InfiniHuman: Realistic 3D Human Creation with Precise Control

YUXUAN XUE, University of Tuebingen, Germany
XIANGHUI XIE, University of Tuebingen, Germany and Max Planck Institute for Informatics, Germany
MARGARET KOSTYRKO, University of Tuebingen, Germany
GERARD PONS-MOLL, University of Tuebingen, Germany and Max Planck Institute for Informatics, Germany

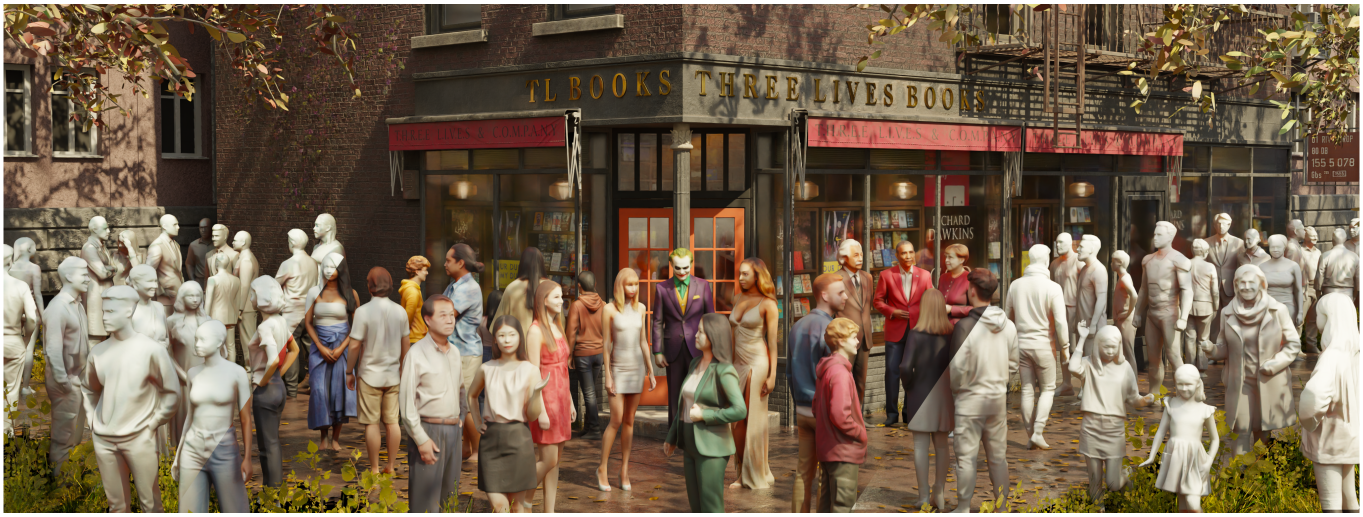

Fig. 1. Using *text description*, *explicit body shape*, *cloth image* as input, our 3D human generative method, **InfiniHuman**, can automatically create a variety of realistic 3D humans with high-fidelity texture and geometry. Our InfiniHuman allows for generating infinite 3D humans with precise user control.

Generating realistic and controllable 3D human avatars is a long-standing challenge. The difficulty increases when covering a broad range of attributes such as ethnicity, age, clothing styles, and detailed body shapes. Capturing and annotating large-scale human datasets for training generative models is prohibitively expensive and limited in both scale and diversity. The central question we address in this paper is: *Can we distill existing foundation models to generate theoretically unbounded richly annotated 3D human data?* We introduce **InfiniHuman**, a novel framework to distill these models synergistically, to generate richly annotated human data with minimal cost and theoretically unlimited scalability. Specifically, we propose **InfiniHuman-Data**, a fully automatic pipeline that leverages vision-language and image generation models to create a large-scale multi-modal dataset. Remarkably, users cannot distinguish our automatically generated identities from scan renderings. InfiniHumanData contains **111K identities** and covers unprecedented diversity in ethnicity, age, clothing styles, and more. Each identity is annotated with multi-granularity text descriptions, multi-view RGB images, detailed clothing images, and SMPL body shape parameters. Based on this, we learn **InfiniHumanGen**, a diffusion-based generative pipeline conditioned on text, body shape, and clothing assets. InfiniHumanGen enables fast, realistic, and precisely controllable avatar generation. Extensive experiments demonstrate that InfiniHuman significantly surpasses existing state-of-the-art methods in terms of visual quality, generation speed, and controllability. Importantly, our approach democratizes high-quality avatar generation with fine-grained control at infinite scale through a practical and affordable solution. To facilitate future research, we will publicly release our automatic data generation pipeline and the comprehensive dataset **InfiniHuman-Data**, and the generative models **InfiniHumanGen**. The code and data of InfiniHuman is publicly available at https://yuxuan-xue.com/infini-human.

CCS Concepts: • **Computing methodologies** → *Appearance and texture representations*; **Shape Inference**; *Machine learning approaches.*

Additional Key Words and Phrases: Text-guided 3D Generation, Digital Human, Text-to-Image Diffusion Model, Image-based Modeling

ACM Reference Format:
Yuxuan Xue, Xianghui Xie, Margaret Kostyrko, and Gerard Pons-Moll. 2025. InfiniHuman: Realistic 3D Human Creation with Precise Control. In *SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers '25), December 15–18, 2025, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3757377.3763815

1 Introduction

Creating realistic and controllable 3D human avatars is a fundamental problem of growing significance in virtual reality, digital fashion, gaming, and social telepresence. Applications increasingly

Authors' Contact Information: Yuxuan Xue, University of Tuebingen, Tuebingen, Germany, yuxuan.xue@uni-tuebingen.de; Xianghui Xie, University of Tuebingen, Tuebingen, Germany and Max Planck Institute for Informatics, Tuebingen, Germany, xianghui.xie@uni-tuebingen.de; Margaret Kostyrko, University of Tuebingen, Tuebingen, Germany, margaret.kostyrko@student.uni-tuebingen.de; Gerard Pons-Moll, University of Tuebingen, Tuebingen, Germany and Max Planck Institute for Informatics, Tuebingen, Germany, gerard.pons-moll@uni-tuebingen.de.

SA Conference Papers '25, Hong Kong, Hong Kong

ACM ISBN 979-8-4007-2137-3/25/12
https://doi.org/10.1145/3757377.3763815

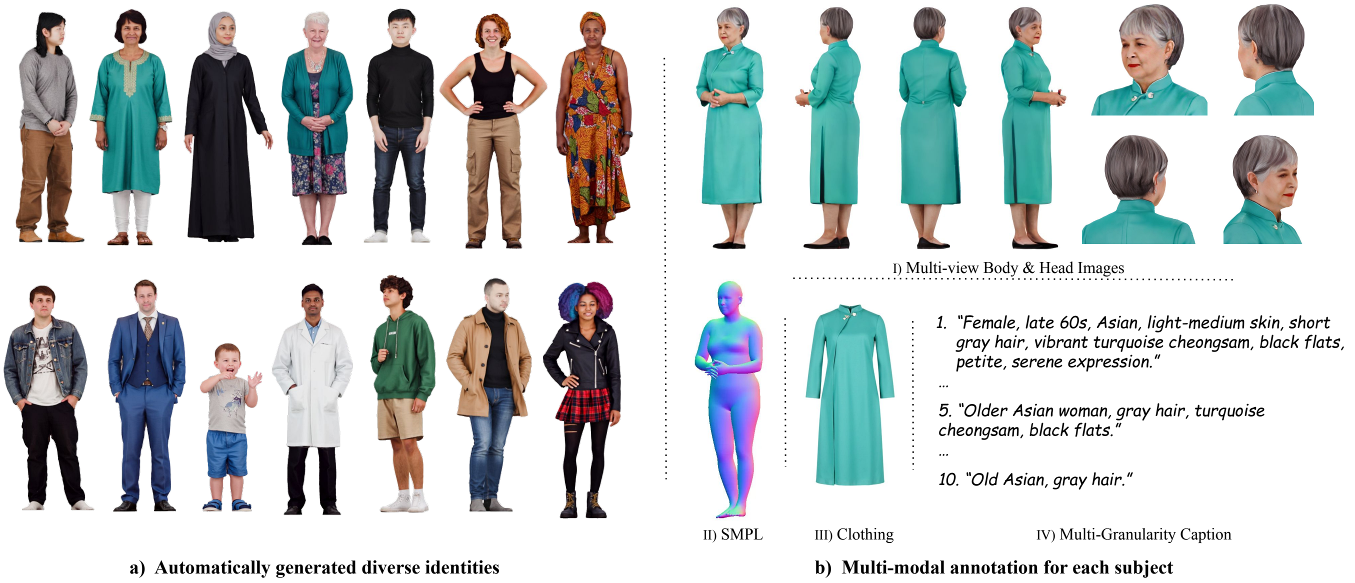
I) Multi-view Body & Head Images
1. "Female, late 60s, Asian, light-medium skin, short gray hair, vibrant turquoise cheongsam, black flats, petite, serene expression."
...
5. "Older Asian woman, gray hair, turquoise cheongsam, black flats."
...
10. "Old Asian, gray hair."
II) SMPL
III) Clothing
IV) Multi-Granularity Caption
a) Automatically generated diverse identities
b) Multi-modal annotation for each subject

Fig. 2. Examples from InfiniHumanData. a) Diverse human identities covering a wide range of ethnicities, age groups (including children), clothing styles, hair types, and skin tones, which are visually indistinguishable from real scans rendering (Sec. 4.2). **b)** Multi-modal annotations per each subject, including I) multi-view RGB images (full-body and head), II) SMPL parameters, III) clothing asset images, and IV) multi-granularity text descriptions.

demand photorealistic avatars that can be personalized to match textual descriptions, specific body shapes, and user-provided clothing. However, the limitations of existing generation techniques have become increasingly apparent. In particular, generating diverse and semantically rich 3D humans, varying in clothing, ethnicity, age, gender, and shape, remains difficult due to the high cost and limited diversity of manually captured datasets.

Recent training-free approaches such as Score Distillation Sampling (SDS) [Poole et al. 2023] have leveraged powerful text-to-image diffusion models to bypass dataset acquisition. However, these methods suffer from long optimization times, limited visual fidelity, and a lack of precise control over attributes like garment appearance or detailed body shape. These limitations motivate a critical research question: *Can we distill the capabilities of foundation models to generate richly annotated 3D human data at theoretically unlimited scale and with precise controllability?*

We propose **InfiniHuman**, a fully automated framework that addresses this question by systematically repurposing and integrating existing vision-language, image synthesis, pose estimation, and diffusion models. Our method produces realistic 3D human identities at unprecedented scale, each annotated with multi-view images, fine-grained textual descriptions, SMPL parameters, and explicit clothing representations. The resulting dataset, **InfiniHumanData**, contains over 111K identities and supports detailed control across age, ethnicity, clothing, and body morphology.

Built upon this dataset, we introduce **InfiniHumanGen**, a pair of generative models capable of synthesizing 3D avatars conditioned jointly on text, clothing image and body shape, giving the user powerful controls. It includes two complementary models: **Gen-Schnell**, which enables rapid 3D generation and produces a Gaussian splatting output, and **Gen-HRes**, which produces high-resolution, photorealistic textured meshes. Our models outperform prior works on visual quality, speed, and attribute controllability, achieving state-of-the-art results with significantly lower computational cost.

In summary, the main technical contributions of our work include:

- **InfiniHuman**, a framework to generate virtually unlimited richly annotated data of humans by distilling existing foundation models. The framework is fully automatic and generates identities indistinguishable from real scans.
- **InfiniHumanData**, the first large-scale multi-modal human dataset comprising 111K diverse identities with rich multi-modal annotations essential for precise avatar generation.
- **InfiniHumanGen**, a novel generative framework supporting two distinct models: Gen-Schnell for fast and interactive 3D human generation and Gen-HRes for high-resolution and visually detailed 3D human creation; both from various user-specified inputs such as text, clothing, or body shape.

By removing the need for costly scans, our method democratizes high-quality avatar creation, empowering applications in fashion, gaming, AR/VR, and beyond.

2 Related work

2.1 3D Human Generation.

The creation of 3D human avatars from user-defined conditions is a long-standing problem in vision and graphics, with most prior works falling into two categories: reconstruction from images [Liao et al. 2025; Saito et al. 2019; Xiu et al. 2023, 2022; Zheng et al. 2021], and generation from text [Cao et al. 2023; Han et al. 2023a; Hong et al. 2022; Kim et al. 2022; Kolotouros et al. 2023; Liao et al. 2023;

Table 1. Comparison of related datasets. Most existing human datasets are limited at scale and none of them provide detailed identity annotation like fine-grained text and clothing image.

Type	Dataset	IDs	Multi-Text	Cloth Assets
3D Scans	CustomHuman [Ho et al. 2023]	80	✗	✗
	Sizer [Tiwari et al. 2020]	97	✗	✗
	2K2K [Han et al. 2023b]	2050	✗	✗
	THuman2.1 [Yu et al. 2021]	2500	✗	✗
Multi-view Images	ActorsHQ [Isik et al. 2023]	8	✗	✗
	ZJU-MoCap [Peng et al. 2021]	10	✗	✗
	DNA-Rendering [Cheng et al. 2023]	500	✗	✗
	HUMBI [Yu et al. 2020]	772	✗	✗
	HuMMan [Cai et al. 2022]	1000	✗	✗
	MVHumanNet [Xiong et al. 2024]	4500	✗	✗
	IDOL [Zhuang et al. 2025]	100K	✗	✗
Ours	**InfiniHumanData**	**111K**	✓	✓

Liu et al. 2024; Wang et al. 2024; Yuan et al. 2024; Zhang et al. 2023]. Recent methods have also explored learning avatars from large-scale 2D image collections [Dong et al. 2023; Hong et al. 2023; Xiu et al. 2024].

A key limitation in existing works is controllability: prior approaches support conditioning on either text or body shape, but none allow direct, explicit control over detailed clothing items in addition to text and shape. This restricts their application in domains requiring personalized appearance, such as digital fashion or virtual fitting rooms. We fill this gap by introducing a scalable and fully automatic data generation pipeline that enables the training of generative models conditioned on text, SMPL body shape, and specific clothing images. Our models achieve high-quality 3D human synthesis consistent with all these modalities, offering unprecedented fine-grained control and realism.

2.2 Large-Scale 3D Datasets.

The availability of high-quality, large-scale 3D datasets is a key driver of progress in generative 3D modeling. While object-centric datasets like Objaverse [Deitke et al. 2023] and ShapeNet [Chang et al. 2015] have enabled remarkable advances for general object synthesis and reconstruction, 3D human datasets pose unique challenges. Commercial human scan repositories such as RenderPeople, Twindom, and Axyz provide highly realistic scans, but are expensive (often ~100 USD per identity). Publicly available 3D human datasets (Tab. 1) are often constrained by participant recruitment, scanning logistics, and privacy considerations, resulting in limited scale, demographic diversity, and coverage of clothing, age, and body morphology.

Some alternatives use multi-view image capture to reduce costs, but these datasets are typically restricted to fixed camera viewpoints and controlled lighting, limiting their generalizability and value for generative tasks. Recent innovations, such as the IDOL dataset [Zhuang et al. 2025], leverage video diffusion models to synthesize 360-degree images from a single 2D input. However, video diffusion often introduces view inconsistencies and lacks true 3D geometry (see Supp. Mat.), due to neighbor-only attention mechanisms and the absence of explicit 3D supervision.

Critically, existing datasets rarely provide fine-grained annotations of identity level that are essential for training generative models capable of precise control over appearance attributes. Our **InfiniHumanData** addresses all these limitations by using multimodal foundation models to generate a large-scale, richly annotated dataset, covering unprecedented diversity across age, ethnicity, body shape, and clothing style, and providing annotations that support high-fidelity, controllable 3D human generation. To accelerate research and enable further expansion, we publicly release our fully automatic data generation pipeline and dataset, empowering the community to create virtually unlimited, realistic human identities.

3 Method

Our objective is to generate highly realistic 3D avatars that allow precise and flexible control based on multiple user-specified conditions. These conditions include (i) natural language descriptions to define the subject's appearance, (ii) SMPL parameters to govern body shape and pose, and (iii) reference images to specify clothing style. To enable such fine-grained generation, we must model the joint conditional distribution $P(\boldsymbol{y}|\boldsymbol{c}^{\text{text}}, \boldsymbol{c}^{\text{SMPL}}, \boldsymbol{c}^{\text{cloth}})$, where $\mathbf{y}$ represents the generated avatars, and $\mathbf{c}$ terms represent the conditioning signals specified by users.

This task requires a large, diverse, and richly annotated dataset of 3D human avatars, which is costly and impractical to collect and annotate manually. Instead, we present **InfiniHuman**, a fully automated framework that synthesizes such a dataset by distilling existing foundation models across multiple domains. We first detail the construction of our dataset, **InfiniHumanData**, in Sec.3.1, and then describe our controllable generative models, **InfiniHumanGen**, in Sec.3.2.

3.1 InfiniHumanData - Generation by Reconstruction

To enable highly controllable 3D avatar generation, we first construct a large-scale, richly annotated dataset, **InfiniHumanData**. Our data generator produces multi-modal outputs for each identity, including structured text descriptions, clothing style images, SMPL body shape and keypoints, and orthographic multi-view images with controlled lighting suitable for 3D lifting (see Fig. 3 for visualization and detailed breakdown). In the following, we describe the major components of our data generation pipeline.

A) Multi-Granularity Text Description. To encode diverse semantic concepts, we design a captioning system that generates both detailed and progressively abstracted descriptions. We first caption existing human scan datasets [Han et al. 2023b; Ho et al. 2023; Yu et al. 2021] using the protocol from Trellis [Xiang et al. 2024]. Next, we randomly sample ten captions and provide them as in-context examples to GPT-4o, prompting it to generate new variations. These generated captions maintain similar lengths and formats, while diversifying attributes such as ethnicity, age group, and clothing style. Each caption is then summarized into ten levels of granularity, ranging from 40 words to 5 words. This hierarchical annotation enriches training by exposing models to both coarse (*e.g.*, *old*) and fine-grained (*e.g.*, *late sixties to early seventies*) semantic cues.

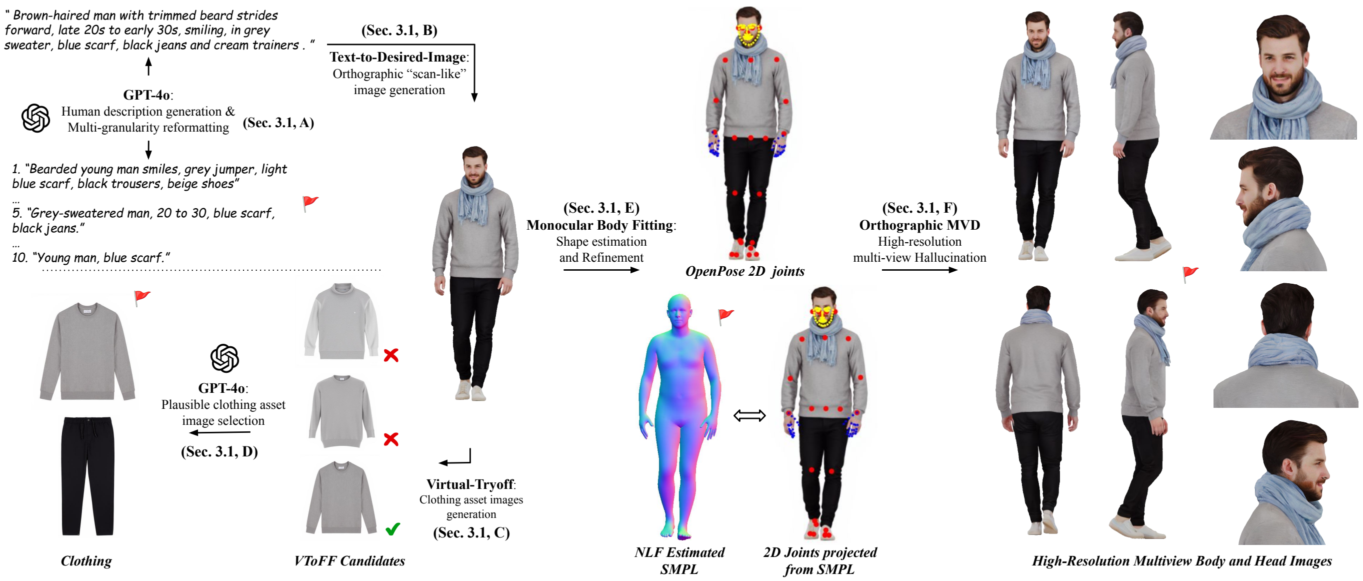
" Brown-haired man with trimmed beard strides forward, late 20s to early 30s, smiling, in grey sweater, blue scarf, black jeans and cream trainers . "
GPT-4o: Human description generation & Multi-granularity reformatting
(Sec. 3.1, A)
1. "Bearded young man smiles, grey jumper, light blue scarf, black trousers, beige shoes"
...
5. "Grey-sweatered man, 20 to 30, blue scarf, black jeans."
...
10. "Young man, blue scarf."
(Sec. 3.1, B)
Text-to-Desired-Image: Orthographic "scan-like" image generation
(Sec. 3.1, E)
Monocular Body Fitting: Shape estimation and Refinement
OpenPose 2D joints
(Sec. 3.1, F)
Orthographic MVD High-resolution multi-view Hallucination
GPT-4o: Plausible clothing asset image selection
(Sec. 3.1, D)
Virtual-Tryoff: Clothing asset images generation
(Sec. 3.1, C)
Clothing
VToFF Candidates
NLF Estimated SMPL
2D Joints projected from SMPL
High-Resolution Multiview Body and Head Images

Fig. 3. **Overview of data generation framework** in **InfiniHumanData**. The process is fully automated by leveraging foundation models. Desired outputs are marked with flags: **A)** Structured text descriptions, **C)** Clothing style images, **E)** Body shape in SMPL format plus face and hand keypoints, **F)** Orthographic multi-view images with controlled lighting conditions suitable for 3D lifting.

B) Orthographic Text-to-Image. Most text-to-image models (e.g., FLUX) produce images with dramatic perspective and complex lighting, which are suboptimal for 3D reconstruction tasks. To address this, we fine-tune FLUX with a LoRA adapter [Hu et al. 2022] on orthographic renderings of a few thousand scans under uniform lighting, enabling generation of "scan-like" images (see Fig. 3). This stylization step ensures compatibility with downstream 3D lifting processes. In particular, orthographic views are essential for our multi-view diffusion, which relies on simplified epipolar attention [Li et al. 2024a]. Importantly, this approach preserves the inherent diversity of FLUX while aligning the image domain for reconstruction. A challenging discriminative user study (Sec. 4.2) further demonstrates that our generated identities achieve visual realism on par with scan renderings.

C) Virtual-TryOff for Clothing Control. Because a single image can convey garment appearance more precisely than any text description, we provide direct clothing control by reversing the try-on process. Given a full-body image, we fine-tune OminiControl [Tan et al. 2024] to extract a clean garment image via text-based image-to-image translation. This task, termed *Instruct-Virtual-TryOff*, is trained using garment-actor pairs from existing Virtual-TryOn datasets [Choi et al. 2021; Morelli et al. 2022] and prompts like "<Please extract *{Garment}* for this person>".

Each training instance consists of a garment image I^{cloth}, a corresponding try-on image I^{vton}, and a textual prompt e^{text}. The model parameters $\boldsymbol{\theta}$ are optimized via the flow-matching objective:

$$\mathcal{L}_{\text{VToFF}}(\boldsymbol{\theta}) = \mathbb{E}_{t,\boldsymbol{\varepsilon}} \left\| v_{\boldsymbol{\theta}}\left(\mathbf{x}_t,\, I^{\text{vton}},\, e^{\text{text}},\, t\right) - \left(\boldsymbol{\varepsilon} - I^{\text{cloth}}\right) \right\|^2, \quad (1)$$

$$\text{where} \quad \mathbf{x}_t = (1-t)\, I^{\text{cloth}} + t\boldsymbol{\varepsilon},\ \ \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I}). \quad (2)$$

Here, $v_{\boldsymbol{\theta}}$ denotes the network, $\mathbf{x}_t$ is a noisy version of the garment image, and e^{text} provides the instruction (see Fig. 3). The network learns to synthesize clean garment images conditioned on full-body images and textual instructions. This enables users to specify clothing via image, without requiring paired image-scan training data.

D) Negative Samples Rejection. To remove occasionally wrongly generated images, we use the sampling rejection strategy: first generate four garment images per subject and then employ GPT-4o to select the best match based on considerations like color, texture, length, and detailed features (e.g. zippers, pockets). The detailed prompt for sampling rejection can be found in Supp. Mat.

E) Monocular Body Fitting for Shape and Pose Control. We use NLF [Sárándi and Pons-Moll 2024] to regress SMPL parameters from orthographic views by setting FoV to 0.1, followed by refinement via OpenPose 2D joint alignment [Cao et al. 2019]. This two-step process ensures that SMPL parameters align accurately with both overall pose and pixel-level features (particularly at face), which is crucial for consistent multi-view generation conditioned on SMPL. More specifically, we optimize the SMPL body pose parameters w.r.t. the reprojection error between the orthographically projected SMPL joints and 2D joints estimated by OpenPose:

$$\mathcal{L}_{\text{reproj}}(\boldsymbol{\theta}) = \sum_{k=1}^{K} w_{ik} \left\| \pi_{\text{ortho}}\!\left(J_k(\text{SMPL}(\boldsymbol{\theta_i}, \boldsymbol{\beta}))\right) - J_k^{\text{OpenPose}} \right\|_2^2 \quad (3)$$

We carefully tweak the per-joint weights and the regularization to achieve the best pixel-level matching between 3D SMPL and 2D images. Please refer to Fig. 14 and supplementary material for qualitative visual examples and ablation studies.

F) Orthographic MV-Diffusion. To produce high-resolution, consistent multi-views, we train a diffusion model on orthographic

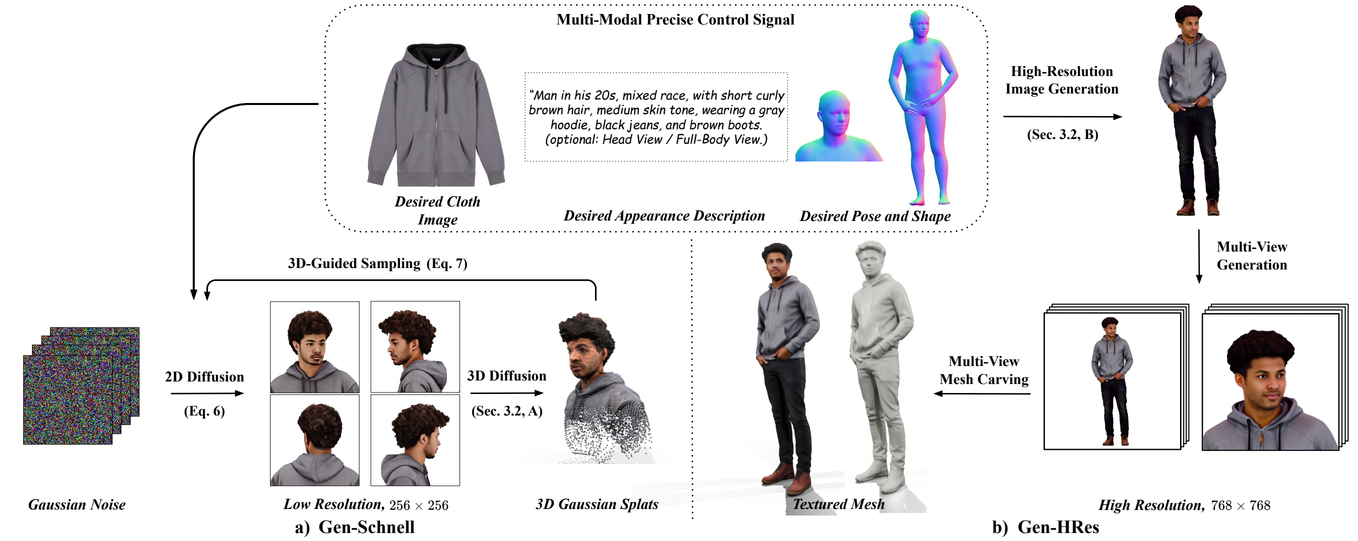
Multi-Modal Precise Control Signal
"Man in his 20s, mixed race, with short curly brown hair, medium skin tone, wearing a gray hoodie, black jeans, and brown boots. (optional: Head View / Full-Body View.)
Desired Cloth Image
Desired Appearance Description
Desired Pose and Shape
High-Resolution Image Generation
(Sec. 3.2, B)
Multi-View Generation
3D-Guided Sampling (Eq. 7)
2D Diffusion
(Eq. 6)
3D Diffusion
(Sec. 3.2, A)
Multi-View Mesh Carving
Gaussian Noise
Low Resolution, 256 × 256
3D Gaussian Splats
Textured Mesh
High Resolution, 768 × 768
a) Gen-Schnell
b) Gen-HRes

Fig. 4. Overview of a) Gen-Schnell and b) Gen-HRes in InfiniHumanGen. Taking text description, explicit SMPL shape, and a cloth image as input, Gen-Schnell generates 3D-GS end-to-end, while Gen-HRes generates high-resolution textured mesh, both matched to input conditions.

projections with uniform lighting. Orthographic views have horizontal epipoles, enabling efficient row-wise attention across views.

Given an orthographic RGB image $I^{\text{in}} \in \mathbb{R}^{H\times W\times C}$, our multi-view diffusion (MVD) model generates N views of high-resolution full-body images $I^{\text{body}} \in \mathbb{R}^{N\times H\times W\times C}$ and head images $I^{\text{head}} \in \mathbb{R}^{N\times H\times W\times C}$ from the front, left, right, and back directions. We provide geometric guidance by rendering SMPL normal maps I^{SMPL} and encoding them, together with the reference image, into the latent space using a pretrained VAE from PSHuman [Li et al. 2024b]. For multi-view consistency, we apply orthographic multi-view attention separately within the body and head views, where each row of the each view attends to the same row of other views due to the orthographic constraint across views. Please refer to Fig. 15 for visual examples. For body-head consistency, we use dense pixel-level cross-attention between corresponding body and head views, where each pixel of body image attends to pixels of the head image under the same view. The UNet denoiser $\epsilon(\boldsymbol{\theta})$ is fine-tuned using the following objective:

$$\mathcal{L}_{\text{MVD}}(\boldsymbol{\theta}) = \mathbb{E}_{t,\boldsymbol{\varepsilon}} \sum_{p\in\{\text{body},\text{head}\}} \left\| \epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t^p,\, I^{\text{in}},\, I^{\text{SMPL}},\, t) - \boldsymbol{\varepsilon} \right\|^2, \quad (4)$$

$$\text{where} \quad \mathbf{x}_t^p = \sqrt{\bar{\alpha}_t} I^p + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0,\mathbf{I}). \quad (5)$$

Here, $\bar{\alpha}_t$ determines the noise level at each diffusion step t. At inference time, our multi-view diffusion model takes an orthographic input image and SMPL normal maps as input, generating high-resolution multi-view body and head images (see Fig. 3, right).

3.2 InfiniHumanGen - Generation with Precise Control

3.2.1 Joint Conditional Distribution. Leveraging InfiniHumanData, which contains 111K diverse identities each annotated with multi-granularity text captions $\boldsymbol{c}^{\text{text}}$, SMPL parameters $\boldsymbol{c}^{\text{SMPL}}$, corresponding cloth images $\boldsymbol{c}^{\text{cloth}}$, and orthographic multi-view images $\boldsymbol{y}^{mv}$, we learn a joint conditional distribution $P(\boldsymbol{y}|\boldsymbol{c}^{\text{text}}, \boldsymbol{c}^{\text{SMPL}}, \boldsymbol{c}^{\text{cloth}})$ to enable precise avatar generation. We train two complementary models to support both fast and high-fidelity generation:

A) Gen-Schnell: Fast End-to-End Generation. Gen-Schnell is a low-latency model that directly generates 3D avatars as Gaussian splats [Kerbl et al. 2023]. Inspired by Human-3Diffusion [Xue et al. 2024], we combine 2D multi-view generation (from MVDream [Shi et al. 2024]) with a splatting decoder that enforces consistency across views. To inject condition signals, we encode SMPL normal maps $\boldsymbol{c}^{\text{SMPL}}$ and clothing images $\boldsymbol{c}^{\text{cloth}}$ using the MVDream VAE, and concatenate them channel-wise with the initial noise $\mathbf{x}_t$. The 2D diffusion model $\epsilon(\boldsymbol{\theta})$ predicts noise values, which are used to reconstruct clean multi-view images $\tilde{\mathbf{x}}_0$:

$$\tilde{\mathbf{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_{\boldsymbol{\theta}} \left(\mathbf{x}_t, \mathbf{c}^{\text{text}}, \mathbf{c}^{\text{SMPL}}, \mathbf{c}^{\text{cloth}}, t \right) \right). \quad (6)$$

While the resulting multi-view images $\tilde{\mathbf{x}}_0$ provide strong shape priors, they may exhibit inconsistencies across views. To address this, our 3D-GS generator $g(\boldsymbol{\phi})$ takes the predicted multi-view images $\tilde{\mathbf{x}}_0$ and initial noise $\mathbf{x}_t$ to generate consistent 3D Gaussian splats $\hat{\mathcal{G}}_0$, which render consistent multi-view images $\hat{\mathbf{x}}_0$. During each sampling step from t to $t-1$, we replace 2D predictions with 3D-GS rendered images to ensure consistency:

$$\begin{aligned} \mu_{t-1}(\mathbf{x}_t, \hat{\mathbf{x}}_0) &= \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t}\,\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\,\beta_t}{1-\bar{\alpha}_t}\,\hat{\mathbf{x}}_0, \\ \mathbf{x}_{t-1} &\sim \mathcal{N}\!\left(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \hat{\mathbf{x}}_0)\,, \tilde{\beta}_{t-1}\mathbf{I}\right). \end{aligned} \quad (7)$$

At the final timestep $t = 0$, $\hat{\mathcal{G}}_0$ is output as the final 3D-GS, see Fig. 4. Gen-Schnell is highly efficient and produces 3D-GS in about 12 seconds. However, due to the low-resolution constraint of MVDream (256×256), detailed features (e.g., facial textures, textual elements) appear blurry, motivating our high-resolution generator, Gen-HRes.

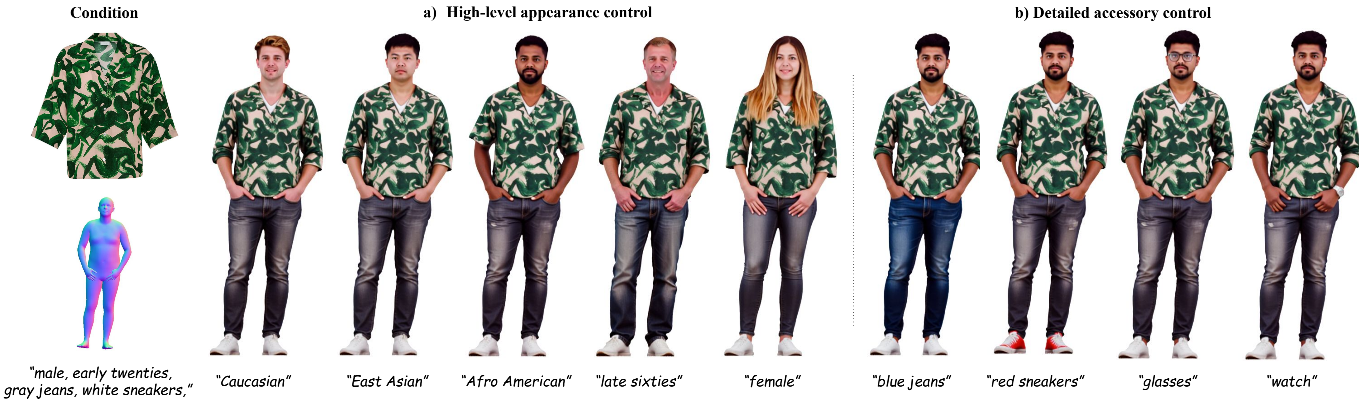
Condition
a) High-level appearance control
b) Detailed accessory control
"male, early twenties, gray jeans, white sneakers,"
"Caucasian"
"East Asian"
"Afro American"
"late sixties"
"female"
"blue jeans"
"red sneakers"
"glasses"
"watch"

Fig. 5. Fine-grained text controllability in Gen-HRes over (a) overall subject identity, such as ethnicity, age, gender, etc. By fixing the initial Gaussian noise, Gen-HRes can generate (b) same identity with different detailed accessory appearance, such as watch, glasses, and colors of wearing assets.

B) Gen-HRes: High-Resolution Generation. For photorealistic avatar generation from multiple conditions, Gen-HRes frames it as a multi-image-to-image translation task, where we fine-tune OminiControl2 [Tan et al. 2025] on InfiniHumanData. Using full-body images $\mathbf{y}^{2D}$ as target, we optimize the flow matching objective:

$$\mathcal{L}_{\text{HRes}}(\boldsymbol{\theta}) = \mathbb{E}_{t,\boldsymbol{\varepsilon}} \left\| v_{\boldsymbol{\theta}}(\mathbf{x}_t,\, \mathbf{c}^{\text{text}},\, \mathbf{c}^{\text{cloth}},\, \mathbf{c}^{\text{SMPL}},\, t) - (\boldsymbol{\varepsilon} - \mathbf{y}^{2D}) \right\|^2, \quad (8)$$

$$\text{where} \quad \mathbf{x}_t = (1-t)\,\mathbf{y}^{2D} + t\boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I}). \quad (9)$$

Our data and model design ensure that the generated multi-view images are well aligned with the conditioning SMPL mesh. This alignment allows us to compute surface normals with Sapiens2B [Khirodkar et al. 2024] and apply SMPL-driven volumetric carving via PSHuman [Li et al. 2024b] for high-fidelity 3D mesh reconstruction.

Compared to Gen-Schnell, Gen-HRes not only achieves higher resolution and visual fidelity, but also supports detailed text prompting. By fixing initial Gaussian noise, Gen-HRes can precisely control fine-grained attributes, such as glasses or garment colors, through detailed text descriptions, as shown in Fig. 5. Gen-HRes enables high-fidelity avatar generation in approximately 4 minutes.

4 Experiments

4.1 Implementation Details

The orthographic multiview diffusion model used in InfiniHuman and Gen-HRes is built upon the pre-trained text-to-image model SD2.1-unclip. We concatenate the input image latents with noise latents along the channel dimension. The noise latents are replicated for each view, and the text embedding is repurposed to generate distinct head and body views, similar to PSHuman [Li et al. 2024b]. Our model generates four orthogonal body views and four head views from single orthographic input body image. For the orthographic multiview diffusion models, we train on 8 H100 with effective batch size of 128 for 2 days on orthographic uniform lighting rendering of 6000 high-quality human scans from Twindom, CustomHuman, and THuman2.1 [twi 2023; Ho et al. 2023; Yu et al. 2021]. We use front-view renders with text labels to fine-tune Flux-Dev [Black Forest Labs 2024] LoRA for the orthographic text-to-image task.

Table 2. Quantitative comparison results. We report user study results for appearance quality and text alignment, where most participants prefer our method. We also achieve SOTA in T2I metrics such as CLIP and FID.

Method	Quality↑ (User Study)	Alignment↑ (User Study)	FID↓	CLIP Score↑	Runtime↓
MVDream	20.83%	20.36%	141.33	30.37	**2.8s**
SPAD	2.02%	1.55%	150.43	28.58	13.9s
Gen-Schnell	**77.14%**	**78.10%**	**100.39**	**30.82**	12.9s
TADA	1.27%	1.27%	129.68	28.84	213m
DreamAvatar	1.27%	1.90%	151.57	28.42	384m
HumanGaussian	2.22%	3.48%	140.24	**30.56**	40 m
HumanNorm	2.54%	3.48%	101.84	28.30	117 m
AvatarVerse	0.32%	0.32%	156.52	28.69	44 m
Gen-HRes	**92.39%**	**89.56%**	**82.28**	30.43	**4 m**

For constructing the InfiniHumanData, we use GPT-4o to enhance correctness of automatic cloth labeling, where each subject takes around $0.03. Based on InfiniHumanData, we train Gen-Schnell on 8 A100 GPUs with effective batch size 256 over approximately 2 days, and Gen-HRes on 2 H100 GPUs with effective batch size of 32 for 2 days. Please refer to Supp. Mat. for implementation details on Gen-Schnell as well as Gen-HRes, and for prompting details on constructing InfiniHumanData.

4.2 Evaluation Benchmark

We compare Gen-Schnell with feed-forward text-based multi-view generation approaches such as MVDream [Shi et al. 2024] and SPAD [Kant et al. 2024], which can generate multi-view images from text prompt within a minute. We compare Gen-HRes with SDS-based text-to-avatar approaches, e.g. DreamAvatar [Cao et al. 2024], AvatarVerse [Zhang et al. 2024], HumanGaussian [Liu et al. 2024], HumanNorm [Huang et al. 2024], and TADA [Liao et al. 2024]. These optimization-based methods achieve higher quality than feed-forward approaches but typically require several hours for generation. Therefore, we also compare with Chupa [Kim et al. 2023], a mesh-based avatar generator directly learned from 3D scans. Furthermore, we evaluate the realism of generated identities in InfiniHumanData.

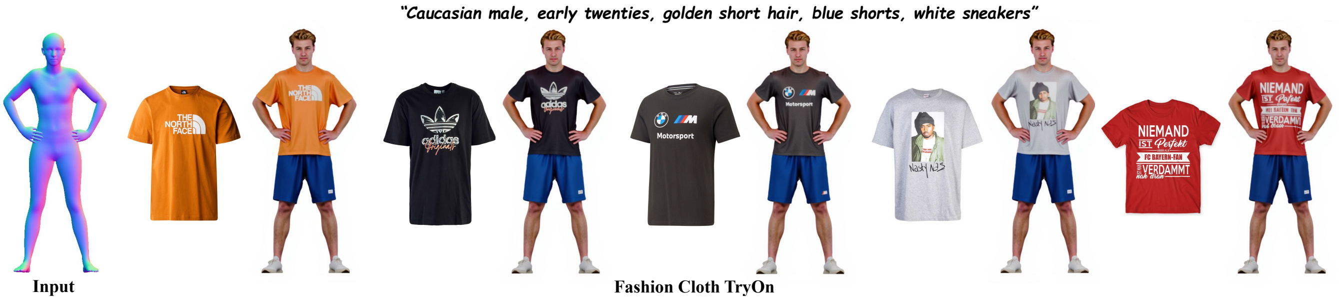
"Caucasian male, early twenties, golden short hair, blue shorts, white sneakers"
Input
Fashion Cloth TryOn

Fig. 6. **Generate avatars with given garment from fashion industry**. The identity is preserved while TryOn garment is changing.

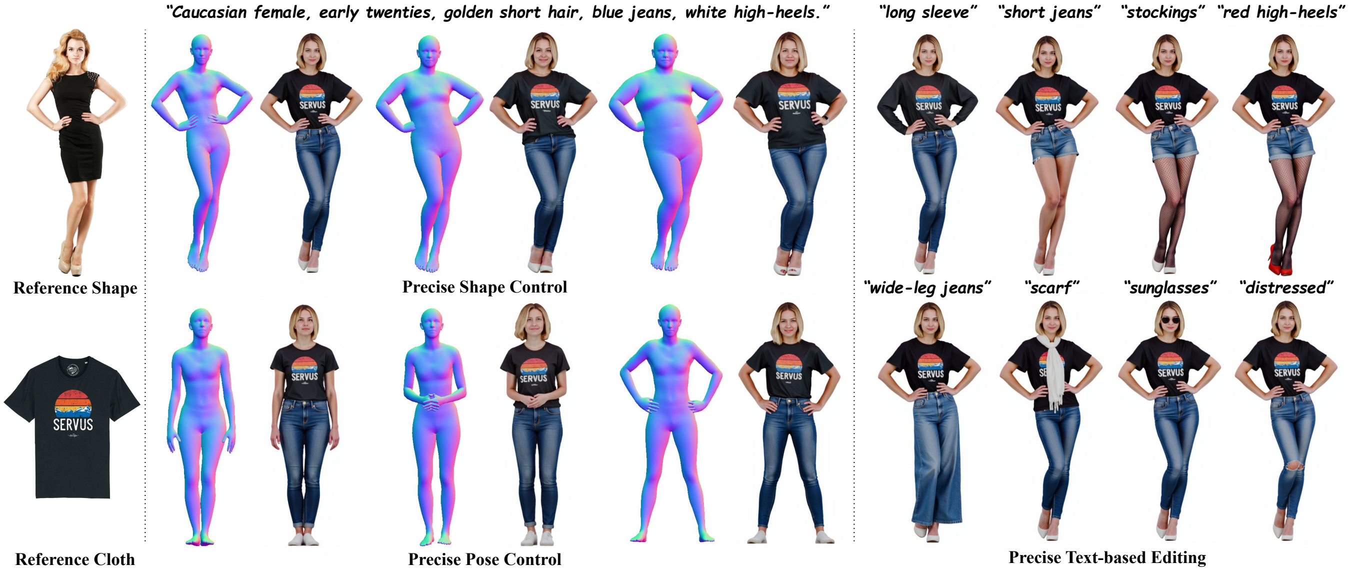

Fig. 7. **Generate avatars with precise pose shape control and text-based editing**. The identity is preserved during shape and text-based editing.

4.2.1 Qualitative Comparison. As depicted in Fig. 11, Gen-HRes has various advantages over baselines: (1) multi-view consistency, because Gen-HRes generates textured mesh as output, while SDS-method optimizes per view given text prompt, which can lead to the Janus problem. (2) enhanced realism, Gen-HRes does not suffer from unnatural saturation, which is a typical problem in SDS-based generation. (3) text-following ability, Gen-HRes leverages foundational text-based generation capability from FLUX, which shows stronger text-following ability than SDS-based methods, especially in details such as color of garments. Gen-Schnell also shows better text-following ability (e.g. head view, color) than previous works. Please refer to Fig.3 and Fig.4 in Supp. Mat. for more comparison.

4.2.2 Quantitative Comparison. We conducted a user study to quantitatively compare with SOTA methods in text-based generation. We asked 42 participants to evaluate videos rendered from generated 3D avatars generated by different methods and to vote for the best methods based on overall appearance quality and alignment with text description. We also report quantitative numbers in FID between generated results and rendered images from human scans. Additionally, we use the CLIP Score to quantify the semantic alignment between the text description and the renderings. Tab. 2 presents average scores across 32 prompts. The results of user studies, FID, and CLIP demonstrate that our Gen-Schnell and Gen-HRes outperforms SOTA feed-forward text-to-3D and SDS-based text-to-3D avatar methods, respectively. Our method achieves the highest overall result quality and the most accurate alignment with the prompt's semantics. More importantly, our Gen-HRes can generate 3D avatars with at least 8 times less computational time than high-resolution baselines, demonstrating that our InfiniHumanGen is the most efficient high-resolution avatar generative model.

4.2.3 InfiniHumanData Evaluation. To assess the realism of InfiniHumanData, we conducted a user study comparing it against renderings from real human scans. The goal was to evaluate whether users could distinguish our generated avatars from those based on actual 3D scan data. Participants were presented with image pairs: one image rendered from a real scan, and the other randomly sampled from InfiniHumanData. For each pair, users were asked to select the more realistic image, or choose "both" if they could not tell the difference. Across all trials, real scan renderings received ***746*** votes, while InfiniHumanData images received ***765*** votes. The small difference in votes indicates that our InfiniHumanData achieves a high degree of visual realism, closely matching the appearance of scans.

4.3 Fine-grained Controllability

4.3.1 Precise Clothing Control. As shown in Fig. 6, Gen-HRes can generate avatars with high fidelity to the input clothing images. By fixing the initial Gaussian noise, we can generate the same subject wearing different garments, preserving identity across try-on results. This demonstrates strong, identity-preserving clothing controllability.

4.3.2 Precise Pose and Shape Control. As illustrated in Fig. 7, Gen-HRes accurately follows the body shape and pose specified by the SMPL condition, faithfully transferring to the generated avatar.

4.3.3 Precise Text-based Generation and Editing. Gen-HRes enables control over high-level human attributes, such as ethnicity, age, and gender, all through text input (see Fig. 5). More importantly, it supports fine-grained text-based editing while maintaining identity consistency. As shown in Fig. 7, we generate the same subject with different accessories, such as stockings, scarves, or sunglasses.

4.4 Application

4.4.1 TryOn from Photographs. Our *Instruct-Virtual-TryOff* module demonstrates strong generalization: it can extract clean garment images directly from real-world photographs. As shown in Fig. 10, we extract clothing assets from photo captures and generate corresponding avatars with user-specified text controls.

4.4.2 Re-animation. Leveraging the underlying SMPL parametric body, our generated 3D avatars can be reanimated using SMPL motion data by barycentric interpolation of SMPL skinning weights onto the generated mesh surface. See Fig. 13 for re-animation examples.

4.4.3 Figurine Fabrication. Gen-HRes produces high-quality, watertight 3D meshes, enabling direct 3D printing of physical figurines. The printed figurines are physically robust and can stand independently, as shown in Fig. 13, demonstrating the real-world physical compatibility [Guo et al. 2024] of generated avatars.

4.5 Ablation Study

We qualitatively ablate different design choices, showcasing the importance of orthographic MVD (Sec. 3.1F, Fig. 15), generating scan-like images (Sec. 3.1B, Figs. 15, 16), additional SMPL fitting (Sec. 3.1E, Fig. 14), and tolerance to inaccurate SMPL for children generation (Fig. 14). Please refer to individual figures for examples.

5 Limitations and Future Works

Although our Gen-HRes can perform high-fidelity generation, it is still slower than the end-to-end 3D generation pipeline, Gen-Schnell. However, Gen-Schnell cannot generate faithful details such as face because of the low-resolution (256×256) of pretrained MVDream. Due to limited training resources, we cannot directly train a higher-resolution Gen-Schnell. However, we publicly release all high-resolution (768×768) InfiniHumanData with multi-modal labels. Future works can consider training a high-resolution text-based 3D-GS model, which achieves fast and high-quality end-to-end multi-modal avatar generation.

As shown in Fig. 9, our pipeline can generate famous people by names. However, GPT-4o refuses to identify unmatched samples because of privacy issues. Future works may adopt a different vision-language model to include famous names in InfiniHumanData. Moreover, our Gen-HRes adopts multi-view mesh carving to obtain textured mesh from orthographic views, which can cause texture artifacts in self-occluded parts of the avatar. Future works may consider a data-driven approach for the mesh reconstruction from multi-view images.

6 Conclusion

In this work, we present InfiniHuman, a novel framework for realistic and highly controllable 3D avatar generation. To overcome the fundamental challenge of scarce and expensive annotated human data, we developed a fully automated data generation framework that repurposes multiple pretrained foundation models. This enables the creation of InfiniHumanData, a large-scale, richly annotated dataset with 111K diverse identities and comprehensive control signals. Building on this foundation, our InfiniHumanGen framework delivers rapid, high-fidelity avatar synthesis with unprecedented fine-grained control, enabling users to specify appearance, shape, pose, and clothing through intuitive multi-modal inputs. Extensive experiments demonstrate that InfiniHuman not only outperforms prior methods in visual quality and speed, but also sets a new standard for precise, attribute-level controllability in 3D human generation. Importantly, our approach democratizes high-quality avatar creation via an accessible and scalable solution. To support further research and broad adoption, we will publicly release InfiniHumanData, InfiniHumanGen, and our automatic data generation pipeline, empowering the community to create unlimited, realistic, and diverse 3D humans with full user control.

Acknowledgments

This work is made possible by funding from the Carl Zeiss Foundation. This work is also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (EmmyNoether Programme, project: Real Virtual Humans) and the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Y.Xue. G. Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

Y.Xue is the first author and corresponding author. Y.Xue initialized the core idea, organized the project, developed the current method, conducted experiments, and wrote the draft. X.Xie contributed to the method development and draft writing. M.Kostyrko contributed to the teaser figure rendering and re-animation of Gen-HRes avatars. All the team members made necessary contributions to the method development and paper writing.

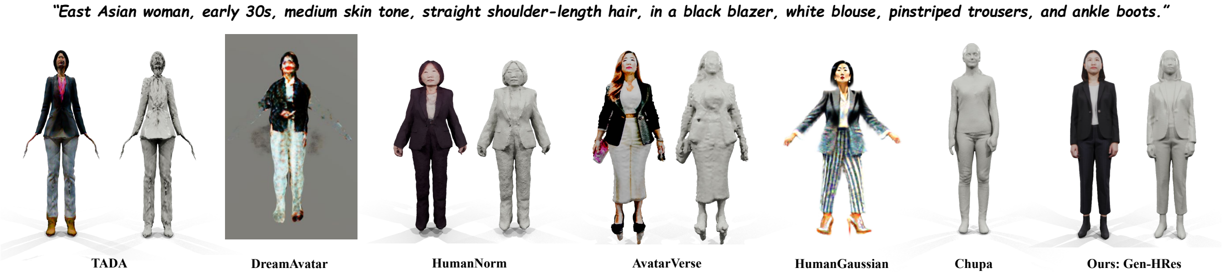
"East Asian woman, early 30s, medium skin tone, straight shoulder-length hair, in a black blazer, white blouse, pinstriped trousers, and ankle boots."
TADA
DreamAvatar
HumanNorm
AvatarVerse
HumanGaussian
Chupa
Ours: Gen-HRes

Fig. 8. Qualitative comparison to SOTA text-to-3D avatar generators. We compare with SDS-based avatar generation methods and a mesh-based avatar generation method Chupa [Kim et al. 2023]. Our generator can follow the text very well and also achieve outstanding generation quality.

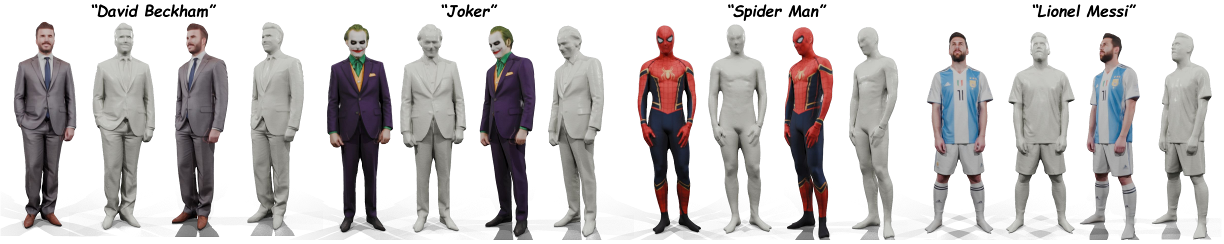

Fig. 9. Generated famous people and characters by names in Gen-HRes. Please zoom in for details.

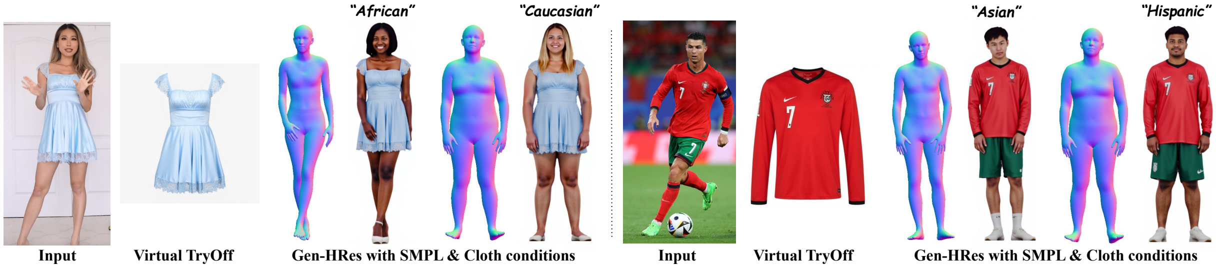

Fig. 10. Generated avatars with garments in real person photos. We can extract clean garments from photos and use for TryOn. Images from Shutterstock.

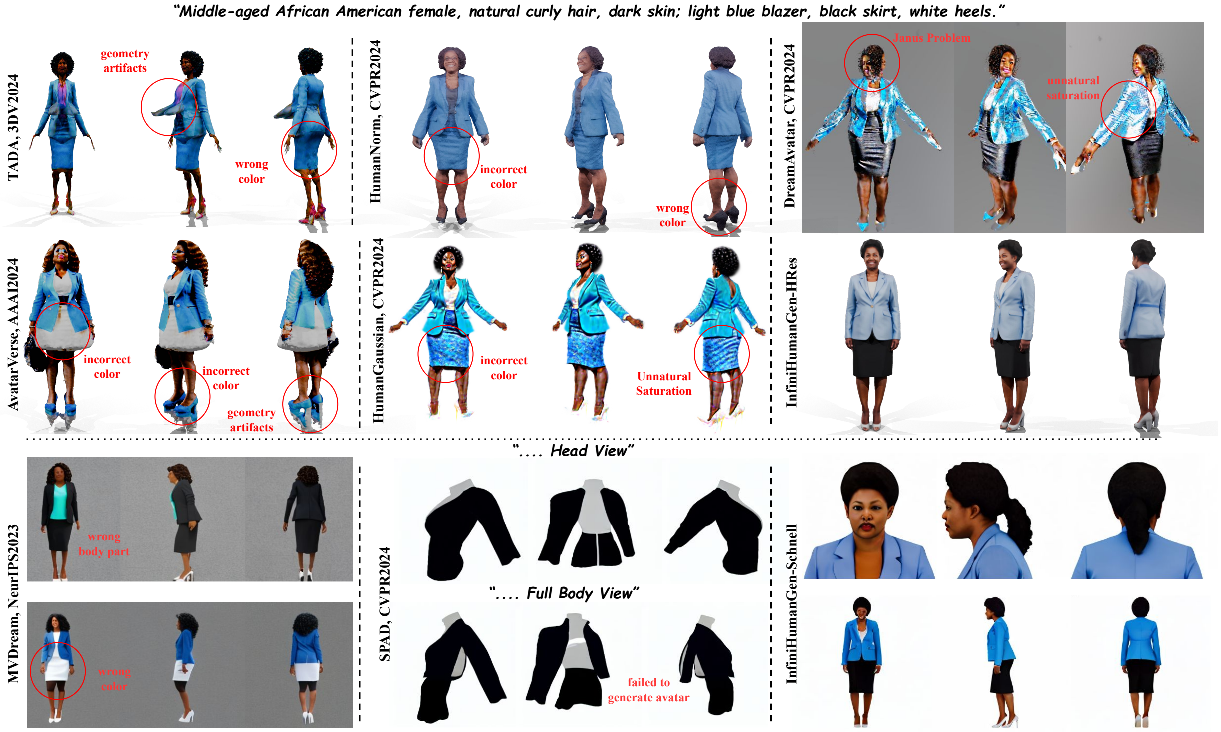

Fig. 11. Qualitative comparison to SOTA text-to-3D avatar approaches. Gen-HRes avoids Janus/artifacts, aligns to prompts, and is 8× faster (Tab. 2).

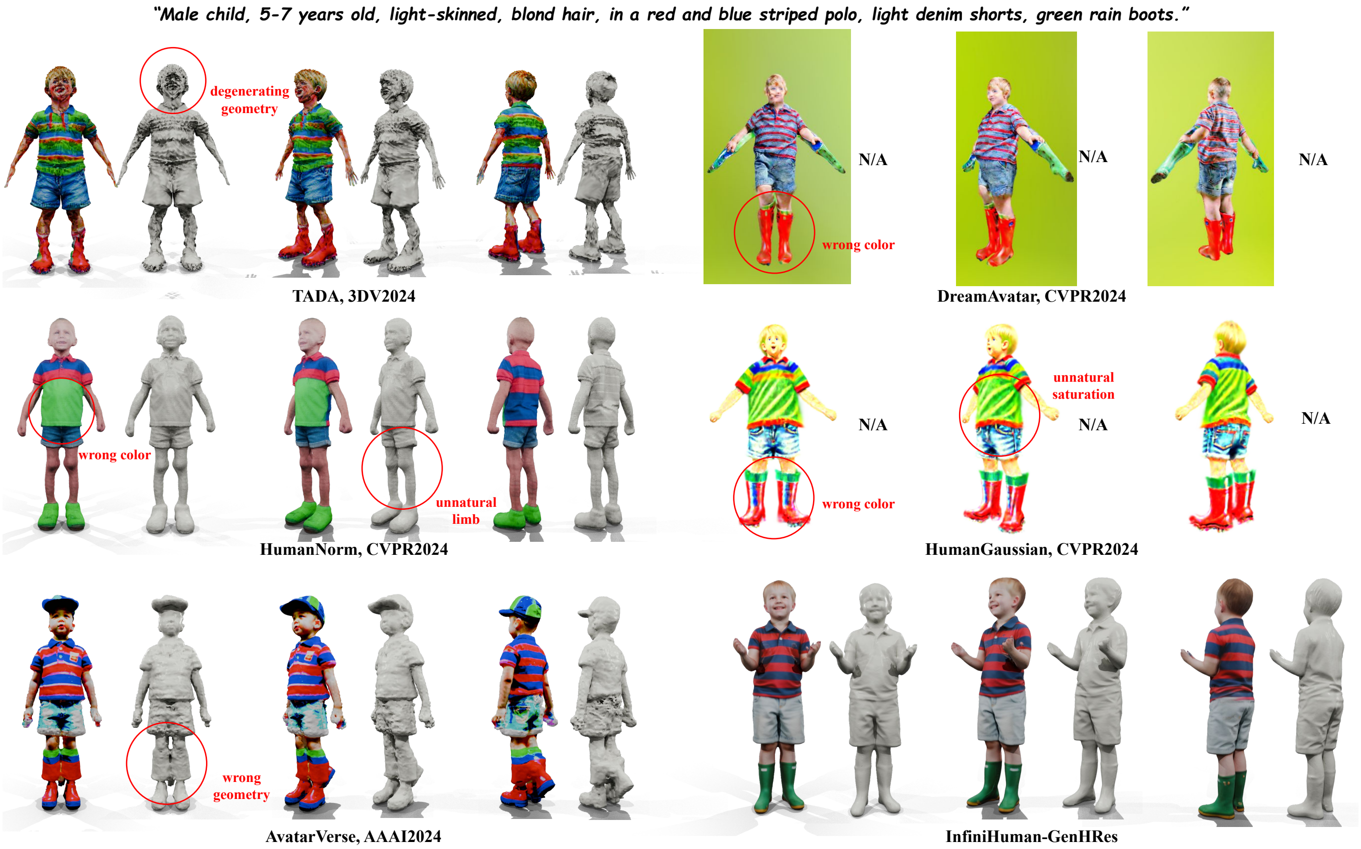

Fig. 12. Qualitative appearance and geometry comparison to SOTA text-to-3D avatar approaches. Please refer to Supp. Mat. for more comparisons.

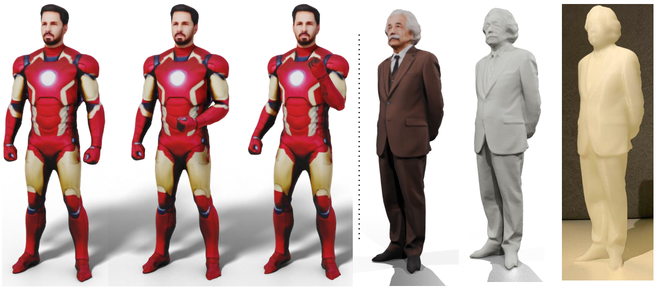

Fig. 13. Re-animation (left) and Fabrication (right) of Gen-HRes avatars.

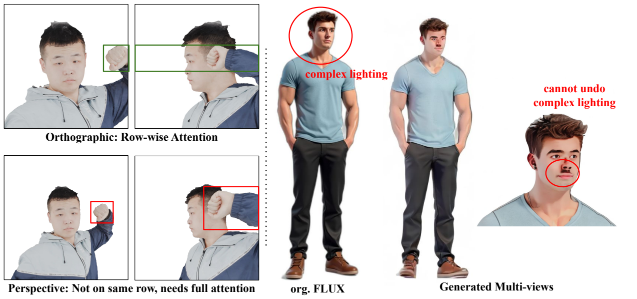
Orthographic: Row-wise Attention
Perspective: Not on same row, needs full attention
complex lighting
org. FLUX
Generated Multi-views
cannot undo complex lighting

Fig. 15. Orthographic and Perspective in Multi-View Attention (left). Org. FLUX gives complex lighting, degrading multi-view generation (right).

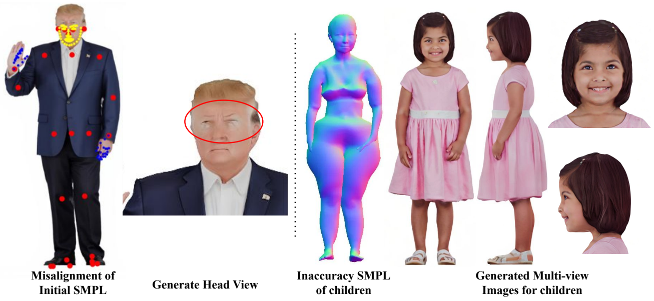
Misalignment of Initial SMPL
Generate Head View
Inaccuracy SMPL of children
Generated Multi-view Images for children

Fig. 14. Misaligned joints cause bad face generation (left). Our pipeline tolerates bad SMPL estimation for children, yielding good multi-views (right).

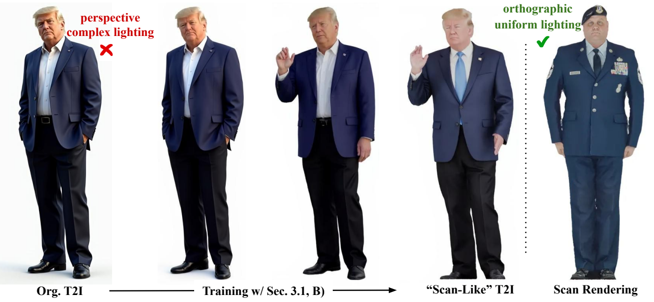

Fig. 16. Our finetuned FLUX can generate desired images from text prompt with orthographic view and uniform lighting, similar to the scan rendering.

References

2023. Twindom. https://web.twindom.com/

BFL Black Forest Labs. 2024. FLUX. https://github.com/black-forest-labs/flux.

Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. 2022. HuMMan: Multi-modal 4D Human Dataset for Versatile Sensing and Modeling. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VII (Lecture Notes in Computer Science, Vol. 13667)*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer, 557–577. doi:10.1007/978-3-031-20071-7_33

Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K. Wong. 2023. DreamAvatar: Text-and-Shape Guided 3D Human Avatar Generation via Diffusion Models. *arXiv preprint arXiv:2304.00916* (2023). https://arxiv.org/abs/2304.00916

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 958–968.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. 2015. ShapeNet: An Information-Rich 3D Model Repository. *CoRR* abs/1512.03012 (2015). arXiv:1512.03012 http://arxiv.org/abs/1512.03012

Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, Daxuan Ren, Lei Yang, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, Bo Dai, and Kwan-Yee Lin. 2023. DNA-Rendering: A Diverse Neural Actor Repository for High-Fidelity Human-centric Rendering. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023.* IEEE, 19925–19936. doi:10.1109/ICCV51070.2023.01829

Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. 2021. VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization. In *Proc. of the IEEE conference on 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. 2023. Objaverse-XL: A Universe of 10M+ 3D Objects. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/70364304877b5e767de4e9a2a511be0c-Abstract-Datasets_and_Benchmarks.html

Zijian Dong, Xu Chen, Jinlong Yang, Michael J. Black, Otmar Hilliges, and Andreas Geiger. 2023. AG3D: Learning to Generate 3D Avatars from 2D Image Collections. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023.* IEEE, 14870–14881. doi:10.1109/ICCV51070.2023.01370

Minghao Guo, Bohan Wang, Pingchuan Ma, Tianyuan Zhang, Crystal Elaine Owens, Chuang Gan, Josh Tenenbaum, Kaiming He, and Wojciech Matusik. 2024. Physically Compatible 3D Object Modeling from a Single Image. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). http://papers.nips.cc/paper_files/paper/2024/hash/d7af02c8a8e26608199c087f50a21d37-Abstract-Conference.html

Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. 2023b. High-fidelity 3D Human Digitization from Single 2K Resolution Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2023).*

Xiao Han, Yukang Cao, Kai Han, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, Tao Xiang, and Kwan-Yee K. Wong. 2023a. HeadSculpt: Crafting 3D Head Avatars with Text. *arXiv preprint arXiv:2306.03038* (2023). https://arxiv.org/abs/2306.03038

Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. 2023. Learning Locally Editable Virtual Humans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023.* IEEE, 21024–21035. doi:10.1109/CVPR52729.2023.02014

Fangzhou Hong, Zhaoxi Chen, Yushi LAN, Liang Pan, and Ziwei Liu. 2023. EVA3D: Compositional 3D Human Generation from 2D Image Collections. In *International Conference on Learning Representations.* https://openreview.net/forum?id=g7U9jD_2CUr

Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. 2022. AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars. In *ACM SIGGRAPH Conference Proceedings.* https://arxiv.org/abs/2205.08535

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022.* OpenReview.net. https://openreview.net/forum?id=nZeVKeeFYf9

Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. 2024. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation.

Mustafa Isik, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. 2023. HumanRF: High-Fidelity Neural Radiance Fields for Humans in Motion. *ACM Trans. Graph.* 42, 4 (2023), 160:1–160:12. doi:10.1145/3592415

Yash Kant, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, Igor Gilitschenski, and Aliaksandr Siarohin. 2024. SPAD : Spatially Aware Multiview Diffusers. arXiv:2402.05235 [cs.CV]

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* 42, 4 (2023), 139:1–139:14. doi:10.1145/3592433

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. 2024. Sapiens: Foundation for Human Vision Models. *arXiv preprint arXiv:2408.12569* (2024).

Byungjun Kim, Patrick Kwon, Kwangho Lee, Myunggi Lee, Sookwan Han, Daesik Kim, and Hanbyul Joo. 2023. Chupa: Carving 3D Clothed Humans from Skinned Shape Priors using 2D Diffusion Probabilistic Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).* 15965–15976.

Youwang Kim, Ji-Yeon Kim, and Tae-Hyun Oh. 2022. CLIP-Actor: Text-Driven Recommendation and Stylization for Animating Human Meshes. In *European Conference on Computer Vision (ECCV).* https://arxiv.org/abs/2206.04382

Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. 2023. DreamHuman: Animatable 3D Avatars from Text. *arXiv preprint arXiv:2306.09329* (2023). https://arxiv.org/abs/2306.09329

Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wei Xue, Wenhan Luo, Ping Tan, Wenping Wang, Qifeng Liu, and Yike Guo. 2024a. Era3D: High-Resolution Multiview Diffusion using Efficient Row-wise Attention. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). http://papers.nips.cc/paper_files/paper/2024/hash/65a723bf7d8dad838c09178270d30e80-Abstract-Conference.html

Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Mengfei Li, Xiaowei Chi, Siyu Xia, Wei Xue, Wenhan Luo, Qifeng Liu, and Yike Guo. 2024b. PSHuman: Photorealistic Single-view Human Reconstruction using Cross-Scale Diffusion. *CoRR* abs/2409.10141 (2024). doi:10.48550/ARXIV.2409.10141 arXiv:2409.10141

Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J. Black. 2023. TADA! Text to Animatable Digital Avatars. *arXiv preprint arXiv:2308.10899* (2023). https://arxiv.org/abs/2308.10899

Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J. Black. 2024. TADA! Text to Animatable Digital Avatars. In *International Conference on 3D Vision (3DV).*

Tingting Liao, Yujian Zheng, Yuliang Xiu, Adilbek Karmanov, Liwen Hu, Leyang Jin, and Hao Li. 2025. SOAP: Style-Omniscient Animatable Portraits. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '25).* Association for Computing Machinery, New York, NY, USA, Article 28, 11 pages. doi:10.1145/3721238.3730691

Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. 2024. HumanGaussian: Text-Driven 3D Human Generation with Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 6646–6657.

Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. 2022. Dress Code: High-Resolution Multi-Category Virtual Try-On. In *Proceedings of the European Conference on Computer Vision.*

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021. Neural Body: Implicit Neural Representations With Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021.* Computer Vision Foundation / IEEE, 9054–9063. doi:10.1109/CVPR46437.2021.00894

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023.* OpenReview.net. https://openreview.net/forum?id=FjNys5c7VyY

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. 2019. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).* 2304–2314. https://arxiv.org/abs/1905.05172

István Sárándi and Gerard Pons-Moll. 2024. Neural Localizer Fields for Continuous 3D Human Pose and Shape Estimation. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems*

2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). http://papers.nips.cc/paper_files/paper/2024/hash/fd23a1f3bc89e042d70960b466dc20e8-Abstract-Conference.html

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. 2024. MVDream: Multi-view Diffusion for 3D Generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=FUgrjq2pbB

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. 2024. OminiControl: Minimal and Universal Control for Diffusion Transformer. *CoRR* abs/2411.15098 (2024). doi:10.48550/ARXIV.2411.15098 arXiv:2411.15098

Zhenxiong Tan, Qiaochu Xue, Xingyi Yang, Songhua Liu, and Xinchao Wang. 2025. OminiControl2: Efficient Conditioning for Diffusion Transformers. *CoRR* abs/2503.08280 (2025). doi:10.48550/ARXIV.2503.08280 arXiv:2503.08280

Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. 2020. SIZER: A Dataset and Model for Parsing 3D Clothing and Learning Size Sensitive 3D Clothing. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 12348)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 1–18. doi:10.1007/978-3-030-58580-8_1

Jionghao Wang, Yuan Liu, Zhiyang Dou, Zhengming Yu, Yongqing Liang, Cheng Lin, Xin Li, Wenping Wang, Rong Xie, and Li Song. 2024. Disentangled Clothed Avatar Generation from Text Descriptions. *arXiv preprint arXiv:2312.05295* (2024). https://arxiv.org/abs/2312.05295

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. 2024. Structured 3D Latents for Scalable and Versatile 3D Generation. *arXiv preprint arXiv:2412.01506* (2024).

Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, Shuguang Cui, and Xiaoguang Han. 2024. MVHumanNet: A Large-Scale Dataset of Multi-View Daily Dressing Human Captures. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 19801–19811. doi:10.1109/CVPR52733.2024.01872

Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. 2023. ECON: Explicit Clothed Humans Optimized via Normal Integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 512–523. https://openaccess.thecvf.com/content/CVPR2023/html/Xiu_ECON_Explicit_Clothed_Humans_Optimized_via_Normal_Integration_CVPR_2023_paper.html

Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. 2022. ICON: Implicit Clothed Humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13296–13306. https://openaccess.thecvf.com/content/CVPR2022/html/Xiu_ICON_Implicit_Clothed_Humans_Obtained_From_Normals_CVPR_2022_paper.html

Yuliang Xiu, Yufei Ye, Zhen Liu, Dimitrios Tzionas, and Michael J Black. 2024. PuzzleAvatar: Assembling 3D Avatars from Personal Albums. *ACM Transactions on Graphics (TOG)* (2024).

Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard Pons-Moll. 2024. Human-3Diffusion: Realistic Avatar Creation via Explicit 3D Consistent Diffusion Models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). http://papers.nips.cc/paper_files/paper/2024/hash/b46aaf640bc8659e65a1a573971ba5a2-Abstract-Conference.html

Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. 2021. Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*.

Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. 2020. HUMBI: A Large Multiview Dataset of Human Body Expressions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2987–2997. doi:10.1109/CVPR42600.2020.00306

Ye Yuan, Xueting Li, Yangyi Huang, Shalini De Mello, Koki Nagano, Jan Kautz, and Umar Iqbal. 2024. GAvatar: Animatable 3D Gaussian Avatars with Implicit Mesh Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/2312.11461

Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Daniel Du, and Min Zheng. 2024. AvatarVerse: High-Quality & Stable 3D Avatar Creation from Text and Pose. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 7 (Mar. 2024), 7124–7132. doi:10.1609/aaai.v38i7.28540

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. *arXiv preprint arXiv:2308.03610* (2023). https://arxiv.org/abs/2308.03610

Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. 2021. PaMIR: Parametric Model-Conditioned Implicit Representation for Image-Based Human Reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–1. doi:10.1109/TPAMI.2021.3050505

Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. 2025. IDOL: Instant Photorealistic 3D Human Creation from a Single Image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/2412.14963