AvatarDynamizer: FROM STATIC TO DYNAMIC HUMAN AVATARS VIA GENERATIVE DYNAMIC TEXTURES

Anonymous authors
Paper under double-blind review

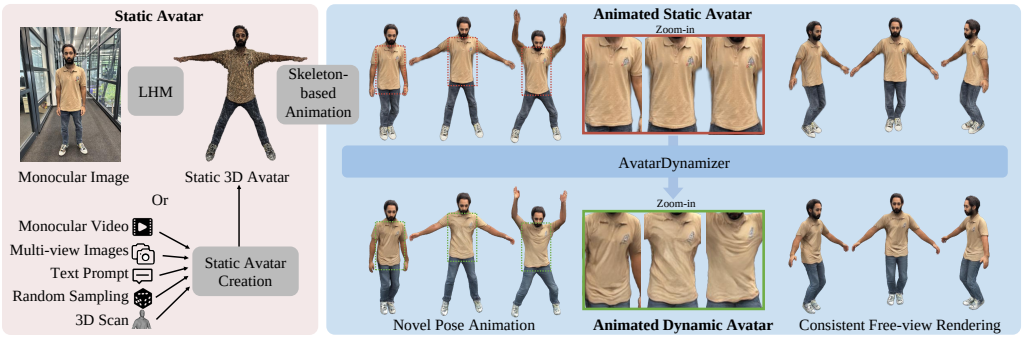


Figure 1: Given a 3D static avatar from any multimodal input, e.g., a monocular image with LHM Qiu et al. (2025a), we introduce a novel method, *AvatarDynamizer*, which converts static avatars into controllable 4D human avatars with pose-dependent surface dynamics (see the shirt).

ABSTRACT

Modeling pose-dependent surface dynamics is essential for realistic full-body avatars. *Person-agnostic* methods recover *static* 3D avatars from monocular images, videos, or text prompts, but their skeleton-driven animations lack realistic surface dynamics such as clothing wrinkles. In contrast, *Person-specific* methods capture such dynamics but require costly multi-view acquisition for each identity. *Generalizable dynamic* methods either produce limited surface dynamics or sacrifice multi-view consistency when introducing them. We present *AvatarDynamizer*, a generative method that transforms an off-the-shelf static avatar into a controllable 4D avatar with realistic surface dynamics and multi-view-consistent rendering. Our key insight is to represent pose-dependent surface dynamics as dynamic textures, bridging pretrained video diffusion models and 3D Gaussian splatting. Specifically, our *Dynamic Texture Generator* produces fine-grained, motion-dependent textures, which are decoded by a *Generalized Gaussian Decoder* into 3D Gaussians for inherently multi-view-consistent rendering. To bridge the gap between identity-rich but motion-limited datasets and motion-rich datasets with few subjects, we collect a high-quality multi-view dataset with dense camera coverage and diverse identities and motions. Experiments across four datasets show that our method outperforms competing generalizable dynamic methods on generative metrics and produces more faithful high-frequency surface dynamics.

1 INTRODUCTION

Creating faithful and controllable digital humans is a longstanding goal in Vision and Graphics, with broad applications in telecommunications, remote coaching, and virtual agents. Recent advances, supported by large-scale datasets Li et al. (2025a); Cheng et al. (2023), enable generalizable avatar reconstruction from sparse views Qiu et al. (2025a); Zubekhin et al. (2026) and generation from text or unconditional noise Liao et al. (2024); Dong et al. (2024); Zhang et al. (2024); Chen et al. (2023). These methods primarily produce *static* avatars that can be animated into arbitrary poses using an skeletal rigs. However, standard skinning-like deformations Lewis et al. (2000) are approximately piecewise rigid and cannot reproduce surface dynamics such as clothing wrinkles and induced shading, making the resulting animations appear unnatural and contributing to an uncanny-valley effect. In this work, we ask: *Can we introduce faithful dynamics to off-the-shelf static 3D avatars?*

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.

Traditionally, modeling avatar surface dynamics requires a manual graphics pipeline involving rigging, skinning, animation, cloth simulation, and rendering. Recent learning-based methods map body motion to dynamic geometry and appearance represented by radiance or SDF fields Zhu et al. (2024a), Gaussian splats Pang et al. (2024); Li et al. (2024); Junkawitsch et al. (2025), light fields Kwon et al. (2023), or conventional textures Liu et al. (2020). While achieving high rendering quality and fine-grained control, these methods require multi-camera capture and costly training for each identity, limiting cross-identity generalization.

More recently, generalizable 2D animation methods leverage foundation models to animate an identity image using pose-derived controls Xu et al. (2024); Chang et al. (2024); Zhu et al. (2024b). Some methods additionally use camera controls Shao et al. (2024); Wang et al. (2026), but their dynamics remain represented in 2D image space without an explicit shared 3D representation. Recent works explore cross-identity surface dynamics for 3D avatars Guo et al. (2025); Lu et al. (2025), with the key distinction being where the dynamics are embedded. GAS Lu et al. (2025) renders a static 3D avatar into 2D views and then uses video diffusion to generate surface dynamics in image space, without explicitly enforcing consistency across views. In contrast, Vid2Avatar-Pro Guo et al. (2025) embeds surface dynamics implicitly through a deterministic feed-forward mapping from identity and motion conditions to 3D Gaussians. Although 3D regression preserves multi-view consistency, it cannot directly exploit pretrained video priors and relies on comparatively limited multi-view 4D human data. Motion-dependent wrinkles are high-frequency and conditionally stochastic under the observed identity and motion signals, deterministic regression may average plausible surface states, whereas generative modeling is better suited to model their conditional distribution. Accordingly, the regression baseline Vid2Avatar-Pro produces over-smoothed or nearly static surface dynamics.

These observations suggest that the key bottleneck lies in how surface dynamics are represented and embedded. A desired representation should enable temporal surface-dynamics generation with pretrained video diffusion priors while supporting consistent 3D rendering. Inspired by latent diffusion Rombach et al. (2022), we explicitly represent pose-dependent surface dynamics as compact 2D dynamic textures within an encoder–decoder avatar representation. These dynamic textures serve as a temporally smooth generative intermediate representation, connecting pretrained video diffusion models with multi-view-consistent 3D Gaussian rendering.

More precisely, given a static 3D human avatar from diverse multi-modal sources, such as a monocular image, text prompt, 3D scan, or multi-view images, we aim to build a generalizable avatar dynamics model, dubbed *AvatarDynamizer*, which transforms the static avatar into a controllable 4D avatar with faithful surface and appearance dynamics (see also Fig. 1). At its core, our method generates motion-conditioned textures that encode pose-dependent surface dynamics.

Our method consists of three stages: 1) multi-view data encoding and *Generalized Gaussian Decoder* learning, 2) dynamics generator learning, and 3) inference with a novel static avatar. In the first stage, we encode multi-view videos as 2D texture videos by non-rigidly registering SMPL-X Pavlakos et al. (2019) to each frame and optimizing the corresponding colored dynamic texture maps. Using these textures, we train a cross-identity *Generalized Gaussian Decoder* to decode the dynamic textures and posed normal maps into 3D Gaussian splats under multi-view image supervision. Frame-wise decoder training allows us to combine multiple multi-view datasets for improved generalization to unseen identities (Tab. 2).

The second stage requires identity- and motion-diverse multi-view videos, but existing datasets are either identity-rich with limited motion or motion-rich with few subjects. To address this gap, we collect *DynaHuman*, comprising 58 actors, each captured for $27{,}000$ frames by 100 4K cameras. Using these data, we train our *Dynamic Texture Generator* to generate pose-dependent dynamic textures from a static identity texture and posed normal maps. We formulate this task as texture-space video generation, leveraging a pretrained video prior Blattmann et al. (2023) to complement the available multi-view 4D human data. At inference, we fit a novel avatar's coarse shape and identity texture, generate its motion-dependent textures, and decode them into dynamic 3D Gaussians.

Our contributions are threefold. First, we introduce a texture-space surface-dynamics embedding that represents 4D humans as compact 2D dynamic texture maps, together with a *Generalized Gaussian Decoder* for multi-view-consistent rendering. Second, we propose a *Dynamic Texture Generator* that leverages pre-trained video diffusion priors to generate highly detailed dynamic tex-

tures conditioned on identity and motion. Third, we present *DynaHuman*, a high-quality multi-view dataset featuring diverse identities, dense camera coverage, and substantial motion diversity.

2 Related Work

We review works on human representations, priors, and multi-view datasets for modeling diverse identities, body motions, and clothing dynamics. Head avatar methods Cao et al. (2013); Thies et al. (2016); Zhou et al. (2023); Qian et al. (2024); Rao et al. (2024); Tran et al. (2024); Li et al. (2025b); Zhou et al. (2025); Gerogiannis et al. (2025); Liu et al. (2025); Gao et al. (2025); He et al. (2025); Wu et al. (2026), which focus on facial expressions, skin materials, and hair, are beyond our scope.

2.1 Human Representations

Developing effective 3D human representations is a fundamental problem for human modeling, reconstruction, and rendering. Early methods commonly rely on mesh-based templates, often combined with motion capture techniques Stoll et al. (2011); Kanazawa et al. (2018), to improve mocap accuracy Joo et al. (2018); Hedlin et al. (2022) or capture human performances Habermann et al. (2020); Guo et al. (2021); Shetty et al. (2024). These representations are efficient and compatible with graphics rendering pipelines, but they struggle to capture fine-grained geometry, appearance details, and complex clothing deformation. With the development of neural rendering, diverse representations have been explored for human avatars, including neural fields Saito et al. (2019); Wang et al. (2021); Sun et al. (2021), point clouds Wu et al. (2020); Ma et al. (2022), NeRF Peng et al. (2021a); Kwon et al. (2021), NeuS Zeng et al. (2023); Sun et al. (2024), 3D Gaussians Hu et al. (2024); Li et al. (2024); Jiang et al. (2025b); Chen et al. (2025); Sun et al. (2025); Zhu et al. (2026) and transformer-based representations Yu et al. (2025). These methods achieve high-quality rendering and controllability through rigged skeletons. However, such 3D representations are usually designed for person-specific reconstruction or animation, making it non-trivial to integrate them with pre-trained 2D generative pipelines for generalizable, pose-dependent surface modeling. Recent works further investigate texture-based representations for human rendering and reconstruction Shetty et al. (2024); Sun et al. (2025); Alldieck et al. (2019). GIGA Zubekhin et al. (2026) unprojects sparse-view images into a texture map and predicts Gaussians in texel space, but it does not adopt a generative formulation. MoGA Dong et al. (2025) learns an implicit texture latent for human inversion; however, its pose-independent design restricts it to static human generation and prevents effective modeling of dynamic surface changes. In contrast, we interpret dynamic textures as an explicit, time-varying representation of surface dynamics that connects video diffusion with multi-view-consistent 3D Gaussian rendering.

2.2 Human Priors

Human priors are widely used to regularize optimization, improve reconstruction, and enable generalizable generation. In motion capture, prior works use hand-crafted constraints or learned statistical priors Stoll et al. (2011); Bogo et al. (2016); Pavlakos et al. (2019); Davydov et al. (2022) for robust body pose estimation. For image-based reconstruction, regression-based methods learn priors to predict neural fields Saito et al. (2019); Xiu et al. (2022); Alldieck et al. (2022), multi-view images Huang et al. (2025); Yu et al. (2025), normal maps Xiu et al. (2023); Li et al. (2025c), or 3D Gaussians Zhuang et al. (2025); Qiu et al. (2025a;b). While generalizing across identities and clothing, they mainly target static or frame-wise reconstruction and struggle with temporally coherent, pose-dependent surface dynamics. Diffusion-based monocular approaches such as AvatarPopUp Kolotouros et al. (2024) exploit generative priors for single-view avatar reconstruction, but controllable 4D modeling with temporal wrinkles and multi-view consistency remains challenging. Another line of work adopts image or video diffusion models to represent humans in 2D image space Wang et al. (2025); Xue et al. (2025); Xu et al. (2024); Zhu et al. (2024b); Chang et al. (2025); Li et al. (2025c); Huang et al. (2025). These methods benefit from strong identity generalization enabled by large-scale generative priors, but they often lack multi-view consistency. Toward consistent 4D avatar generation, HDA Cao et al. (2025) learns a generative model in the hyperspace of an animatable avatar representation Pang et al. (2024), though its quality is limited by the fitting-based preprocessing of network weights. Most closely related to our work, Vid2AvatarPro Guo et al. (2025) and GAS Lu et al. (2025) aim to model generalizable human surface and appearance

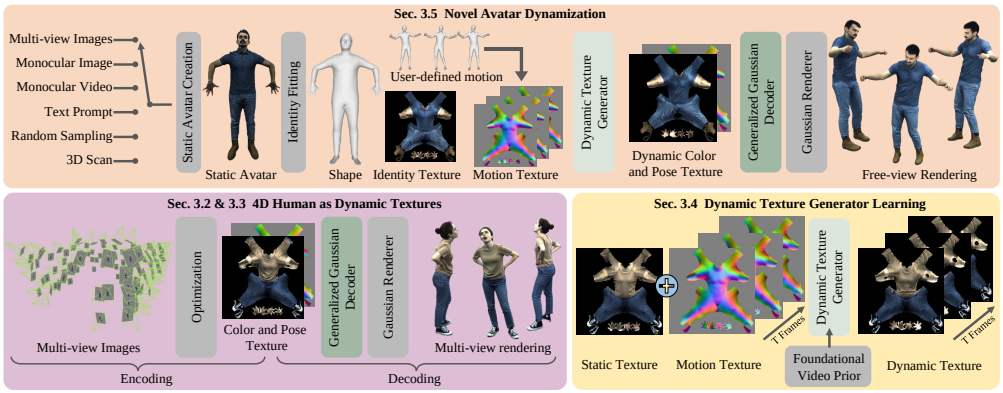


Figure 2: **Overview**. Given a static human avatar reconstructed from multimodal inputs, e.g., multi-view images, we first perform identity fitting to obtain its coarse shape and identity texture. Our *Dynamic Texture Generator* predicts dynamic texture maps conditioned on identity and novel pose. Then, our *Generalized Gaussian Decoder* recovers Gaussian splats for faithful and view-consistent renderings.

dynamics. Vid2AvatarPro Guo et al. (2025) conditions on both identity and pose for dynamic avatar reconstruction, but jointly learning identity and pose control through a regression model is challenging and often results in static appearances even when pose input is provided. Moreover, directly decoding 3D Gaussians makes it difficult to leverage pre-trained foundation video models. GAS Lu et al. (2025) learns video generation priors with a switcher to produce dynamic wrinkles for a static 3D avatar in image space, but this comes at the cost of multi-view consistency. In this work, we embed surface dynamics into texture space: a video diffusion prior models motion-dependent texture dynamics, while a reconstruction prior decodes the generated textures into 3D Gaussians for multi-view consistent rendering.

2.3 Human Datasets

High-quality multi-view human datasets are essential for learning generalizable and dynamic human avatars. Existing datasets mainly fall into two categories, each with distinct limitations for modeling both identity diversity and pose-dependent surface dynamics. On the one hand, dynamic datasets of individual subjects, such as DDC Habermann et al. (2021) and DUT Sun et al. (2025), capture detailed body motions and surface deformations. These datasets are valuable for learning precise dynamics, but they contain very few identities, which limits cross-subject generalization. Conversely, identity-rich datasets such as MVHumanNet++ Li et al. (2025a) and DNA-Rendering Cheng et al. (2023) provide diverse human appearances, clothing styles, and body shapes. However, they generally contain very limited motion ranges, making them insufficient for learning complex dynamic surface priors such as temporal cloth wrinkling. This dichotomy motivates us to leverage pre-trained 2D video diffusion priors to bridge the gap between large-scale identity diversity and realistic human dynamics, and to collect a large-scale dataset with diverse identities and body motions.

3 Method

We introduce *AvatarDynamizer*, which transforms an off-the-shelf static human avatar obtained from diverse input modalities, such as monocular, multi-view images or texts, into a controllable 4D avatar with pose-dependent surface dynamics, including wrinkles and self-shadowing (Fig. 2). To better embed 4D human dynamics, we represent multi-view human videos as compact dynamic textures via multi-stage optimization. We train our *Generalized Gaussian Decoder* to reconstruct texel-aligned 3D Gaussian splats from dynamic color and pose textures, supervised by multi-view images (Sec. 3.3). Representing each dynamic frame with 3D Gaussian splats enables inherently multi-view-consistent rendering. With the resulting dynamic textures, our *Dynamic Texture Generator* models dynamics learning as conditional video generation in texture space, using posed normal maps and identity textures as inputs (Sec. 3.4). This formulation captures the stochastic nature of surface dynamics and enables initialization from a pretrained video foundation model Blattmann et al. (2023). At inference, we fit the coarse shape and identity texture of the static avatar, generate

motion-dependent textures from the identity texture and posed normal map, and decode them into 3D Gaussians for novel-view rendering (Sec. 3.5). We next review the background (Sec. 3.1) and present our representation design (Sec. 3.2).

3.1 PRELIMINARIES

Latent Video Diffusion. LVD Blattmann et al. (2023) learns to generate videos by modeling the reverse diffusion process in a compressed latent space. A pre-trained 3D variational autoencoder (VAE) $\mathcal{E}(\mathcal{V}) = \mathbf{z}_0$ first embeds a T-frame video sequence $\mathcal{V} \in \mathbb{R}^{T\times H\times W\times 3}$ into a latent $\mathbf{z}_0$. With a predefined noise schedule, we compute a noisy latent $\mathbf{z}_t$ at timestep t as $\mathbf{z}_t = \alpha_t \mathbf{z}_0 + \sigma_t \epsilon$, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ denotes the Gaussian noise added to the latent representation. At each timestep t, a denoising network $\epsilon_\theta(\cdot)$ predicts the noise, and the training objective is formulated as:

$$\mathcal{L}_{LVD} = \mathbb{E}_{\mathbf{z}_0,\epsilon\sim\mathcal{N}(0,\mathbf{I}),t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{C})\|_2^2\right] \tag{1}$$

where $\mathbf{C}$ represents the conditioning signals, which could be text prompts, depth maps, or normal maps, and θ refers to the trainable network weights.

Texel Gaussian Avatar. We use the parametric model SMPL-X Pavlakos et al. (2019) as the base template $\bar{\mathbf{V}} \in \mathbb{R}^{N_\mathrm{V}\times 3}$ and bind 3D Gaussians to its texture Pang et al. (2024). The coarse shape is defined by the shape parameter $\boldsymbol{\beta}$, cloth displacements $\mathbf{d}_{\mathrm{cloth}}$ and hair displacements $\mathbf{d}_{\mathrm{hair}}$. Given a skeletal motion $\mathcal{M}$, i.e., joint rotations, as input, we animate it through LBS Lewis et al. (2000):

$$\mathbf{V} = T_{\mathrm{LBS}}(\mathcal{M}, T_{\mathrm{D}}(\bar{\mathbf{V}}, \boldsymbol{\beta}, \mathbf{d}_{\mathrm{hair}}, \mathbf{d}_{\mathrm{cloth}})), \tag{2}$$

where we deform the base mesh in the canonical pose using $T_{\mathrm{D}}(\cdot)$. For each valid texel i of the texture map, we parameterize the corresponding Gaussian as $\mathcal{T}_{\mathcal{G}_i} = \{\boldsymbol{\mu}_i, \mathbf{h}_i, \mathbf{s}_i, \mathbf{r}_i, \boldsymbol{\alpha}_i\}$, including 3D positions in world space, $\boldsymbol{\mu}_i = \mathbf{T}_{\mathrm{LBS}}(\mathcal{M}, \mathbf{T}_{\mathrm{D}}(\cdot) + \boldsymbol{\delta}_i)$, spherical harmonics, scaling, rotations, and opacity values, respectively. $\boldsymbol{\delta}_i$ denotes a motion-dependent per-Gaussian displacement that is later predicted by our *Generalized Gaussian Decoder* alongside the other Gaussian parameters. Given a virtual camera view matrix $\mathbf{W}$, Gaussians are rendered into a 2D image with a tile-based 3DGS rasterizer Kerbl et al. (2023): $\mathcal{R}(\mathcal{G}, \mathbf{W}) = \hat{\mathcal{I}}$.

3.2 EMBEDDING 4D HUMANS AS DYNAMIC TEXTURES

Motivation. For faithful and controllable 4D human avatars, how surface dynamics are represented and embedded is crucial. Existing representations either lack compatibility with powerful diffusion priors Guo et al. (2025); Dong et al. (2025) or sacrifice multi-view consistency Xue et al. (2025); Lu et al. (2025). We therefore seek a representation that supports generative modeling while enabling consistent 3D rendering. Inspired by latent diffusion Rombach et al. (2022), we represent surface dynamics as compact 2D dynamic textures and decode them into 3D Gaussians using our *Generalized Gaussian Decoder*. Unlike GIGA Zubekhin et al. (2026), which uses textures to fuse multi-view RGB inputs without explicit non-rigid alignment, we treat dynamic textures as an explicit, time-varying representation of surface dynamics. They are optimized from multi-view videos, generated under motion control, and decoded into 3D Gaussians for consistent rendering. This shifts texture maps from reconstruction features to a generative dynamics embedding, allowing us to leverage video diffusion priors Rombach et al. (2022); Blattmann et al. (2023). Unlike image-space generation Xue et al. (2025); Lu et al. (2025), decoding the generated textures into 3D Gaussians preserves multi-view consistency.

Dynamic Texture Embedding. Embedding multi-view videos into dynamic textures involves multi-stage optimization. For clarity, we omit the temporal smoothing used during batch optimization. Given multi-view images $\{\mathcal{I}^k\}_{k\in[1,N]}$ captured from N camera views with paired foreground segmentations $\{\mathcal{S}^k\}_{k\in[1,N]}$ of each frame, we first reconstruct point cloud geometry $\mathbf{p}$ Wang et al. (2023) and run 2D marker detection Cao et al. (2019) and triangulation to recover 3D markers $\mathbf{m}$. With $\mathbf{p}$ and $\mathbf{m}$ as reference, we optimize the pose parameter $\mathcal{M}$ and shape parameter $\boldsymbol{\beta}$ of SMPL-X Pavlakos et al. (2019), yielding coarse geometric tracking. To capture fine textures, we further optimize time-varying cloth displacements $\tilde{\mathbf{d}}_{\mathrm{cloth}}$ in the canonical pose on top of the coarsely tracked SMPL-X template, with $\mathbf{p}$ as the reference geometry. Next, we recover the RGB texture map $\mathcal{T}$

through differentiable rasterization $\mathcal{DR}$ Laine et al. (2020) from multi-view images:

$$\min_{\mathcal{T}} \sum_{k=1}^{N} \left\| \hat{\mathcal{I}}_k - \mathcal{DR}(\mathbf{V}(\mathcal{M}, \boldsymbol{\beta}, \mathbf{d}_{\text{hair}}, \tilde{\mathbf{d}}_{\text{cloth}}), \mathcal{T}, \mathbf{W}_{\mathbf{k}}) \right\|_2^2 \tag{3}$$

Thus, human information is encoded in textures $\mathcal{T}$ for all subjects and frames.

3.3 Generalized Gaussian Decoder

With the multi-view data compressed into dynamic textures, the decoder takes them as input to reconstruct 3D Gaussian primitives and render images accordingly. Specifically, the Generalized Gaussian Decoder Φ_{GGD} takes the texture map $\mathcal{T}$ and the root-normalized normal map $\mathcal{T}_{\bar{\text{N}}}$ as inputs and predicts the texel-aligned 3D Gaussian parameters

$$\mathcal{T}_{\boldsymbol{\mathcal{G}}} = \Phi_{\text{GGD}}(\mathcal{T}, \mathcal{T}_{\bar{\text{N}}}) \, . \tag{4}$$

We supervise Φ_{GGD} with multi-view image losses, including L1, SSIM Wang et al. (2004), and IDMRF Wang et al. (2018), and geometric regularization on the displacements of the Gaussians:

$$\mathcal{L}_{\text{GGD}} = \mathcal{L}_{\text{L1}} + \lambda_{\text{SSIM}}\mathcal{L}_{\text{SSIM}} + \lambda_{\text{IDM}}\mathcal{L}_{\text{IDM}} + \lambda_{\text{Reg}}\mathcal{L}_{\text{Reg}}. \tag{5}$$

Φ_{GGD} therefore lifts our texture-space surface-dynamics embedding to 3D Gaussians, enabling video diffusion priors to model dynamics in texture space while preserving multi-view consistency after decoding.

3.4 Dynamic Texture Generator

Our texture-space embedding compresses 4D human surface dynamics as dynamic RGB texture videos, providing a compact generative space that can directly leverage pre-trained video diffusion priors. Next, we introduce the *Dynamic Texture Generator* Φ_{DTG} to model motion-dependent surface dynamics directly in texture space. At this stage, 3D Gaussians are not involved; instead, the generated dynamic textures are later decoded by the *Generalized Gaussian Decoder* for multi-view consistent rendering. To faithfully model the wrinkles on the dynamic textures, unlike Vid2Avatar-Pro Guo et al. (2025), which models the prior through regression, we formulate our generator Φ_{DTG} as a skeletal motion-conditioned and identity-conditioned video generation task in texel space. Specifically, we adopt a stable video diffusion model Blattmann et al. (2023) as the base model and condition **character identity** on the RGB texture map $\mathcal{T}_{\text{id}}$ capturing the character appearance in 'A pose', and **character motion** on a stack of normal maps $\mathcal{T}_{\bar{\text{N}}}$ rendered from the SMPL-X template mesh with the root translation and rotation factored out. Under this conditioning scheme, we formulate the denoising prior model Φ_{DTG} as

$$\epsilon_\theta(\mathbf{z}_t, t, \mathbf{C}) = \Phi_{\text{DTG}}(\mathbf{z}_t, t, \mathcal{E}(\mathcal{T}_{\text{id}}), \mathcal{E}(\{\mathcal{T}_{\bar{\text{N}}}^j\}_{j\in[1,T]})), \tag{6}$$

which is trained with the denoising loss in Eq. 1 using dynamic textures from multiple subjects. Once we obtain generated dynamic textures $\{\hat{\mathcal{T}}^j\}_{j\in[1,T]}$ that describe the character under the given motions, we feed them into the Φ_{GGD} to reconstruct the 4D dynamic Gaussian Splats.

3.5 Novel Avatar Dynamization

Existing methods reconstruct static avatars from text Liao et al. (2024), monocular or multi-view images Zhuang et al. (2025); Qiu et al. (2025a); Pang et al. (2024), videos Guo et al. (2023), or 3D scans Zheng et al. (2022). We dynamize them in the following steps.

Identity Fitting. Given a static avatar, we render multi-view images of it and follow Sec. 3.2 to optimize rest body pose $\mathcal{M}$, shape parameter $\boldsymbol{\beta}$, cloth displacements $\mathbf{d}_{\text{cloth}}$, hair displacements $\mathbf{d}_{\text{hair}}$, and identity texture $\mathcal{T}_{\text{id}}$. **Dynamic Texture Generation.** We feed the novel identity texture $\mathcal{T}_{\text{id}}$ and control poses $\{\mathcal{T}_{\bar{\text{N}}}^t\}_{t\in[1,M]}$ to the Φ_{DTG} to generate dynamic textures $\{\hat{\mathcal{T}}^t\}_{t\in[1,M]}$. For arbitrary-length inference, we use sliding-window blending of overlapping VAE latents for temporal consistency. **Avatar Rendering.** Our Φ_{GGD} decodes $\{\mathcal{T}_{\bar{\text{N}}}^t\}_{t=1}^M$ and $\{\hat{\mathcal{T}}^t\}_{t=1}^M$ into 3D Gaussians, which are rendered with the camera parameters using the Gaussian rasterizer Kerbl et al. (2023).

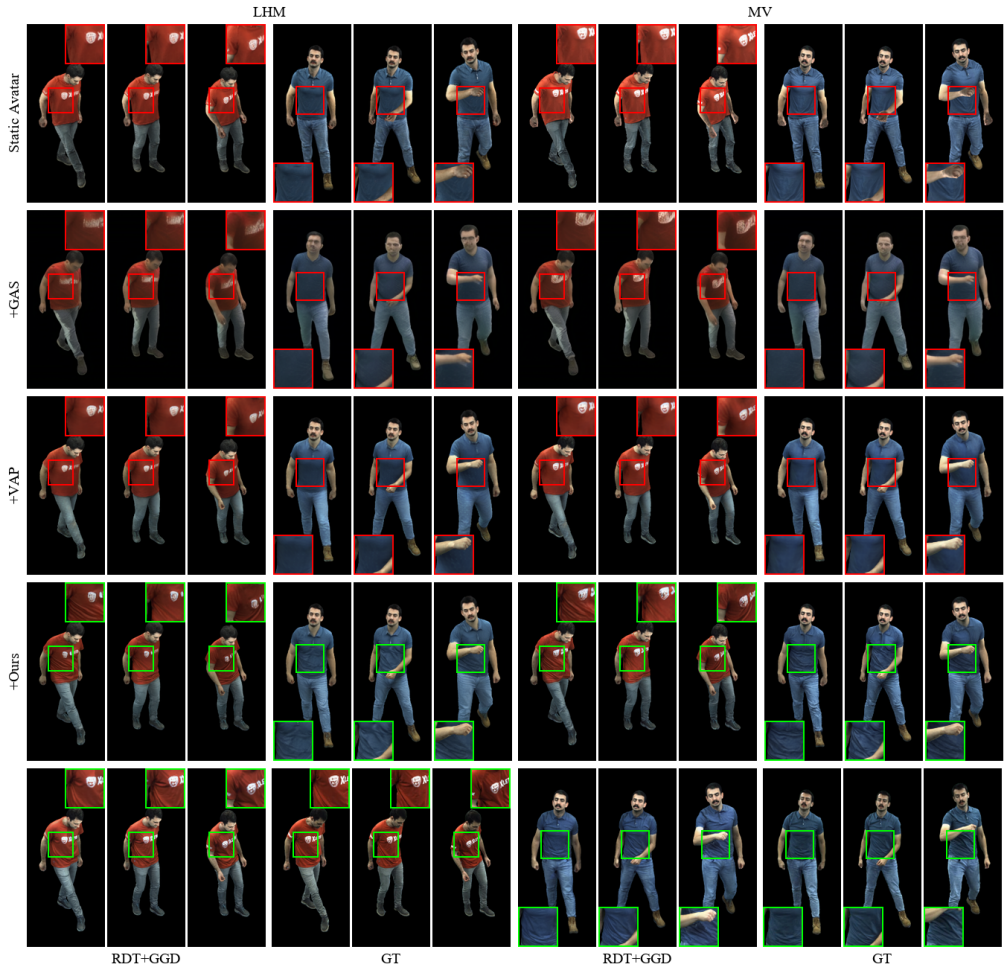


Figure 3: **Qualitative Comparison on DynaHuman**. Given a static avatar reconstructed from a monocular image using LHM Qiu et al. (2025a) or from multi-view images (MV), we compare our method with state-of-the-art generalizable surface dynamics methods. Our AvatarDynamizer produces faithful wrinkles while preserving both 3D geometry and identity consistency.

4 RESULTS

Datasets. We collected a large-scale multi-view human dataset named *DynaHuman*, which consists of 58 subjects. Each subject is captured for 27,000 frames by 100 cameras at 4K resolution. In terms of motion diversity, our dataset substantially surpasses existing ones. We isolate two subjects for evaluation. For the Φ_{GGD}, we use 56 subjects from *DynaHuman* and 478 subjects from MVHPP Li et al. (2025a) for better identity and clothing generalization. We use registered texture maps from *DynaHuman* to train the Φ_{DTG} because of the temporal smoothness and diversity of identity and motion. In addition to *DynaHuman*, we also evaluate our method on public datasets, DeepCap Habermann et al. (2020), DDC Habermann et al. (2021), and MetaCap Sun et al. (2024) as reported in Tab. 1. For each dataset, we use two subjects and compute the average metric scores.

Implementation Details. For texture embedding, we first optimize SMPL-X parameters for each frame using the point cloud and 3D markers, and then refine them by processing 120-frame chunks with 20-frame overlap. For deformation optimization, the sequence is divided into 20-frame chunks with an overlap of 4 frames. The texture resolution is set to 256, and we uniformly sample 30 to 32 cameras for texture optimization. For the Φ_{GGD}, we first train it on MVHPP for 1M steps, and then fine-tune it for another 600k steps using 90% from *DynaHuman* and 10% from MVHPP at 2K resolution. We use AdamW Loshchilov & Hutter (2019) with batch size 1 and learning rate 10^{-4}. For Φ_{DTG}, we use LoRA Hu et al. (2022) to fine-tune Hu et al. (2022) the 1.3B Wan video diffusion model Wan et al. (2025) with batch size 4 and learning rate 10^{-4}.

Table 1: **Quantitative comparison across datasets** using LHM (left block) and MV (right block) as static avatars. Results are averaged over two novel subjects. Higher is better for PSNR and SSIM, and lower for LPIPS, FID, and FVD. The best and second-best results are highlighted.

Dataset	Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓	Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓
DeepCap	LHM	30.61	0.715	0.201	50.10	150.12	MV	30.86	0.721	0.191	39.27	142.18
	+GAS	30.14	0.700	0.248	57.14	223.82	+GAS	30.48	0.705	0.242	54.28	222.78
	+VAP	31.31	0.730	0.221	101.73	206.88	+VAP	32.64	0.752	0.197	73.47	178.08
	+Ours	31.19	0.723	0.200	48.69	160.75	**+Ours**	31.95	0.735	0.185	37.05	134.26
DDC	LHM	29.65	0.696	0.206	42.64	108.00	MV	29.52	0.698	0.189	36.14	103.10
	+GAS	29.40	0.676	0.281	44.32	167.20	+GAS	29.64	0.680	0.279	44.35	164.61
	+VAP	30.81	0.729	0.224	77.90	199.85	+VAP	31.97	0.750	0.206	64.51	165.77
	+Ours	30.70	0.716	0.199	34.97	104.72	**+Ours**	31.26	0.728	0.186	30.43	93.95
MetaCap	LHM	28.11	0.704	0.173	80.19	80.19	MV	30.06	0.719	0.157	23.24	69.69
	+GAS	29.07	0.685	0.275	100.37	100.36	+GAS	29.33	0.689	0.273	44.50	99.48
	+VAP	30.88	0.727	0.199	115.71	115.71	+VAP	32.28	0.751	0.177	43.68	106.36
	+Ours	30.36	0.713	0.177	25.68	60.14	**+Ours**	31.46	0.728	0.158	18.31	48.59
DynaHuman	LHM	25.32	0.719	0.193	31.13	75.39	MV	25.20	0.727	0.181	25.83	52.79
	+GAS	25.74	0.723	0.254	50.03	124.51	+GAS	25.97	0.725	0.248	49.53	107.15
	+VAP	26.81	0.751	0.211	39.34	79.01	+VAP	27.64	0.771	0.190	37.51	54.97
	+Ours	26.43	0.729	0.202	22.58	56.72	**+Ours**	26.56	0.736	0.186	15.28	37.00

Evaluation Metrics. We use PSNR, SSIM Wang et al. (2004), and LPIPS Zhang et al. (2018) to evaluate the image reconstruction quality. To evaluate generative quality at both the image and video levels, we report FID Heusel et al. (2017) and FVD Unterthiner et al. (2018).

4.1 Comparison

To examine our central hypothesis, we compare methods according to where surface dynamics are embedded: no learned dynamics for static avatars, image space for GAS Lu et al. (2025), a deterministic regression network for VAP Guo et al. (2025), and dynamic texture space for our method.

Competing Methods. *Static methods:* LHM Qiu et al. (2025a) reconstructs a static human avatar from a monocular image, while MV denotes a static Gaussian avatar optimized from multi-view images. Both static avatars are animated using skeletal motion as described in Sec. 3.1. *Dynamic methods:* GAS is a video diffusion-based model that first renders a static avatar to 2D images and then generates surface details in image space. Vid2Avatar-Pro (VAP) is a regression-based method that maps identity and pose information to 3D Gaussians in texel space. We evaluate all dynamic methods using both LHM and MV static avatars.

Tab. 1 shows quantitative comparisons on four datasets, and Fig. 3 visualizes qualitative comparisons on DynaHuman. Compared with LHM, MV generally improves perceptual and generative metrics by leveraging additional input views. GAS generates surface dynamics in image space, which can introduce identity and cross-view inconsistencies, resulting in generally weaker generative performance. VAP achieves strong reconstruction metrics but produces smooth, nearly uniform wrinkles with limited high-frequency, motion-dependent details. Our method achieves the best FID and FVD among the dynamic methods across all datasets and input settings, ranks second in reconstruction quality, and consistently produces more faithful high-frequency surface dynamics in qualitative comparisons. Because generation is probabilistic, plausible high-frequency dynamics may differ from the ground truth and be penalized by reference-based reconstruction metrics. In contrast, regression-based methods such as VAP Guo et al. (2025) tend to conservatively predict low-frequency appearance changes, yielding strong reconstruction metrics but weak surface dynamics. For this generative task, FID and FVD complement reconstruction metrics by capturing image- and video-level generative quality.

4.2 Ablation Studies

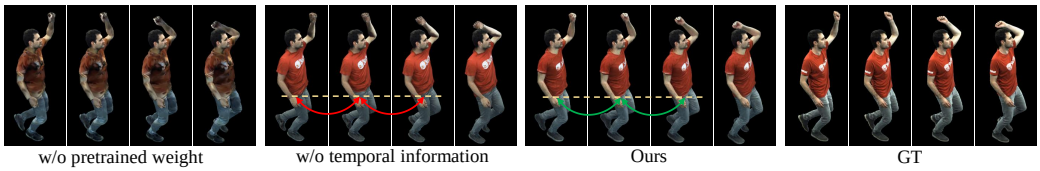
Dynamic-Texture Representation Capacity. To assess representation capacity independently of generation, we compare VAP Guo et al. (2025) with our optimized reference dynamic textures (RDT), decoded by the GGD, on one training subject (S59). While VAP struggles to recover high-frequency surface dynamics, RDT explicitly embeds these dynamics in texture space and achieves better reconstruction: 32.02 PSNR, 0.874 SSIM, and 0.103 LPIPS, compared with 30.36, 0.808,

Table 2: **Quantitative Ablation on Training Datasets of Generalized Gaussian Decoder.** Training with both datasets results in improved generalization and higher reconstruction accuracy.

Train Set	Test Set	PSNR	LPIPS	FVD
w/o DynaHuman	DynaHuman	28.77	0.142	33.60
w/o MVHPP	DynaHuman	29.70	0.124	25.77
Ours	DynaHuman	29.71	0.122	22.97
w/o DynaHuman	MVHPP	29.45	0.190	-
w/o MVHPP	MVHPP	27.22	0.213	-
Ours	MVHPP	29.51	0.190	-

Table 3: **Quantitative Ablation on Dynamics Prior** with Subject71. Without pretrained diffusion weight prior, our *Dynamic Texture Generator* has limited performance. Removing the temporal information results in degraded temporal performance.

Baselines	PSNR	SSIM	LPIPS	FID	FVD
w/o PP	24.70	0.700	0.240	60.42	228.21
w/o TI	26.61	0.748	0.187	16.91	55.20
Ours	26.61	0.745	0.182	13.84	**37.74**
w/ RDT	29.89	0.847	0.116	8.47	20.73



w/o pretrained weight | w/o temporal information | Ours | GT

Figure 4: **Qualitative Ablation on Dynamics Prior.** Training the *Dynamic Texture Generator* from scratch without pretrained diffusion weights yields limited quality. Removing the temporal block reduces temporal consistency and causes frame-to-frame variations (red arrows). The full model achieves the best generation quality and temporal consistency.

and 0.160 for VAP. RDT requires only 110 MB of texture storage, compared with 67.6 GB for the original multi-view data. Together with Sec. 4.1, these results suggest that dynamic textures provide a compact 2D space for generating fine-grained, motion-dependent dynamics, whereas direct Gaussian regression tends to smooth them.

Generation vs. Reconstruction. To disentangle dynamic-texture generation from Gaussian decoding, we replace the generated textures with optimized RDT while keeping the Φ_{GGD} fixed. As shown in Tab. 3, RDT improves all metrics and serves as an upper bound on the reconstruction capacity of the Φ_{GGD}.

Training-Data Composition. We evaluate the Φ_{GGD} trained on MVHPP only, *DynaHuman* only, or both datasets. MVHPP Li et al. (2025a) offers more identities but less motion variation and slightly lower image quality, whereas *DynaHuman* provides higher-quality captures and richer motion diversity while covering fewer identities. Joint training achieves the best results across both test sets, indicating that the datasets provide complementary training signals (Tab. 2).

Pretrained Video Model Weights. Given the limited availability of dynamic multi-view human data, we evaluate the benefit of pretrained video priors by training the *Dynamic Texture Generator* from scratch under the same setting (w/o PP). Without pretrained weights, both reconstruction and generative quality degrade substantially, as shown in Tab. 3 and Fig. 4. This validates a central advantage of our texture representation: compatibility with pretrained video priors.

Temporal Model vs. Image Model. To evaluate temporal modeling, we set the temporal window to 1 for the *Dynamic Texture Generator*, removing temporal information (TI). While per-frame quality remains similar, temporal quality degrades substantially, as reflected by the FVD in Tab. 3. The model also exhibits abrupt changes in cloth appearance, as highlighted in Fig. 4, demonstrating the benefit of temporal video modeling.

5 Conclusion

We presented *AvatarDynamizer*, a method that transforms off-the-shelf static avatars into controllable 4D avatars. Our main insight is to embed surface dynamics into dynamic textures, bridging video diffusion models and 3D Gaussian splatting. This design leverages pretrained video priors to generate fine-grained, motion-dependent details while maintaining multi-view-consistent rendering. Experiments with static avatars obtained from multiple modalities, including monocular and multi-view images show improved generative metrics and more faithful surface dynamics than competing generalizable dynamic methods. A current limitation is the fixed template, which restricts loose garments and topology changes; relaxing this assumption is an important direction for future work.

AI USE STATEMENT

In this work, we did not use generative AI tools for tasks requiring disclosure, including formulating or refining research hypotheses, designing or providing feedback on the research methodology or experiments, implementing the proposed method, formulating or proving mathematical claims, generating or processing datasets, conducting qualitative analysis, or interpreting experimental results. Other required-disclosure tasks are not applicable to this work.

Additionally, we used generative AI tools to assist with language editing and the refinement of wording in the manuscript, as well as with the presentation and visualization of the supplementary video. We have reviewed all AI-assisted work to ensure that it accurately reflects the intended content of the paper and supplementary materials. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2293–2303, 2019.

Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pp. 561–578. Springer, 2016.

Chen Cao, Yanlin Weng, Stephen Lin, and Kun Zhou. 3d shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):1–10, 2013.

Dongliang Cao, Guoxing Sun, Marc Habermann, and Florian Bernard. Hyper diffusion avatars: Dynamic human avatar generation using network weight space diffusion. In *Thirteenth International Conference on 3D Vision*, 2025.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *International Conference on Machine Learning*, pp. 6263–6285. PMLR, 2024.

Di Chang, Hongyi Xu, You Xie, Yipeng Gao, Zhengfei Kuang, Shengqu Cai, Chenxu Zhang, Guoxian Song, Chao Wang, Yichun Shi, et al. X-dyna: Expressive dynamic human image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5499–5509, 2025.

Jianchuan Chen, Jingchuan Hu, Gaige Wang, Zhonghua Jiang, Tiansong Zhou, Zhiwen Chen, and Chengfei Lv. Taoavatar: Real-time lifelike full-body talking avatars for augmented reality via 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10723–10734, 2025.

Zhaoxi Chen, Fangzhou Hong, Haiyi Mei, Guangcong Wang, Lei Yang, and Ziwei Liu. Primdiffusion: Volumetric primitives diffusion for 3d human generation. *Advances in Neural Information Processing Systems*, 36:13664–13677, 2023.

Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, et al. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19982–19993, 2023.

Andrey Davydov, Anastasia Remizova, Victor Constantin, Sina Honari, Mathieu Salzmann, and Pascal Fua. Adversarial parametric pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10997–11005, 2022.

Junting Dong, Qi Fang, Zehuan Huang, Xudong Xu, Jingbo Wang, Sida Peng, and Bo Dai. Tela: Text to layer-wise 3d clothed human generation. In *European Conference on Computer Vision*, pp. 19–36. Springer, 2024.

Zijian Dong, Longteng Duan, Jie Song, Michael J.Black, and Andreas Geiger. MoGA: 3d Generative Avatar Prior for Monocular Gaussian Avatar Reconstruction. In *International Conference on Computer Vision (ICCV)*, 2025.

EasyMoCap. Easymocap - make human motion capture easier. Github, 2021. URL `https://github.com/zju3dv/EasyMocap`.

Xuan Gao, Jingtao Zhou, Dongyu Liu, Yuqi Zhou, and Juyong Zhang. Constructing diffusion avatar with learnable embeddings. In *ACM SIGGRAPH Asia Conference Proceedings*, 2025.

Dimitrios Gerogiannis, Foivos Paraperas Papantoniou, Rolandos Alexandros Potamias, Alexandros Lattas, and Stefanos Zafeiriou. Arc2avatar: Generating expressive 3d avatars from a single image via id guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10770–10782, 2025.

Chen Guo, Xu Chen, Jie Song, and Otmar Hilliges. Human performance capture from monocular video in the wild. In *2021 International Conference on 3D Vision (3DV)*, pp. 889–898. IEEE, 2021.

Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

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

Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020.

Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Transactions on Graphics*, 40(4), aug 2021.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12869–12879, 2023.

Yisheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Weihao Yuan, Zilong Dong, and Liefeng Bo. Lam: Large avatar model for one-shot animatable gaussian head. *arXiv preprint arXiv:2502.17796*, 2025.

Eric Hedlin, Helge Rhodin, and Kwang Moo Yi. A simple method to boost human pose estimation accuracy by correcting the joint regressor for the human3. 6m dataset. In *2022 19th Conference on Robots and Vision (CRV)*, pp. 1–7. IEEE, 2022.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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

Yangyi Huang, Ye Yuan, Xueting Li, Jan Kautz, and Umar Iqbal. Adahuman: Animatable detailed 3d human generation with compositional multiview diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13533–13543, 2025.

Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12753–12762, 2021.

Yuheng Jiang, Chengcheng Guo, Yize Wu, Yu Hong, Shengkun Zhu, Zhehao Shen, Yingliang Zhang, Shaohui Jiao, Zhuo Su, Lan Xu, Marc Habermann, and Christian Theobalt. Topology-aware optimization of gaussian primitives for human-centric volumetric videos. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, SA Conference Papers '25, New York, NY, USA, 2025a. Association for Computing Machinery. ISBN 9798400721373. doi: 10.1145/3757377.3763975. URL https://doi.org/10.1145/3757377.3763975.

Yuheng Jiang, Zhehao Shen, Chengcheng Guo, Yu Hong, Zhuo Su, Yingliang Zhang, Marc Habermann, and Lan Xu. Reperformer: Immersive human-centric volumetric videos from playback to photoreal reperformance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11349–11360, 2025b.

Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8320–8329, 2018.

Hendrik Junkawitsch, Guoxing Sun, Heming Zhu, Christian Theobalt, and Marc Habermann. Eva: Expressive virtual avatars from multi-view videos. In *SIGGRAPH 2025 Conference Papers*, pp. 1–11, 2025.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7122–7131, 2018.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Nikos Kolotouros, Thiemo Alldieck, Enric Corona, Eduard Gabriel Bazavan, and Cristian Sminchisescu. Instant 3d human avatar generation using image diffusion models. In *European Conference on Computer Vision (ECCV)*, 2024.

Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34:24741–24752, 2021.

Youngjoong Kwon, Lingjie Liu, Henry Fuchs, Marc Habermann, and Christian Theobalt. Deliffas: Deformable light fields for fast avatar synthesis. *Advances in Neural Information Processing Systems*, 2023.

Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020.

JP Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 165–172, 2000.

Chenghong Li, Hongjie Liao, Yihao Zhi, Xihe Yang, Zhengwentai Sun, Jiahao Chang, Shuguang Cui, and Xiaoguang Han. Mvhumannet++: A large-scale dataset of multi-view daily dressing human captures with richer annotations for 3d human digitization. *arXiv preprint arXiv:2505.01838*, 2025a.

Linzhou Li, Yumeng Li, Yanlin Weng, Youyi Zheng, and Kun Zhou. Rgbavatar: Reduced gaussian blendshapes for online modeling of head avatars. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

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 *Proceedings of the computer vision and pattern recognition conference*, pp. 16008–16018, 2025c.

Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19711–19722, 2024.

Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. In *2024 International Conference on 3D Vision (3DV)*, pp. 1508–1519. IEEE, 2024.

Di Liu, Teng Deng, Giljoo Nam, Yu Rong, Stanislav Pidhorskyi, Junxuan Li, Jason Saragih, Dimitris N. Metaxas, and Chen Cao. Lucas: Layered universal codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

Lingjie Liu, Weipeng Xu, Marc Habermann, Michael Zollhöfer, Florian Bernard, Hyeongwoo Kim, Wenping Wang, and Christian Theobalt. Neural human video rendering by learning dynamic textures and rendering-to-video translation. *arXiv preprint arXiv:2001.04947*, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Yixing Lu, Junting Dong, Youngjoong Kwon, Qin Zhao, Bo Dai, and Fernando De la Torre. Gas: Generative avatar synthesis from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12883–12893, 2025.

Qianli Ma, Jinlong Yang, Michael J. Black, and Siyu Tang. Neural point-based shape modeling of humans in challenging clothing. In *2022 International Conference on 3D Vision (3DV)*, 2022.

OpenAI. Dall·e, 2024. URL `https://openai.com/dall-e`.

Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1165–1175, June 2024.

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

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

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9054–9063, 2021b.

Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20299–20309, 2024.

Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, et al. Lhm: Large animatable human reconstruction model for single image to 3d in seconds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14184–14194, 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.

Pramod Rao, Gereon Fox, Abhimitra Meka, Mallikarjun B R, Fangneng Zhan, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, and Christian Theobalt. Lite2relight: 3d-aware single image portrait relighting. In *ACM SIGGRAPH 2024 Conference Proceedings*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2304–2314, 2019.

Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. Human4dit: 360-degree human video generation with 4d diffusion transformer. *ACM Transactions on Graphics (TOG)*, 43(6), 2024.

Ashwath Shetty, Marc Habermann, Guoxing Sun, Diogo Luvizon, Vladislav Golyanik, and Christian Theobalt. Holoported characters: Real-time free-viewpoint rendering of humans from sparse rgb cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1206–1215, 2024.

Carsten Stoll, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *2011 international conference on computer vision*, pp. 951–958. IEEE, 2011.

Guoxing Sun, Xin Chen, Yizhang Chen, Anqi Pang, Pei Lin, Yuheng Jiang, Lan Xu, Jingyi Yu, and Jingya Wang. Neural free-viewpoint performance rendering under complex human-object interactions. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4651–4660, 2021.

Guoxing Sun, Rishabh Dabral, Pascal Fua, Christian Theobalt, and Marc Habermann. Metacap: Meta-learning priors from multi-view imagery for sparse-view human performance capture and rendering. In *ECCV*, 2024.

Guoxing Sun, Rishabh Dabral, Heming Zhu, Pascal Fua, Christian Theobalt, and Marc Habermann. Real-time free-view human rendering from sparse-view rgb videos using double unprojected textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

Timo Teufel, Pulkit Gera, Xilong Zhou, Umar Iqbal, Pramod Rao, Jan Kautz, Vladislav Golyanik, and Christian Theobalt. Humanolat: A large-scale dataset for full-body human relighting and novel-view synthesis. In *International Conference on Computer Vision (ICCV)*, 2025.

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2387–2395, 2016.

Phong Tran, Egor Zakharov, Long-Nhat Ho, Liwen Hu, Adilbek Karmanov, Aviral Agarwal, McLean Goldwhite, Ariana Bermudez Venegas, Anh Tuan Tran, and Hao Li. Voodoo xp: Expressive one-shot head reenactment for vr telepresence. *ACM Transactions on Graphics, Proceedings of the 17th ACM SIGGRAPH Conference and Exhibition in Asia 2024, (SIGGRAPH Asia 2024), 12/2024*, 2024.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Guangyuan Wang, Li Hu, Dechao Meng, Zhongyi Zhang, Peng Zhang, Mingyang Huang, Ruoshi Zhang, Ke Sun, Zhe Zhang, Xingjun Wang, Gang Cheng, and Bang Zhang. Wan-animate-2: Real-time end-to-end character animation via diffusion transformer. *arXiv preprint arXiv:TODO.06009*, 2026. URL https://arxiv.org/abs/TODO.06009.

Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. In *Advances in Neural Information Processing Systems*, 2021.

Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 331–340, 2018.

Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Yuhan Wang, Fangzhou Hong, Shuai Yang, Liming Jiang, Wayne Wu, and Chen Change Loy. Meat: Multiview diffusion model for human generation on megapixels with mesh attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11297–11306, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1682–1691, 2020.

Yue Wu, Xuanhong Chen, Yufan Wu, Wen Li, Yuxi Lu, and Kairui Feng. Fastavatar: Towards unified and fast 3d avatar reconstruction with large gaussian reconstruction transformers. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13296–13306, June 2022.

Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1481–1490, 2024.

Yuxuan Xue, Xianghui Xie, Margaret Kostyrko, and Gerard Pons-Moll. Infinihuman: Realistic 3d human creation with precise control. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5746–5756, 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.

Yifei Zeng, Yuanxun Lu, Xinya Ji, Yao Yao, Hao Zhu, and Xun Cao. Avatarbooth: High-quality and customizable 3d human avatar generation. *arXiv preprint arXiv:2306.09864*, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Weitian Zhang, Yichao Yan, Yunhui Liu, Xingdong Sheng, and Xiaokang Yang. E3gen: Efficient, expressive and editable avatars generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 6860–6869, 2024.

Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

Yuxiao Zhou, Menglei Chai, Alessandro Pepe, Markus Gross, and Thabo Beeler. Groomgen: A high-quality generative hair model using hierarchical latent representations. *ACM Transactions on Graphics (TOG)*, 42(6):1–16, 2023.

Zhenglin Zhou, Fan Ma, Hehe Fan, and Tat-Seng Chua. Zero-1-to-a: Zero-shot one image to animatable head avatars using video diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Heming Zhu, Fangneng Zhan, Christian Theobalt, and Marc Habermann. Trihuman: a real-time and controllable tri-plane representation for detailed human geometry and appearance synthesis. *ACM Transactions on Graphics*, 44(1):1–17, 2024a.

Heming Zhu, Guoxing Sun, Christian Theobalt, and Marc Habermann. Uma: Ultra-detailed human avatars via multi-level surface alignment. *ACM Transactions on Graphics*, 2026. doi: 10.1145/3829365.

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision (ECCV)*, 2024b.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26308–26319, 2025.

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

In this supplemental material, we present more qualitative (Sec. B and Sec. G) and ablative (Sec. C) results. We discuss the drawbacks of big 2D-only method and compare our method with them (Sec. I). Then we provide more details regarding the large-scale DynaHuman dataset (Sec. D). Further, we provide more implementation details regarding texture optimization (Sec. E) and network structures (Sec. F). We perform user study on comparison methods (Sec. H). We discuss the limitations of our method (Sec. J). Lastly, we carefully discuss the potential ethical issues and concerns (Sec. K).

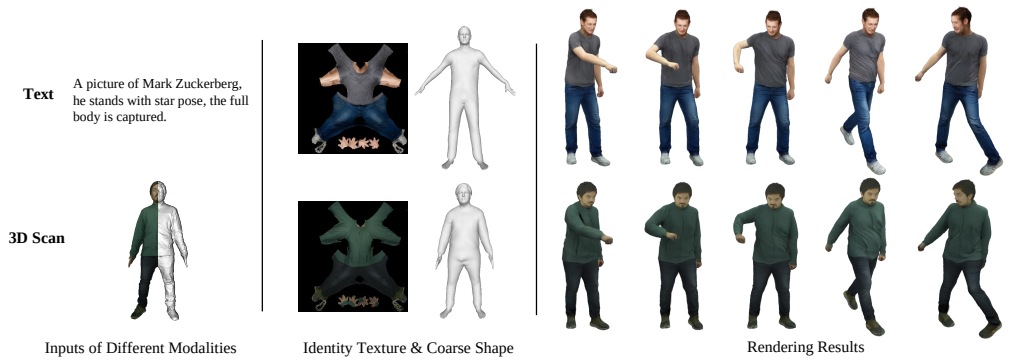


Figure 5: **Multi-modal Avatar Generation**. Our method transforms static avatars that are acquired from multi-modal inputs into dynamic ones. Here, we show static avatars that are created from text and 3D scans. Next, the static avatar is converted into identity texture and coarse shape. Finally, we can drive the avatar using novel skeletal motions. We highlight the different wrinkle patters and dynamics that are induced by the motion and that our method can faithfully generate.

Table 4: **Dataset Comparison.** Statistics are based on the publicly available versions of the datasets (e.g., MVHPP Li et al. (2025a) only releases part of the data). Our dataset matches or surpasses person-specific approaches Peng et al. (2021a); Habermann et al. (2021); Sun et al. (2025) in motion diversity and number of views while providing significantly more identities. Compared to large-scale identity datasets Cheng et al. (2023); Li et al. (2025a), our dataset has significantly higher motion diversity, views, and frames per subject. Besides, we provide high-precision geometry and registered textures. Thus, our dataset bridges an important gap in literature to cover surface dynamics.

Dataset	Frames Per Subject	Views	Identities	Motion Diversity	Geometry	Texture	4K Resolution	Images (In total)
DDC Habermann et al. (2021)	$\sim$ 26,000	$\sim$ 90	5	✓	✗	✗	✓	11.8M
DUT Sun et al. (2025)	$\sim$ 35,000	$\sim$ 116	2	✓	✓	✗	✓	8.1M
ZJU Peng et al. (2021b)	$\sim$ 500-2,000	24	9	✗	✗	✗	✗	180K
DNA Cheng et al. (2023)	$\sim$ 200	60	500	✗	✓	✗	(✓)	67.5M
MVHPP Li et al. (2025a)	$\sim$ 60	16	4,500	✗	✓	✗	✗	645.1M
Ours	27,000	100	58	✓	✓	✓	✓	156.6M

A Conceptual Comparison of Dynamic Human Representation Paradigms

Fig. 6 compares three design choices for dynamic human representation. Regression-based approaches are effective for static 3D reconstruction, but their extension to dynamic humans remains difficult. Deterministic regression is not well suited for modeling fine-grained surface dynamics under limited data, as motion-dependent details such as wrinkles are high-frequency and inherently ambiguous: the same identity and pose can correspond to multiple plausible surface states depending on unobserved physical factors. As a result, these methods often produce over-smoothed or nearly static dynamics, even when motion conditions are provided. To alleviate this limitation, they typically require large-scale multi-view motion data, which is expensive to collect and often not publicly available. Another option is to condition a video model on a static human rendering, allowing the model to generate plausible dynamics from learned 2D video priors. However, since the generation is performed in image space, the results do not explicitly guarantee 3D consistency across time and viewpoints. Our representation is designed to address this gap. It leverages video priors for dynamic generation while using the Generalized Gaussian Decoder to produce a dynamic 3D human representation that remains consistent in 3D.

B Multi-modal Input Results

We present additional qualitative results in Fig. 5, showing that our method can generate dynamic avatars from multi-modal inputs, like text OpenAI (2024) and 3D scans Teufel et al. (2025). Given

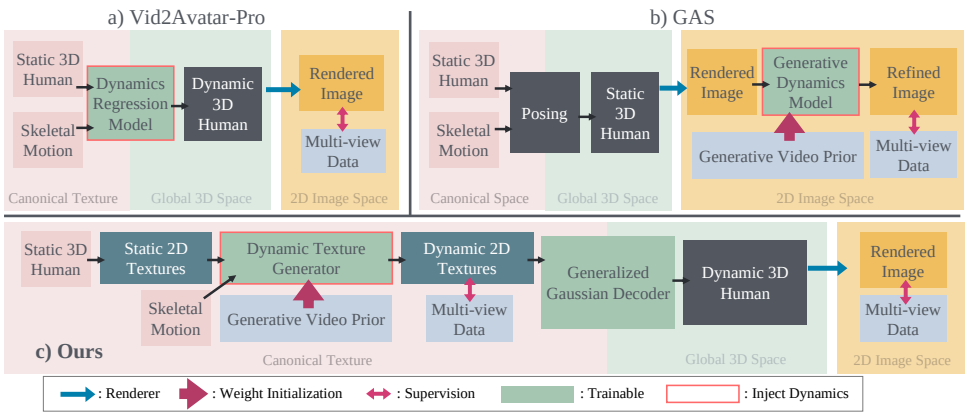


Figure 6: **Conceptual Comparison of Generalized and Dynamic Human Representations**. a) Regressing 3D human dynamics directly cannot leverage foundational video priors. Thus, such methods rely on large-scale (closed-source) multi-view data. While regression-based models perform well for static reconstruction, they struggle with dynamic reconstruction. b) Leveraging a static rendering as conditioning for a video model allows generative modeling of dynamics. However, since the model operates in 2D image space, 3D consistency is not ensured. c) Our human representation unites the best of both worlds and enables generative modeling and video prior usage for dynamics while our *Generalized Gaussian Decoder* predicts a dynamic 3D human that is 3D consistent.



Figure 7: **Representation Ablation.** The qualitative ablation over the avatar representation on the training split. Given the same multi-view videos, compared with the baseline approach, i.e., VAP Guo et al. (2025), our proposed Generalized Gaussian Decoder (GGD), which takes the reference dynamic texture (RDT) as input, reconstructs the pose-dependent wrinkles more accurately.

either textual input or a static 3D scan, we first generate the identity texture and a coarse body shape, represented by an SMPL-X template with personalized offsets. Through our proposed pipeline, we synthesize clothing dynamics induced by the subject's motion, enriching the static rigged avatar with realistic and expressive motion-aware wrinkle details.

C Qualitative Ablations

Representation Ablation. We compare our dynamic-texture representation against the static-texture representation of VAP Guo et al. (2025). Both methods have access to the full multi-view videos of the training subject. However, VAP implicitly encodes dynamics within the regression model through a static identity texture, whereas *AvatarDynamizer* supplies an explicit per-frame dynamic texture to the *Generalized Gaussian Decoder* (GGD). As shown in Fig. 7, the GGD faithfully reconstructs fine-grained wrinkles when they are explicitly represented in the dynamic texture. In contrast, the static-texture baseline results in a flat and blurry appearance.

Training Dataset Ablation. Fig. 8 presents qualitative ablations analyzing the influence of the training datasets on the *Generalized Gaussian Decoder*. Removing either DynaHuman or MVHPP Li et al. (2025a) from training leads to inaccurate coarse shape estimation, uneven lighting, and ghost-

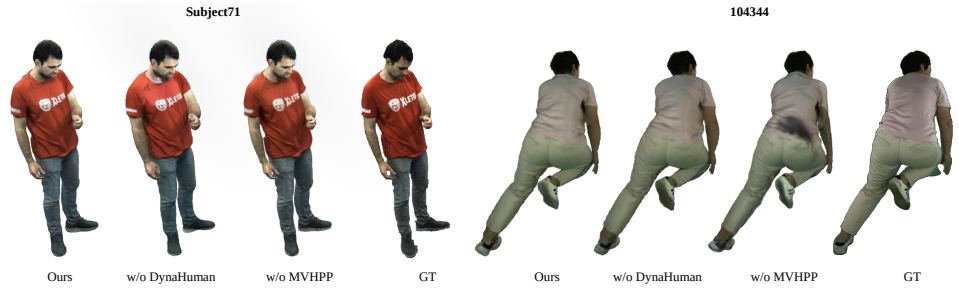


Figure 8: **Dataset Ablation.** The qualitative ablation over the training datasets. Removing the DynaHuman dataset from training leads to uneven lighting, while excluding the MVHPP dataset Li et al. (2025a) from training could cause severe artifacts. By combining our DynaHuman dataset and the MVHPP dataset, our Generalized Gaussian Decoder achieves better reconstruction.

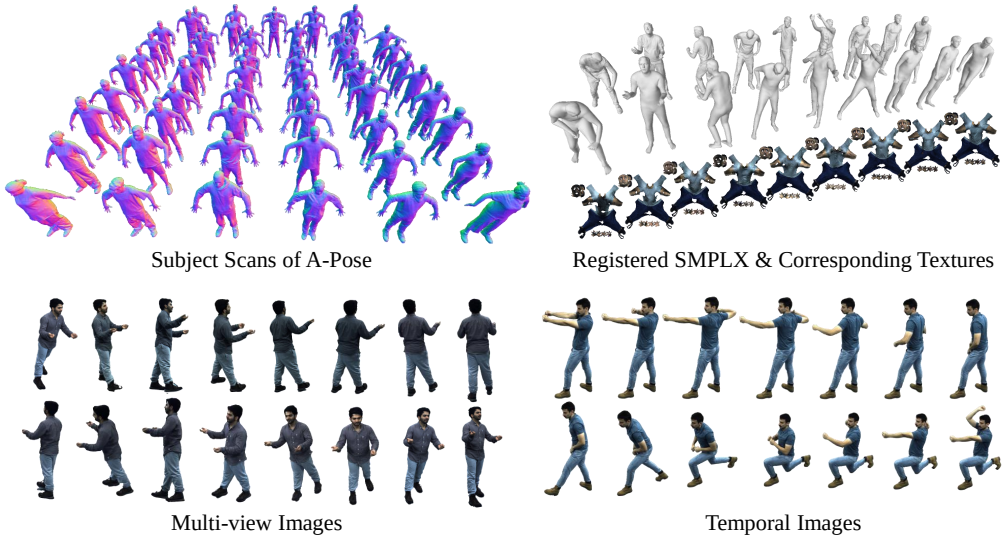


Figure 9: **Dataset Visualization**. For each subject, we provide an A-pose 3D scan, multi-view videos, and registered and deformed SMPL-X as well as dynamic texture annotations.

ing artifacts. In contrast, leveraging both datasets leads to significantly improved robustness and reconstruction quality.

D DynaHuman Dataset

Capture setup. *DynaHuman* is captured in a dedicated multi-camera studio equipped with 100 synchronized, calibrated $4K$ cameras arranged in a dome configuration that covers the subject from all elevations. Notably, the studio is equipped with uniform lighting, which facilitates accurate photometric geometry reconstruction. As shown in Fig. 9 and Tab. 4, subjects in *DynaHuman* are dressed in diverse clothing styles, such as casual, formal, and sportswear, with varying degrees of tightness, which enhances the diversity of clothing deformation and dynamics.

Statistics. The dataset comprises 58 subjects, each recorded for approximately 18 minutes (27,000 frames at 25 fps) of continuous, natural full-body motion. Each subject performs motions including locomotion (walking, running, jogging), dynamic whole-body actions (jumping, squatting, turning), and expressive upper-body gestures (arm raises, waving, reaching), deliberately chosen to elicit large, rapid cloth deformations. Each frame is annotated with SMPL-X tracking and geometry reconstructed using NeuS2 Wang et al. (2023). Additionally, we provide the corresponding texture maps as described in the main paper.

Comparison with existing datasets. As summarized in Tab. 1 of the main paper, existing datasets fall into two extremes: DDC Habermann et al. (2021) and DUT Sun et al. (2025) provide rich clothing dynamics but cover very few identities. In contrast, MVHumanNet++ Li et al. (2025a) and DNA-Rendering Cheng et al. (2023) include many subjects, but these subjects mostly perform slow or static motions. *DynaHuman* bridges this gap, combining 58 identities with long sequences, dense camera coverage, and diverse motions.

E Texture Embedding via Multi-stage Optimization

For each frame, we first obtain point cloud geometry $\mathbf{p}$ and 3D markers $\mathbf{m}$ from multi-view images as illustrated in Sec. 3.3 of the main paper. Next, we introduce how to register SMPL-X Pavlakos et al. (2019) with coarse shape, body motion, and deformations.

SMPL-X Tracking. We split SMPL-X tracking into three steps. In the first step, we follow EasyMocap EasyMoCap, which estimates initial SMPL-X parameters from $\mathbf{m}$ as

$$\begin{aligned}\hat{\mathcal{M}}^{\text{step1}}, \hat{\boldsymbol{\beta}}^{\text{step1}} = \arg\min_{\mathcal{M}, \boldsymbol{\beta}} \\ \lambda_m \left\|\mathbf{X}(\mathbf{V}(\mathcal{M}, \boldsymbol{\beta})) - \mathbf{m}\right\|_2^2 \\ + \lambda_{\mathcal{M}} \sum_{t=2}^{T} \left\|\mathcal{M}^t - \mathcal{M}^{t-1}\right\|_2^2 + \lambda_\beta \|\boldsymbol{\beta}\|_2^2,\end{aligned} \tag{7}$$

where $\mathbf{X}(\cdot)$ is the marker regressor.

In the second step, we refine the SMPL-X parameters per frame with $\mathbf{p}$ and $\mathbf{m}$ for better alignment, initialized by $\hat{\mathcal{M}}^{\text{step1}}$ and $\hat{\boldsymbol{\beta}}^{\text{step1}}$:

$$\begin{aligned}\hat{\mathcal{M}}^{\text{step2}}, \hat{\boldsymbol{\beta}}^{\text{step2}} = \arg\min_{\mathcal{M}^t, \boldsymbol{\beta}} \lambda_m \left\|\mathbf{X}\big(\mathbf{V}(\mathcal{M}^t, \boldsymbol{\beta})\big) - \mathbf{m}^t\right\|_2^2 \\ + \lambda_p \left\|\mathbf{V}(\mathcal{M}^t, \boldsymbol{\beta}) - \mathbf{p}^t\right\|_2^2.\end{aligned} \tag{8}$$

In the third step, we smooth the SMPL-X parameters with $\mathbf{p}$ and $\mathbf{m}$, initialized by $\hat{\mathcal{M}}^{\text{step2}}$ and $\hat{\boldsymbol{\beta}}^{\text{step2}}$:

$$\begin{aligned}\hat{\mathcal{M}}^{\text{step3}} = \arg\min_{\mathcal{M}} \lambda_m \sum_{t=1}^{T} \left\|\mathbf{X}\big(\mathbf{V}(\mathcal{M}^t, \boldsymbol{\beta})\big) - \mathbf{m}^t\right\|_2^2 \\ + \lambda_p \sum_{t=1}^{T} \left\|\mathbf{V}(\mathcal{M}^t, \boldsymbol{\beta}) - \mathbf{p}^t\right\|_2^2 \\ + \lambda_{\mathcal{M}} \sum_{t=2}^{T} \left\|\mathcal{M}^t - \mathcal{M}^{t-1}\right\|_2^2.\end{aligned} \tag{9}$$

Deformation Optimization. With optimized SMPL-X body motion $\mathcal{M}$ and coarse shape $\boldsymbol{\beta}$, we optimize time-varying cloth displacements $\tilde{\mathbf{d}}_{\text{cloth}}$:

$$\begin{aligned}\hat{\tilde{\mathbf{d}}}_{\text{cloth}} = \min_{\tilde{\mathbf{d}}_{\text{cloth}}} \sum_{t=1}^{T} \left\|\mathbf{V}(\mathcal{M}^t, \boldsymbol{\beta}, \mathbf{d}_{\text{hair}}, \tilde{\mathbf{d}}_{\text{cloth}}^t) - \mathbf{p}^t\right\|_2^2 \\ + \sum_{t=1}^{T-1} \Big\|\mathbf{V}(\mathcal{M}^t, \boldsymbol{\beta}, \mathbf{d}_{\text{hair}}, \tilde{\mathbf{d}}_{\text{cloth}}^t) \\ - \mathbf{V}(\mathcal{M}^{t+1}, \boldsymbol{\beta}, \mathbf{d}_{\text{hair}}, \tilde{\mathbf{d}}_{\text{cloth}}^{t+1})\Big\|_2^2,\end{aligned} \tag{10}$$

where $\mathbf{d}_{\text{hair}}$ comes from the coarse shape of the initial static pose.

Texture Optimization. Given the deformed human geometry from the previous step, for each frame, we then optimize the texture of the avatar, i.e., unproject the images onto the texture space, using differentiable rasterization. The exact formulation is provided in Eq. 3 in the main paper.

F More Implementation Details

As head modeling is beyond the main scope of body avatar dynamization in this work, we employ a head replacement strategy in VAP Guo et al. (2025) and ours to ensure stable head rendering. Specifically, we replace the head Gaussian splats of the dynamic avatars with those from the corresponding static avatars.

Evaluation. For image reconstruction evaluation, we uniformly sample every 10 frames from the sequences of the testing subjects and select four camera views that are uniformly placed around the subject. The metrics are computed at a resolution of $2K$. For generative quality evaluation, we use $26{,}800$ frames from each test subject sequence and select a camera view facing the subject at the first frame. The images are cropped and resized to 512×512, with the subject centered in each frame.

LHM. Since LHM Qiu et al. (2025a) takes a monocular image as input, we provide a front-facing 'A'-pose image and multi-view estimated SMPL-X Pavlakos et al. (2019) shape parameters $\boldsymbol{\beta}$ for fair comparison. We use the official 'LHM-500M' checkpoint for all experiments. We additionally optimize a static Gaussian avatar from LHM outputs with the same number and ordering of points as the 3D methods for head replacement.

MV. MV denotes reconstructing an initial static human avatar from multi-view images. Instead of optimizing from scratch, we feed the identity map into our *Generalized Gaussian Decoder* and optimize it for $5k$ steps with a cropped head region loss weight of 1.0 and a regularization weight of 0.01.

GAS. Since GAS Lu et al. (2025) does not provide training code, we use the official checkpoint and perform inference in the 'novel pose' mode. Note that although we do not fine-tune GAS on our *DynaHuman* dataset, it has already been trained on several other datasets in addition to MVHPP Li et al. (2025a), including THuman2.1 Yu et al. (2021), 2K2K Han et al. (2023), TikTok Jafarian & Park (2021), and additional internet videos, which are not used for training our approach.

VAP. Because the VAP Guo et al. (2025) code and training data are unavailable, we implement a controlled regression baseline using the same datasets, UNet architecture, and training strategy as our GGD. The two models differ primarily in their input representations: the regression baseline uses a static identity texture and pose controls, whereas the GGD receives per-frame dynamic textures. Our re-implementation slightly differs from their original one – representing the canonical space in texture space rather than orthogonal projection space, and predicting Gaussian deltas versus directly predicting Gaussian parameters – to better evaluate the core difference in representation, i.e., dynamic vs. static textures. The original VAP was trained on 1,000 dynamic-human sequences that are not publicly available, motivating our controlled re-implementation using the same training data as our GGD. Under this matched training setting, our dynamic-texture representation achieves better reconstruction and preserves finer surface dynamics than the static-texture regression baseline.

Ours. For the *Generalized Gaussian Decoder*, we adopt a network architecture similar to DUT Sun et al. (2025). The temporal window size of our *Dynamic Texture Generator* is 49, with an overlap of 5 frames during inference. We fine-tune the video diffusion model for 100k steps on 4 NVIDIA H100 GPUs. We set the LoRA weight to 1.0 for LHM-created avatars and 0.8 for MV-created avatars.

RDT+GGD. In the reference dynamic texture (RDT) and *Generalized Gaussian Decoder* (GGD) experiment, we use reference textures optimized from multi-view images of the test subjects and feed them into the *Generalized Gaussian Decoder*. Thus, the model operates in a reconstruction setting, i.e., reconstructing Gaussians from the reference dynamic textures.

G More Qualitative Results.

Fig. 12 and Fig. 13 shows additional qualitative results on DeepCap Habermann et al. (2020), DDC Habermann et al. (2021), MetaCap Sun et al. (2024) and in-the-wild monocular images. Our method successfully generates pose-dependent wrinkles across different motions and clothing types. For in-the-wild subjects, we drive them with the same body motion. The wrinkles evolve with the motion and vary across different clothing types.

	Identity Consistency (Q1)	Physical Plausibility (Q2)	View Consistency (Q3)	Overall Score (Q4)
Static Avatar	3.944/5.0	3.289/5.0	4.089/5.0	3.644/5.0
+GAS	1.233/5.0	1.289/5.0	1.356/5.0	1.256/5.0
+VAP	3.511/5.0	2.622/5.0	3.689/5.0	2.878/5.0
+Ours	4.000/5.0	4.003/5.0	3.789/5.0	3.947/5.0

Table 5: **User-study results.** Mean ratings from 18 participants on five subjects; 1 is poor and 5 is excellent.

Table 6: **Quantitative comparison on DynaHuman.** Here, we compare Wan-Animate 2 Wang et al. (2026) with our method. Our method achieves consistent better results.

	Subject71		**Subject76**	
	FID	FVD	FID	FVD
Wan-Animate 2	31.83	185.09	28.20	179.98
LHM+Ours	28.20	72.66	19.65	40.78

H User Study

Beyond qualitative and quantitative evaluations, we conduct a user study to assess human preferences on the perceptual quality of different methods. A total of 18 participants took part in the study and were asked to compare the results on five subjects according to the following questions:

- Q1. How well do the methods **preserve the identity** of the original subject?
- Q2. How do these methods present **physically plausible** surface dynamics?
- Q3. Are these results **consistent across different views**?
- Q4. The **overall quality** considering all technical aspects.

As shown in Tab. 5, our method is preferred in terms of identity preservation, physical plausibility, and overall quality, and achieves the second-best score in view consistency. Although view consistency is included in the study, it is worth noting that methods based on explicit 3D primitives, including the static avatar baseline, VAP, and ours, are naturally consistent when rendering novel views.

I Discussions and Comparisons with Large-scale 2D animation models.

Inspired by the success of scaling laws Kaplan et al. (2020), Wan-Animate 2 Wang et al. (2026) scales up the 2D-based paradigm exemplified by MagicPose Chang et al. (2024), leveraging substantially larger training data and a more advanced diffusion architecture. Given a reference image and a conditional motion video, these methods generate an animated 2D video of the reference subject following the prescribed motion. However, they do not employ explicit 3D representations. As a result, their temporal and spatial consistency must be learned implicitly through attention mechanisms Vaswani et al. (2017) and large-scale training data, which does not guarantee consistent appearance across time or viewpoints. To run Wan-Animate 2 efficiently on a single GPU, we use its official distilled inference setting. We evaluate 26,800-frame sequences at 512×512 resolution using 25-frame clips with a stride of 25 frames (no overlap), as overlapping clips caused progressive color saturation in our preliminary runs. For a fair comparison, we use the same reference image provided to LHM as the identity condition and render the prescribed motion from each corresponding test-camera viewpoint to construct the motion-conditioning video. Fig. 10 illustrates typical failures in maintaining temporal and spatial consistency. Fig. 11 shows additional failure cases of Wan-Animate 2 on a DynaHuman subject, including incorrect motion embedding, pose-to-appearance ambiguity, inconsistent lighting, inaccurate body shape, and background artifacts. We quantitatively compare our method with Wan-Animate 2 Wang et al. (2026) in Tab. 6, where our method consistently achieves better performance. Please refer to the main video for qualitative comparisons.

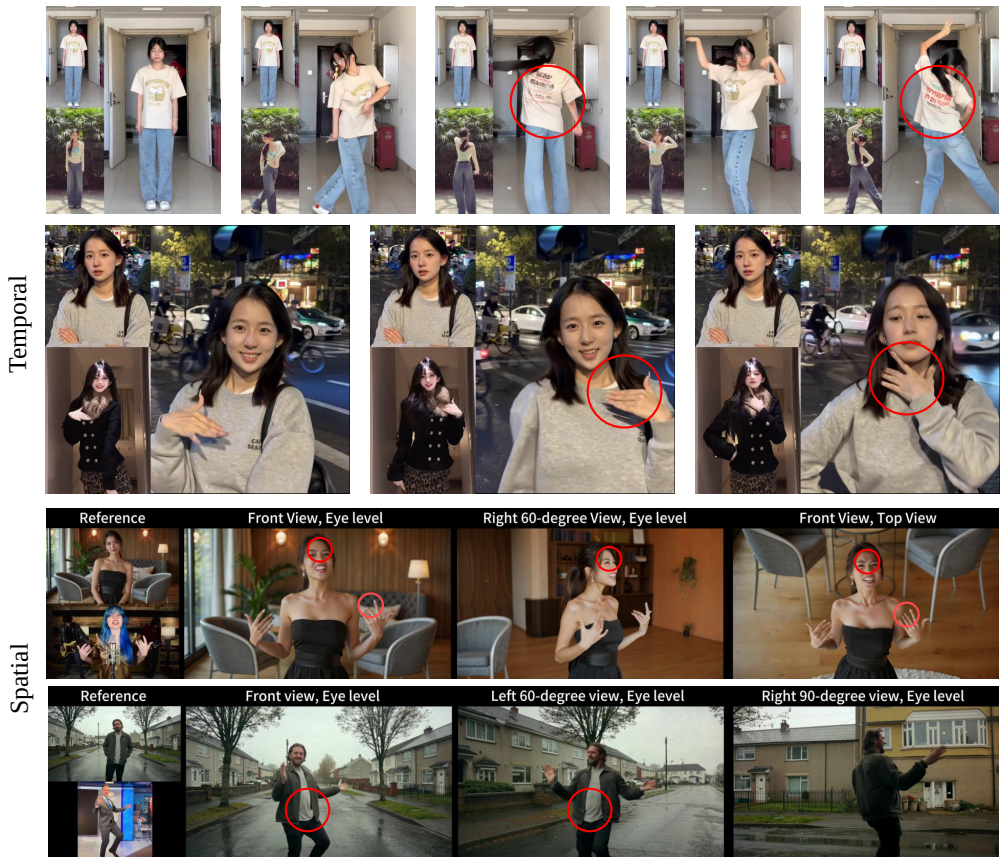


Figure 10: **Failure of Wan-Animate 2**. State-of-the-art 2D-only animation models fail to preserve spatial and temporal consistency. Note how the back texture and nail colors change over time (temporal inconsistency), and how facial wrinkles, clothing wrinkles, and nail shapes change across different camera views (spatial inconsistency). All results shown above are from the official Wan-Animate 2 results.



Figure 11: **Additional Failure of Wan-Animate 2**. Given a DynaHuman subject, Wan-Animate 2 exhibits several limitations, including incorrect motion embedding, pose-to-appearance ambiguity, inconsistent lighting, distorted body shape, and background artifacts.

J LIMITATIONS.

Our method inherits several limitations common to template-based avatar representations. First, extreme non-rigid garment deformation challenges registration to the underlying SMPL-X template, limiting generalization to complex loose clothing. More flexible canonical representations could improve tracking and dynamics modeling in these cases. Second, our representation relies on a fixed mesh template, i.e., SMPL-X, which limits its ability to explicitly model topological changes such as



Figure 12: **Qualitative Comparisons on DeepCap, DDC and MetaCap**. We show additional qualitative comparisons on public datasets, i.e. DeepCap Habermann et al. (2020), DDC Habermann et al. (2021) and MetaCap Sun et al. (2024). The results are shown in the order of ground truth, MV (static avatar), and our method. Our method successfully generates pose-dependent wrinkles across different motions and clothing types.

opening a jacket. While recent work explores Gaussian-based tracking of such changes Jiang et al. (2025a), generalizable tracking across subjects is still challenging. Third, self-occlusion leaves parts of the explicit texture representation unobserved during embedding. Finally, the generator learns perceptually plausible surface dynamics from data without explicit physical constraints. Incorporating cloth-physics priors could improve physical accuracy and generalization to unseen garments and motions.

K Ethical Discussions

Our method enables the generation of vivid, temporally coherent surface dynamics for animatable static human avatars, while supporting a wide range of input modalities. By explicitly modeling fine-grained dynamic textures, it enhances realism, generates subtle surface details such as wrinkles and cloth deformation, and allows for controllable avatar animation. These capabilities open up promising applications in areas such as digital communication, virtual reality, telepresence, enter-



Figure 13: **Qualitative Comparisons on In-the-wild Images**. The leftmost image is the input to LHM Qiu et al. (2025a). The results are shown in the order of LHM (static avatar), and our method. We drive them with the same body motion. Notice that how the wrinkles evolve with the motion and vary across different clothing types.'

tainment, accessibility technologies, and personalized digital humans. In particular, the method can facilitate more natural remote interaction, immersive storytelling, and inclusive human–computer interfaces by bridging the gap between static identity capture and dynamic expressiveness.

As with many powerful generative technologies, the method could be misused. For instance, it may be employed to create fabricated or misleading videos in which an individual's identity is reconstructed from publicly available social media images and subsequently animated according to artificially designed scenarios. Such misuse raises concerns regarding consent and authenticity.

To mitigate these risks, responsible deployment is essential. Appropriate safeguards should therefore be considered prior to practical adoption. These may, for example, include robust identity verification mechanisms, explicit consent requirements for identity reconstruction, or watermarking. By integrating such technical protections along with ethical guidelines and governance measures, the technology can be directed toward socially beneficial use while minimizing potential harm.