skinning with dual quaternions. In *Proceedings of the 2007 symposium on Interactive 3D graphics and games*, 39–46.

Kavan, L.; Collins, S.; Žára, J.; and O'Sullivan, C. 2008. Geometric skinning with approximate dual quaternion blending. *ACM Transactions on Graphics (TOG)*, 27(4): 1–23.

Kavan, L.; Sloan, P.-P.; and O'Sullivan, C. 2010. Fast and efficient skinning of animated meshes. In *Computer Graphics Forum*, volume 29, 327–336. Wiley Online Library.

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.

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.

Kwon, Y.; Fang, B.; Lu, Y.; Dong, H.; Zhang, C.; Carrasco, F. V.; Mosella-Montoro, A.; Xu, J.; Takagi, S.; Kim, D.; et al. 2024. Generalizable human gaussians for sparse view synthesis. In *European Conference on Computer Vision*, 451–468. Springer.

Lewis, J. P.; Cordner, M.; and Fong, N. 2023. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 811–818.

Li, Z.; Chen, L.; Liu, C.; Gao, Y.; Ha, Y.; Xu, C.; Quan, S.; and Xu, Y. 2019. 3d human avatar digitization from a single image. In *Proceedings of the 17th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*, 1–8.

Liu, X.; Zhan, X.; Tang, J.; Shan, Y.; Zeng, G.; Lin, D.; Liu, X.; and Liu, Z. 2024a. Humangaussian: Text-driven 3d human generation with gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6646–6657.

Liu, Y.; Huang, X.; Qin, M.; Lin, Q.; and Wang, H. 2024b. Animatable 3d gaussian: Fast and high-quality reconstruction of multiple human avatars. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1120–1129.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 851–866.

Lu, F.; Dong, Z.; Song, J.; and Hilliges, O. 2024. AvatarPose: Avatar-guided 3D Pose Estimation of Close Human Interaction from Sparse Multi-view Videos. In *European Conference on Computer Vision*, 215–233. Springer.

McManus, E. A.; Bodenheimer, B.; Streuber, S.; De La Rosa, S.; Bülthoff, H. H.; and Mohler, B. J. 2011. The influence of avatar (self and character) animations on distance estimation, object interaction and locomotion in immersive virtual environments. In *Proceedings of the ACM SIGGRAPH Symposium on applied perception in graphics and visualization*, 37–44.

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

Moreau, A.; Song, J.; Dhamo, H.; Shaw, R.; Zhou, Y.; and Pérez-Pellitero, E. 2024. Human gaussian splatting: Real-time rendering of animatable avatars. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, 788–798.

Nuvoli, S.; Pietroni, N.; Cignoni, P.; Scateni, R.; and Tarini, M. 2022. SkinMixer: Blending 3D animated models. *ACM Transactions on Graphics (TOG)*, 41(6): 1–15.

Pang, H.; Zhu, H.; Kortylewski, A.; Theobalt, C.; and Habermann, M. 2024. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1165–1175.

Pantuwong, N.; and Sugimoto, M. 2012. A novel template-based automatic rigging algorithm for articulated-character animation. *Computer Animation and Virtual Worlds*, 23(2): 125–141.

Pons-Moll, G.; Romero, J.; Mahmood, N.; and Black, M. J. 2015. Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics (TOG)*, 34(4): 1–14.

Saito, S.; Yang, J.; Ma, Q.; and Black, M. J. 2021. SCANimate: Weakly Supervised Learning of Skinned Clothed Avatar Networks. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.

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.

Shen, W.; Zhang, G.; Zhang, J.; Feng, Y.; Yao, N.; Zhang, X.; and Wang, H. 2025. SMPL Normal Map Is All You Need for Single-view Textured Human Reconstruction. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.

Tang, J.; Chen, Z.; Chen, X.; Wang, T.; Zeng, G.; and Liu, Z. 2024. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, 1–18. Springer.

Tatarchenko, M.; Richter, S. R.; Ranftl, R.; Li, Z.; Koltun, V.; and Brox, T. 2019. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3405–3414.

Varol, G.; Ceylan, D.; Russell, B.; Yang, J.; Yumer, E.; Laptev, I.; and Schmid, C. 2018. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European conference on computer vision (ECCV)*, 20–36.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wen, J.; Schwing, A. G.; and Wang, S. 2025. LIFe-GoM: Generalizable Human Rendering with Learned Iterative Feedback Over Multi-Resolution Gaussians-on-Mesh. *arXiv preprint arXiv:2502.09617*.

Xu, Z.; Peng, S.; Geng, C.; Mou, L.; Yan, Z.; Sun, J.; Bao, H.; and Zhou, X. 2024. Relightable and animatable neural avatar from sparse-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 990–1000.

Yang, X.; Somasekharan, A.; and Zhang, J. J. 2006. Curve skeleton skinning for human and creature characters. *Computer Animation and Virtual Worlds*, 17(3-4): 281–292.

Yuan, Y.; Li, X.; Huang, Y.; De Mello, S.; Nagano, K.; Kautz, J.; and Iqbal, U. 2024. Gavatar: Animatable 3d gaussian avatars with implicit mesh learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 896–905.

Zhang, G.; Shu, J.; Yao, N.; and Wang, H. 2025. SAT: Supervisor Regularization and Animation Augmentation for Two-process Monocular Texture 3D Human Reconstruction. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 10563–10572.

Zhang, G.; Yao, N.; Zhang, S.; Zhao, H.; Pang, G.; Shu, J.; and Wang, H. 2024. MultiGO: Towards Multi-level Geometry Learning for Monocular 3D Textured Human Reconstruction. *arXiv preprint arXiv:2412.03103*.

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.

Zheng, S.; Zhou, B.; Shao, R.; Liu, B.; Zhang, S.; Nie, L.; and Liu, Y. 2024. 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*, 19680–19690.