Single-Image 3D Human Reconstruction with 3D-Aware Diffusion Priors and Facial Enhancement

JIE YANG, Institute of Computing Technology, Chinese Academy of Sciences, China
BO-TAO ZHANG, Institute of Computing Technology, Chinese Academy of Sciences, China and University of Chinese Academy of Sciences, China
FENG-LIN LIU, Institute of Computing Technology, Chinese Academy of Sciences, China and University of Chinese Academy of Sciences, China
HONGBO FU, Hong Kong University of Science and Technology, Hong Kong
YU-KUN LAI, Cardiff University, United Kingdom
LIN GAO*, Institute of Computing Technology, Chinese Academy of Sciences, China and University of Chinese Academy of Sciences, China

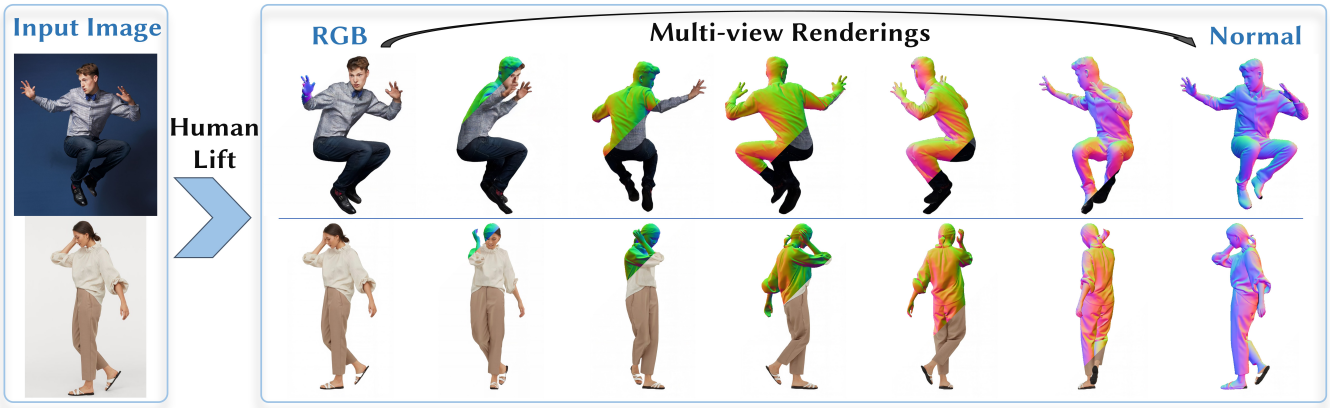

Fig. 1. Our method, HumanLift, reconstructs a photorealistic 3D human in 3D-GS from a single reference image. We enhance multi-view consistency by leveraging the generalization of video-based diffusion transformers and introducing a 3D-aware prior. We further boost facial details through a face generative prior and normal supervision during 3D Gaussian optimization. As the figure illustrates, the left is the input image, and the right part is the multi-view normal and RGB images rendered from our reconstructed human 3D Gaussians. The visual results demonstrate our method synthesizes richer garment details while maintaining excellent facial consistency with the reference image.

Creating high-quality, photorealistic 3D digital humans from a single image remains challenging. While existing methods can generate visually appealing multi-view outputs, they often suffer from inconsistencies in viewpoints and camera poses, resulting in suboptimal 3D reconstructions with reduced realism. Furthermore, most approaches focus on body generation while overlooking facial consistency – a perceptually critical issue caused by the fact that the face occupies only a small area in a full-body image (e.g., ~ 80×80 pixels out of a 512 × 512 image). This limited resolution and low weight for the facial regions during optimization leads to insufficient facial details and inconsistent facial identity features across multiple views. To address these challenges, we leverage the powerful capabilities of 2D video diffusion models for consistent multi-view *RGB* and *Normal* human image generation, combined with the 3D SMPL-X representation to enable spatial consistency and geometrical details. By fine-tuning the DiT models (HumanWan-DiTs) on realistic 3D human datasets using the LoRA technique, our method ensures both generalizability and 3D visual consistency on realistic multi-view human image generation. The proposed facial enhancement is integrated into 3D Gaussian optimization to enhance facial details. To further refine results, we apply super-resolution and generative priors to reduce facial blurring alongside SMPL-X parameter tuning and the assistance of generated multi-view normal images, achieving photorealistic and consistent rendering from a single image. Extensive experiments demonstrate that our approach

*Corresponding author: Lin Gao (gaolin@ict.ac.cn).

Authors' Contact Information: Jie Yang, Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, yangjie01@ict.ac.cn; Bo-Tao Zhang, Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China and University of Chinese Academy of Sciences, Beijing, China, zhangbotao24s@ict.ac.cn; Feng-Lin Liu, Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China and University of Chinese Academy of Sciences, Beijing, China, liufenglin21s@ict.ac.cn; Hongbo Fu, Hong Kong University of Science and Technology, Hong Kong, fuplus@gmail.com; Yu-Kun Lai, Cardiff University, United Kingdom, laiy4@cardiff.ac.uk; Lin Gao, Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China and University of Chinese Academy of Sciences, Beijing, China, gaolin@ict.ac.cn.

SA Conference Papers '25, Hong Kong, Hong Kong

ACM ISBN 979-8-4007-2137-3/25/12
https://doi.org/10.1145/3757377.3763839

outperforms existing methods, producing photorealistic, consistent, and fine-detailed human renderings.

CCS Concepts: • **Computing methodologies** → **Reconstruction**; **Computer vision**.

Additional Key Words and Phrases: Multiview Image Diffusion Model, Image to 3D Human Generation, 3D Gaussian Splatting, Digital Human

ACM Reference Format:
Jie Yang, Bo-Tao Zhang, Feng-Lin Liu, Hongbo Fu, Yu-Kun Lai, and Lin Gao. 2025. Single-Image 3D Human Reconstruction with 3D-Aware Diffusion Priors and Facial Enhancement. In *SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers '25), December 15–18, 2025, Hong Kong, Hong Kong.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3757377.3763839

1 Introduction

Creating photorealistic 3D digital humans is a crucial task in computer vision and graphics and offers benefits across various domains, including 3D films, gaming, and virtual/augmented reality. The majority of techniques [Balan et al. 2007; Bhatnagar et al. 2019; Gao et al. 2022; Noguchi et al. 2021; Peng et al. 2021a,b; Wu et al. 2020] reconstruct a high-quality, animatable 3D human model using multi-view calibrated cameras. Nonetheless, reconstructing such a model from a single reference image continues to be labor-intensive and technically challenging. This task involves estimating the visual appearance and clothed geometry for unseen views while also preserving multi-view consistency and ensuring that the 3D model conforms to the priors derived from the reference image.

The rapid advancement of generative priors (e.g. Stable Diffusion [Rombach et al. 2022]) and differentiable inverse rendering methods (e.g. NeRF [Mildenhall et al. 2021], 3D Gaussian Splatting [Kerbl et al. 2023]) has bridged the gap between 2D images and the 3D domain, enabling the transfer of 2D generative capabilities into 3D generation. Recent works [Lin et al. 2023; Long et al. 2024; Shi et al. 2024; Zhang et al. 2024b] have shown promising results in 3D generation from single images of general objects. However, efforts to extend this to high-quality 3D human models [Huang et al. 2024; Li et al. 2024b; Saito et al. 2019; Zhang et al. 2024c] often suffer from limited realism due to the scarcity of photorealistic 3D human data. Video diffusion models [Wang et al. 2025a, 2024c] have achieved remarkable realism in dynamic human generation [Hu 2024; Shao et al. 2024a; Wang et al. 2024b], though they may exhibit inconsistency at the different views across frames.

Multi-view photorealistic generation methods [He et al. 2024; Li et al. 2024b; Xue et al. 2024; Zhang et al. 2024c] still suffer from challenges such as view inconsistency, inaccurate camera poses, and poor 3D reconstruction. Most approaches also prioritize body generation while overlooking facial consistency, which is essential for realism. Face occupies only a tiny region (~80×80 pixels) in a full-body image (512×512), making it difficult to preserve fine facial details and maintain identity across views, as subtle geometry and high-frequency features are often lost during optimization.

To address these challenges, we propose a novel approach (named HumanLift) that leverages a pre-trained large-scale 2D video generation model (Wan2.1) [Wang et al. 2025a] as a generative prior. We enhance this model by injecting 3D consistency from SMPL-X [Pavlakos et al. 2019] and camera poses via fine-tuning it on realistic 3D human datasets [Shao et al. 2024a]. This integration ensures the generative model (named HumanWan-DiT) produces highly lifelike and consistent 3D human multi-view RGB photos (named HumanWan-DiT (RGB)) while preserving generalization across various human poses and camera viewpoints. Furthermore, a key strength of our HumanWan-DiT is its capability to predict highly consistent multiview normals (named HumanWan-DiT (Normal)) by leveraging the injected 3D priors. Conditioned directly on the generated multiview RGB images, HumanWan-DiT simultaneously predicts corresponding normal maps that exhibit exceptional consistency across different viewpoints. This robust normal prediction capability is paramount for accurately capturing the intricate surface details and underlying geometric structure of the human body, providing essential information beyond just color and texture. Unlike traditional 2D generative approaches that often struggle to maintain geometric fidelity across views, leading to artifacts and inconsistencies in inferred 3D, our method explicitly enforces 3D consistency.

To enable the downstream 3D tasks, we further propose a technique for lifting the 2D multi-view images (inferred from a single image) into the 3D domain using 3D Gaussian Splatting (3D-GS) [Kerbl et al. 2023] with facial detail enhancement. During the Gaussian learning, the camera poses of generated multi-view face images are corrected by transformation between the FLAME [Li et al. 2017] and head part of the optimized SMPL-X, while the Gaussian distribution is supervised by human body multiview images (RGB and Normal), and face multiview images. This method facilitates free-view rendering with consistency and fine details while enabling precise SMPL-X pose optimization and enhancement of facial details via super-resolution and generative priors [Li et al. 2025]. Finally, our approach not only permits the photorealistic rendering of 3D humans from arbitrary views but also supports further applications in future (e.g., realistic animation). An example of our HumanLift is shown in Figure 1. In summary, our core contributions are listed:

- A novel method, called HumanLift, which elevates a single reference image to the 3D domain, facilitating view-consistent and photorealistic full-body image synthesis with high-quality facial details.
- A novel diffusion transformer (HumanWan-DiT) based on Wan2.1 that incorporates 3D awareness from various 3D human models, successfully synthesizing photorealistic multi-view images/normals with coherent facial details.
- A Gaussian optimization framework that integrates SMPL-X pose optimization with facial detail enhancement from generative priors, designed to elevate multi-view images into the 3D domain for further applications.

Extensive experiments demonstrate that our approach can synthesize photorealistic, consistent, and free-view human renderings, outperforming existing methods quantitatively and qualitatively. By integrating the strengths of 2D diffusion transformer, 3D SMPL-X pose optimization, and advanced face enhancement techniques, our method offers a reliable solution to high-quality 3D human modeling from a single image. Our code will be made publicly accessible for research purposes.

2 Related Work

2.1 Single-View 3D Human Creation

Significant progress has been made in 3D human reconstruction from multi-view inputs [Lin et al. 2024b; Peng et al. 2023; Saito et al. 2019] and monocular input [Alldieck et al. 2018a,b; Li et al. 2020; Weng et al. 2022], leveraging visual cues to infer 3D geometry. Recently, the focus has shifted to recovering complete human avatars—including body shape, garments, and textures—from a single image with the aid of priors. Early parametric approaches estimated coarse shapes [Alldieck et al. 2018a, 2019; Bhatnagar et al. 2019; Choutas et al. 2022; Kanazawa et al. 2018; Kocabas et al. 2020; Kolotouros et al. 2019; Saito et al. 2021; Smith et al. 2019; Sun et al. 2022; Zhu et al. 2019] but struggled with clothed humans due to fixed topology. Implicit methods [Alldieck et al. 2022; Cao et al. 2022; Corona et al. 2023; Saito et al. 2019, 2020; Xiu et al. 2023, 2022; Zhang et al. 2024c] improved flexibility by modeling complex topologies with implicit functions (SDF/UDF, signed/unsigned distance fields), though at the cost of heavy sampling.

Head avatar creation has been explored through a variety of approaches. Video-driven approaches [Shao et al. 2024b; Thies et al. 2016; Zheng et al. 2022, 2023; Zielonka et al. 2023] only reconstruct facial expressions and poses, while recent advances [An et al. 2023; Li et al. 2025] synthesize static full-head avatars by combining GAN (Generative Adversarial Network) priors with triplane representations [Chan et al. 2022]. Multi-camera systems, such as GaussianHeadAvatar [Xu et al. 2023] and GaussianAvatars [Qian et al. 2023], deliver high-fidelity and view-consistent results but rely on complex capture setups. NeRSemble [Kirschstein et al. 2023] leverages a single camera to model temporal dynamics, though it generalizes poorly across subjects. Image-based methods [Liu and Hao 2022; Yu et al. 2023] exploit GAN priors to construct avatars from one or a few inputs, achieving identity preservation but often lacking cross-view consistency.

Recent diffusion-based generative models have shown promising results with given conditional inputs, especially for single-view to multi-view generation, enabling 3D reconstruction using differentiable inverse rendering (e.g., NeRF, 3D-GS). The foundation models [Liu et al. 2023b,a; Long et al. 2024; Voleti et al. 2025; Yang et al. 2024], trained on Objaverse(-XL) dataset [2023a; 2023b], have made significant progress in 3D object generation. Due to the limited high-quality realistic 3D human data, some methods [Cao et al. 2024; Huang et al. 2024; Jiang et al. 2023; Svitov et al. 2023; Zhang et al. 2023] use Score Distillation Sampling [Poole et al. 2022] to distill 3D humans (e.g., NeRF, SDF, DMtet [Shen et al. 2021]) from the diffusion priors. However, the ambiguity of text conditions makes coarse human clothing details difficult and time-consuming to optimize. In contrast, we present 3D-aware conditions to direct multi-view generation, thereby ensuring enhanced view consistency.

Inspired by diffusion-based multi-view generations [Li et al. 2024a; Liu et al. 2023a; Long et al. 2024; Tang et al. 2024; Voleti et al. 2025], recent works directly produce multi-view human images from a single image without optimization. For example, Human-VDM [Liu et al. 2024] fine-tunes a multi-view diffusion model on realistic 3D humans, combining vanilla 3D-GS with super-resolution [Zhou et al. 2022] for detail enhancement. In contrast, Human 3Diffusion [Xue et al. 2024] unifies multi-view generation and 3D-GS within a single module for joint optimization. Both approaches, however, suffer from multi-view inconsistency due to the absence of explicit 3D priors. Recent works address this limitation by incorporating structured human priors (e.g., SMPL, SMPL-X) to improve consistency. SiTH [Ho et al. 2024] employs a diffusion model and the SMPL model prior to predict front and back human body images and then adds normal guidance for improved geometric reconstruction. HumanSplat [Pan et al. 2024] uses human semantic priors for high-quality texture modeling in 3D-GS reconstruction from multi-view images. PSHuman [Li et al. 2024b] employs separate diffusion models for the face and body to enhance facial details but struggles with challenging poses when SMPL-X pose estimation fails. MagicMan [He et al. 2024] integrates 3D awareness with pose optimization in 2D multi-view generation, substantially improving pose accuracy. LHM [Qiu et al. 2025] and IDOL [Zhuang et al. 2024] infer the high-fidelity 3D avatar in 3D-GS via a feed-forward pass.

While existing methods enhance view consistency for human generation, they often fail to preserve facial details due to the face's small pixel area in the whole human image, leading to unrealistic results. Unlike MagicMan, which focuses on generating 2D multi-view body images, our method achieves full 3D reconstruction with facial consistency by leveraging a video-diffusion model and converting temporal consistency into spatial consistency through SMPL-X 3D awareness and 3D training data. PSHuman [Li et al. 2024b] applies separate diffusion models for the face and body but remains limited by low-resolution face crops, leading to insufficient detail. In contrast, our approach incorporates facial priors to enforce identity consistency and enhance detail during both generation and reconstruction, yielding higher-quality avatars.

2.2 Human Video Diffusion

Building a unified diffusion model for generating consistent multi-view human images is challenging due to the lack of realistic 3D human data. However, human video data enables diffusion-based methods to generate realistic dynamic videos efficiently. DreamPose [Karras et al. 2023] and DISCO [Wang et al. 2023] apply diffusion to human image animation but struggle to retain fine-grained details only with CLIP features as guidance. Recent methods like AnimateAnyone [Hu 2024], MagicAnimate [Xu et al. 2024], and MagicPose [Chang et al. 2023] use ReferenceNet with spatial self-attention to maintain reference image consistency and employ pose-guided networks (e.g., DWPose 2D skeleton [Yang et al. 2023]) for animation guidance. Sketch2HumanVideo[Qu et al. 2025] proposed a sparse sketch encoder to achieve sketch-controllable human video generation. UniAnimate [Wang et al. 2024b] introduces Mamba [Gu and Dao 2023] into the diffusion model to improve the efficiency. Its extension (UniAnimate-DiT [Wang et al. 2025b]) on large-scale video diffusion transformer (Wan2.1 [Wang et al. 2025a]) achieves visually appealing and temporally coherent 2D animations. Champ [Zhu et al. 2025], MIMO [Men et al. 2024], and Human4DiT [Shao et al. 2024a] employ 3D-aware shape guidance [Loper et al. 2015; Pavlakos et al. 2019] rather than a sparse 2D skeleton, enabling more accurate and controllable image generation. These methods enhance

clothing textures and backgrounds but struggle with facial identity consistency and distortion. MimicMotion [Zhang et al. 2024a] uses FaceFusion [Henry 2024] for postprocessing to reduce face distortion, while VividPose [Wang et al. 2024a] and StableAnimator [Tu et al. 2024] employ ArcFace [Deng et al. 2019] to ensure face consistency in an end-to-end manner. While these methods improve animation realism and consistency, the 2D skeleton is unable to perceive the view changes. In contrast, we integrate 3D awareness into the diffusion model for multi-view image generation, leveraging realistic human priors from video diffusion models. Our approach utilizes a Gaussian optimization framework and enables photorealistic rendering with fine details.

3 Methodology

Given a single image $\mathcal{I}$ of a person, we aim to create a 3D digital avatar that captures the photorealistic appearance and fine details (e.g., clothing), enabling free-view exploration. This task is challenging due to the use of unconstrained daily captured photos, including varied poses, self-occlusion, unknown viewpoints, and cluttered backgrounds–especially with a single image as input. Additionally, small facial regions in full-body images result in insufficient facial details that degrade avatar quality. Photorealistic reconstruction under these conditions thus remains highly challenging.

Our key insight is to introduce 3D awareness (multi-view consistency and geometrical details) during the multi-view generation and perform the photorealistic 3D avatar reconstruction with a flexible 3D-GS representation with facial detail enhancement. As illustrated in Figure 2, our method employs two main stages for generating lifelike 3D avatars. Specifically, we first fine-tune a video diffusion transformer model [Wang et al. 2025a] injected with 3D awareness from the SMPL-X model [Pavlakos et al. 2019] to generate multi-view RGB images (HumanWan-DiT (RGB)) while maintaining consistency. Further, another fine-tuned Wan-DiT model (named HumanWan-DiT (Normal)) converts the generated multi-view RGB images to multi-view Normal images, injecting geometrical details. Both components collaboratively turn a single image to multi-view images with 3D consistency and geometrical details (Sec. 3.1). Then, we lift the multi-view images into a 3D full avatar with highly photorealistic and high-fidelity facial details (Sec. 3.2).

3.1 3D-Aware HumanWan-DiT

Our first step is to generate photorealistic multi-view images from a single daily-captured personal image. We develop a multi-view human diffusion transformer that enables 3D awareness, i.e., 3D consistency and geometrical details. For a better generalization to in-the-wild images, our backbone is based on advanced video Diffusion Transformer (DiT)-based model [Wang et al. 2025a] that is trained on numerous 2D general dynamic videos. It provides a powerful capability to infer the high-fidelity videos. We employ the similar network configurations with its extension [Wang et al. 2025b] on the 2D human animation task, with the pretrained weights inherited.

We propose to design two models, HumanWan-DiT (RGB) and HumanWan-DiT (Normal), which are tailored for the human body to enhance multi-view consistency and geometrical details across different views. The normal and semantic images of SMPL-X are introduced as 3D priors, which are then incorporated into the Wan2.1 to offer 3D guidance. Then, LoRA is employed for memory-efficient parameter fine-tuning, reducing training memory overhead while preserving generative potency. Driving motion information is encoded by a lightweight condition encoder comprising stacked 3D convolutional layers. The detailed structure is presented in Figure 3.

3D-Aware Multi-View RGB/Normal HumanWan-DiT. Building on recent advancements in general video DiT models [Wang et al. 2025a,b], Wan-VAE is a novel 3D causal VAE architecture specifically designed for efficient and scalable video generation by employing strategies in IV-VAE [Wu et al. 2024] to improve spatiotemporal compression and ensure temporal causality, making it well-suited for integration with diffusion models such as DiT. We utilize Wan-Encoder Enc_{Wan} to extract features $f_{\mathcal{I}}^{Ref}$ from a single reference image $\mathcal{I}$, which are concatenated with noise and fed into the Video DiT Model (Wan2.1), ensuring the consistency between the final generated multi-view images and the reference image at the appearance level.

Wan-DiT mainly consists of a patchifying module, transformer blocks and an unpatchifying module; please refer to Wan2.1 [Wang et al. 2025a] for the detailed architecture. The patchifier uses 3D convolutions with kernel size of (1,2,2) to convert video latents into a sequence of video tokens. The Wan-DiT block accepts the patchified embeddings of Wan-Encoder Enc_{Wan} and 3D-aware conditions as inputs. Then, the Wan-Decoder Dec_{Wan} decodes the unpatchified outputs of Wan-DiT block to the consistent multi-view results $\{\mathcal{I}_1, \mathcal{I}_2, \dots \mathcal{I}_N\}$ (in our implementation, $N = 81$ views, fitting the video DiT model). We propose to inject the 3D priors into the video model. The 3D-aware priors include the SMPL-X semantic images $\{\mathcal{I}_i^{Sem}, 1 \le i \le N\}$ and normal images $\{\mathcal{I}_i^{Nor}, 1 \le i \le N\}$.

In Figure 2, we design two DiT models, HumanWan-DiT (RGB) and HumanWan-DiT (Normal), to generate high-quality multi-view RGB and normal images. HumanWan-DiT (RGB) takes SMPL-X multi-view semantic images $\{\mathcal{I}_i^{Sem}, 1 \le i \le N\}$ as input and produces spatially consistent multi-view RGB images. HumanWan-DiT (Normal) then uses these RGB images as conditions to infer the detailed multi-view normal images, supervised by normal prior $\{\mathcal{I}_i^{Nor}, 1 \le i \le N\}$ from SMPL-X.

For the conditional encoder, we adopt a 7-layer 3D convolutional architecture [Wang et al. 2025b] to effectively capture spatial features of the conditioning inputs (SMPL-X semantic rendering for RGB branch, and RGB images for normal branch). This achieves a balance between receptive field size and computational efficiency. The resulting features are projected into a high-dimensional patchified token space (5120-D), enabling effective integration of pose information. The combination of 3D convolutional encoding and token-space representation enhances consistency and preserves fine details during multi-view generation.

Training Objective. We fine-tune two DiT models by leveraging the flow matching framework [Esser et al. 2024], following Wan2.1 [Wang et al. 2025a]. The training input x_t, follows Rectified Flows [Liu et al. 2022], is defined as a linear interpolation between random noise $x_0 \sim \mathcal{N}(0, I)$ and video latent x_1 at timestep t: $x_t = tx_1 + (1 -$

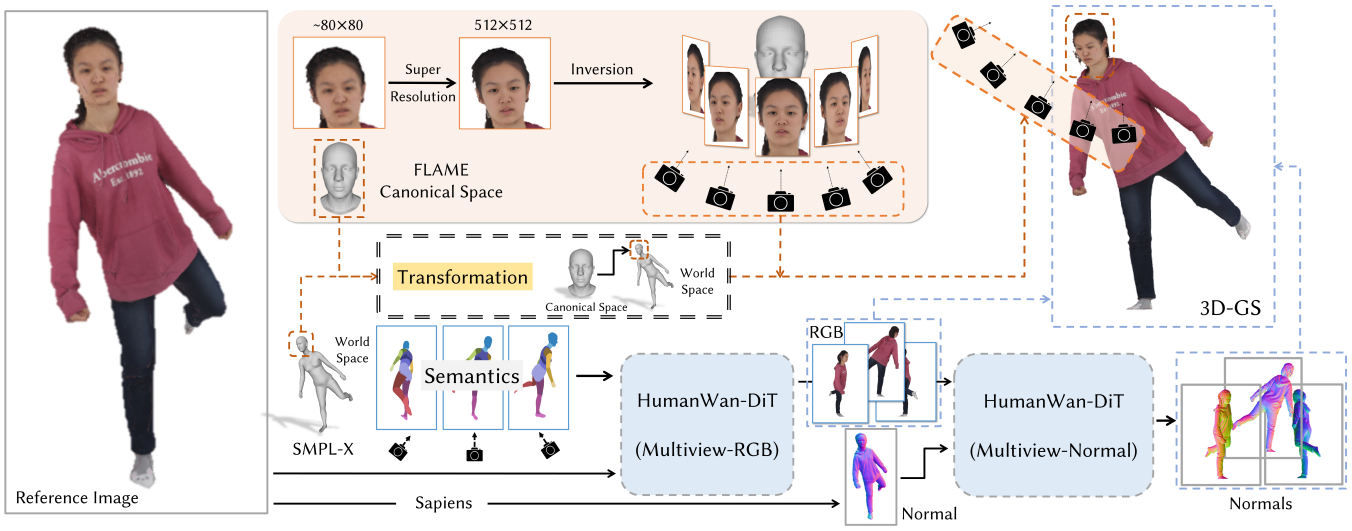

Fig. 2. The pipeline of our method HumanLift. Our reconstruction lifts the generated multi-view images to a 3D-GS representation. The human face occupies a small area within the entire image, leading to the absence of details and blurring. We propose face enhancement with SMPL-X pose optimization during the 3D-GS training. The super-resolved facial image is inverted using the existing generative model [Li et al. 2025] to produce high-quality multi-view facial images with camera poses in canonical space. The 3D-GS is supervised by the multi-view face and human images. It should be noted that the camera poses (canonical space) of the face images are transformed to the SMPL-X head (world space) to monitor the face of 3D-GS. Since the initial SMPL-X is not aligned with the 3D Gaussian points accurately, we also optimize the SMPL-X pose parameters to adjust face camera pose according to the optimized SMPL-X for each iteration, ensuring consistent alignment between face camera pose and 3D Gaussian points of the head.

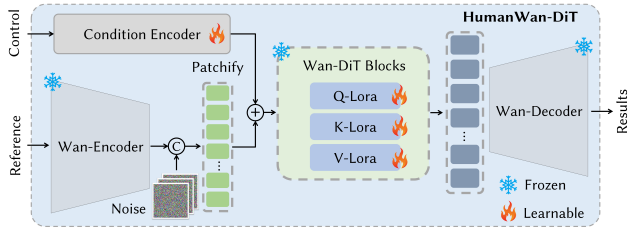

Fig. 3. Our HumanWan-DiT network architecture adapts the Wan2.1 video diffusion transformer (DiT) model [Wang et al. 2025a] for static human multi-view image generation, building upon concepts similar to UniAnimate-DiT. Its key components include the base diffusion transformer enhanced with LoRA fine-tuning and a versatile condition encoder that accepts various inputs for predicting multi-view images. To produce high-quality RGB and Normal outputs, we fine-tune two distinct models: HumanWan-DiT (RGB) and HumanWan-DiT (Normal).

$t)x_0$. The groundtruth velocity v_t is

$$v_t = \frac{dx_t}{dt} = x_1 - x_0 \tag{1}$$

The loss function can be formulated as:

$$\mathcal{L}_{DiT} = \mathbb{E}_{x_0, x_1, c_{\text{txt}}, t} \left\| u(x_t, c_{\text{txt}}, t; \theta) - v_t \right\|^2 . \tag{2}$$

The model is trained by minimizing the mean squared error between predicted velocity $u(x_t, c_{\text{txt}}, t; \theta)$ by our model and v_t, where θ is the tunable parameters, and c_{txt} is the umT5 [Chung et al. 2023] text embedding sequence of 512 tokens. In our implementation, we use the same text description ("A camera smoothly circles around the person horizontally, completing a full rotation, capturing them from all sides. The person remains still throughout the video.") for all videos.

3.2 Human 3D Gaussian with Face Enhancement

The second step involves converting two multi-view image sets: multi-view RGB: $\mathcal{S}_C = \{\tilde{I}_i^C, 1 \le i \le N\}$, and multi-view Normal: $\mathcal{S}_N = \{\tilde{I}_i^N, 1 \le i \le N\}$ into 3D for downstream tasks (e.g., geometry reconstruction, human animation). Multi-view human DiT models cannot achieve perfect consistency of generated images across different views because it just introduces 3D priors, making generated images as consistent as possible. Considering that the face only takes a small area of a human image, the existing approach creates the entire body and ignores facial details. To address this issue, we use a 3D-GS representation to learn the human and face Gaussian distributions with face enhancement, whose goal is to supervise the high-quality face learning using the inversion techniques [Guo et al. 2024a] on the super-resolved face image (512×512). Finally, a photorealistic human with a high-quality face can be rendered and animated in free views.

3D-GS [Kerbl et al. 2023] achieves impressive results on Novel View Synthesis (NVS) with efficient computation cost, while keeping explicit point clouds for other 3D applications. It can eliminate inconsistency from the diffusion model and further benefit downstream applications. Motivated by the work [Lin et al. 2024a], we leverage the layered 3D-GS representation of the face and body to reconstruct a high-quality 3D Gaussian representation. The Gaussian distributions attributed to the face and body are optimized by the different sets of images with the corresponding camera poses.

We reconstruct the 3D-GS representation of the human body using two generated multi-view image sets $\mathcal{S}_C$ and $\mathcal{S}_N$. For the

face, we employ the face generation method [Li et al. 2025] to invert the super-resolved face image, resulting in high-quality multi-view images in the 90° range in front of the face $\{\tilde{I}_i^{face}, 1 \le i \le N_f\}$ (i.e., $N_f = 40$ views). Since the head of SMPL-X model is the same as FLAME (vertices and face connection), we follow the correspondence provided by SMPL-X [Pavlakos et al. 2019] to align the FLAME to SMPL-X head using the ICP (Iterative Closest Point) algorithm. Calculating the global rigid transformation between the head of SMPL-X and FLAME [Li et al. 2017] transforms N_f face camera poses (FLAME canonical space) to supervise 3D-GS (world space) learning of the face component. However, the FLAME model cannot be aligned well with the Gaussians of the face directly due to inaccurate SMPL-X estimation from a single image, resulting in artifacts in the final rendering result, especially for the face. We thus propose SMPL-X pose optimization-based face enhancement to fix this issue. To supervise face 3D-GS learning, we continually correct the SMPL-X shape and update the head camera pose during training, resulting in high-quality and reasonable rendering outputs.

The Gaussian distribution learning uses the SMPL-X model to initialize Gaussian kernels for quick convergence and further animation. Since there are not enough views to supervise 3D-GS learning, we scale the multi-view images into four resolutions (832×480, 416×240, 208×120, 104×60) to train the 3D-GS distribution to distribute Gaussian kernels uniformly. For the losses, two groups of loss terms collaborate to optimize 3D-GS and SMPL-X parameters: $\mathbb{L}_{GS}$ and $\mathbb{L}_{SMPLX}$. $\mathbb{L}_{GS}$ includes the RGB loss for human body $\mathbb{L}_{body}^{RGB}$ and face $\mathbb{L}_{face}^{RGB}$, Normal loss for human body $\mathbb{L}_{body}^{Nor}$, Mask Loss $\mathbb{L}_{body}^{Mask}$, and Gaussian kernel constraint $\mathbb{L}^{Scale}$ (adopted from [Gao et al. 2024]) to optimize our layered 3D-GS representation:

$$\mathbb{L}_{GS} = \lambda_b \mathbb{L}_{body}^{RGB} + \lambda_n \mathbb{L}_{body}^{Nor} + \lambda_f \mathbb{L}_{face}^{RGB} + \lambda_m \mathbb{L}_{body}^{Mask} + \lambda_S \mathbb{L}^{Scale}, \quad (3)$$

where the hyper-parameters for 3D-GS optimization, $\lambda_b, \lambda_n, \lambda_f, \lambda_m, \lambda_S$, are empirically set to 1.0, 0.5, 5.0, 10.0, 100.0, respectively. For each iteration, we randomly choose a camera to optimize the human body or face. If the camera view of the human body includes the face, we will use the face camera adjacent to the human camera to create a mask to optimize the body. Another loss $\mathbb{L}_{SMPLX}$ also works together to correct the face camera pose for face 3D-GS optimization, including LandMark Loss $\mathbb{L}_{face}^{LandMark}$, SMPL-X Joints loss $\mathbb{L}_{body}^{Joints}$, and distance from Gaussian points to SMPL-X surface $\mathbb{L}_{p2s}$:

$$\mathbb{L}_{SMPL-X} = \lambda_{LM} \mathbb{L}_{face}^{LandMark} + \lambda_J \mathbb{L}_{body}^{Joints} + \lambda_{p2s} \mathbb{L}_{p2s}, \quad (4)$$

where the GT Face landmarks and SMPL-X joints are detected by [Cao et al. 2019] and optimized in the five front views. All our tests use default hyperparameters of 100.0, 1.0, 50.0 for $\lambda_{LM}, \lambda_J, \lambda_{p2s}$.

4 Experiments

In this section, we present qualitative and quantitative evaluations of our method on novel view synthesis of 3D-GS reconstruction, including comparisons with existing approaches, user study, and ablation studies to analyze the importance of our key designs. For data preparation and implementations, please refer to our supplementary material.

4.1 Evaluations

4.1.1 Baselines and Metrics. We evaluate the performance of our proposed method HumanLift and alternative baselines on the Human4DiT Dataset[Shao et al. 2024a] and various in-the-wild examples qualitatively and quantitatively. We consider comparing with two types of methods: one is based on single view human reconstruction, and the other is based on 2D human animation. We consider five state-of-the-art methods for single view human reconstruction: Human 3Diffusion [Xue et al. 2024], SiTH [Ho et al. 2024], PSHuman [Li et al. 2024b], MagicMan [He et al. 2024], and IDOL [Zhuang et al. 2024]. Human 3Diffusion, combined with 3D-GS focuses on human multi-view image generation. SiTH and PSHuman utilize the diffusion model to generate images from extra views and reconstruct a 3D human avatar using the SDF representation and [Palfinger 2022]. MagicMan is a recent work that achieves impressive results in human multi-view generation with SMPL-X pose optimization. IDOL is a recent human 3D Gaussian inference method using a single image in a feed-forward way.

In addition, we compare with several methods on 2D human animation: MagicAnimate [Xu et al. 2024], AnimateAnyone [Hu 2024], Champ [Zhu et al. 2025], Human4DiT [Shao et al. 2024a]. All these method can synthesis the realistic 2D human animation via different motion images (e.g. 2D skeleton, SMPL semantic images), which can achieve the static multi-view human image synthesis by changing the conditional motion images. We follow the same setting (fine-tuned MagicAnimate, AnimateAnyone and Champ on Human4DiT dataset) as the Human4DiT to enable them to generate the multi-view video for static 3D humans, performing the comparison as fair as possible. We adopt three commonly used metrics (i.e., Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM) [Wang et al. 2004], and Learned Perceptual Image Patch Similarity (LPIPS) [Zhang et al. 2018]) to evaluate the multi-view generation performance of each method qualitatively. CSIM [Guo et al. 2024b] assesses face consistency by comparing the cosine similarity of facial embeddings from two images. Since the GT images are strictly identity-consistent, the CSIM scores inherently reflect the identity coherence across views and the consistency with the reference image. Furthermore, to explicitly assess identity drift across views, we compute the average CSIM between consecutive frames.

4.1.2 Results. We have validated the performance on test set of Human4DiT dataset and various in-the-wild images. Figure 4 presents novel-view synthesis results on four cases. For each example, three views are displayed. We highlighted the details with colored box for detailed comparison, focusing on clothing details and multi-view consistency. From the results, it is obvious that our method can synthesize higher-fidelity details exceptionally and preserve consistency with the reference image better. Human 3Diffusion and MagicMan use SMPL-X priors for consistency but cannot preserve facial details well (the first row in Figure 4). We also compare with the recent impressive work IDOL, which efficiently infers a 3D-GS human in a feed-forward manner. In Figure 6, we evaluate the performance on two given examples in their official code repository. From the visual results, it is clear that our method can maintain the ultra-well consistency with the input image and can synthesize complex and fine garment details well. In contrast, the baselines

Fig. 4. **Evaluations on Novel View Synthesis of 3D-GS Reconstruction.** We evaluate novel view synthesis and compare our method with four alternatives, including Human 3Diffusion [Xue et al. 2024], SiTH [Ho et al. 2024], MagicMan [He et al. 2024], and PSHuman [Li et al. 2024b]. The facial and body details are highlighted by the colored boxes. The results indicate that our method achieves the best visual performance with high-quality facial details, while preserving more clothing details. Zoom in to see the details.

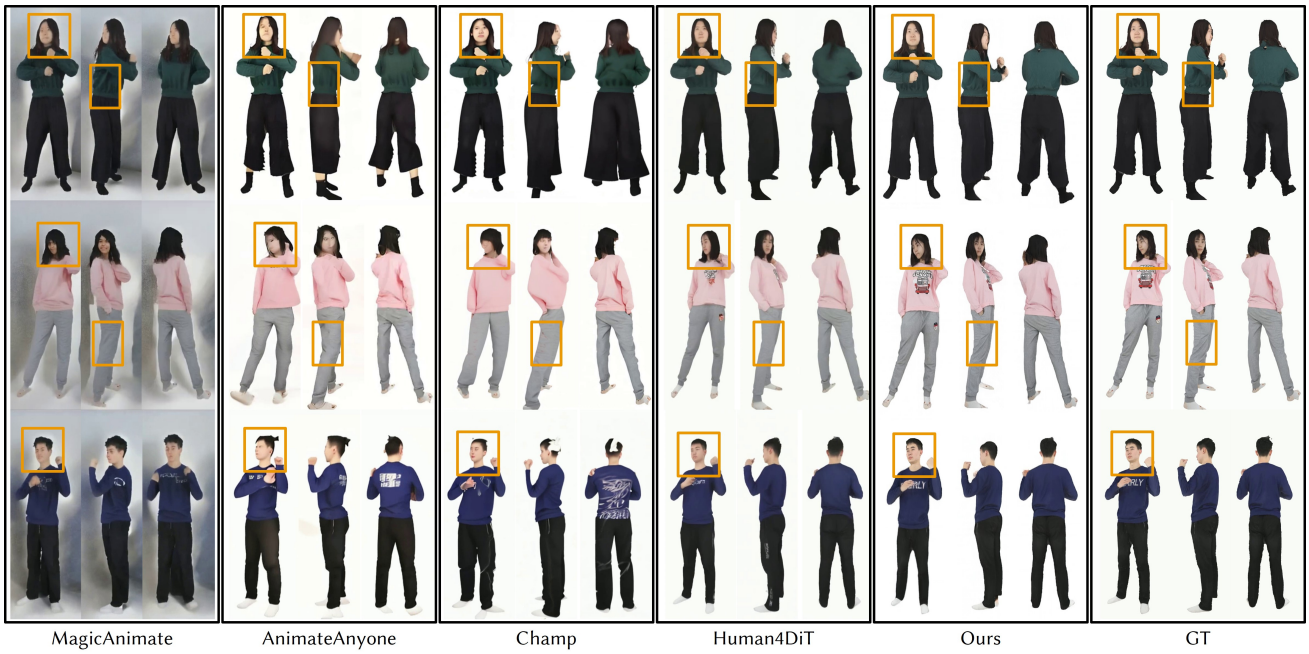

Fig. 5. **Comparison with Methods based on 2D Human Animation.** We evaluate novel view synthesis and compare our method with four alternatives, including MagicAnimate [Xu et al. 2024], AnimateAnyone [Hu 2024], Champ [Zhu et al. 2025], Human4DiT [Shao et al. 2024a]. We fine-tune MagicAnimate, AnimateAnyone, and Champ on THuman2.0 dataset to enable them to generate the multi-view image for static 3D humans. The facial and cloth details are highlighted by the colored boxes. It is clear that our method achieves the best visual performance, while capturing more details on clothing and preserving the garment wrinkles from reference images. Zoom in to see the details.

Table 1. **Quantitative comparison.** In this table, we evaluate the performance of all baselines in our paper. Four metrics on image quality and face consistency are reported. CSIM-GT scores inherently reflect the identity coherence across views and the consistency with the GT images. CSIM-Adj scores measure the facial similarity between consecutive frames, reporting the average CSIM. For single-view human reconstruction, we evaluate the performance of our method compared with Human 3Diffusion, SiTH, MagicMan PSHuman, and IDOL. For multi-view human generation, we evaluate the performance on multi-view renderings with MagicAnimate, AnimateAnyone, Champ, and Human4DiT. The results clearly demonstrate that our method achieves the best performance over recent alternatives in image quality metrics (PSNR, SSIM, LPIPS), and face consistency. The top two results are colored as first and second. Mark * means that we fine-tuned them on static 3D THuman2.0 dataset to enable them to have the ability of multi-view generation for static 3D humans.

Methods	Multi-View Human Image Quality			Face Consistency	
	PSNR ↑	SSIM ↑	LPIPS ↓	CSIM-GT ↑	CSIM-Adj ↑
Human 3Diffusion	21.49	0.918	0.060	0.316	0.579
MagicMan	23.78	0.921	0.057	0.611	0.730
SiTH	18.45	0.820	0.100	0.457	0.601
PSHuman	20.76	0.850	0.079	0.701	0.737
IDOL	20.90	0.872	0.104	0.502	0.698
MagicAnimate*	20.16	0.910	0.141	0.417	0.537
AnimateAnyone*	21.92	0.937	0.068	0.355	0.479
Champ*	22.18	0.940	0.060	0.503	0.615
Human4DiT	23.37	0.962	0.045	0.730	0.690
Ours w/o Face Enhancement	21.33	0.850	0.049	0.606	0.742
Ours w/o SMPL-X Optimization	19.71	0.797	0.053	0.534	0.701
Ours w/o Normal Supervision	23.60	0.947	0.045	0.740	0.778
Ours	**24.17**	**0.967**	**0.042**	**0.781**	**0.802**

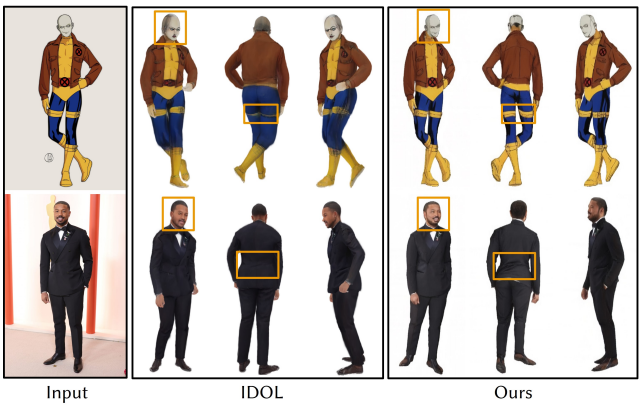

Fig. 6. **Comparison with IDOL.** We compare with IDOL [Zhuang et al. 2024], a feed-forward method to predict 3D Gaussians. We choose two demo examples in their official code repository, a cartoon character and a real human. The difference on facial and body details are highlighted by the colored boxes. The results indicate that our method achieves the best visual performance and consistent multi-view images with high-quality details. Zoom in to see the details.

struggle to catch details, notably garment details and consistency. While PSHuman employs the body-face cross-scale diffusion model to improve facial details, it still struggles to capture high-frequency details like garment wrinkles (see Figure 4).

Further, we follow the same setting as Human4DiT to perform the evaluation with 2D human animation methods, which also support the multi-view image synthesis for 3D static human using the suitable motion control. We select 100 3D humans as test set from THUman2.0 and render them to multi-view image set for comparison. The frontal view is used as the reference image and others are the test images for numerical evaluation. For each reference image, we render the estimated SMPL-X model to 2D skeleton, depth, semantic, normal images in 360° as condition, feeding them into the above 2D human animation method (MagicAnimate, AnimateAnyone, Champ, Human4DiT). Qualitative comparison is presented in Figure 5. Our methods outperforms others, demonstrating the ability to learn the spatial consistency and details preservation, as well as the quantitative evaluation in Table 1.

Table 2. **User Study.** We conducted a user study to compare our method against existing techniques using in-the-wild examples, evaluating them from a human perspective under four criterias. Our method achieves the best in all criterias. Top two results are colored as first and second.

Methods	Human Preference			
	3D-C ↓	CR ↓	FD ↓	RQ ↓
Human 3Diffusion	3.67	2.65	4.73	3.46
MagicMan	3.13	2.00	2.31	4.58
SiTH	2.64	4.46	3.00	3.50
PSHuman	3.56	4.23	3.15	1.92
Ours	**2.00**	**1.65**	**1.81**	**1.54**

Additionally, we evaluate the performance numerically in Table 1 by calculating the average score across views. Table 1 reports the results and clearly shows our method outperforms the all baselines in four metrics (PSNR, SSIM, LPIPS, and face consistency). The results reveal that our method can generate photorealistic and view-consistency multi-view images, achieving the best consistency across different views and between reference images. Table 1 also reports the performance on 2D human animation methods, regarding multi-view renderings. It is obvious that our method achieves the best performance on the three commonly used metrics, while maintaining the consistent face identity. Our approach can synthesize realistic and high-quality face details thanks to 3D-GS representation with face enhancement.

Figure 7 present the multi-view rendering results on two cases. To verify our generalization ability, we present additional diverse results in our supplementary material and video, involving various styles of human images. As is clearly visible, our results achieve photorealism and consistency while preserving more complex details from reference images. Especially for some decorations or pendants, our method can retain the features into multi-view images with high consistency.

4.1.3 User Study. We conducted a user study to compare our method against existing techniques using in-the-wild examples, evaluating them from a human perspective. We randomly selected 15 examples and invited 20 participants to complete our questionnaires. Participants ranked all alternatives based on four criteria: 3D spatial consistency (3D-C), consistency with reference image (CR), facial details (FD), and overall rendering quality (RQ). The rankings provided by participants served as scores. In Table 2, our method consistently outperformed others across all criteria, demonstrating its superior performance in 3D human generation from a single image.

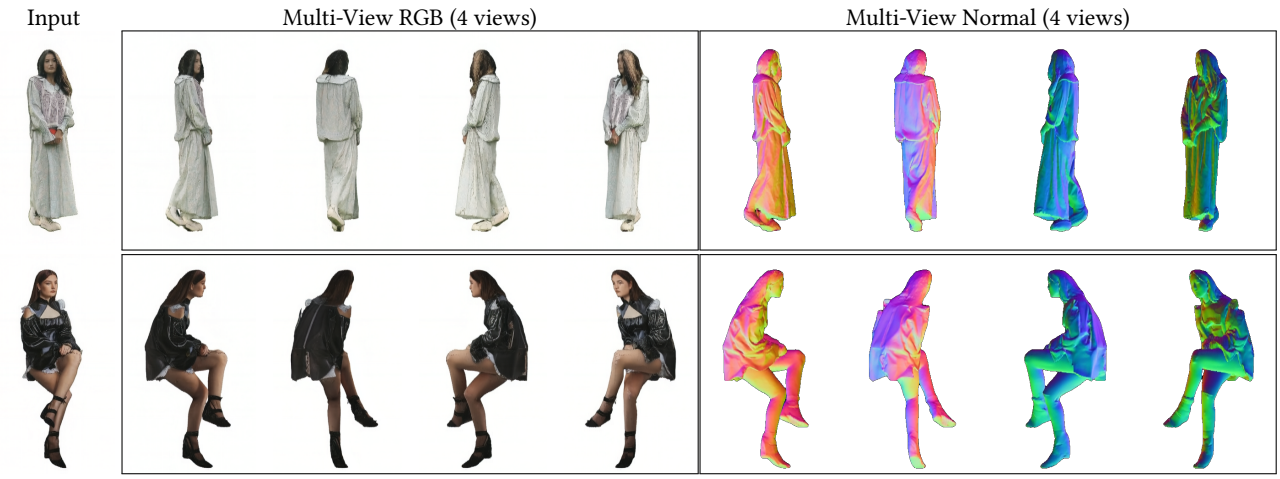

Fig. 7. **More Results on Various Cloth Styles and Human Poses.** Here are three in-the-wild examples. For each reference image, we present four multiview generated RGB and Normal images. It is evident that our results are photorealistic, with spatial consistency and fine geometrical details. Please refer to our supplementary and video for more multi-view generation results.

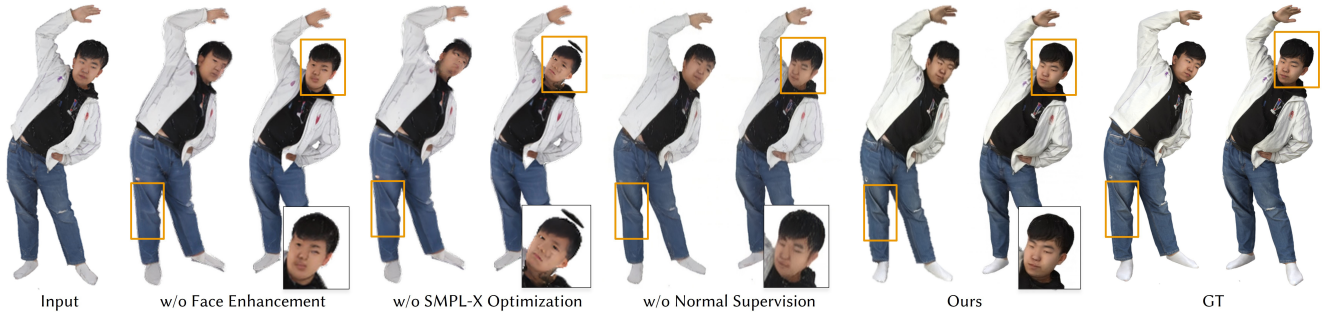

Fig. 8. **Ablations.** We conduct a qualitative comparison of our three ablated versions: without Face Enhancement, without SMPL-X pose optimization, and without normal supervision. The results indicate that face enhancement markedly enhances facial details through generative priors. Additionally, SMPL-X optimization effectively adjusts the SMPL-X pose and subsequently updates the face camera pose to guide the 3D-GS on the head, resulting in more plausible rendering results. The differences are highlighted with colored boxes. The extra normal supervision from our fine-tuned HumanWan-DiT Normal model indicates that our 3D-GS representation can maintain more clothing details under the guidance of generated consistent multi-view normal images. Zoom in to see the details.

4.2 Ablations

In this section, we demonstrate the effectiveness of our key designs, which play an important role in the photorealistic human reconstruction. Specifically, we evaluate the effectiveness of Normal Supervision, Face Enhancement, and SMPL-X optimization during 3D-GS optimization.

4.2.1 Normal Supervision. To further improve the quality of 3D-GS reconstruction, we fine-tune a multi-view normal prediction model using our HumanWan-DiT, which takes the multi-view RGB images as control inputs and under the guidance of reference normal image extracted by Sapiens. The model aims to predict the multi-view normal images that is consistent as generated multi-view RGB images, while capturing more geometrical details from reference image. To validate the effectiveness of Normal Supervision on 3D-GS learning, we remove the supervision to train the 3D-GS under the guidance of multi-view RGB images and detailed facial images, using PSNR, SSIM, LPIPS and CSIM to measure the quality of multi view images and face consistency across the views. Figure 8 illustrates that Normal Supervision can successfully improve the rendering quality both quantitatively and qualitatively, and preserve more details from the reference image.

4.2.2 Face Enhancement. Although we tried our best to preserve the facial details during multi-view generation, the human face still occupies a small area in the reference image, which makes it hard to capture sufficient details for high-quality and photorealistic 3D-GS reconstruction. To address this, we introduce the Face Enhancement, which leverages the face generation prior and super-resolution technique to improve the human face quality. In Figure 8, it is evident that the ablated version leads to blurred facial results, while our methods can capture more details.

4.2.3 SMPL-X Optimization. Since the SMPL-X pose estimation from the single image is an ill-posed problem, it often results in inaccurate pose estimation when the reference image contains more challenging poses. This will cause the camera pose of the face image to be incorrectly positioned during 3D-GS reconstruction, which is misaligned with the trained 3D-GS supervised by multi-view human

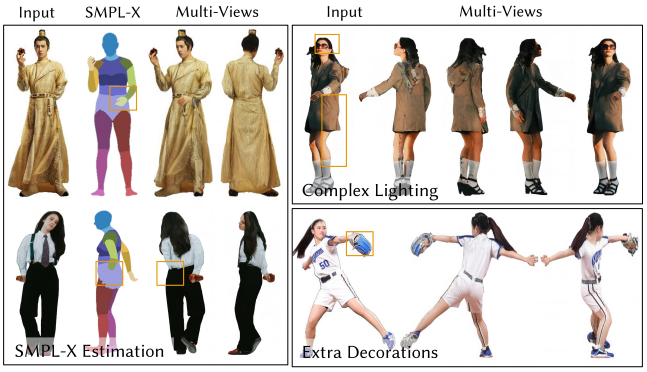

Fig. 9. **Failure Cases.** We show three types of failure cases to illustrate the limitations of our method. For complex clothing or human poses, inaccurate SMPL-X pose prediction leads to inconsistent rendering results from reference images (such as the hands in left part). For the complex lighting and other unseen object attached on human, it still results in unrealistic rendering results and unreasonable multi-view images for objects. Zoom in to see the details.

images, resulting in the head being inaccurately placed in the final rendered image. Figure 8 illustrates the artifact clearly. Disabling SMPL-X pose optimization during 3D-GS training leads to incorrect head poses and facial artifacts.

5 Conclusion and Discussion

In this work, we introduced HumanLift, a novel approach for 3D human reconstruction from a single reference image using video DiT priors and facial detail enhancement. Our HumanLift aims to recover a 3D-GS represented human from a single RGB image. The 3D human is reconstructed under the three supervisions: 1. multi-view RGB images, 2. multi-view normal images, and 3. multi-view facial images. To generate the high-quality and consistent multi-view images, we build our model upon powerful DiT model (Wan2.1) for the consistent and high-fidelity multi-view RGB/Normal images, which leverages 3D awareness from SMPL-X and geometrical details from 3D human scan data to preserve coherence and details across views. To enable downstream 3D applications requiring high-quality facial details, we raise the multi-view images to the 3D-GS representation based on SMPL-X pose optimization and face enhancement. Overall, we offer a viable method for photorealistic 3D person modeling from a single image.

Limitations & Future Works. Our approach achieves realistic 3D human rendering with high-quality facial details, yet it has several drawbacks: 1. Inaccurate SMPL-X and Inconsistent Multi-view Generation: Single-image pose estimation remains an ill-posed problem, leading to inaccurate SMPL-X estimations for complex poses and clothing, especially for severe occlusion of human body. This inconsistency propagates to the generated multi-view images (see left in Figure 9). Inspired by [He et al. 2024], we are integrating multi-view generation and 3D-GS optimization into a unified, end-to-end framework. This allows for iterative refinement, where generated multi-view images correct SMPL-X poses, which then guide subsequent image generation. 2. Challenges with Novel Decorations: Since our training data does not cover more diverse decorations for human body, our method struggles to generate accurate multi-view images for unusual personal decorations (e.g. a baseball glove), resulting in artifacts and ambiguous representations. Augmenting datasets might improve the generalization capabilities in these challenging cases. 3. Lighting Condition Limitations: Our human DiT model can sometimes bake challenging lighting conditions directly into the multi-view images, leading to unrealistic renderings (see right in Figure 9). 4. Weak Supervision for Unseen Regions: Direct optimization of 3D-GS from generated multi-view images results in weak supervision for self-occluded and unseen regions. Inspired by Pippo [Kant et al. 2025], a unified large model for photorealistic human body and face is a promising direction for 3D digital human creation. Also, we are interested in exploring multi-view generation and dynamic generation for humans by designing a 4D multi-view generation framework and introducing more priors during 3D-GS reconstruction in the future.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62322210 and No.62302484), the Innovation Funding of ICT, CAS (No. E461020), the Beijing Municipal Science and Technology Commission (No. Z231100005923031), the China Postdoctoral Science Foundation (No. BX20230377 and No. 2023M743568), and the Engineering and Physical Sciences Research Council (No. EP/Y028805/1). The authors would like to acknowledge the Nanjing Institute of InforSuperBahn OneAiNexus for providing the training and evaluation platform.

References

Thiemo Alldieck, Marcus A. Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018a. Detailed Human Avatars from Monocular Video. In *International Conference on 3D Vision (3DV)*.

Thiemo Alldieck, Marcus A. Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018b. Video Based Reconstruction of 3D People Models. In *Computer Vision and Pattern Recognition (CVPR)*.

Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. 2019. Tex2Shape: Detailed Full Human Body Geometry From a Single Image. In *Computer Vision and Pattern Recognition (CVPR)*.

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

Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y Ogras, and Linjie Luo. 2023. PanoHead: Geometry-aware 3D full-head synthesis in 360°. In *Computer Vision and Pattern Recognition (CVPR)*. 20950–20959.

Alexandru O. Balan, Leonid Sigal, Michael J. Black, James E. Davis, and Horst W. Haussecker. 2007. Detailed human shape and pose from images. In *Computer Vision and Pattern Recognition (CVPR)*.

Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. 2019. Multi-garment net: Learning to dress 3d people from images. In *International Conference on Computer Vision (ICCV)*. 5420–5430.

Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. 2024. DreamAvatar: Text-and-Shape Guided 3D Human Avatar Generation via Diffusion Models. *Computer Vision and Pattern Recognition (CVPR)* (2024).

Yukang Cao, Guanying Chen, Kai Han, Wenqi Yang, and Kwan-Yee K. Wong. 2022. JIFF: Jointly-aligned Implicit Face Function for High Quality Single View Clothed Human Reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*.

Zhe Cao, Gines Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2019).

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. 2022. Efficient Geometry-Aware 3D Generative Adversarial Networks. In *Computer Vision and Pattern Recognition (CVPR)*.

Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. 2023. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *International Conference on Machine Learning (ICML)*.

Vasileios Choutas, Lea Muller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, and Michael J. Black. 2022. Accurate 3D Body Shape Regression using Metric and Semantic Attributes. In *Computer Vision and Pattern Recognition (CVPR)*.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *arXiv preprint arXiv:2304.09151* (2023).

Enric Corona, Mihai Zanfir, Thiemo Alldieck, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. 2023. Structured 3D Features for Reconstructing Relightable and Animatable Avatars. In *Computer Vision and Pattern Recognition (CVPR)*.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. 2023a. Objaverse-XL: A Universe of 10M+ 3D Objects. *arXiv preprint arXiv:2307.05663* (2023).

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023b. Objaverse: A universe of annotated 3d objects. In *Computer Vision and Pattern Recognition (CVPR)*.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Computer Vision and Pattern Recognition (CVPR)*. 4690–4699.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *International Conference on Machine Learning (ICML)*.

Lin Gao, Jie Yang, Bo-Tao Zhang, Jia-Mu Sun, Yu-Jie Yuan, Hongbo Fu, and Yu-Kun Lai. 2024. Real-time large-scale deformation of Gaussian Splatting. *Transactions on Graphics (TOG)* 43, 6 (2024), 1–17.

Xiangjun Gao, Jiaolong Yang, Jongyoo Kim, Sida Peng, Zicheng Liu, and Xin Tong. 2022. Mps-nerf: Generalizable 3d human rendering from multiview images. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2022).

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* (2023).

Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. 2024b. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168* (2024).

Minghao Guo, Bohan Wang, Kaiming He, and Wojciech Matusik. 2024a. TetSphere Splatting: Representing High-Quality Geometry with Lagrangian Volumetric Meshes. *arXiv preprint arXiv:2405.20283* (2024).

Xu He, Xiaoyu Li, Di Kang, Jiangnan Ye, Chaopeng Zhang, Liyang Chen, Xiangjun Gao, Han Zhang, Zhiyong Wu, and Haolin Zhuang. 2024. Magicman: Generative novel view synthesis of humans with 3d-aware diffusion and iterative refinement. *arXiv preprint arXiv:2408.14211* (2024).

Ruhs Henry. 2024. FaceFusion. https://github.com/facefusion/facefusion.

I Ho, Jie Song, Otmar Hilliges, et al. 2024. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Computer Vision and Pattern Recognition (CVPR)*. 538–549.

Li Hu. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Computer Vision and Pattern Recognition (CVPR)*. 8153–8163.

Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. 2024. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *International Conference on 3D Vision (3DV)*.

Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2023. AvatarCraft: Transforming Text into Neural Human Avatars with Parameterized Shape and Pose Control. In *International Conference on Computer Vision (ICCV)*.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2018. End-to-end Recovery of Human Shape and Pose. In *Computer Vision and Pattern Recognition (CVPR)*.

Yash Kant, Ethan Weber, Jin Kyu Kim, Rawal Khirodkar, Su Zhaoen, Julieta Martinez, Igor Gilitschenski, Shunsuke Saito, and Timur Bagautdinov. 2025. Pippo: High-Resolution Multi-View Humans from a Single Image. *arXiv preprint arXiv:2502.07785* (2025).

Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. 2023. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *International Conference on Computer Vision (ICCV)*. IEEE, 22623–22633.

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

Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. 2023. NeRSemble: Multi-view Radiance Field Reconstruction of Human Heads. *arXiv preprint arXiv:2305.03027* (2023).

Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. 2020. VIBE: Video Inference for Human Body Pose and Shape Estimation. In *Computer Vision and Pattern Recognition (CVPR)*.

Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. 2019. Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop. In *International Conference on Computer Vision (ICCV)*.

Heyuan Li, Ce Chen, Tianhao Shi, Yuda Qiu, Sizhe An, Guanying Chen, and Xiaoguang Han. 2025. Spherehead: stable 3d full-head synthesis with spherical tri-plane representation. In *European Conference on Computer Vision (ECCV)*. Springer, 324–341.

Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. 2024a. Era3D: High-Resolution Multiview Diffusion using Efficient Row-wise Attention. *Conference on Neural Information Processing Systems (NeurIPS)* (2024).

Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Mengfei Li, Xiaowei Chi, Siyu Xia, Wei Xue, et al. 2024b. PSHuman: Photorealistic Single-view Human Reconstruction using Cross-Scale Diffusion. *arXiv preprint arXiv:2409.10141* (2024).

Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. 2020. Monocular real-time volumetric performance capture. In *European Conference on Computer Vision (ECCV)*.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *Transactions on Graphics (TOG)* 36, 6 (2017), 194–1.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2023. Magic3D: High-Resolution Text-to-3D Content Creation. In *Computer Vision and Pattern Recognition (CVPR)*.

Lixiang Lin, Songyou Peng, Qijun Gan, and Jianke Zhu. 2024b. FastHuman: Reconstructing High-Quality Clothed Human in Minutes. In *International Conference on 3D Vision, 3DV*.

Siyou Lin, Zhe Li, Zhaoqi Su, Zerong Zheng, Hongwen Zhang, and Yebin Liu. 2024a. Layga: Layered gaussian avatars for animatable clothing transfer. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*. 1–11.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023b. Zero-1-to-3: Zero-shot One Image to 3D Object. In *International Conference on Computer Vision (ICCV)*.

Shiguang Liu and Jiaqi Hao. 2022. Generating talking face with controllable eye movements by disentangled blinking feature. *Transactions on Visualization and Computer Graphics (TVCG)* 29, 12 (2022), 5050–5061.

Xingchao Liu, Chengyue Gong, and Qiang Liu. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003* (2022).

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. 2023a. SyncDreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453* (2023).

Zhibin Liu, Haoye Dong, Aviral Chharia, and Hefeng Wu. 2024. Human-vdm: Learning single-image 3d human gaussian splatting from video diffusion models. *arXiv preprint arXiv:2409.02851* (2024).

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. 2024. Wonder3d: Single image to 3d using cross-domain diffusion. In *Computer Vision and Pattern Recognition (CVPR)*. 9970–9980.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *Transactions on Graphics (TOG)* (2015).

Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. 2024. Mimo: Controllable character video synthesis with spatial decomposed modeling. *arXiv preprint arXiv:2409.16160* (2024).

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. 2021. Neural Articulated Radiance Field. In *International Conference on Computer Vision (ICCV)*.

Werner Palfinger. 2022. Continuous remeshing for inverse rendering. *Computer Animation and Virtual Worlds* 33, 5 (2022), e2101.

Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. 2024. HumanSplat: Generalizable Single-Image Human Gaussian Splatting with Structure Priors. *Conference on Neural Information Processing Systems (NeurIPS)* (2024).

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

Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. 2021a. Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies. In *International Conference on Computer Vision (ICCV)*. 14314–14323.

Sida Peng, Chen Geng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. 2023. Implicit Neural Representations with Structured Latent Codes for Human Body Modeling. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2023).

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021b. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Computer Vision and Pattern Recognition (CVPR)*. 9054–9063.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. DreamFusion: Text-to-3D using 2D diffusion. *arXiv preprint arXiv:2209.14988* (2022).

Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. 2023. GaussianAvatars: Photorealistic head avatars with rigged 3D Gaussians. *arXiv preprint arXiv:2312.02069* (2023).

Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, et al. 2025. LHM: Large Animatable Human Reconstruction Model from a Single Image in Seconds. *arXiv preprint arXiv:2503.10625* (2025).

Linzi Qu, Jiaxiang Shang, Miu-Ling Lam, and Hongbo Fu. 2025. Controllable Human Video Generation from Sparse Sketches. *IEEE Transactions on Visualization and Computer Graphics* (2025).

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*.

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. 2019. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In *International Conference on Computer Vision (ICCV)*.

Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 2020. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *Computer Vision and Pattern Recognition (CVPR)*.

Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. 2021. SCANimate: Weakly Supervised Learning of Skinned Clothed Avatar Networks. In *Computer Vision and Pattern Recognition (CVPR)*.

Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. 2024a. Human4DiT: Free-view Human Video Generation with 4D Diffusion Transformer. *arXiv preprint arXiv:2405.17405* (2024).

Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. 2024b. SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *Computer Vision and Pattern Recognition (CVPR)*.

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. 2021. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Conference on Neural Information Processing Systems (NeurIPS)* 34 (2021), 6087–6101.

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. 2024. MVDream: Multi-view Diffusion for 3D Generation. *International Conference on Learning Representations (ICLR)* (2024).

David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. 2019. FACSIMILE: Fast and Accurate Scans From an Image in Less Than a Second. In *International Conference on Computer Vision (ICCV)*.

Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J. Black. 2022. Putting People in their Place: Monocular Regression of 3D People in Depth. In *Computer Vision and Pattern Recognition (CVPR)*.

David Svitov, Dmitrii Gudkov, Renat Bashirov, and Victor Lemptisky. 2023. DINAR: Diffusion Inpainting of Neural Textures for One-Shot Human Avatars. In *International Conference on Computer Vision (ICCV)*.

Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. 2024. MVDiffusion++: A Dense High-resolution Multi-view Diffusion Model for Single or Sparse-view 3D Object Reconstruction. *arXiv preprint arXiv:2402.12712* (2024).

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of RGB videos. In *Computer Vision and Pattern Recognition (CVPR)*. 2387–2395.

Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. 2024. StableAnimator: High-Quality Identity-Preserving Human Image Animation. *arXiv preprint arXiv:2411.17697* (2024).

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. 2025. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision (ECCV)*. Springer, 439–457.

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

Qilin Wang, Zhengkai Jiang, Chengming Xu, Jiangning Zhang, Yabiao Wang, Xinyi Zhang, Yun Cao, Weijian Cao, Chengjie Wang, and Yanwei Fu. 2024a. VividPose: Advancing Stable Video Diffusion for Realistic Human Image Animation. *arXiv preprint arXiv:2405.18156* (2024).

Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. 2023. Disco: Disentangled control for referring human dance generation in real world. *arXiv e-prints* (2023), arXiv–2307.

Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. 2024b. UniAnimate: Taming Unified Video Diffusion Models for Consistent Human Image Animation. *arXiv preprint arXiv:2406.01188* (2024).

Xiang Wang, Shiwei Zhang, Longxiang Tang, Yingya Zhang, Changxin Gao, Yuehuan Wang, and Nong Sang. 2025b. UniAnimate-DiT: Human Image Animation with Large-Scale Video Diffusion Transformer. *arXiv preprint arXiv:2504.11289* (2025).

Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. 2024c. A recipe for scaling up text-to-video generation with text-free videos. In *Computer Vision and Pattern Recognition (CVPR)*. 6572–6582.

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

Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. 2022. HumanNeRF: Free-Viewpoint Rendering of Moving People From Monocular Video. In *Computer Vision and Pattern Recognition (CVPR)*.

Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. 2020. Multi-view neural human rendering. In *Computer Vision and Pattern Recognition (CVPR)*. 1682–1691.

Pingyu Wu, Kai Zhu, Yu Liu, Liming Zhao, Wei Zhai, Yang Cao, and Zheng-Jun Zha. 2024. Improved Video VAE for Latent Video Diffusion Model. *arXiv preprint arXiv:2411.06449* (2024).

Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. 2023. ECON: Explicit Clothed humans Optimized via Normal integration. In *Computer Vision and Pattern Recognition (CVPR)*.

Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. 2022. ICON: Implicit Clothed humans Obtained from Normals. In *Computer Vision and Pattern Recognition (CVPR)*.

Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. 2023. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. *arXiv preprint arXiv:2312.03029* (2023).

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. 2024. Magicanimate: Temporally consistent human image animation using diffusion model. In *Computer Vision and Pattern Recognition (CVPR)*. 1481–1490.

Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard Pons-Moll. 2024. Human 3Diffusion: Realistic Avatar Creation via Explicit 3D Consistent Diffusion Models. *arXiv preprint arXiv:2406.08475* (2024).

Haibo Yang, Yang Chen, Yingwei Pan, Ting Yao, Zhineng Chen, Chong-Wah Ngo, and Tao Mei. 2024. Hi3D: Pursuing High-Resolution Image-to-3D Generation with Video Diffusion Models. In *ACM International Conference on Multimedia (MM)*. 6870–6879.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. 2023. Effective whole-body pose estimation with two-stages distillation. In *International Conference on Computer Vision (ICCV)*. 4210–4220.

Wangbo Yu, Yanbo Fan, Yong Zhang, Xuan Wang, Fei Yin, Yunpeng Bai, Yan-Pei Cao, Ying Shan, Yang Wu, Zhongqian Sun, et al. 2023. NOFA: NeRF-based one-shot facial avatar reconstruction. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*. 1–12.

Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Kang Du, and Min Zheng. 2023. AvatarVerse: High-quality & Stable 3D Avatar Creation from Text and Pose. *Conference on Artificial Intelligence (AAAI)* (2023).

Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. 2024b. CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets. *Transactions on Graphics (TOG)* 43, 4 (2024), 1–20.

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

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. 2024a. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680* (2024).

Zechuan Zhang, Zongxin Yang, and Yi Yang. 2024c. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*. 9936–9947.

Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. 2022. I M avatar: Implicit morphable head avatars from videos. In *Computer Vision and Pattern Recognition (CVPR)*. 13545–13555.

Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. 2023. PointAvatar: Deformable point-based head avatars from videos. In *Computer*

Vision and Pattern Recognition (CVPR). 21057–21067.

Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. 2022. Towards robust blind face restoration with codebook lookup transformer. *Conference on Neural Information Processing Systems (NeurIPS)* 35 (2022), 30599–30611.

Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. 2019. Detailed Human Shape Estimation from a Single Image by Hierarchical Mesh Deformation. In *Computer Vision and Pattern Recognition (CVPR)*.

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. 2025. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision (ECCV)*. Springer, 145–162.

Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. 2024. IDOL: Instant Photorealistic 3D Human Creation from a Single Image. *arXiv preprint arXiv:2412.14963* (2024).

Wojciech Zielonka, Timo Bolkart, and Justus Thies. 2023. Instant volumetric head avatars. In *Computer Vision and Pattern Recognition (CVPR)*. 4574–4584.