FACt: Feed-forward Animatable Clothed Human Reconstruction

Anonymous authors
Paper under double-blind review

Abstract

We present FACt, a novel feed-forward framework for animatable clothed human reconstruction from sparse-view images. Unlike optimization-based methods requiring per-subject refinement or feed-forward approaches limited to static capture, FACt leverages 3D Gaussian Splatting to enable high-fidelity avatar reconstruction and adaptive clothing deformation in a single forward pass. Our model's architecture efficiently fuses 3D representation with locally aligned image features via a deformable transformer, effectively preserving geometric and appearance fidelity in the canonical space. To accurately model dynamic clothing, we introduce an adaptive cloth joint learning strategy. This strategy extends the standard Linear Blend Skinning kinematic chain beyond the SMPL-X template, enabling realistic motion control across diverse clothing styles and poses. Experiments on real-world datasets demonstrate that FACt produces realistic and animatable human avatars across novel views and poses, outperforming state-of-the-art methods.

1 Introduction

Animatable 3D human avatar reconstruction from sparse-view images is crucial for diverse downstream applications, including virtual try-on, interactive gaming, and industrial digital human synthesis. However, recovering high-fidelity clothed humans with diverse garment topologies and dynamic articulation remains a long-standing challenge due to spatial ambiguities and complex non-rigid deformations. While recent progress has shifted from NeRF- or SDF-based implicit representations (Weng et al., 2022; Zhao et al., 2022; Peng et al., 2021; Jiang et al., 2022; Zhang et al., 2024b; 2023) toward explicit 3D Gaussian Splatting (3DGS) (Hu et al., 2024; Qian et al., 2024; Kocabas et al., 2024; Qiu et al., 2025a; Kwon et al., 2024; Zhuang et al., 2025; Pan et al., 2024; Li et al., 2026), novel-view and novel-pose rendering under sparse view input settings still yields suboptimal fidelity, particularly for loose garments that deviate significantly from the underlying body topology.

Early 3DGS models rely on parametric templates (*e.g.*, SMPL/SMPL-X (Loper et al., 2015; Pavlakos et al., 2019)) to initialize Gaussian primitives and perform per-subject optimization from monocular video using Linear Blend Skinning (LBS) for deformation (Hu et al., 2024; Qian et al., 2024; Kocabas et al., 2024). Although capable of capturing fine-grained details, these approaches suffer from poor generalizability and time-consuming per-scene optimization. Furthermore, standard LBS is inherently tailored to body-surface deformation and fails to model unconstrained, loose garment dynamics. To address these limitations, recent efforts have shifted toward feed-forward reconstruction models that avoid per-subject optimization. Building on the success of feed-forward pipelines in reconstructing general 3D assets from sparse views (Charatan et al., 2024; Chen et al., 2024b; Szymanowicz et al., 2024), and inspired by the strong scaling capability of transformers in 3D representation learning (Hong et al., 2023; Wang et al., 2025a;c), several recent works (Qiu et al., 2025a;c; Zhuang et al., 2025; Zhang et al., 2023; Li et al., 2026) have enabled high-quality reconstruction of diverse human avatars from sparse view images.

Despite these advances, two critical challenges remain. First, sparse image features observed in posed views should be reliably associated with a canonical 3D representation. Dense global cross-attention between thousands of Gaussian primitives and high-dimensional features from 2D foundation models (*e.g.*, DINO (Oquab et al., 2023) and Sapiens (Khirodkar et al., 2026)) is computationally redundant and does not explicitly model these local correspondences. Reliable local feature aggregation is therefore essential for constructing cloth-aware canonical Gaussian features. Second, standard LBS

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

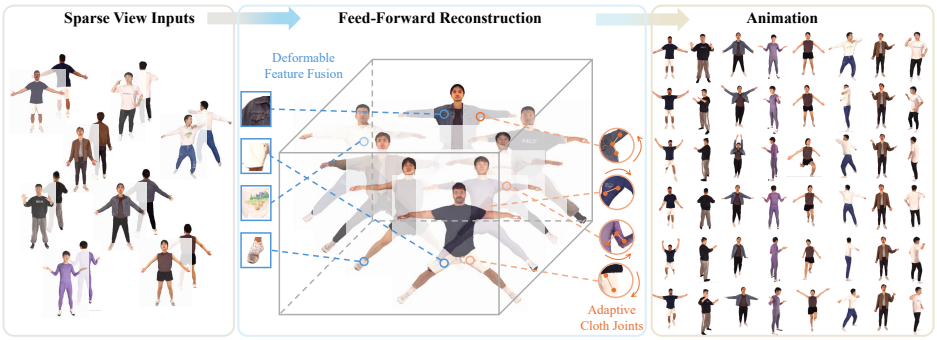


Figure 1: Overview of FACt. Our method reconstructs animatable clothed human avatars in a feed-forward manner, integrating deformable feature fusion from sparse-view images and a cloth joint-driven kinematic chain.

alone cannot capture garment motion that deviates substantially from body-surface deformation, particularly for loose clothing. Existing methods commonly address this with unconstrained MLP-predicted residual offsets. Although flexible, this black-box formulation provides limited interpretable control over garment motion. It contradicts standard industrial graphics workflows where technical directors require explicit handle controls to artistically edit and tweak garment dynamics. In fact, garment deformation can be viewed as a combination of low-frequency quasi-rigid motion and high-frequency non-rigid residuals, where the former part can be modeled with LBS-like virtual joints (Pan et al., 2022; Lei et al., 2024; Zhan et al., 2025; Guo et al., 2024). However, such ideas are often restricted to per-subject optimization (Zhan et al., 2025), require ground-truth garment meshes (Pan et al., 2022), or lack a unified kinematic chain propagating motion from body joints to virtual joints (Lei et al., 2024; Guo et al., 2024). This makes them inapplicable to feed-forward settings driven solely by photometric supervision.

To address these challenges, we introduce **FACt**, a framework for feed-forward animatable clothed human reconstruction. FACt factorizes the task into canonical representation learning and pose-conditioned garment control. First, we design a *deformable transformer* that grounds each Gaussian in locally aligned image features from sparse views and pre-trained 2D foundation models. This local aggregation enriches canonical geometry and cloth-related appearance while avoiding redundant global cross-attention. Later, we introduce *adaptive cloth joint learning*, which infers cloth-joint positions and rotations from the cloth-aware Gaussian representation and the target pose. The inferred joints form a body-linked hierarchical kinematic chain that propagates motion from body joints to cloth joints and subsequently to cloth Gaussians through extended LBS. Given sparse-view images, FACt reconstructs a canonical Gaussian avatar and produces interpretable, cloth-aware deformations under novel poses without test-time per-subject optimization.

In summary, our main contributions are threefold. First, we propose a feed-forward framework capable of reconstructing high-fidelity, animatable 3D human avatars from sparse-view images. Second, we design a deformable transformer for feature aggregation that significantly mitigates global cross-attention redundancy while enhancing canonical appearance and geometric fidelity. Third, we introduce an adaptive cloth joint learning mechanism along with an extended LBS formulation that establishes a hierarchical body-to-garment kinematic chain, enabling interpretable garment dynamics across diverse clothing styles and complex human poses.

2 Related Work

3D human reconstruction Recent 3D Gaussian Splatting (3DGS) methods have enabled efficient animatable human reconstruction from images (Kerbl et al., 2023). Optimization-based approaches initialize Gaussian primitives with parametric human templates and refine them for each subject (Hu et al., 2024; Qian et al., 2024; Kocabas et al., 2024; Tiruneh et al., 2026). Although they achieve high visual fidelity, their test-time optimization limits scalability. In contrast, feed-forward methods directly predict human representations from sparse images. GPS-Gaussian (Zheng et al., 2024) predicts pixel-aligned Gaussians, whereas GHG (Kwon et al., 2024), RoGSplat (Xiao et al., 2025), and GIGA (Zubekhin et al., 2026) employ UV-space representations. Beyond 3DGS representation, HumanRAM (Yu et al., 2025) and FRESA (Wang et al., 2025b) explore implicit and mesh-based

reconstruction. More recent large reconstruction models, including LHM (Qiu et al., 2025a), PF-LHM (Qiu et al., 2025c), IDOL (Zhuang et al., 2025), LHM++ (Qiu et al., 2025b), and LCA (Li et al., 2026), use multimodal transformers to integrate image feature and human-body priors for canonical avatar reconstruction. To compensate for the lack of information in sparse-view inputs, diffusion models are applied to synthesize novel views (Li et al., 2025b; Cai et al., 2026; Buehler et al., 2026) or distill generative priors (Pan et al., 2024; Jin et al., 2025; Chen et al., 2024a; Weng et al., 2024). However, the denoising or distillation remains computationally demanding, impeding real-time deployment.

FACt belongs to the non-generative feed-forward paradigm but addresses two limitations of existing feature aggregation and animation schemes. Rather than globally associating every Gaussian with image tokens, FACt uses deformable attention to aggregate locally aligned, visibility-aware image features for each canonical Gaussian. Moreover, beyond reconstructing canonical Gaussian attributes, FACt infers a garment-specific, body-linked kinematic chain to model pose-conditioned clothing deformation.

3D cloth deformation Modeling realistic clothing dynamics remains a key challenge in 3D human avatar reconstruction. Early methods reconstruct clothing through mesh-based optimization or implicit surface representations, often prioritizing geometric fidelity over detailed texture synthesis (Liao et al., 2023; Saito et al., 2019; 2020; Huang et al., 2020; He et al., 2021; Shin et al., 2024). More recent approaches explicitly separate cloth reconstruction from the underlying body (Pang et al., 2025; Zhang et al., 2025) or represent low-frequency garment motion using latent bones and LBS-like deformation (Pan et al., 2022; Lei et al., 2024; Zhan et al., 2025; Guo et al., 2024). These methods demonstrate the value of structured garment motion, but are commonly designed for per-subject optimization, require garment geometry supervision, or do not infer an explicit body-linked hierarchical cloth kinematic chain from sparse image observations. In contrast, FACt predicts cloth-joint states from cloth-aware Gaussian features and the target pose under image-space supervision. The inferred joints extend the body skeleton through a hierarchical body-to-cloth chain, providing structured and interpretable garment control.

3 METHOD

3.1 OVERVIEW

Given sparse-view RGB images and associated pose annotations, FACt is a feed-forward transformer framework for reconstructing high-fidelity, animatable clothed avatars. We initialize body and cloth Gaussians separately from the canonical SMPL-X body and offline fitted cloth templates (Sec. 3.2). Each Gaussian carries a learnable feature that encodes its geometric and appearance properties. To construct a canonical avatar, a deformable transformer (Sec. 3.3) fuses Gaussian features with locally aligned image features. The fused features are decoded into canonical Gaussian primitives, including position offsets, rotations, scales, opacities, and colors. For pose-driven animation, we introduce adaptive cloth joint learning (Sec. 3.4), which predicts body-linked pseudo cloth joints conditioned on canonical Gaussian features and target pose. These joints extend the standard LBS hierarchy and propagate motion from body joints to cloth joints and subsequently to cloth Gaussians, enabling garment motion that departs from body-surface deformation. The posed avatar is rendered differentiably, and the entire framework is trained with image-space reconstruction and geometric regularization losses (Sec. 3.5).

3.2 GAUSSIANS INITIALIZATION

Body Gaussians are initialized by uniformly sampling the canonical SMPL-X mesh surface. Cloth Gaussians are initialized from offline prepared cloth templates constructed from DeepFashion3D (Heming et al., 2020). Specifically, we define four major garment categories, upper garments, pants, skirts, and dresses, with further subdivisions based on sleeve and garment length. For each category, a canonical template is initialized from the corresponding SMPL body regions. For each DeepFashion3D sample, we first deform the canonical template into the posed space through forward LBS and refine it against the observed garment mesh with non-rigid iterative closest point registration (NRICP) (Amberg et al., 2007). The fitted deformation is then mapped back to the canonical space

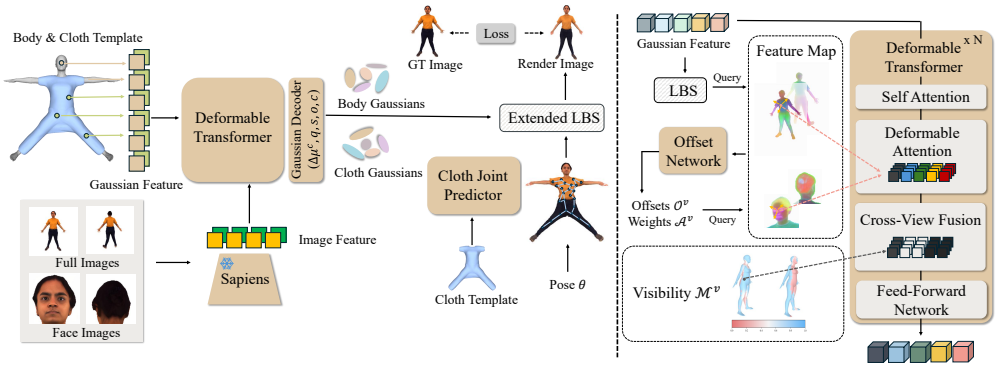


Figure 2: Architecture of FACt. (*Left*) Sparse-view inputs initialize body and cloth Gaussians from SMPL-X (Pavlakos et al., 2019) and cloth templates. Full and face image features are extracted by Sapiens (Khirodkar et al., 2026) and fused with Gaussian features to predict canonical Gaussian primitives. Pose-conditioned cloth-joint states are further predicted and combined with body joints in an extended LBS to model garment dynamics. (*Right*) Deformable transformer. An offset network samples locally aligned image features, while SMPL depth-based visibility preserves cross-view consistency for feature fusion.

through inverse LBS. Such procedure is aggregated across different poses of the same garment and across garments within the same category. Finally, surface smoothing is applied to produce a unified cloth template for each category. Notably, this cloth template preparation is independent of any target subject for reconstruction and does not require ground-truth garment meshes during later feed-forward reconstruction process. Additional details are in App. A.1 and B.1.

3.3 Deformable Transformer

To effectively learn the geometric and appearance properties of canonical Gaussians, we design a deformable transformer architecture that integrates deformable attention into point transformer (Wu et al., 2024) backbone. This mechanism adaptively associates local image features extracted from Sapiens (Khirodkar et al., 2026) encoder with the corresponding Gaussian features, enabling reliable cross-modal feature fusion.

Dual-level deformable attention As shown in Fig. 2, we draw inspiration from GeoLRM (Zhang et al., 2024a) but significantly adapt the deformable attention for articulated human modeling. To ensure pose-space image features are accurately anchored to the canonical 3D representation, LBS deforms the input Gaussians from canonical space to the target pose space of view v. Concurrently, Sapiens (Khirodkar et al., 2026) is employed to perform dual-level feature extraction on both the full image and a cropped face image, yielding hierarchical image features $\mathcal{F}^v_{\text{full}}$ and $\mathcal{F}^v_{\text{face}}$. An offset network further predicts per-Gaussian offsets $\mathcal{O}^v$ and attention weights $\mathcal{A}^v$ for both the full and face levels. They are explicitly conditioned on the Gaussian feature f_i and the corresponding image feature $\mathcal{F}^v(\pi(\mu_i))$ queried at the projected Gaussian location in pose space. To further enhance representational capacity, we extend the attention mechanism to a multi-head and multi-offset formulation. This design effectively broadens the spatial receptive field across diverse directions and scales, enabling the capture of fine-grained local 2D features. Finally, the per-Gaussians offsets and attention weights are used to aggregate locally aligned image features.

$$(\mathcal{O}^v_{i,h,o}, \mathcal{A}^v_{i,h,o}) = \text{OffsetNet}(\mathcal{F}^v(\pi(\mu_i)), f_i), \quad z^v_i = \frac{1}{|H| \times |O|} \sum_{h,o} \mathcal{A}^v_{i,h,o}\, \mathcal{F}^v\big(\pi(\mu_i) + \mathcal{O}^v_{i,h,o}\big) \tag{1}$$

Visibility-aware cross-view fusion Human avatars frequently self-occlude under varied poses, collapsing front- and back-facing 3D points onto the same 2D projection and making per-view features ambiguous. To resolve this, we design a visibility-aware cross-view attention to fuse dual-level, multi-view deformable attention features into Gaussian feature. We first estimate a soft visibility probability $\mathcal{M}^v_i$ for each Gaussian in view v by comparing its camera-space depth with a reference depth queried from a rendered SMPL-X depth map. Fusion then proceeds in two stages. In *level-wise fusion*, the face feature $z^v_{\text{face},i}$ is preferred over the full one $z^v_{\text{full},i}$ only when the Gaussian is more reliably visible at the face level (indicated by $\kappa^v_i = \mathbf{I}[\mathcal{M}^v_{\text{face},i} > 0.5 \vee \mathcal{M}^v_{\text{face},i} > \mathcal{M}^v_{\text{full},i}]$). Otherwise

the full-image feature is kept, yielding a merged feature z_i^v and visibility $\mathcal{M}_i^v$ for each view.

$$z_i^v = \kappa_i^v z_{\text{face},i}^v + (1-\kappa_i^v) z_{\text{full},i}^v, \quad \mathcal{M}_i^v = \kappa_i^v \mathcal{M}_{\text{face},i}^v + (1-\kappa_i^v)\mathcal{M}_{\text{full},i}^v, \tag{2}$$

In *view-wise fusion*, the Gaussian feature is updated via cross-attention with z_i^v. The attention is masked by $\mathcal{M}_i^v$, and a Gaussian invisible from all views receives zero attention so that no invalid residuals are injected.

Unlike global cross-attention, our module attends only to visible views and locally sampled deformable regions, making the cross-modal fusion occlusion-robust and efficient. Finally, an MLP head is employed to decode Gaussian primitives in canonical space based on updated Gaussian feature f_i, including the residual position relative to the position of Gaussians initialized from the canonical template $\Delta\mu_i^c$, rotation q_i, scale s_i, opacity o_i, and color c_i.

3.4 Cloth Joint Deformation Learning

To capture non-rigid clothing dynamics beyond standard LBS, we introduce a cloth joint-driven extended LBS, as shown in Fig. 3. This approach is based on the principle that clothing motion generally follows the body's dynamics but diverges with distance from the surface. We augment the body kinematic chain with learned cloth joints, enabling unified body and cloth deformation in a quasi-rigid manner. Specifically, the cloth joint positions and relative rotations are learned based on Gaussian feature and the target pose. The LBS transformation chain is extended to propagate motion sequentially: body joints → cloth joints → surface vertices. Finally, a deformation MLP refines high-frequency non-rigid residuals that cannot be explained by extended LBS.

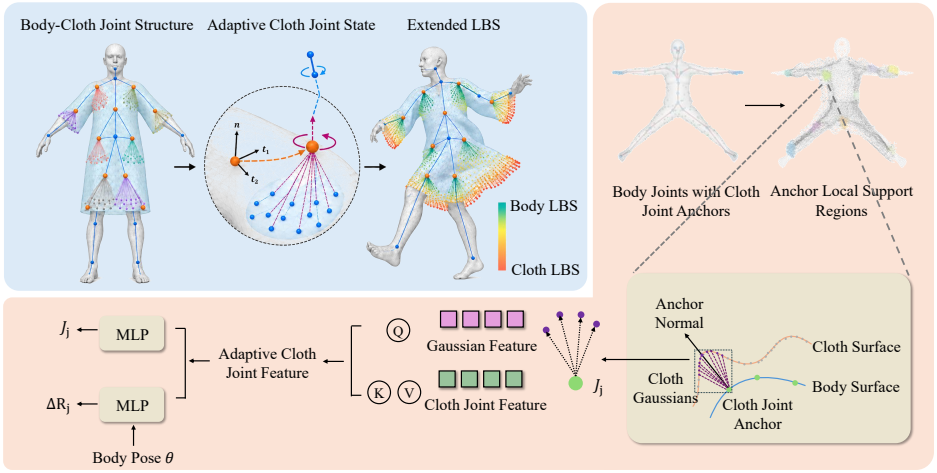


Figure 3: Adaptive cloth joint learning. Cloth joint anchors are initialized from the canonical SMPL-X mesh, and anchor support regions associate cloth Gaussians with corresponding cloth joints. Cross-attention aggregates Gaussian features within each region, while two MLP heads predict joint offsets and rotations relative to the parent body joints. Extended LBS further combines body- and cloth-joint influences, allowing cloth Gaussians farther from the body to rely more on cloth joints.

Adaptive cloth joint learning Our design of cloth joints relies on three core assumptions. (1) Cloth joints are constrained to the SMPL(-X) surface, serving as kinematic connectors between the inner body and the outer garment; (2) Their spatial locations dynamically adapt to varying clothing geometries; (3) Their rotation dynamics are hierarchically driven by the parent body joint decided by largest SMPL skinning weight.

Specifically, a set of cloth joint anchors is uniformly initialized from the canonical SMPL-X surface via Farthest Point Sampling (FPS). Inspired by ETCH (Li et al., 2025a), we associate each anchor with a localized support region on the canonical cloth template by casting rays along the anchor normals. Garment template points are grouped in anchor support regions based on how close they are to the anchor and how similar their surface normals are. Computed once in the canonical space, these correspondences remain fixed across identities and poses, endowing each cloth joint with a stable semantic domain.

Given the predicted cloth Gaussian primitives, we retrieve the Gaussian features within each joint's support region and aggregate them via cross-attention to yield an adaptive joint feature. Conditioning on this feature, we predict both the positional displacement of the joint along the SMPL-X surface relative to its canonical anchor and a residual rotation $\Delta R_j \in SO(3)$ relative to its parent body joint. Consequently, each cloth joint inherits the primary motion of the body joint while capturing garment-specific dynamics such as bending, twisting, and swinging. This design forms an extended body-to-cloth hierarchy that facilitates flexible control over loose garment deformations.

Extended LBS Given the updated position J_j and relative rotation ΔR_j of a cloth joint, the kinematic chain can be extended from body joints to cloth joints, enabling each cloth joint to inherit the underlying body motion while introducing a learned residual transformation for quasi-rigid garment dynamics. Eq. 3 reveals the connection between the rest-normalized skinning transform of cloth joint A_j^C and the parent body joint $A_{p(j)}^B$.

$$A_j^C = A_{p(j)}^B \, T_j^C \begin{bmatrix} \Delta R_j & 0 \\ 0 & 1 \end{bmatrix} (T_j^C)^{-1} \tag{3}$$

To smoothly integrate body- and cloth-joint deformation, we introduce a blend distance r_{blend} based on the distance γ_i from each cloth Gaussian to the SMPL-X body surface. Therefore, cloth Gaussians close to the body primarily inherit stable body motion, while loose regions farther from the body are progressively controlled by the learned cloth joints. More details on cloth joint deformation learning are in App. A.3.

$$\alpha_i = \text{clip}\left(\frac{\gamma_i}{r_{\text{blend}}}, 0, 1\right)^2, \quad \mu_i = (1-\alpha_i) \sum_{b \in \text{body}} w_{i,b}^B A_b^B \mu_i^c + \alpha_i \sum_{j \in \text{cloth}} w_{i,j}^C A_j^C \mu_i^c \tag{4}$$

Finally, to capture the high-frequency clothing dynamic such as wrinkles which cloth joints' rigid deformation may not handle, we integrate a deformation MLP to predict the position offsets of deformed Gaussians, rotations and scales after extended-LBS is applied in pose space.

3.5 Training Objectives

The overall training objective bridges photometric supervision with geometric regularization, jointly ensuring high-fidelity visual reconstruction and adaptive cloth motion control. The photometric supervision uses differentiable rasterization to optimize the rendered RGB images and masks against ground truth. The geometric regularization stabilizes the optimization process by constraining Gaussian offsets and scales, while enforcing non-rigid regularization on cloth joint locations and structural deformations. Details are in App. A.4.

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{2D}}}_{\text{image}} + \underbrace{\lambda_{\text{offset}} \mathcal{L}_{\text{offset}} + \lambda_{\text{scale}} \mathcal{L}_{\text{scale}} + \lambda_{\text{nonrigid}} \mathcal{L}_{\text{nonrigid}}}_{\text{geometry}}. \tag{5}$$

4 Experiments

4.1 Datasets and Implementation Details

Dataset preprocessing and metrics Our model is trained on two real-world datasets: 4D-DRESS (Wang et al., 2024) and THuman2.1 (Yu et al., 2021). During training, each batch takes 2 source views (frontal and back) together with 50K Gaussian points as input. These points are initialized from SMPL-X (Loper et al., 2015) and our fitted cloth template to serve as a geometric prior, and each batch is supervised by 6 target views. The rendered images are quantitatively evaluated against the ground-truth images using standard metrics: Peak Signal-to-Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS), and Structural Similarity (SSIM). Refer to App. B.3 for more dataset processing and split details.

Implementation Training is conducted in a distributed manner over 100 epochs using NVIDIA L40 GPUs, where both datasets are trained jointly. A two-stage strategy is employed to gradually increase the rendering resolution to learn coarse-to-fine details. Sapiens2-1B is used for image feature

Table 1: Quantitative comparison on THuman2.1 and 4D-DRESS. Our method achieves the best overall performance with fewer parameters and Gaussian primitives. Best and second-best results are shown in **bold** and underlined, respectively. For GHG, only in-shell Gaussians are counted because outer-shell Gaussians depend on garment geometry and visibility.

Method	#Params	#Gaussians	#Views	THuman2.1 PSNR ↑	THuman2.1 LPIPS ↓	THuman2.1 SSIM ↑	4D-DRESS PSNR ↑	4D-DRESS LPIPS ↓	4D-DRESS SSIM ↑
LHM	442M	20K	1	19.36	0.201	**0.807**	17.92	0.182	0.793
GHG	18M	>700K	3	20.06	0.251	0.801	17.56	0.196	**0.829**
NoPo-Avatar	636M	300K	3	18.60	0.254	0.709	16.12	0.290	0.687
GIGA	205M	260K	3-4	11.94	0.370	0.584	17.66	0.227	0.765
Ours	133M	50K	2	**20.53**	**0.198**	0.802	**19.61**	**0.166**	0.808

extraction. AdamW (Kingma, 2014) optimizer is applied with weight decay of 5×10^{-4}. The learning rate follows a cosine schedule, from 5×10^{-4} to 1×10^{-5} in the coarse stage and from 5×10^{-5} to 1×10^{-6} in the fine stage. See App. B.4 for other configurations. During training and inference, we require sparse-view images with known or estimated SMPL(-X) parameters as inputs. The cloth type can be automatically inferred from images to match a suitable prepared cloth template (App. B.2).



Figure 4: Qualitative comparison of FACt with other SOTAs.

4.2 Comparison with Existing Methods

We evaluate our method against four state-of-the-art (SOTA) feed-forward baselines: GHG (Kwon et al., 2024), GIGA (Zubekhin et al., 2026), LHM (Qiu et al., 2025a), and NoPo-Avatar (Wen et al., 2026). Specifically, GHG and GIGA predict UV-space-aligned Gaussian primitives directly from images, requiring the ground-truth mesh texture map to be pre-mapped onto the human template UV space. Furthermore, GHG is trained on a small-scale dataset without generalizability evaluation. LHM employs a canonical SMPL-X template as a geometric prior similar to our design, but relies on dense cross-attention between Sapiens-based image features and Gaussians to predict Gaussian primitives within the SMPL structure. NoPo-Avatar combines UV-space geometric priors with image-space features via cross-attention, maintaining multiple distinct sets of Gaussians that are later merged into an excessively large set (∼300K), which often introduces optimization redundancy.

For fair comparison, we evaluate all baselines using their pretrained checkpoints on our processed 4D-DRESS and THuman2.1 test sets, where the latter falls within the in-distribution training domain of GHG, GIGA, and NoPo-Avatar. Further baseline implementation details are provided in App. B.5.

Novel view synthesis The evaluation is conducted by feeding sparse-view images into each method to generate 3D human avatars. The reconstructed avatars are then rendered from 24 orbit target views for direct comparison with the ground truth. Fig. 4 presents qualitative comparisons with the selected baselines across both datasets, showing reconstruction results from source views. Specifically, GHG

and GIGA exhibit noticeable artifacts when reconstructing loose dresses in 4D-DRESS, as the fabric regions between the legs are mistakenly baked into the leg textures. NoPo-Avatar suffers from distinct rendering artifacts caused by under-optimized Gaussian scatterings, which may be related to its extremely large Gaussian sets to be optimized. Although LHM produces competitive results, it tends to blur fine details along garment boundaries. In contrast, our method faithfully reconstructs intricate garment geometry and preserves cloth boundaries across views, yielding visual results closest to the ground truth.

Quantitatively, in Tab. 1, our method achieves superior quantitative performance across THuman2.1 and 4D-DRESS. Specifically, on THuman2.1, our approach achieves a PSNR of 20.53, outperforming the strongest GHG and other feed-forward baselines while maintaining the lowest LPIPS (0.198). Furthermore, when evaluated on 4D-DRESS, our model delivers SOTA results with a PSNR of 19.61 and an LPIPS of 0.166. Compared to dense cross-attention mechanisms like LHM, NoPo-Avatar and GIGA, our approach achieves superior rendering quality under a significantly smaller network. This advantage stems from deformable attention, which selectively aggregates locally aligned image features to Gaussians and avoids the redundant computational overhead of global cross-attention. Additional results on novel view synthesis are provided in App. C.

Temporal pose evaluation We evaluate reconstruction quality on temporal pose clips from 4D-DRESS under two evaluation modes: *fixed source* (estimating canonical Gaussians and cloth joints only from the first frame) and *per frame* (re-estimating canonical assets for each frame). In Tab. 2, our method significantly outperforms LHM and GIGA in PSNR for both modes. Despite the additional difficulty of the *fixed source* setup, our method retains the highest PSNR and exhibits a performance drop comparable to other methods, supporting the temporal robustness of its canonical asset estimation. Fig. 5 further shows qualitative reconstruction on pose clips by comparing with LHM. Notably, our method produces clear cloth boundaries during motion benefiting from the cloth joint learning. Refer to App. C.1 for evaluation settings and detailed visual comparison.

Table 2: Reconstruction on temporal pose clips from 4D-DRESS.

Method	PSNR ↑	LPIPS ↓	SSIM ↑
Per Frame			
LHM	<u>17.60</u>	**0.195**	**0.795**
GIGA	15.56	0.202	0.791
Ours	**19.25**	<u>0.201</u>	<u>0.793</u>
Fixed Source			
LHM	<u>17.19</u>	**0.199**	**0.791**
GIGA	14.11	0.236	0.749
Ours	**17.95**	<u>0.210</u>	<u>0.773</u>

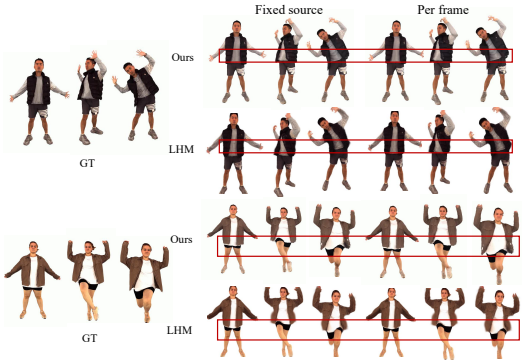


Figure 5: Temporal pose sequence reconstruction compared with LHM.

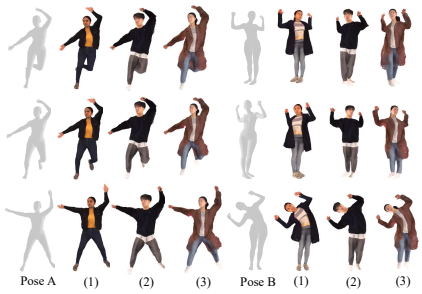


Figure 6: Novel pose synthesis. (1) poses extracted from 4D-DRESS samples; (2, 3) the driven THuman samples.

Novel pose synthesis We use pose series from 4D-DRESS to drive the motion of samples from THuman in Fig. 6. Our results visually demonstrate reasonable motion control on THuman samples in either tight or loose clothing. More examples are in App. C.3.

4.3 Ablation Study

We conduct ablation studies on the effectiveness of core modules, including dual-level deformable attention, visibility-aware cross-view fusion and cloth joint-based extended LBS.

Effect of core components In Tab. 3, the combination of deformable attention and cross-view fusion yields a modest PSNR improvement over the cross-attention baseline.

Notably, we isolate the deformation MLP from the extended LBS to explicitly disentangle the contribution of a residual non-rigid deformation from the quasi-rigid deformation provided by our extended LBS. The deformation MLP brings consistent gains under both settings, raising PSNR from 17.67 to 19.50 with standard LBS and from 18.67 to 21.03 with extended LBS. Combining the deformation MLP with our cloth template and extended LBS achieves the best performance across all metrics. This confirms that extended LBS provides a structural constraint that complements the residual deformation field to achieve high-fidelity garment reconstruction and motion synthesis.

Table 3: Ablation of core components on 4D-DRESS small split. Cross-attention and standard LBS replace deformable attention and extended LBS when disabled.

deformable attention	cross-view fusion	cloth template	extended LBS	deform MLP	PSNR ↑	LPIPS ↓	SSIM ↑
✗	✗	✗	✗	✗	17.51	0.231	0.740
✓	✓	✗	✗	✗	17.67	0.233	0.727
✓	✓	✗	✗	✓	19.50	0.216	0.771
✓	✓	✓	✓	✗	18.67	0.224	0.757
✓	✓	✓	✓	✓	**21.03**	**0.200**	**0.816**

Table 4: Ablation of cross-view fusion on THuman small split. Multi-view disabled and enabled denote single- and two-view inputs.

Multi-View	Cross-View Fusion	PSNR ↑	LPIPS ↓	SSIM ↑
✗	✗	22.19	0.172	0.830
✗	✓	21.78	0.184	0.811
✓	✗	22.22	0.175	0.830
✓	✓	**22.73**	**0.167**	**0.836**

Effect of cross-view fusion To validate the benefits of our visibility-aware cross-view fusion module, we conduct an ablation study under both single-view and two-view setups in Tab. 4. Combining two source views with cross-view fusion improves PSNR/LPIPS/SSIM by +0.54 / -0.005 / +0.006 over the single-view setting without fusion. However, simply adding a second source view without cross-view fusion provides only a marginal PSNR gain of 0.03. In Fig. 7, using two views enhances the capture of finer details, such as shadow variations, against the single-view input. The auxiliary views complement regions occluded or poorly observed in the primary view. Furthermore, our visibility gating filters out misleading features during fusion, effectively handling self-occlusion.

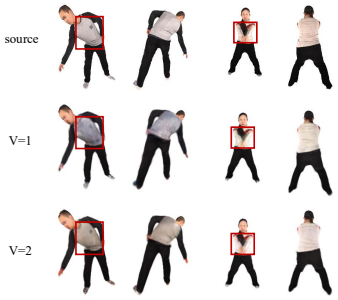


Figure 7: Ablation of source views.

Visualization of cloth joints We select some cloth joints and visualize their trajectories and rotations during motion in Fig. 8. We further render the LBS weights of both body and cloth joints via Gaussian rasterization, allowing us to inspect the region controlled by each joint set individually. Notably, loose garments exhibit high cloth LBS activity for non-rigid dynamics (*e.g.*, red regions between the legs of dresses and on flying outer coats), whereas tight-fitting cloth parts rely predominantly on body joints. Together, body and cloth joints provide complementary control over Gaussian primitives across cloth types.

5 Conclusion

We present FACt, a feed-forward framework for high-fidelity animatable clothed human reconstruction. Dual-level multi-view deformable attention maps locally aligned image features to 3D space, while SMPL-X depth-guided cross-view fusion exploits visible sparse-view inputs for stable canonical-space representation learning. We further propose adaptive cloth-joint learning to build a body-to-cloth kinematic chain. Combined with extended LBS, it enables interpretable control of low-frequency, quasi-rigid cloth dynamics, offering insights for future digital human modeling.

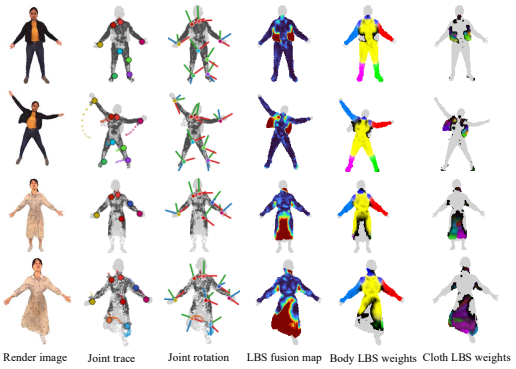


Figure 8: Visualization of cloth joints on 4D-DRESS subjects.

AI USE STATEMENT

In this work, we used generative AI tools to polish the textual expression and improve writing quality. They are also used to refine the pipeline figure design, summarize existing research literature, and engineering implementation. We have not used generative AI tools for core methodology design, critical mathematical formulation, or primary experiment design. We have thoroughly reviewed, verified, and edited all AI-assisted content for factual accuracy, consistency, and correctness. We take full responsibility for the final content of this work, including all written texts, images, videos and scientific claims.

ETHICS STATEMENT

This work strictly adheres to high standards of research integrity and ethical guidelines. All experiments conducted in this paper exclusively utilize publicly available human avatar datasets (e.g., THuman and 4D-DRESS). These datasets were obtained with proper permissions and are used in full compliance with their respective original licenses and terms of use for academic research. Our study does not involve new human subject data collection, personal identifiable information leakage, or privacy concerns. Furthermore, the methodologies presented have no foreseeable negative societal impacts or safety risks.

REPRODUCIBILITY STATEMENT

To ensure full reproducibility of our work, we provide comprehensive implementation and experimental details in the Appendix. Specifically, we elaborate on the hyperparameter configurations, network architectures, and dataset preprocessing protocols. In addition, the Appendix details the evaluation methodologies used for all baseline comparisons as well as the setup for extended comparative and ablation studies. Upon acceptance of the paper, we will release the complete codebase, data processing scripts, and pre-trained model weights to the research community to support further research.

REFERENCES

Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid icp algorithms for surface registration. In *CVPR*, 2007.

Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *ECCV*, 2024.

Marcel C Buehler, Ye Yuan, Xueting Li, Yangyi Huang, Koki Nagano, and Umar Iqbal. Dream, lift, animate: From single images to animatable gaussian avatars. In *International Conference on 3D Vision (3DV)*, 2026.

Zeyu Cai, Ziyang Li, Xiaoben Li, Boqian Li, Zeyu Wang, Zhenyu Zhang, and Yuliang Xiu. Up2you: Fast reconstruction of yourself from unconstrained photo collections. In *ICLR*, 2026.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, pp. 19457–19467, 2024.

Jinnan Chen, Chen Li, Jianfeng Zhang, Lingting Zhu, Buzhen Huang, Hanlin Chen, and Gim Hee Lee. Generalizable human gaussians from single-view image. In *ICLR*, 2024a.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *ECCV*, pp. 370–386, 2024b.

Aaron Ferguson, Ahmed AA Osman, Berta Bescos, Carsten Stoll, Chris Twigg, Christoph Lassner, David Otte, Eric Vignola, Fabian Prada, Federica Bogo, et al. Mhr: Momentum human rig. *arXiv preprint arXiv:2511.15586*, 2025.

Guénolé Fiche, Philippe Weinzaepfel, Romain Brégier, and Fabien Baradel. Multi-hmr 2: Multi-person camera-centric human detection, mesh recovery and tracking. In *ECCV*, 2026.

Chen Guo, Tianjian Jiang, Manuel Kaufmann, Chengwei Zheng, Julien Valentin, Jie Song, and Otmar Hilliges. Reloo: Reconstructing humans dressed in loose garments from monocular video in the wild. In *ECCV*, 2024.

Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *CVPR*, 2023.

Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *CVPR*, 2021.

Zhu Heming, Cao Yu, Jin Hang, Chen Weikai, Du Dong, Wang Zhangye, Cui Shuguang, and Han Xiaoguang. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *ECCV*, 2020.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.

Shoukang Hu, Tao Hu, and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *CVPR*, pp. 20418–20431, 2024.

Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *CVPR*, 2020.

Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *ECCV*, pp. 402–418, 2022.

Yudong Jin, Sida Peng, Xuan Wang, Tao Xie, Zhen Xu, Yifan Yang, Yujun Shen, Hujun Bao, and Xiaowei Zhou. Diffuman4d: 4d consistent human view synthesis from sparse-view videos with spatio-temporal diffusion models. In *ICCV*, pp. 11047–11057, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4):139–1, 2023.

Rawal Khirodkar, He Wen, Julieta Martinez, Yuan Dong, Zhaoen Su, and Shunsuke Saito. Sapiens2. In *ICLR*, 2026.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. In *CVPR*, pp. 505–515, 2024.

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 *ECCV*, pp. 451–468. Springer, 2024.

Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *CVPR*, pp. 19876–19887, 2024.

Boqian Li, Haiwen Feng, Zeyu Cai, Michael J Black, and Yuliang Xiu. Etch: Generalizing body fitting to clothed humans via equivariant tightness. In *ICCV*, 2025a.

Junxuan Li, Rawal Khirodkar, Chengan He, Zhongshi Jiang, Giljoo Nam, Lingchen Yang, Jihyun Lee, Egor Zakharov, Zhaoen Su, Rinat Abdrashitov, et al. Large-scale codec avatars: The unreasonable effectiveness of large-scale avatar pretraining. In *CVPR*, 2026.

Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE TPAMI*, 2020.

Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Xiaowei Chi, Siyu Xia, Yan-Pei Cao, Wei Xue, et al. Pshuman: Photorealistic single-image 3d human reconstruction using cross-scale multiview diffusion and explicit remeshing. In *CVPR*, pp. 16008–16018, 2025b.

Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, and Zhen Lei. High-fidelity clothed avatar reconstruction from a single image. In *CVPR*, 2023.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 34(6):248:1–248:16, 2015.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. Humansplat: Generalizable single-image human gaussian splatting with structure priors. In *NeurIPS*, pp. 74383–74410, 2024.

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*, 2022.

Hui En Pang, Shuai Liu, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu. Disco4d: Disentangled 4d human generation and animation from a single image. In *CVPR*, pp. 26331–26344, 2025.

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 *CVPR*, 2019.

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*, pp. 9054–9063, 2021.

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

Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, and Liefeng Bo. Lhm: Large animatable human reconstruction model from a single image in seconds. In *ICCV*, 2025a.

Lingteng Qiu, Peihao Li, Heyuan Li, Qi Zuo, Xiaodong Gu, Yuan Dong, Weihao Yuan, Rui Peng, Siyu Zhu, Xiaoguang Han, Guanying Chen, and Zilong Dong. Lhm++: An efficient large human reconstruction model for pose-free images to 3d. *arXiv preprint arXiv:2503.10625*, 2025b.

Lingteng Qiu, Peihao Li, Qi Zuo, Xiaodong Gu, Yuan Dong, Weihao Yuan, Siyu Zhu, Xiaoguang Han, Guanying Chen, and Zilong Dong. Pf-lhm: 3d animatable avatar reconstruction from pose-free articulated human images. *arXiv preprint arXiv:2506.13766*, 2025c.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 2017.

Jun Saito, Jiefeng Li, Michael de Ruyter, Miguel Guerrero, Edy Lim, Ehsan Hassani, Roger Blanco Ribera, Hyejin Moon, Magdalena Dadela, Marco Di Lucca, et al. Soma: Unifying parametric human body models. *arXiv preprint arXiv:2603.16858*, 2026.

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *CVPR*, 2019.

Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020.

Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. 360-degree human video generation with 4d diffusion transformer. *ACM Transactions on Graphics (ToG)*, 2024.

Jisu Shin, Junmyeong Lee, Seongmin Lee, Min-Gyu Park, Ju-Mi Kang, Ju Hong Yoon, and Hae-Gon Jeon. Canonicalfusion: Generating drivable 3d human avatars from multiple images. In *ECCV*, 2024.

Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *CVPR*, pp. 10208–10217, 2024.

Yihalem Yimolal Tiruneh, Muhammad Salman Ali, Uyoung Jeong, Muneeb A Khan, MD Khalequzzaman Chowdhury Sayem, Allanur Bayramgeldiyev, Binod Bhattarai, and Seungryul Baek. Flexiavatar: Unified 3d gaussian human avatars under arbitrary body visibility. In *ECCV*, 2026.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025a.

Rong Wang, Fabian Prada, Ziyan Wang, Zhongshi Jiang, Chengxiang Yin, Junxuan Li, Shunsuke Saito, Igor Santesteban, Javier Romero, Rohan Joshi, et al. Fresa: Feedforward reconstruction of personalized skinned avatars from few images. In *CVPR*, 2025b.

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 *CVPR*, 2024.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025c.

Jing Wen, Alex Schwing, and Shenlong Wang. Nopo-avatar: Generalizable and animatable avatars from sparse inputs without human poses. 2026.

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 *CVPR*, pp. 16210–16220, 2022.

Zhenzhen Weng, Jingyuan Liu, Hao Tan, Zhan Xu, Yang Zhou, Serena Yeung-Levy, and Jimei Yang. Template-free single-view 3d human digitalization with diffusion-guided lrm. *arXiv preprint arXiv:2401.12175*, 2024.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *CVPR*, pp. 4840–4851, 2024.

Junjin Xiao, Qing Zhang, Yonewei Nie, Lei Zhu, and Wei-Shi Zheng. Rogsplat: Learning robust generalizable human gaussian splatting from sparse multi-view images. In *CVPR*, pp. 5980–5990, 2025.

Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, et al. Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19801–19811, 2024.

Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *CVPR*, 2021.

Zhiyuan Yu, Zhe Li, Hujun Bao, Can Yang, and Xiaowei Zhou. Humanram: Feed-forward human reconstruction and animation model using transformers. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–13, 2025.

Yifan Zhan, Qingtian Zhu, Muyao Niu, Mingze Ma, Jiancheng Zhao, Zhihang Zhong, Xiao Sun, Yu Qiao, and Yinqiang Zheng. Tomie: Towards modular growth in enhanced smpl skeleton for 3d human with animatable garments. In *ICCV*, 2025.

Chubin Zhang, Hongliang Song, Yi Wei, Chen Yu, Jiwen Lu, and Yansong Tang. Geolrm: Geometry-aware large reconstruction model for high-quality 3d gaussian generation. In *NeurIPS*, pp. 55761–55784, 2024a.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.

Weitian Zhang, Yichao Yan, Sijing Wu, Manwen Liao, and Xiaokang Yang. Disentangled clothed avatar generation with layered representation. In *ICCV*, 2025.

Zechuan Zhang, Li Sun, Zongxin Yang, Ling Chen, and Yi Yang. Global-correlated 3d-decoupling transformer for clothed avatar reconstruction. *NeurIPS*, 36:7818–7830, 2023.

Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *CVPR*, pp. 9936–9947, 2024b.

Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *CVPR*, pp. 7743–7753, 2022.

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 *CVPR*, pp. 19680–19690, 2024.

Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3d human creation from a single image. In *CVPR*, pp. 26308–26319, 2025.

Anton Zubekhin, Heming Zhu, Paulo Gotardo, Thabo Beeler, Marc Habermann, and Christian Theobalt. Giga: Generalizable sparse image-driven gaussian humans. In *International Conference on 3D Vision (3DV)*, 2026.

APPENDIX

Supplementary Video is available at https://anonymous.4open.science/w/anonymous_2944/. It presents full 360° novel view synthesis, temporal sequences from four orthogonal viewpoints, novel pose synthesis, and visualizations of the learned cloth joints together with the resulting body–cloth LBS weight maps.

This appendix is organized as follows.

App. A Model Design Details includes fitting of the offline canonical cloth templates, the deformable transformer, the cloth joint deformation module with its extended-LBS formulation, and the training losses.

App. B Implementation Details includes Gaussian initialization, cloth template identification, dataset preprocessing, the two-stage training procedure, the evaluation setting, and the baseline configurations.

App. C Additional Visualization Results includes further qualitative comparisons with baselines, an analysis of the learned deformable-attention offsets, and LBS weight maps showing how body and cloth joints share control over the Gaussians.

App. D Efficiency Evaluation shows trainable parameters, per-iteration latency, and CUDA memory of our method against LHM and LHM++.

App. E Additional Ablations shows sensitivity of extended LBS to the number of cloth joint anchors, the cloth joint feature dimension, and the blend distance.

App. F Limitations and Broader Impacts discusses potential limitations of our method and the broader implications of digital human technology.

A MODEL DESIGN DETAILS

A.1 CLOTH TEMPLATE FITTING

We construct topology-consistent cloth base templates from DeepFashion3D (DF3D) (Heming et al., 2020), which provides registered garment meshes, SMPL poses, and cloth edge annotations. Garments are grouped into upper, pants, skirt, and dress categories, with finer subdivisions by sleeve and garment length (Fig. A1). Template construction comprises initialization, pose-space fitting and canonical fusion, and post-processing. For each garment type, we initialize a shell in canonical SMPL Vitruvian space. Upper garments and pants are initialized from expanded SMPL body regions, whereas skirts and dresses are initialized by inverse-LBS canonicalization of representative DF3D garments. Dresses additionally combine the canonicalized garment with a body-derived upper shell. The mesh topology remains fixed throughout fitting and fusion.

For each DF3D sample, we transform the template to pose space with forward LBS and align it to the target garment using NRICP (Amberg et al., 2007), with cloth edge annotations constraining characteristic boundaries. The fitted deformation is then mapped back to canonical space through inverse LBS. We average results across poses of the same garment and fuse them across garments in each category, iteratively using the fused template for the next fitting round. Finally, we filter disconnected components, smooth the surface, and resolve body–cloth intersections. The resulting templates provide stable canonical geometry for cloth Gaussian initialization (App. B.1).

A.2 DEFORMABLE TRANSFORMER

We adopt Point Transformer v3 (Wu et al., 2024) as the backbone of our deformable transformer, using four encoder blocks and three decoder blocks. In each block, our deformable attention module is inserted after the self-attention layer. As the encoder depth increases, we progressively expand the Gaussian feature embedding dimension from $48 \rightarrow 96 \rightarrow 192 \rightarrow 384$ and simultaneously increase the number of attention heads and learned offsets from $4 \rightarrow 8 \rightarrow 16$. These adjustments align with the downsampling hierarchy of the encoder and provide deeper layers with larger capacity and broader receptive fields. Conversely, the decoder mirrors this process in reverse: the embedding dimension is gradually reduced back to 48, and the number of heads and offsets is similarly decreased during

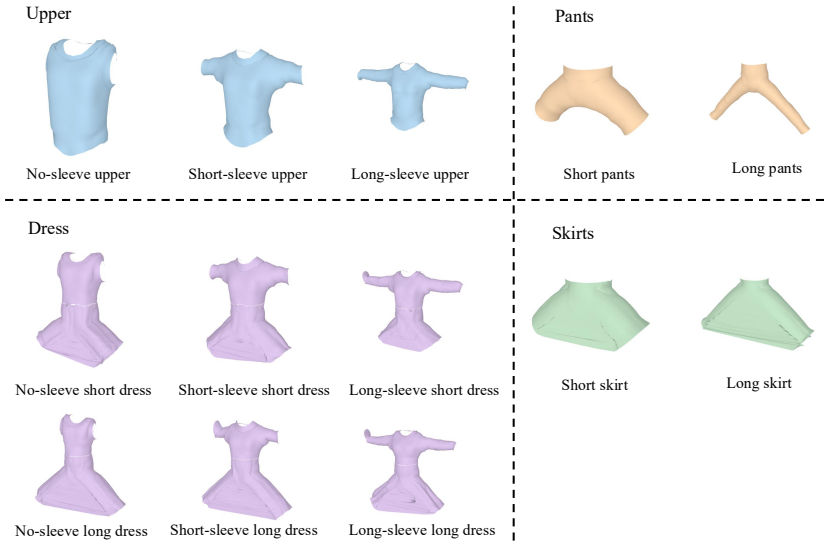


Figure A1: Category-specific cloth templates construction using DeepFashion3D. We define four garment categories (upper garments, pants, skirts, and dresses), with 13 template types determined by sleeve and garment length.

upsampling. To enhance feature extraction in the facial region, we employ the Sapiens segmentation model (Khirodkar et al., 2026) to crop faces from the original images, followed by super-resolution to a uniform 512×512 resolution.

A.3 Cloth Joint Deformation Learning

Cloth joint initialization We initialize cloth joint anchors on valid regions of the canonical SMPL-X body surface using FPS. Specifically, we place 200 candidate anchors uniformly across the SMPL body mesh, excluding the head, hands, and feet. Each anchor a_j is parametrized by its host triangle index and barycentric coordinates. The interpolated SMPL-X LBS weights at each anchor determine its fixed parent body joint via

$$\text{parent}(j) = \arg\max_b w^B_{j,b}. \tag{A6}$$

To establish reliable support regions on garments, inspired by ETCH (Li et al., 2025a), we cast rays along the anchor normals to intersect the canonical cloth template. Garment points are then grouped into support sets $\mathcal{S}_j$ based on spatial proximity and surface-normal consistency. Anchors without valid cloth correspondences are discarded for that specific garment category. These anchor-garment associations are computed once in the canonical space and remain fixed across identities and poses.

Cloth joint offset learning To learn the positional offset and relative rotation of each cloth joint with respect to its parent body joint, we construct a joint feature h_j by applying cross-attention over its support set $\mathcal{S}_j$. The geometric prior of each anchor j is defined as $g_j = (a_j, n_j, \tau_j)$, where a_j denotes its normalized coordinates, n_j its surface normal from the SMPL-X mesh, and τ_j its distance to the cloth template. For each Gaussian primitive $i \in \mathcal{S}_j$, its feature is represented as $z_{ij} = (\mu_i^c - a_j, f_i)$, combining its relative position $\mu_i^c - a_j$ with its intrinsic feature f_i. We treat the geometric prior g_j as the query, and the node features z_{ij} of the corresponding support set as keys and values to perform cross-attention, yielding the adaptive joint feature h_j. This joint feature incorporates both the canonical geometric priors and the current identity-specific features, from which the joint offset and relative rotation are subsequently decoded.

An MLP head predicts the positional shift p_j of the cloth joint relative to its initial anchor a_j. We further disentangle this shift into a direction unit vector d_j and a surface offset distance δ_j as

$$p_j = \phi_{\text{loc}}(h_j), \quad d_j = \frac{p_j}{\|p_j\| + \epsilon}, \quad \delta_j = \tanh(\|p_j\|). \tag{A7}$$

To ensure the cloth joints remain constrained to the SMPL-X body surface, we conduct geodesic surface walk along the SMPL-X body mesh. Starting from the initial face of the anchor a_j and its barycentric coordinates, we project the predicted direction d_j onto the local tangent plane, crossing face boundaries through shared edges with geodesic distance of δ_j. The surface point reached by this walk defines the updated joint position J_j.

Furthermore, we introduce a location regularization loss to penalize large surface displacements for cloth joints located near tight-fitting garment regions. Specifically, the local cloth tightness η_j of a joint j is computed as:

$$\gamma_i = \mathrm{dist}(\mu_i^c, \mathbf{M}_{\mathrm{SMPL}}), \quad \eta_i = \left[1 - \frac{\gamma_i}{r_{\mathrm{blend}}}\right]_+^2, \quad \eta_j = \frac{1}{|\mathcal{S}_j|}\sum_{i\in\mathcal{S}_j}\eta_i, \tag{A8}$$

where μ_i^c is the canonical position of cloth Gaussian i, $\mathbf{M}_{\mathrm{SMPL}}$ denotes the canonical SMPL-X body surface mesh, and $r_{\mathrm{blend}} \in (0, 1)$ is a hyperparameter controlling the effective blend radius. The location regularization loss is formulated as

$$\mathcal{L}_{\mathrm{loc}} = \frac{\sum_j \eta_j \delta_j^2}{\sum_j \eta_j + \epsilon}. \tag{A9}$$

Intuitively, a larger r_{blend} expands the region considered tight, thereby imposing a stronger penalty on joint offsets in loose regions.

Cloth joint rotation learning To learn the residual rotation of cloth joints relative to their parent body joints, we first compute the local body pose context $\rho_j = \sum_b w_{j,b}^B R_b$, where $w_{j,b}^B$ are the barycentrically interpolated LBS weights. Another MLP then predicts the residual axis-angle vector ω_j using the local pose context ρ_j, joint feature h_j, and positional displacement p_j:

$$\hat{\omega}_j = \phi_{\mathrm{rot}}(\rho_j, h_j, p_j), \quad \omega_j = \pi \tanh(\|\hat{\omega}_j\|)\frac{\hat{\omega}_j}{\|\hat{\omega}_j\| + \epsilon}, \quad R_j^{\mathrm{local}} = \exp([\omega_j]_\times). \tag{A10}$$

Each cloth joint j predicts its residual rotation within a local coordinate frame defined by the body surface at its updated location J_j. Let $F_j(J_j) = [t_j, b_j, n_j] \in SO(3)$ denote the local frame composed of the tangent, bitangent, and normal vectors on the body surface. The global residual rotation matrix ΔR_j is obtained via conjugation:

$$\Delta R_j = F_j(J_j) R_j^{\mathrm{local}} F_j(J_j)^\top. \tag{A11}$$

This formulation guarantees $\Delta R_j \in SO(3)$ and assigns a clear physical interpretation (bending and twisting relative to the body surface) to the predicted axis-angle vector ω_j. Initializing the rotation head to zero yields $\Delta R_j = I$ at the start of training, enabling the cloth joints to initially follow the body motion and progressively learn fine garment-specific dynamics.

Extended-LBS assignment and body-cloth blending We explicitly construct an extended kinematic chain that propagates motion seamlessly from body joints to cloth joints, and subsequently to cloth Gaussian primitives. For a cloth joint j anchored to a fixed parent body joint $p(j) = \mathrm{parent}(j)$, let $G_{p(j)}^B$ denote the global transformation of the parent joint derived from forward kinematics. The cloth joint first inherits this parent motion and then applies a residual rotation ΔR_j centered at its canonical position c_j, yielding its current global transform G_j^C.

Since LBS is defined relative to the canonical pose, each current transformation matrix must be normalized by its rest-pose counterpart. Let T_b^B and G_b^B be the rest-pose and current global transforms of body joint b, respectively, where the translational component of T_b^B corresponds to the canonical joint coordinate J_b^B. By evaluating G_j^C at $\Delta R_j = I$, the rest-pose transform of the cloth joint, denoted as T_j^C, is defined as

$$G_j^C = G_{p(j)}^B \begin{bmatrix} \Delta R_j & c_j - J_{p(j)}^B \\ 0 & 1 \end{bmatrix}, \quad T_j^C = T_{p(j)}^B \begin{bmatrix} I & c_j - J_{p(j)}^B \\ 0 & 1 \end{bmatrix}. \tag{A12}$$

At rest, the cloth joint undergoes only a translation to its canonical location c_j without relative rotation. The rest-normalized skinning matrices for body and cloth joints are then formulated as

$$A_b^B = G_b^B \left(T_b^B\right)^{-1}, \qquad A_j^C = G_j^C \left(T_j^C\right)^{-1}. \tag{A13}$$

This construction guarantees that $A_b^B = A_j^C = I$ in the canonical pose when $\Delta R_j = I$, thereby perfectly preserving the canonical cloth template. Under non-rest body poses, A_j^C isolates the local motion of the cloth joint relative to its parent body joint.

To determine which cloth joints drive a given cloth Gaussian primitive i, we compute pseudo-LBS weights over its K-nearest effective cloth joints $\mathcal{N}_K(i)$. We initialize a geometric prior score ξ_{ij} based on the squared Euclidean distance between the canonical Gaussian position μ_i^c and the joint position c_j. An MLP then predicts a residual offset δ_{ij} to refine this geometric prior:

$$\xi_{ij} = -\frac{\|\mu_i^c - c_j\|^2}{2\sigma^2}, \qquad w_{ij}^C = \mathrm{softmax}_{j\in\mathcal{N}_K(i)}\big(\xi_{ij} + \delta_{ij}\big). \tag{A14}$$

At the start of training, initializing $\delta_{ij} = 0$ ensures that the learnable pseudo-LBS strictly matches Gaussian spatial distance weighting.

Finally, the body LBS weights W_i^B and pseudo-LBS weights W_i^C are continuously blended according to the distance $\gamma_i = \mathrm{dist}(\mu_i^c, \mathbf{M}_{\mathrm{SMPL}})$ from the cloth point to the canonical SMPL body surface $\mathbf{M}_{\mathrm{SMPL}}$:

$$\alpha_i = \mathrm{clip}\left(\frac{\gamma_i}{r_{\mathrm{blend}}}, 0, 1\right)^2, \qquad W_i = \left[(1-\alpha_i)W_i^B,\ \alpha_i W_i^C\right]. \tag{A15}$$

The deformed position μ_i of cloth point i is obtained by applying the blended transformation:

$$\mu_i = (1-\alpha_i)\sum_{b\in\mathcal{B}} w_{i,b}^B A_b^B \mu_i^c + \alpha_i \sum_{j\in\mathcal{N}_K(i)} w_{i,j}^C A_j^C \mu_i^c. \tag{A16}$$

When a cloth point is adjacent to the body surface ($\alpha_i \approx 0$), its motion is strictly governed by rigid body skinning. For loose garment regions (e.g., skirts, skirts' hems, and wide sleeves), $\alpha_i \approx 1$, allowing local deformation to be dominated by the accumulated bending, twisting, and swinging motions of neighboring cloth joints via extended LBS. This continuous blending scheme eliminates the need for explicit classification or heuristic rules to delineate tight and loose garments, creating a unified and interpretable deformation chain: body joint $\rightarrow$ cloth joint $\rightarrow$ cloth point.

A.4 Loss Functions

Loss functions can be divided into photometric loss and geometric regularization loss. Photometric loss $\mathcal{L}_{\mathrm{2D}}$ is composed of RGB loss $\mathcal{L}_{\mathrm{rgb}}$, mask loss $\mathcal{L}_{\mathrm{mask}}$, and perceptual loss $\mathcal{L}_{\mathrm{lpips}}$. The geometric regularization loss is composed of regularizations on learned offsets and scales of Gaussian primitives, as well as regularizations on cloth joints and extended-LBS to ensure a reasonable cloth joint distribution and deformation.

Photometric loss RGB loss measures the L1 distance between the rendered and ground-truth images, as well as the L1 distance between the facial crop from the rendering and the super-resolved facial crop from the ground-truth image. Mask loss measures the L1 distance between rendered opacity and ground-truth foreground mask. Perceptual loss is based on a pretrained VGG network (LPIPS Zhang et al., 2018) applied between the rendered and ground-truth images.

Geometric regularization loss Similar to LHM (Qiu et al., 2025a), we apply $\mathcal{L}_{\mathrm{offset}}$ and $\mathcal{L}_{\mathrm{scale}}$ to constrain large drifting of Gaussians' positions from the sampled template during optimization, and Gaussians to stay near-spherical for rasterization. The cloth pseudo-joints are constrained by two coupled terms. First, we apply $\mathcal{L}_{\mathrm{loc}}$ to constrain cloth joints close to the body surface to be rigidly attached, as described in App. A.3. Second, the extended-LBS posed position of each cloth joint J_j^{pose} is required to stay consistent with the body LBS driven baseline J_j^{LBS} obtained by skinning the same joint with the underlying body weights only.

$$\mathcal{L}_{\mathrm{LBS}} = \frac{1}{N}\sum_j \left\| J_j^{\mathrm{pose}} - J_j^{\mathrm{LBS}} \right\| \tag{A17}$$

These two terms are combined as $\mathcal{L}_{\mathrm{nonrigid}} = \lambda_{\mathrm{loc}}\mathcal{L}_{\mathrm{loc}} + \mathcal{L}_{\mathrm{LBS}}$ with $\lambda_{\mathrm{loc}} = 0.01$, which keeps the garment deformation faithful to the body motion while allowing pose-dependent cloth details.

Overall, the loss functions during training can be expressed as

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{2D}}}_{\text{image}} + \underbrace{\lambda_{\text{offset}}\,\mathcal{L}_{\text{offset}} + \lambda_{\text{scale}}\,\mathcal{L}_{\text{scale}} + \lambda_{\text{nonrigid}}\,\mathcal{L}_{\text{nonrigid}}}_{\text{geometry}}. \tag{A18}$$

B Implementation Details

B.1 Gaussian Primitive Initialization and Encoding

For Gaussian initialization, we area-uniformly sample surface points on the canonical SMPL-X (Pavlakos et al., 2019) template: $15,000$ Gaussians are drawn from the body region (including hands and feet), and $5,000$ points are allocated to the head region to achieve a higher density for fine facial details. The cloth Gaussians are initialized from fitted canonical garment templates. Given the input image, we first identify suitable garment type among our fitted 13 templates (detailed in App. B.2). For the identified garment type, we uniformly sample the surface of the corresponding cloth template, splitting the total $30{,}000$ Gaussians into two halves when the garment is composed of separate upper and lower parts, and using the full set for dress garment type.

Each Gaussian primitive is parameterized by its center position $\mu_i^c \in \mathbb{R}^3$, an anisotropic 3D covariance encoded by a scale $s_i \in \mathbb{R}^3$, a rotation (unit quaternion) $q_i \in \mathbb{H}$, an opacity $o_i \in [0,1]$, and an RGB color $c_i \in [0,1]^3$, together with a discrete semantic label l_i and a unit surface normal $\mathbf{n}_i$.

During initialization, μ_i^c is set exactly to the template-sampled surface point, and the offset prediction head of Gaussian decoder is zero-initialized so that every Gaussian stays on the template at the first iteration. q_i is the identity quaternion, s_i is initialized to a small near-spherical value, and the opacity o_i is initialized to 0.1. The color is predicted directly as RGB via a sigmoid head without spherical harmonics.

The initial per-Gaussian feature is formed by the surface location and the surface normal at that location, *i.e.* $\mathbf{f}_i^0 = \left[\,\hat{\mu}_i^c,\, n_i\,\right] \in \mathbb{R}^6$, where n_i is the unit normal obtained by barycentric interpolation of the three vertex normals of the template face, and $\hat{\mu}_i^c \in [-1,1]^3$ denotes the normalized canonical position. A Fourier positional embedding of $\hat{\mu}_i^c$ is concatenated to n_i, and further projected by a shared MLP to the Deformable Transformer embedding width.

B.2 Cloth Template Identification

To obtain cloth Gaussians from the corresponding cloth template type summarized in Fig. A1, we need to identify the correct cloth template with a suitable category and sleeve length based on the input image. To achieve this, a human-parsing model (Li et al., 2020) first yields per-pixel segmentation, whose garment labels are collapsed into four families — upper, dress, pants, and skirt. The dominant cloth category is decided by pixel counts. In parallel, we partition the SMPL mesh into semantic regions (torso, hips, upper arms, forearms, upper legs, lower legs), project their vertices into each source view, and keep only those that are inside the image and consistent with a rasterized SMPL depth map, thereby discarding out-of-view and occluded samples. Each semantic region's garment coverage is further calculated to be the fraction of visible vertices covered by the candidate garment mask. Coverage of the forearms/upper arms yields long-sleeve, short-sleeve, or sleeveless, and coverage of the lower legs separates long vs. short pants, skirts, and dresses. Finally, the cloth category and the coverage-derived sleeve/length attributes are combined to output a cloth type (*e.g.* long-sleeve dress, short-sleeve upper).

B.3 Dataset Processing

Rendering Each frame scan is placed at the center of a fixed orbit and rendered from 24 evenly spaced viewpoints (15° apart in azimuth) at 1024×1024 resolution against a white background. The intrinsics and extrinsics of all 24 views are known and saved per sample.

4D-DRESS 4D-DRESS collection covers 32 identities wearing 61 subject–outfit combinations (31 inner-garment and 30 outer-garment outfits), yielding 487 pose sequences (takes). Each pose sequence contains 70–250 frames. Within each subject–outfit–pose sequence, frames are first decimated with a

minimum frame interval of 30, retaining on average 5.3 frames per pose sequence (ranging from 3 to 9). 4D-DRESS is split at the pose sequence level, where whole takes are assigned to either the training or the test set, and no take appears in both. The training set contains 2,085 frames from 390 takes spanning all 32 identities and all 61 outfits, while the test set contains 517 frames from 97 held-out takes covering 29 identities and 52 outfits.

THuman2.1 THuman2.1 provides 2,445 static scans of clothed human subjects, one posed scan per identity. All subjects are shuffled and divided at the subject level into 80% training and 20% test, *i.e.* 1,956 training identities and 489 test identities, and the two sets share no identities.

Source/target view grouping During training, we fix the two evenly spaced views, the frontal view and the back view, 180° apart on the orbit, as source views, from which the network extracts image features and fuses them with the Gaussian features to predict Gaussian primitives in the canonical space. The camera circle is enumerated in 4 rotational groupings of 6 evenly distributed target views (spaced every 60°), so each scan yields 4 disjoint source–target tuples that augment the effective training set. Target views do not participate in the Gaussian-primitive prediction. Instead, they are used in the rendering stage to provide supervision for the reconstruction of Gaussians through 2D photometric, mask, and perceptual losses.

B.4 Training Procedures

Table A1: Two-stage training configurations. S1 and S2 denote the coarse and fine stages, over which the two datasets are trained jointly. Batch size is summed over all GPUs.

	S1 (coarse)	S2 (fine)
deform MLP	✗	✓
image resolution	256×256	512×512
epochs	60	40
peak learning rate	5×10^{-4}	5×10^{-5}
min learning rate	1×10^{-5}	1×10^{-6}
warmup ratio	0.01	0
source / target views	2 / 6	2 / 6
batch size	8	8
Loss weights		
λ_{rgb}	1.0	
λ_{lpips}	0.2	
λ_{mask}	0.5	
λ_{offset}	1.0	
λ_{scale}	1.0	
$\lambda_{\text{nonrigid}}$	1.0	

We adopt a two-stage coarse-to-fine training scheme built upon a shared base architecture, which is applied to the joint training of 4D-DRESS (Wang et al., 2024) and THuman2.1 (Yu et al., 2021) datasets. In both stages, the source-view images are kept at a resolution of 1024×1024 for feature extraction and deformable attention, whereas the rendering resolution varies by stage. In Stage 1, the model renders images at a lower resolution of 256×256 with the pose-conditioned deformation MLP disabled. This coarse resolution accelerates training while reducing peak CUDA memory overhead. Furthermore, disabling the deformation MLP forces the model to learn cloth joints and kinematic deformations from scratch, preventing shortcuts through residual predictions. In Stage 2, we enable the pose-conditioned deformation MLP. Taking the root-excluded SMPL body pose alongside the canonical positions of points as input, this module predicts point-wise offsets $\Delta\mu^c$, Δs, and Δq on top of the Stage-1 extended-LBS deformation. Additionally, Stage 2 increases the rendered image resolution to 512×512 for finer synthesis. Primary training configurations are summarized in Tab. A1.

B.5 Baselines

We provide detailed configurations for the four baseline methods and elaborate on their evaluation setup on our processed 4D-DRESS and THuman2.1 datasets. As all baselines are feed-forward and generalizable, we directly evaluate their official pretrained checkpoints without fine-tuning. For quantitative evaluation, we follow the same metric calculation protocol as our method, computing PSNR, LPIPS, and SSIM within the bounding boxes of the foreground masks. All rendered images are set at a resolution of 512×512 for metric calculation. While each baseline utilizes a different number of source input views according to its default setup, target views are fixed to 24 views across all models for a fair comparison.

LHM LHM (Qiu et al., 2025a) is originally trained on dynamic single-person video sequences and synthetic human scans from 2K2K (Han et al., 2023), Human4DiT (Shao et al., 2024), and RenderPeople. We evaluate its pretrained `LHM-MINI` checkpoint on our datasets under a single-view reconstruction setting using the frontal view. Following the LHM pipeline, an auxiliary 112×112 head crop—extracted from the 1024×1024 source image via tracked FLAME/face bounding boxes—is fed into the specialized head branch.

GHG GHG (Kwon et al., 2024) is trained on THuman2.0 (Yu et al., 2021) and RenderPeople. Following its official implementation, 5-layer shell offset maps and UV visibility maps are pre-rendered at a resolution of 1024×1024 for both THuman2.1 and 4D-DRESS. Because GHG is not trained on 4D-DRESS, the structural discrepancy between the camera settings in THuman and 4D-DRESS induces a potential domain shift during cross-dataset evaluation. During inference, GHG takes three sparse context views as input. For THuman2.1, we use the default fixed camera indices $[0, 6, 11]$. For 4D-DRESS, we use non-uniform camera indices $[0, 9, 17]$ to better match the triangular input-view layout used during THuman training.

GIGA GIGA (Zubekhin et al., 2026) is originally trained on MVHumanNet (Xiong et al., 2024) and THuman2.1 (Yu et al., 2021). Multi-view RGB images are masked onto a white background, converted to RGBA, and resized to 512×512 alongside adjusted camera intrinsics. Following its original pipeline, we extract per-subject static A-pose textures by projecting multi-view images onto the SMPL-X UV layout. The model is conditioned on four input views evenly distributed around the camera rig (indices 00, 06, 12, and 18) and renders target avatars across all 24 views.

NoPo-Avatar NoPo-Avatar (Wen et al., 2026) is trained on THuman2.1 (Yu et al., 2021) and HuGe100K (Zhuang et al., 2025). We evaluate the released checkpoint configured for three-view 1024×1024 inputs trained on THuman2.1. Following its official release, the model takes three context views (camera indices 00, 08, and 16 along the 24-view ring) as input and renders novel views across all 24 camera positions.

C Additional Visualization Results

Figs. A2 and A3 present additional results on 4D-DRESS (Wang et al., 2024) and THuman2.1 (Yu et al., 2021). Across diverse identities, garments, and poses, FACt produces coherent geometry and appearance, while remaining effective for loose clothing with large non-rigid deformation.

C.1 Additional Qualitative Comparisons with Baselines

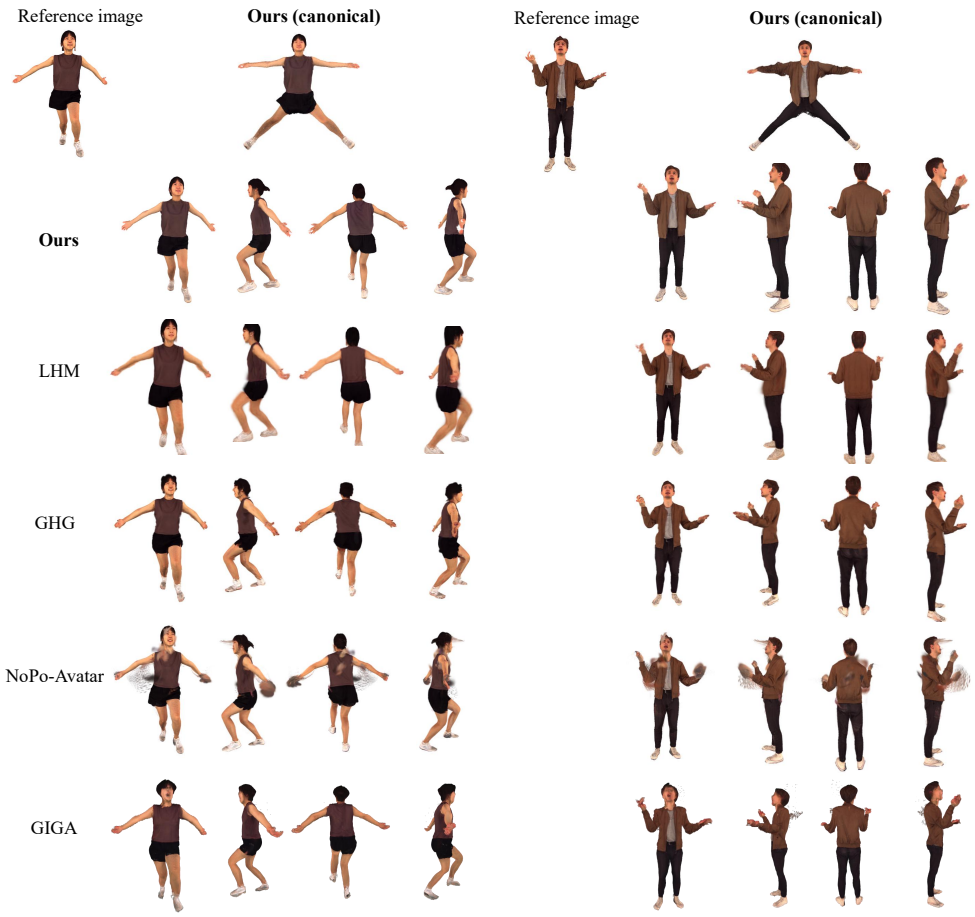
Comparisons with SOTAs on novel view synthesis In addition to Fig. 4, Fig. A4 compares novel-view synthesis on 4D-DRESS and THuman2.1 using target views different from the input observations.



Figure A2: Additional reconstruction results on 4D-DRESS. FACt reconstructs diverse subjects, garments, and poses, including loose clothing with substantial non-rigid deformation.

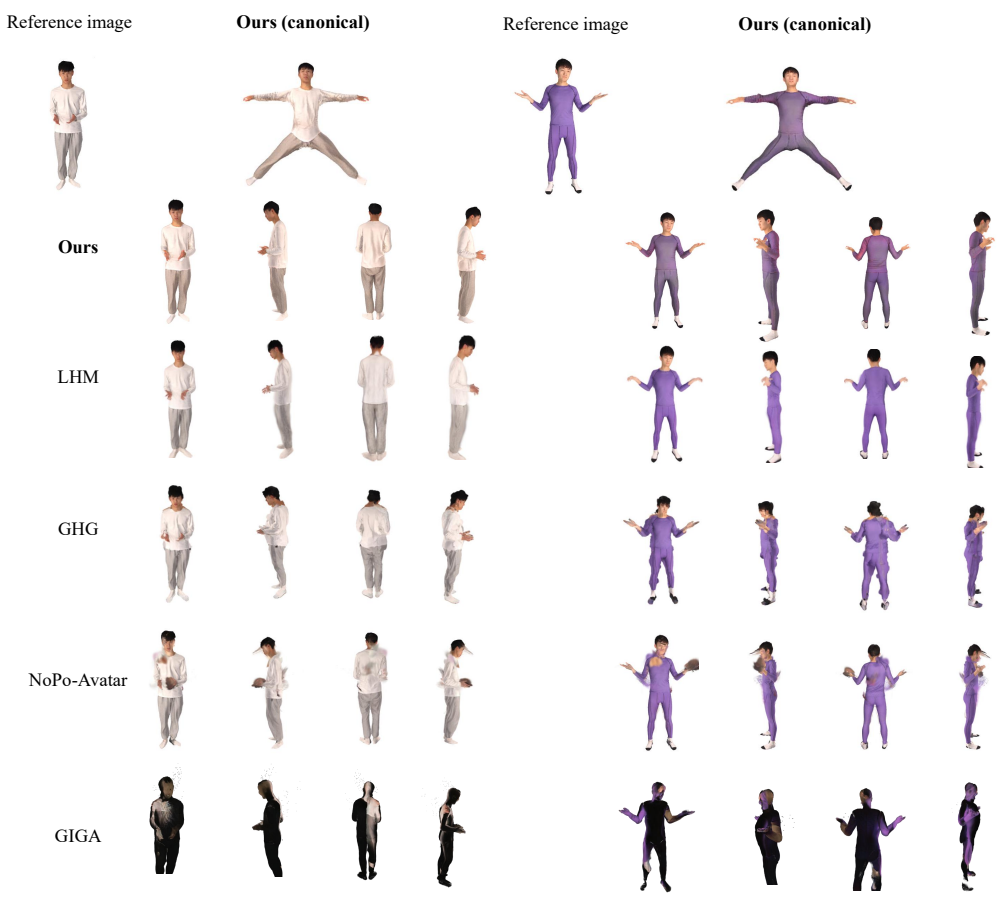


Figure A3: Additional reconstruction results on THuman2.1. FACt preserves the overall appearance and geometry of subjects with diverse identities and clothing.



(a) Samples from 4D-DRESS.

Figure A4: Additional qualitative comparisons on novel view synthesis. FACt produces coherent geometry and appearance in target views that are not included in the input observations, across both dynamic 4D-DRESS and static THuman2.1 subjects.



(b) Samples from THuman2.1.

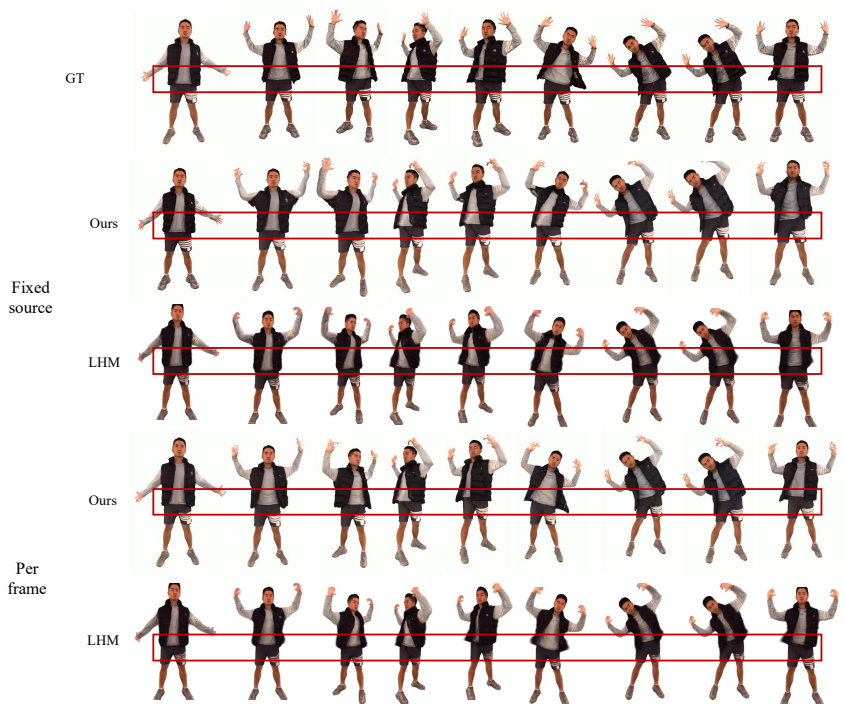
Figure A4: Additional qualitative comparisons on novel view synthesis (continued).

Comparisons with LHM on temporal pose reconstruction For the quantitative temporal pose evaluation in Sec. 4.2, different from the test set split mentioned in App. B.3 which is used for evaluation in Tab. 1, we set the sample frame rate to 1 to reflect the temporal consistency during reconstruction. For the sampled video clips from 4D-DRESS, we select 20 takes from the test set. Then within each take, we uniformly sample 3 clips of length $L = 32$ frames, with some temporal stride between the starting frames of each clip. During evaluation we consider two modes: *per-frame* and *fixed-source*. The latter is clearly harder, since the canonical reconstruction is inferred only once, on the first frame of each clip, and the resulting Gaussian features are then kept fixed. Only pose-dependent quantities are re-predicted per frame, including the body LBS transforms, and each cloth joint's position and residual rotation. Cloth Gaussians are finally skinned with these pose-conditioned joint transforms.

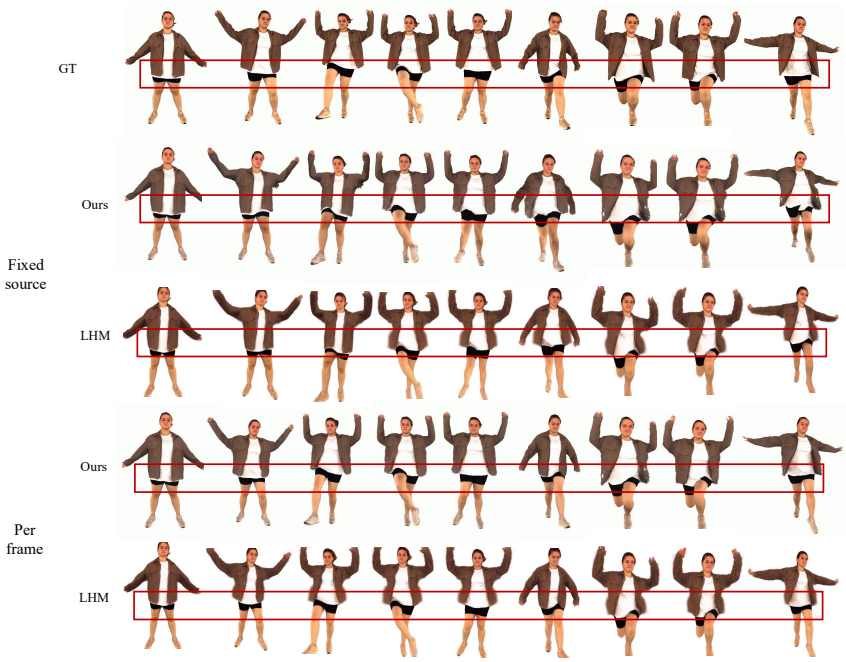
In addition to Fig. 5, we provide other time pose sequences in Fig. A5. It further confirms our benefit of applying cloth joints to obtain clearer cloth boundaries especially in dynamic motion scenarios.

C.2 Visualization of Offsets in Deformable Transformer

We visualize the learned offsets in our deformable attention module to probe how the offset network tracks geometry and localizes corresponding features in the image space. Specifically, we extract the offsets from the deformable attention module in the final decoder block (containing 2 attention heads and 2 sampling offsets per head) and convert the normalized offsets to pixel coordinates at the corresponding feature-map resolution. The distribution of learned offsets for each Gaussian primitive, categorized by semantic labels: full-image vs. face-image features, and cloth Gaussians vs. body



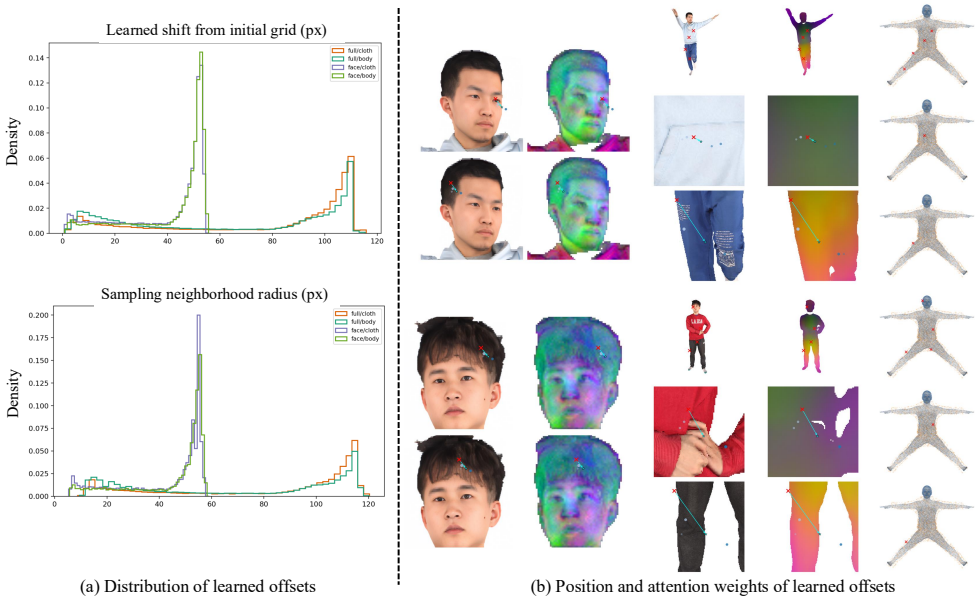
(a) Subject 1.



(b) Subject 2.

Figure A5: Qualitative temporal evaluation on 4D-DRESS. Each sequence compares ground truth (GT) and the performance of our method and LHM under the *Fixed Source* and *Per Frame* settings.

Gaussians is shown in Fig. A6 (a). Furthermore, we randomly sample a subset of Gaussians and plot their corresponding offset locations in the image space, as shown in Fig. A6 (b). To facilitate comparison across domains, source image features extracted by the Sapiens (Khirodkar et al., 2026)



(a) Distribution of learned offsets

(b) Position and attention weights of learned offsets

Figure A6: Visualization of learned offsets in deformable attention. (a) Offset distributions and mean neighborhood radii relative to the initial grid locations. (b) Sampling positions and attention weights in facial and full-body regions. Red crosses denote reference locations in pose space; cyan stars denote attention-weighted sampling centroids. Blue dots denote sampling locations, with size and opacity proportional to their attention weights. Sapiens (Khirodkar et al., 2026) feature maps provide the corresponding image features.

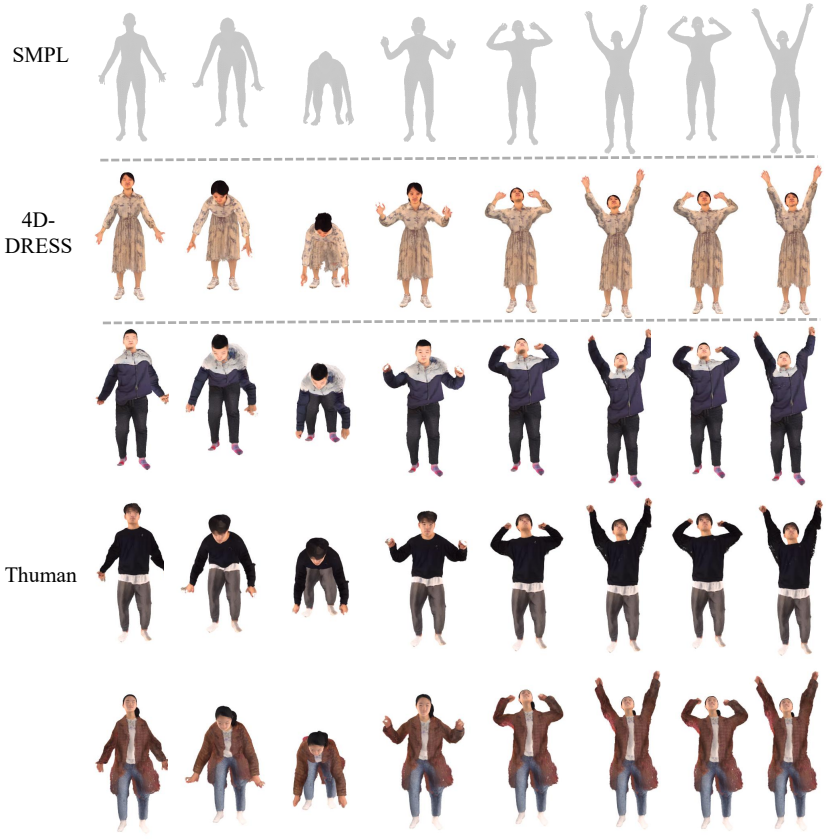
backbone are projected into an RGB space using PCA fitted on foreground pixels, allowing the same panel to be interpreted seamlessly in both image and feature spaces. Since the offsets of deformable attention are initialized across various ranges and directions across different heads and points, we plot their learned distributions relative to the initial grid to analyze the learning patterns. As expected, the offset magnitude is noticeably smaller for facial regions than for the full body, owing to the higher spatial density and localized nature of facial features. In the full-body scenario, the reference Gaussians sampled from the canonical template effectively track corresponding image features in pose space, demonstrating the meaningful geometric tracking process of our deformable attention. Nevertheless, some offsets fall outside the human foreground, particularly near body boundaries. In such cases, the attention mechanism dynamically reweights feature aggregation, prioritizing valid and informative offsets.

C.3 Additional Results of Novel Pose Synthesis

In addition to Fig. 6, Fig. A7 presents further THuman examples driven by poses from 4D-DRESS.

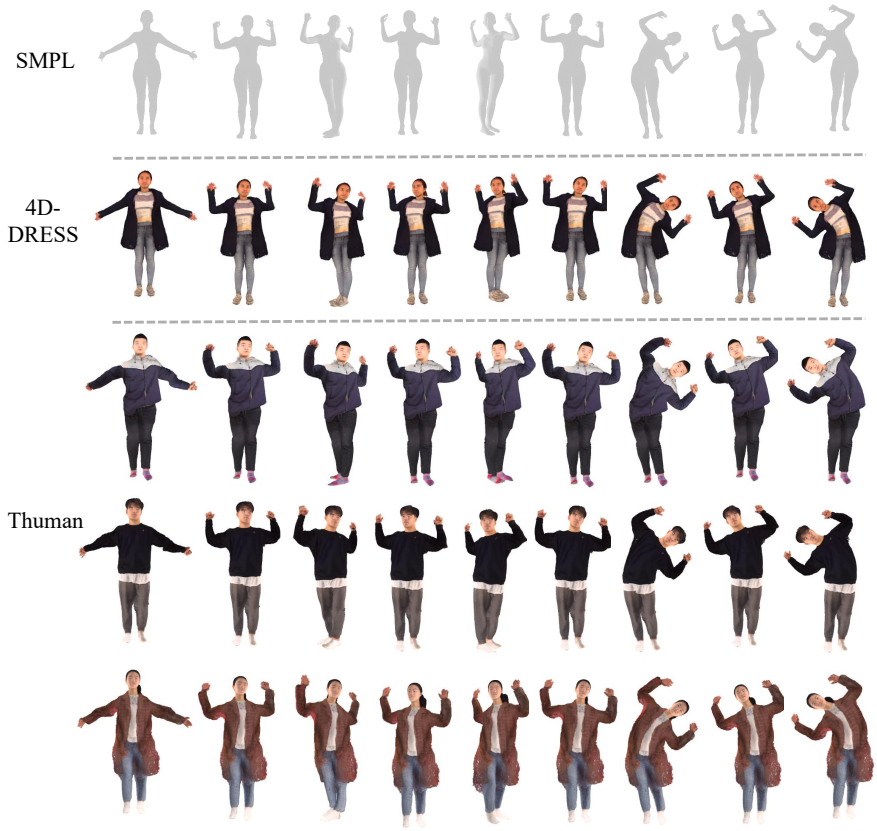


(a) Pose sequence 1.



(b) Pose sequence 2.

Figure A7: Additional results of novel pose synthesis. Given a THuman subject reconstructed from sparse input views, FACt synthesizes the subject under novel poses from 4D-DRESS, preserving identity, garment appearance, and plausible pose-dependent cloth deformation.



(c) Pose sequence 3.

Figure A7: Additional results of novel pose synthesis (continued).

C.4 VIDEO RESULTS

We provide supplementary video results to demonstrate the dynamic performance of our method on both 4D-DRESS and THuman2.1 datasets. The video includes novel view synthesis, temporal sequence visualization, and cloth-joint visualization. Details can be found in our video supplementary https://anonymous.4open.science/w/anonymous_2944/.

Novel view synthesis We render full 360° camera orbits (24 uniformly distributed viewpoints) for both static subjects in THuman2.1 and selected dynamic pose frames in 4D-DRESS to demonstrate our high-fidelity view synthesis.

Temporal sequence visualization Four orthogonal viewpoints (front, back, left, right) are provided for 4D-DRESS sequences to highlight both spatial view consistency and dynamic temporal motion.

Visual validation of learned cloth joints For each 4D-DRESS sequence, we perform per-frame inference to estimate canonical Gaussians and cloth-joint states. We select eight representative joints from 200 for visualization, requiring them to be valid in at least half of the frames and to have non-negligible influence. The selected joints cover all body parts and are spatially distributed using FPS, with a preference for higher-motion joints. We visualize their 3D trajectories in pose space and final orientations using RGB axes, which combine parent body-joint rotations with learned relative cloth-joint rotations. The LBS fusion map rasterizes α_i in Eq. 4 to show the relative influence of cloth joints. Loose regions, such as the area between the legs of a dress, receive higher cloth-joint influence, whereas tight garments remain primarily controlled by body joints. For the body LBS

map, we group the 22 SMPL joints into five anatomical regions and color each Gaussian according to $(1-\alpha_i)\sum_{b\in\mathcal{G}} w_{ib}^{B}$. For the cloth LBS map, each Gaussian is assigned the color of its dominant cloth joint, with brightness weighted by $\alpha_i \max_j w_{ij}^{C}$. Together, these maps visualize the complementary control of body and cloth joints.

D Efficiency Evaluation

Table A2: Efficiency evaluation. We compare trainable parameters, Gaussian primitive counts, input views, and the average per-iteration cost of training and inference, measured as wall-clock time and CUDA memory. Dashes denote settings that were not measured.

				Training		Inference	
Model	#Params	#Gaussians	#Views	Time (s) ↓	Memory (GB) ↓	Time (ms) ↓	Memory (GB) ↓
LHM-MINI	442M	20K	1	–	–	1.21	10.2
LHM-700M	700M	40K	1	–	–	5.73	19.6
			8	81.22	80.3	24.76	25.4
LHM++	700M	160K	1	–	–	0.91	6.78
			8	6.35	48.7	2.09	7.35
Ours	133M	50K	1	2.10	29.5	1.56	10.8
			2	2.88	38.0	2.18	11.1

Tab. A2 reports the parameter count, number of Gaussian primitives, input views, and measured training and inference cost for FACt, LHM, and LHM++. FACt uses 133M parameters, which is $3.3\times$ fewer than LHM-MINI and $5.3\times$ fewer than LHM-700M and LHM++, while using 50K Gaussians and supporting one- and two-view inputs.

FACt requires 2.10 s and 29.5 GB for one-view training, and 2.88 s and 38.0 GB for two-view training. These costs are lower than the reported eight-view training costs of LHM-700M and LHM++, although the view counts are not directly matched. For one-view inference, FACt is faster and uses less memory than LHM-700M, but LHM-MINI and LHM++ have lower latency and memory usage.

E Additional Ablations

Table A3: Ablation of cloth joint settings. We vary the number of joints, feature dimension, and blend distance. The blend distance controls the strength of cloth-joint influence in extended LBS.

joint num	feature dim	blend dist	PSNR ↑	LPIPS ↓	SSIM ↑
50	32	0.05	19.16	0.226	0.807
400	128	0.05	19.19	0.226	0.808
200	64	0.02	19.19	0.226	0.807
200	64	0.08	19.14	0.240	0.807
200	64	0.05	19.17	0.226	0.807

Effect of cloth joint in extended LBS We conduct an ablation study in Tab. A3 to analyze the key hyperparameter choices of the extended LBS module, specifically investigating the number of initialized cloth joint anchors, the dimension of cloth joint features, and the blend distance controlling the influence range of cloth joints. Notably, variations in these parameters lead to no significant performance differences, demonstrating that our adaptive cloth joint learning strategy is robust to hyperparameter fluctuations within a reasonable range.

F Limitations and Broader Impacts

Despite the promising results, there remains room for further enhancement in canonical space reconstruction accuracy, especially regarding the detailed finger structure. Incorporating a hand prior such as MANO (Romero et al., 2017) or applying the same super-resolution supervision as the face crop would be a natural remedy. Furthermore, our method can be adapted to synthetic data such as HuGe100K (Zhuang et al., 2025), and the cloth joints can be generalized from the SMPL kinematic chain to alternative human parametric templates such as SOMA (Saito et al., 2026) and MHR (Ferguson et al., 2025). Finally, our method is pose-dependent, which we regard as a deliberate

trade-off: an explicit pose supplies locally aligned and reliable spatial image features and stabilizes the learning of cloth-joint dynamics, and such pose information can be estimated reliably from images (Baradel et al., 2024; Fiche et al., 2026). By democratizing the creation of digital avatars, this work fosters innovation in virtual production and application in e-commerce, where realistic virtual try-ons can substantially decrease return rates. Furthermore, we recognize the ethical implications of this technology, particularly regarding identity privacy and the potential for synthetic media misuse. We advocate for the strict adherence to informed consent protocols for human digitalization, ensuring that the advancement of digital human technology remains socially responsible and inclusive.