Bringing Your Portrait to 3D Presence

Jiawei Zhang[1,2 †] Lei Chu[2] Jiahao Li[2] Zhenyu Zang[2] Chong Li[2]
Xiao Li[2] Xun Cao[1] Hao Zhu[1] Yan Lu[2]

[1]Nanjing University [2]Microsoft Research Asia

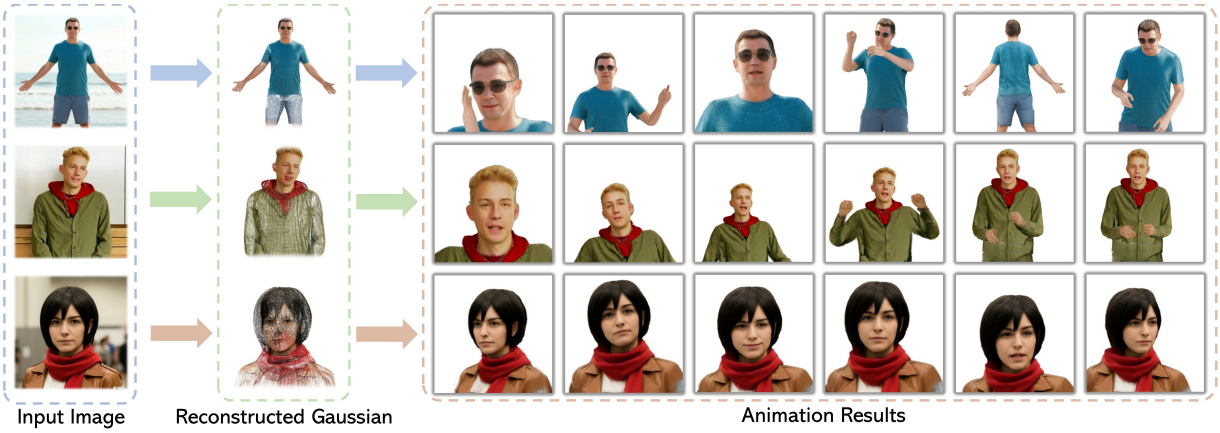

Figure 1. Our method uses a dual-UV formulation to represent 3D avatars, enabling reconstruction from full-body, half-body, and headshot portraits while capturing off-body textures. Trained entirely on synthetic data, it generalizes effectively to in-the-wild images.

Abstract

We present a unified framework for reconstructing animatable 3D human avatars from a single portrait across head, half-body, and full-body inputs. Our method tackles three bottlenecks: pose- and framing-sensitive feature representations, limited scalable data, and unreliable proxy-mesh estimation. We introduce a Dual-UV representation that maps image features to a canonical UV space via Core-UV and Shell-UV branches, eliminating pose- and framing-induced token shifts. We also build a factorized synthetic data manifold combining 2D generative diversity with geometry-consistent 3D renderings, supported by a training scheme that improves realism and identity consistency. A robust proxy-mesh tracker maintains stability under partial visibility. Together, these components enable strong in-the-wild generalization. Trained only on half-body synthetic data, our model achieves state-of-the-art head and upper-body reconstruction and competitive full-body results. Extensive experiments and analyses further validate the effectiveness of our approach.

1. Introduction

Creating *animatable* 3D human avatars is central to telepresence and virtual reality. While high-quality avatars usually rely on multi-view capture or depth sensors, these setups limit scalability; reconstructing an animatable avatar from a *single portrait* offers a far more accessible solution.

Despite rapid progress in 3D human reconstruction, most existing methods are designed either for head-only or for full-body avatars, and often rely on specific input assumptions. In particular, many pipelines assume full-body visibility (including both hands and feet) to obtain stable proxy mesh (*e.g.*, SMPL-X [64], FLAME [50]) fitting, an assumption that is not always satisfied in real scenarios, where upper-body or partial views are more typical. Recent transformer-based frameworks, including the Large Avatar Model [26] (LAM) and Large Human Model [69] (LHM), follow the Large Reconstruction Model [30] (LRM) paradigm by encoding input images into patch-level features and using learnable tokens to query them through cross-attention. This design enables fast single-image reconstruction without explicit geometry or texture optimization but constrains the representation to the image feature space, making it difficult to generalize

[†]This work was done during Jiawei Zhang's internship at Microsoft Research Asia.

across incomplete inputs. As shown in LHM, performance degrades under pose variation or partial-body inputs, and even when trained on half-body data, the ambiguous definition of "half-body," ranging from shoulder to waist or thigh crops, leads to inconsistent spatial correspondence and noticeable quality drops compared to full-body cases. Our goal is to relax both input and data requirements, advancing toward in-the-wild 3D avatar reconstruction from everyday captures such as webcams or phone portraits, thereby extending the scalability of animatable avatar reconstruction beyond controlled environments.

Our investigation reveals that the fundamental obstacles to advancing single-image 3D avatar reconstruction arise primarily from three aspects. First, **representation design**. Most existing pipelines inherit ViT-based pretrained encoders, which lack strict translation invariance and thus require the input image to be spatially aligned to a fixed reference. Unlike general object reconstruction, human images exhibit large pose variations and frequent partial-body visibility, making such alignment inherently unstable. Consequently, the decoder must learn to correlate image patches with 3D representations while adapting to token distribution shifts induced by pose and alignment inconsistencies, often leading to identity drift and texture distortion. Second, **data scalability**. High-quality multi-view human datasets require expensive studio setups with synchronized cameras, whereas real monocular videos demand intensive manual cleaning to ensure temporal and pose consistency. Synthetic data from traditional rendering engines offer controllable geometry but limited appearance diversity and a large domain gap to real imagery. Although 2D generative models can produce photorealistic humans with diverse appearances, their results generally lack identity and cross-view consistency, making them unsuitable for 3D supervision without further processing. Third, **robust body estimation**. Reliable proxy mesh tracking remains a key bottleneck. Existing trackers often assume full-body visibility, some even require both hands or the entire silhouette, to stabilize optimization, which rarely holds in in-the-wild captures dominated by upper-body views.

To address these challenges, we present a unified pipeline that integrates representation design, data construction, and proxy mesh estimation in a coherent framework. (1) **Dual-UV Representation**. At the core of our system is a Dual-UV representation that rearranges image features into a continuous, geometry-aligned UV space. It comprises two complementary branches: a Core-UV that encodes on-surface, geometry-anchored features, and a Shell-UV that captures off-surface details such as hair, clothing, and accessories by sampling features on an offset mesh shell. By anchoring tokens to a canonical surface rather than image coordinates, it eliminates token-distribution shifts caused by pose and alignment variations, which often lead to identity drift and texture distortion. This enables a single model to robustly handle head-only, half-body, and full-body inputs within one framework. (2) **Factorized Synthetic Data Manifold**. Our model is trained entirely on synthetic data that combines 2D generative and 3D rendered sources to achieve both appearance diversity and geometric reliability. Rather than enforcing multi-view consistency on 2D generative models, we leverage their ability to produce diverse, photorealistic appearances resembling real imagery. The 3D renderings, though less realistic, provide consistent geometric supervision that anchors reconstruction. All data are organized within a factorized, controllable manifold defined by semantically interpretable factors. A tailored training scheme mitigates identity and cross-view inconsistencies in the 2D data, while a realism regularizer projects each sample into a physically coherent, filmic space, preserving diversity while enhancing plausibility. Together, these designs yield a scalable synthetic corpus that enables stable training and strong generalization to in-the-wild captures. (3) **Proxy Mesh Estimation**. We empirically identify a stable configuration for human proxy mesh tracking under varying input completeness. Unlike previous approaches requiring full-body or both-hand visibility, our tracker maintains reliable performance across head-only, half-body, and full-body inputs, lowering capture demands and improving data scalability.

In summary, our work advances single-image 3D avatar reconstruction through the following contributions:

- We propose a simple yet effective Dual-UV representation that maps all inputs into a continuous, geometry-aligned UV space, enabling a single model to handle head-only, half-body, and full-body views.
- We construct a factorized synthetic data manifold with rich appearance diversity and controllable structure, supported by a training scheme that reduces identity inconsistency and cross-view mismatch, enabling strong real-world generalization from synthetic data alone.
- We develop a robust proxy-mesh tracker that remains stable across varying input completeness, reducing dependency on full-body visibility and improving scalability for in-the-wild reconstruction.
- Trained solely on half-body data, our approach achieves state-of-the-art head and upper-body results and competitive full-body performance.

2. Related Work

2.1. Human Datasets

Large-scale datasets underpin modern human modeling and fall into two categories. Structured datasets [25, 107] use controlled multi-view capture and provide diverse expressions, poses, and lighting [6, 11, 13, 32, 41, 57, 80, 84, 99, 104]. They support factor disentanglement but are

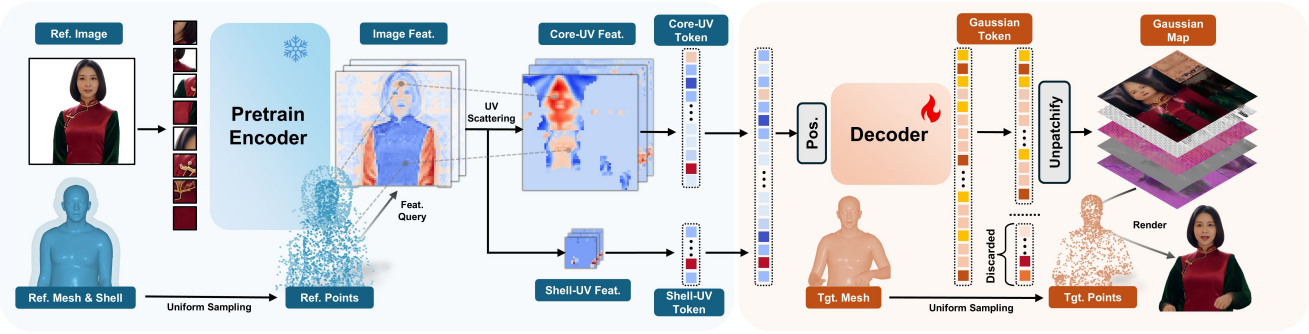

Figure 2. **Reconstruction Pipeline**. Given a reference image and its tracked proxy mesh, dense features from a frozen encoder are sampled along visible rays and scattered into canonical UV space to form the Core-UV map, while an offset shell captures off-surface regions such as hair and clothing. The Core-UV and Shell-UV tokens are fused and decoded by a lightweight transformer to reconstruct UV-space Gaussian attributes, which are then rigged to a target mesh and rendered from arbitrary viewpoints.

costly and limited to studio environments. Unstructured datasets [10, 19, 33, 36, 47, 98] collect in-the-wild images and videos with broad identity coverage but suffer from strong viewpoint bias toward near-frontal shots.

A complementary direction builds synthetic datasets [27, 89, 105] by rendering detailed assets with accurate annotations. They offer high controllability but exhibit realism and texture gaps relative to real imagery. Recent generative models [18, 21, 72, 117] help narrow this gap, enabling large-scale, realistic, and diverse synthetic human data. IDOL [117] also leverages 2D generative models for data curation by finetuning them for multi-view synthesis, but this introduces bias to the data diversity and still lacks reliable geometric consistency. In contrast, we retain the raw texture richness and diversity of 2D generative data without relying on its multi-view validity.

2.2. Human Modeling

3D human modeling spans head and full-body reconstruction, following a similar evolution. Early methods recovered geometry with explicit meshes or implicit fields [73, 100, 101, 113], and neural rendering [37, 58] shifted the focus to appearance. 3D-aware GANs [8, 9] enabled controllable head [2, 42, 45, 46, 106, 108] and body [1, 17, 28, 112] synthesis, often leveraging 3DMM priors [3]. Score-distillation [66] further allowed diffusion-driven generation of diverse 3D humans [43, 55, 76, 90, 97, 116].

A complementary line learns person-specific avatars from studio captures [14] or monocular videos [7]. For heads, studio setups enable generalizable models [24, 29, 48, 95], while single-view methods have advanced controllable 3D head synthesis [15, 16, 38, 52, 67, 75, 81, 109] and per-subject optimization achieves high fidelity [5, 68, 74, 96, 97, 102, 111, 114, 115]. For bodies, appearance and clothing variability make generalization harder, so many works rely on per-subject fitting from monocular [23, 59, 65] or multi-view captures [4, 51, 63, 86, 88, 103].

More recent efforts use multi-view [12, 31, 35, 49] and video-diffusion models [34, 56, 70] to generate multi-view imagery and reconstruct 3D avatars. In parallel, methods [26, 69] based on large reconstruction models [30, 87] provide single-forward 3D inference.

Recent works [1, 42, 48, 117] also use UV space for avatar attribute, but rely on StyleGAN or cross-attention to modulate UV attributes. We instead employ a non-learnable, closed-form projection from image space to UV space, enabling faithful texture transfer and strong identity preservation without generative hallucination.

Upper- or half-body avatar modeling from casual inputs is relatively underexplored and remains challenging. Existing half-body approaches are limited: many depend on fixed multi-view capture for real-time systems [44, 82], while others only extend head models slightly toward the shoulders [20, 91–94]. Diffusion-based methods [22, 53, 54] generate person-specific upper-body avatars but are generally constrained to frontal poses and do not generalize well to wider viewpoints. Recently, GUAVA [110] achieved a generalizable upper-body Gaussian avatar. GUAVA also uses a UV branch for appearance, but unlike our dual-UV design, it relies on a separate template branch to encode the input, and stabilize the training process, thus requires a later refining module to blend the Gaussians from two branches.

3. Method

Given a single RGB portrait I, our goal is to reconstruct an *animatable* 3D human avatar represented by a set of Gaussians $\mathcal{G} = \{g_i = (\boldsymbol{\mu}_i, \Sigma_i, \mathbf{c}_i, \alpha_i)\}_{i=1}^{N}$. We decompose the latent space into two complementary parts: $\mathbf{z} = \{\mathbf{z}_{\text{uv}}, \mathbf{z}_{\text{mesh}}\}$, where $\mathbf{z}_{\text{uv}}$ is a *dual-UV representation* encoding geometry-aligned appearance and visibility in a canonical UV space, and $\mathbf{z}_{\text{mesh}}$ is a *proxy mesh latent* parameterized by proxy mesh to capture pose-dependent deformation. The overall mapping is formulated as

$$f_\theta : I \rightarrow \mathbf{z}_{\text{uv}}, \quad \mathcal{G} = \Phi(\mathbf{z}_{\text{uv}}, \mathbf{z}_{\text{mesh}}),$$

where f_θ denotes the reconstruction network and Φ converts latent codes into posed 3D Gaussians.

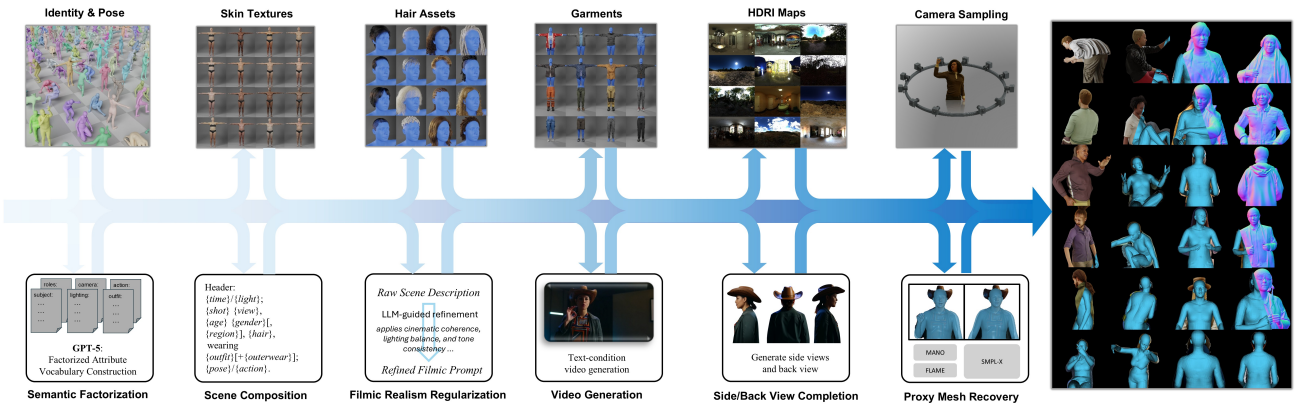

Figure 3. **Data Curation**. We build a hybrid dataset by combining geometry-anchored 3D rendering with semantics-driven generative synthesis. The *synthetic rendering branch* offers geometry-consistent multi-view supervision through procedural sampling of identity, pose, appearance, illumination, and cameras. The *generative branch* constructs a factorized appearance manifold by decomposing scene attributes, applying LLM-based filmic refinement, generating photorealistic sequences, and completing each sample with side/back views for weakly correlated augmentation.

In Sec. 3.1, we present the mask-based reconstruction model, centered on the dual-UV representation for robust alignment across varying input completeness, as shown in Fig. 2. Sec. 3.2 details the factorized synthetic data manifold curation pipeline that provides scalable and diverse supervision. Finally, Sec. 3.3 introduces the proxy mesh estimation framework for stable proxy mesh tracking. Refer to the supplementary for implementation and loss details.

3.1. Mask-based Reconstruction

Large reconstruction models [30, 87] map images to 3D by querying patch features with learnable tokens, which works for generic objects but entangles pose and identity for animatable humans. Because inputs vary greatly, from head-only to full-body, token correspondences become sensitive to framing and alignment, harming generalization. Moreover, these models must implicitly recover the canonical structure from posed images, even though the proxy mesh already provides the deformation field. This forces the decoder to relearn geometric mappings and wastes capacity.

We instead reconstruct through a geometry-aligned *Dual-UV representation* that deterministically maps image features into canonical UV space. This removes pose ambiguity, and lets the network focus solely on identity and appearance detail. The Dual-UV representation contains two complementary components.

Core-UV feature encoding. The Core-UV branch establishes a deterministic correspondence between image pixels and the canonical mesh surface. Using the UV layout of the human mesh, we define a differentiable unparameterization

$$(u, v) = \mathcal{M}^{-1}(\mathbf{p}; M), \tag{1}$$

where each surface point $\mathbf{p}$ on mesh M is uniquely mapped to its UV coordinate (u, v). We uniformly sample surface points $\{\mathbf{p}_i\}_{i=1}^{N}$ on M and precompute their face indices and barycentric coordinates.

Given an input image I and calibrated camera Π, we follow previous works [69, 117], use a frozen Sapiens-1B encoder [39] to extract image features $\mathbf{F} = \mathcal{E}(I)$. We rasterize M under Π with back-face culling and z-buffering to determine visibility. For each visible point $\mathbf{p}_i$, it is projected to the image $\mathbf{x}_i = \Pi(\mathbf{p}_i)$, and its feature is gathered by:

$$\mathbf{f}_i = \mathcal{S}(\mathbf{F}, \mathbf{x}_i), \quad m_i \in \{0, 1\}, \tag{2}$$

where $\mathcal{S}$ denotes differentiable bilinear sampling that interpolates local features from $\mathbf{F}$ at subpixel coordinate $\mathbf{x}_i$. The sampled features are then scattered onto a regular UV grid $\tilde{\mathbf{U}} \in \mathbb{R}^{H_U \times W_U \times C}$:

$$\tilde{\mathbf{U}}(u, v) = \frac{\sum_i m_i \, k((u, v) - (u_i, v_i)) \, \mathbf{f}_i}{\sum_i m_i \, k((u, v) - (u_i, v_i)) + \varepsilon}, \tag{3}$$

where $k(\cdot)$ is a compact aggregation kernel (nearest-neighbor in practice) and ε ensures stability. This produces a canonical, geometry-aligned Core-UV feature map that provides a one-to-one, differentiable link between image observations and mesh-surface coordinates.

Shell-UV feature encoding. The human mesh M captures only the body surface and cannot represent volumetric details such as hair or loose garments. To encode these off-surface regions, we construct an auxiliary shell M^+ by offsetting each mesh vertex along its outward normal: $M^+ = \{\mathbf{p} + \delta \, \mathbf{n}(\mathbf{p}) \mid \mathbf{p} \in M\}$, where $\mathbf{n}(\mathbf{p})$ is the vertex normal and δ is a small offset. We rasterize both M and M^+ under camera Π to obtain their visibility masks m_M and m_{M^+}. The *shell-only* region is defined as the visible area of S excluding the projection of M, $m_{\text{shell}} = m_{M^+} \cdot (1 - m_M)$, ensuring that only off-surface pixels are used for feature sampling. For m_{shell}, it is multiplied by the body mask rendered from the shell proxy mesh.

For each visible point $\mathbf{p}_j \in M^+$ with $m_{\text{shell},j} = 1$, its projection $\mathbf{x}_j = \Pi(\mathbf{p}_j)$ queries the image feature map $\mathbf{F}$ by differentiable bilinear sampling:

$$\mathbf{f}_j = \mathcal{S}(\mathbf{F}, \mathbf{x}_j). \tag{4}$$

The sampled features are transferred to their corresponding UV coordinates $(u_j, v_j) = \mathcal{M}^{-1}(\mathbf{p}_j)$ and aggregated on a coarse UV grid $\tilde{\mathbf{U}}_{\text{shell}} \in \mathbb{R}^{H'_U \times W'_U \times C}$ using the same kernel aggregation as in the Core-UV branch. Although the mapping from M^+ to UV space is approximate, it maintains local spatial coherence and provides a soft positional prior for encoding off-surface appearance.

Decoder Design. Core-UV and Shell-UV tokens are concatenated, processed by a shallow transformer stack, and unpatched to form the UV attributes. Separate heads predict Gaussian attributes in UV space, including color $\boldsymbol{c}$, opacity o, offset $\boldsymbol{d}$, and rotation $\boldsymbol{r}$ and scale $\boldsymbol{s}$.

Discussion. Transformer-based avatar models [69, 117] typically employ large decoders with learnable queries, treating the pretrained encoder as a passive feature extractor and discarding the masked-autoencoding training asymmetry, where a strong encoder enables a lightweight decoder, as in Sapiens [39]. In contrast, we retain this asymmetry: projected image features fill only visible UV cells, leaving occluded regions blank, analogous to masked tokens. A compact transformer propagates information from visible to missing areas. This MAE-aligned design exploits pretrained reconstruction priors in geometry-aligned space, achieving high-quality avatars with far fewer decoder parameters than query-based approaches.

3.2. Dataset Curation

Training a single-image avatar reconstructor requires large-scale data that jointly cover geometric reliability and photorealistic diversity. Existing multi-view human datasets [32, 35, 57] are limited in identity and appearance due to costly studio capture. We therefore synthesize data from two complementary sources: a geometry-accurate *synthetic rendering branch* and a photorealistic *generative branch* (Fig. 3).

Synthetic Rendering Branch. We render multi-view human images with a parametric body model following [27]. Identity, pose, garment, and lighting are sampled procedurally, and each subject is rendered from multiple calibrated viewpoints under HDRI environments. This provides geometry-consistent supervision without manual annotation, forming the structural backbone of training data.

Generative Branch. While prior efforts [21, 117] fine-tune diffusion models for view-consistent humans, such constraints often degrade realism and diversity. We instead embrace a different philosophy: rather than forcing 2D generators to be multi-view consistent, we exploit their strengths, rich identity variation, natural appearance, and realistic motion, to populate a broad and controllable distribution.

To this end, we define a *factorized data manifold* in which each sample is described by interpretable dimensions such as time of day, lighting, shot size, composition, clothing, hairstyle, role, region, and action. Combinations of these factors are first assembled into concise textual descriptions by GPT-5 [60], then refined by a large language model (Qwen2.5-14B-Instruct [77]) acting as a *realism regularizer*. This refinement step projects each prompt into a physically coherent, filmic space—resolving contradictory attributes and enriching it with cinematographic cues on framing, illumination, and tone. The refined prompts are passed to the text-to-video generator Wan2.2 [79] to produce short, temporally consistent human clips. From each clip, one representative frame is selected and complemented by side and back views synthesized through Qwen-Image-Edit [78]. Unlike prior works that assume perfect multi-view consistency, we treat these generated frames as weakly correlated views rather than ground-truth correspondences. During training, we impose a *directional cross-view consistency* that flows only from more reliable to less reliable views (e.g., side→back, front→back), avoiding cyclic constraints that can amplify identity drift or texture aliasing. This asymmetric design effectively stabilizes training while still encouraging view-aware coherence.

Discussion. Our design deliberately avoids fine-tuning generative models for 3D consistency, focusing instead on realism and diversity as complementary to the geometry-rich synthetic renders. Together, these two branches form a scalable corpus where geometry and photorealism are disentangled yet coherent: (i) the synthetic branch anchors geometric supervision, (ii) the generative branch expands appearance coverage within a factorized manifold, and (iii) the realism regularizer maintains filmic plausibility. This combination enables robust training and strong generalization to in-the-wild portraits. *Implementation details of the control factors, prompt templates, and LLM refinement commands are provided in the supplementary material.*

3.3. Proxy Mesh Estimation

A reliable proxy mesh is essential for canonical avatar reconstruction, yet many pretrained models can provides initial estimates, but each of them assumes its own cropped views and coordinate system. We therefore analyze the stability ranges of multiple estimators and build a unified tracking pipeline rather than relying on a single model.

We benchmark representative estimators across head-only, half-body, and full-body inputs. OSX performs well with half-body visibility; Multi-HMR delivers accurate estimates when the entire body is visible; EMICA remains stable for head-dominant inputs; HaMeR provides accurate hand articulation only when hands are visible. This yields

an empirical map of each model's reliable operating regime. Guided by this map, we design a hierarchical framework that activates and fuses outputs from multiple estimators according to detected visibility, then jointly refine via keypoint reprojection and dense vertex alignment.

This tracker produces stable, anatomically coherent meshes for both real and generated images across all input types, from head-only portraits to full-body captures, and serves as a robust foundation for our reconstruction pipeline. Implementation details and further analyses are provided in the supplementary material.

4. Experiment

4.1. Implementation Details

We train on an upper-body–dominant synthetic dataset, while our data generation and training pipeline also flexibly accommodate full-body samples and head-only data. Our rendering pipeline generates 150K subjects assembled from curated 3D assets, each rendered from 12 upper-body–focused views. An additional 300K portrait video clips with talking or upper-body motion are synthesized, along with side and back views for a random frame. Data are split 19:1 for training and validation.

Our decoder has 8 self-attention layers and fewer than 0.1B parameters. Its outputs are unpatchified into 8×8 patches to form 512×512 Gaussian attribute maps (262K Gaussians). We train with AdamW [40] at a learning rate of 1×10^{-4}, using mixed precision and gradient clipping (norm 1.0). Training runs for three days on 4 NVIDIA A100 80G GPUs with a batch size of 8 per GPU, the overall training cost is *significantly lower* than LRM-based methods.

4.2. Experiment Setup

We compare with LHM, LHM-HF, IDOL, LAM, and GUAVA under head, upper-body, and full-body inputs.

LHM targets full-body reconstruction, encoding facial and body regions separately using DINOv2 [61] and Sapiens backbones for cross-attention between modalities. LHM-HF extends this design by training on half-body data augmented by random cropping from LHM's large-scale in-the-wild video dataset. IDOL applies a large Transformer decoder along with a heavy CNN-based decoder to regress Gaussian attributes from learnable tokens. LAM simplifies LHM's architecture, focusing on head modeling using both studio-captured [41] and in-the-wild data [98]. GUAVA specializes in upper-body reconstruction, featuring a dual-branch Gaussian decoder followed by a screen-space CNN refinement [85] stage for improved fidelity.

For upper-body evaluation, we use 100 real clips from OpenHumanVid [47] and 200 synthesized clips; for head portraits, we sample 50 talking clips from RenderMe360 [62]; for full-body, we select 100 subjects from SHHQ [19]. We compare against the relevant baselines in each setting, using their own preprocessing pipeline.

4.3. Comparison Results

Quantitative results are reported in Tab. 1 and qualitative comparisons in Fig. 4. For upper-body, our method surpasses LRM-based approaches (IDOL, LHM-HF) in texture fidelity and identity preservation. While GUAVA is competitive on visible regions, its strict requirement for visible hands causes unstable poses and artifacts under hand occlusions. In contrast, our method avoids any 2D refinement and remains robust across varying visibility conditions.

For head reconstruction, we outperform the head-specific LAM in identity preservation and additionally reconstruct regions below the shoulders that LAM cannot represent. In the full-body setting, our approach achieves performance comparable to dedicated single-image full-body methods, despite being trained almost exclusively on upper-body data and never seeing lower-body regions. This indicates strong generalization to unseen body parts.

As shown in Fig. 5, across all three settings our method produces novel views with consistent geometry and texture under diverse viewpoints and challenging poses.

4.4. Applications

Versatile Editing Our model exhibits strong generalization ability, allowing seamless integration with outputs from advanced image generation or editing models. As illustrated in Fig. 6, it can transform images synthesized by text-to-image or image-editing models into fully animatable 3D avatars, enabling flexible downstream editing and control.

Multiple Inputs Recent works [71] attempt to handle multi-view human images using computationally alternative attention [83]. In contrast, our method can naturally handle multi-view inputs by simply linearly blending the UV features from each image. As illustrated in Fig. 7, this simple strategy is sufficient to produce coherent reconstructions even in challenging multi-view scenarios.

4.5. Ablation Study

Model Design We ablate the decoder architecture, as summarized in Tab. 2 (a) and (d). Increasing the number of decoder layers leads to a consistent performance gain. In addition, introducing the shell token provides a clear boost, indicating that the extra token effectively enriches local surface modeling and, in turn, improves reconstruction quality.

Dataset Scalability We also study the impact of training data type and scale. As shown in Tab. 2 (b) and (c), model performance improves steadily as the dataset grows, highlighting the benefit of larger and more diverse supervision.

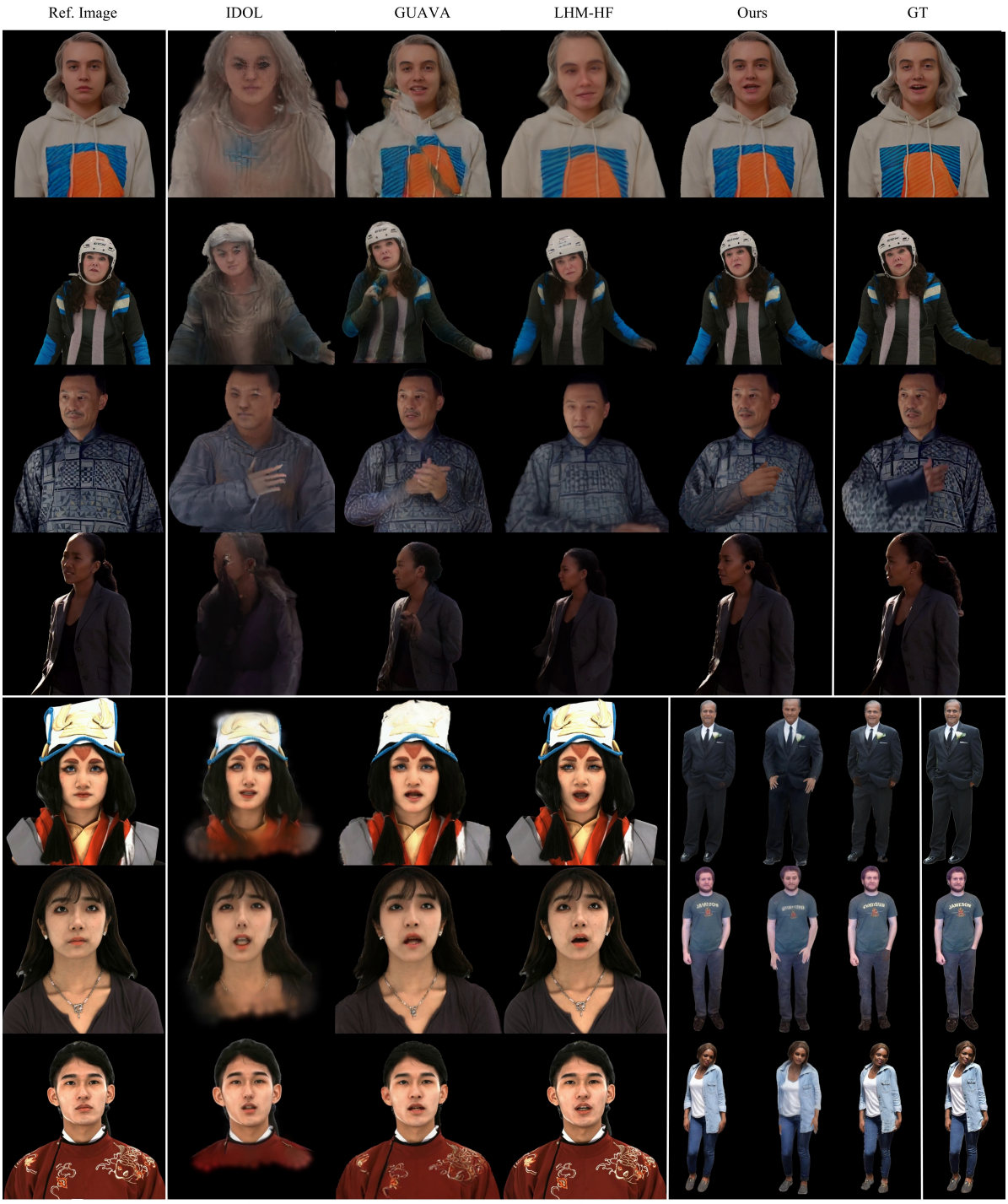

Figure 4. **Reenactment Results**. Our method is trained solely on *upper-body data*, generalizes well to head and full-body inputs.

Table 1. Comparision of quantitative results with state-of-the-art methods.

	Upper-Wild			Upper-Wan				Full-Body				Head		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓		PSNR↑	SSIM↑	LPIPS↓		PSNR↑	SSIM↑	LPIPS↓
Ours	21.09	0.8510	0.1426	20.38	0.7867	0.1635	Ours	24.53	0.8642	0.0916	Ours	19.04	0.8526	0.1613
GUAVA	20.59	0.7864	0.1957	20.24	0.7215	0.1940	IDOL	18.51	0.8753	0.1256	LAM†	17.19	0.7526	0.2207
LHM-HF	13.95	0.7835	0.3335	12.04	0.6664	0.3626	LHM	21.53	0.9151	0.0725	LAM	14.81	0.6789	0.2613
IDOL	12.93	0.7617	0.3465	10.35	0.5802	0.5132	——	——			——	——		

Note: blue and lightblue indicate the best and second-best results. † indicates non-facial parts are parsed out.

Figure 5. **Novel View Synthesis**. Our method generates multi-view human renderings from a single reference image, showing comparatively more consistent appearance, especially in the head and upper-body regions.

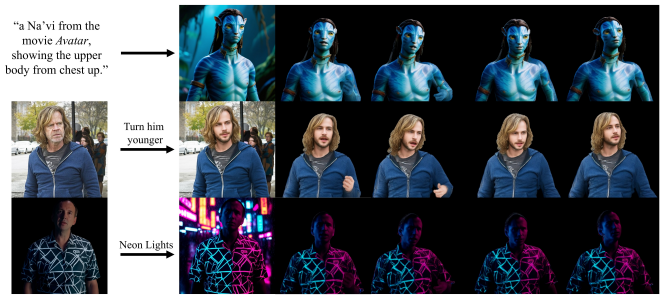

Figure 6. **Editing Results**. Our model supports various appearance edits from a single image, demonstrating its adaptability to diverse visual conditions.

Figure 7. **Multiple Input**. Our model is capable of taking multiple images as input, indicating its potential flexibility in leveraging multi-view information.

When trained only on synthetic data *syn*, the model shows limited generalization. Using only generation data *gen* alleviates this issue, while *gen**, which contains only video-generated data without augmented images, performs worse. The best results are achieved when combining all data sources, suggesting that accurate 3D simulations complemented by diverse 2D views are crucial for high-quality reconstruction.

Table 2. Ablation study on model design and dataset scalability, evaluated on upper-body synthetic clips. Blue indicates the best results.

(a) Decoder depth

blocks	PSNR↑	SSIM↑	LPIPS↓
2	20.31	0.7846	0.1812
4	20.36	0.7862	0.1789
8	20.38	0.7867	0.1635

(b) Dataset scale

ratio	PSNR↑	SSIM↑	LPIPS↓
95%	20.38	0.7867	0.1635
33%	20.33	0.7855	0.1801
3%	20.25	0.7827	0.1888

(c) Dataset type

type	PSNR↑	SSIM↑	LPIPS↓
full	20.38	0.7867	0.1635
syn	19.57	0.7579	0.2418
gen	20.38	0.7861	0.1732
gen*	20.34	0.7851	0.1794

(d) Shell token

	PSNR↑	SSIM↑	LPIPS↓
w/ shell token	20.38	0.7867	0.1635
w/o shell token	20.35	0.7845	0.1801

5. Conclusion

We present a unified framework for reconstructing animatable 3D human avatars, spanning head-only, half-body, and full-body inputs within a single model. We train our model entirely on synthetic data, yet it generalizes well to in-the-wild portraits. Extensive experiments across head, upper-body, and full-body benchmarks demonstrate state-of-the-art or competitive performance, along with strong novel view synthesis and versatile applications such as reenactment, appearance editing, and multi-view fusion.

Despite these advances, our framework still has limitations. It relies on a proxy mesh that may fail in extreme or highly articulated poses, leading to noticeable reconstruction errors. Although our synthetic data manifold provides diverse identities and appearances, the views are still sparse, with limited side or rear perspectives, resulting in incomplete viewpoint coverage during training.

Future work will explore weakly pose-dependent representations to mitigate these issues and further enhance robustness and generalization. We also believe that improving viewpoint coverage and reducing the dependence on proxy geometry could further strengthen performance in more unconstrained real-world scenarios.

Acknowledgements

The authors would like to express their sincere gratitude to Paul McIlroy, Tadas Baltrusaitis, and Charlie Hewitt for generously providing the synthetic human rendering pipeline, which played a crucial role in this project. Additionally, we are deeply thankful to Ross Cutler's team for their continuous support and insightful discussions, which greatly enhanced the quality of our work.

References

[1] Rameen Abdal, Wang Yifan, Zifan Shi, Yinghao Xu, Ryan Po, Zhengfei Kuang, Qifeng Chen, Dit-Yan Yeung, and Gordon Wetzstein. Gaussian shell maps for efficient 3d human generation. In *CVPR*, 2024. 3

[2] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *CVPR*, pages 20950–20959, 2023. 3

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *ACM TOG*, page 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 3

[4] Marcel C. Bühler, Ye Yuan, Xueting Li, Yangyi Huang, Koki Nagano, and Umar Iqbal. Dream, lift, animate: From single images to animatable gaussian avatars, 2025. 3

[5] Hongrui Cai, Yuting Xiao, Xuan Wang, Jiafei Li, Yudong Guo, Yanbo Fan, Shenghua Gao, and Juyong Zhang. Hera: Hybrid explicit representation for ultra-realistic head avatars. In *CVPR*, 2025. 3

[6] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE TVCG*, 20(3):413–425, 2014. 2

[7] Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. Real-time facial animation with image-based dynamic avatars. *ACM TOG*, 35(4), 2016. 3

[8] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, 2021. 3

[9] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. 3

[10] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *ICML*, pages 6263–6285, 2024. 3

[11] Jianchuan Chen, Jingchuan Hu, Gaige Wang, Zhonghua Jiang, Tiansong Zhou, Zhiwen Chen, and Chengfei Lv. Taoavatar: Real-time lifelike full-body talking avatars for augmented reality via 3d gaussian splatting. In *CVPR*, pages 10723–10734, 2025. 2

[12] Wenyue Chen, Peng Li, Wangguandong Zheng, Chengfeng Zhao, Mengfei Li, Yaolong Zhu, Zhiyang Dou, Ronggang Wang, and Yuan Liu. Synchuman: Synchronizing 2d and 3d diffusion models for single-view human reconstruction. In *NeurIPS*, 2025. 3

[13] 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. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *ICCV*, pages 19982–19993, 2023. 2

[14] Paul Debevec. The light stages and their applications to photoreal digital actors. *ACM TOG*, 2(4):1–6, 2012. 3

[15] Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data. In *CVPR*, 2024. 3

[16] Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. In *ECCV*, 2024. 3

[17] Zijian Dong, Xu Chen, Jinlong Yang, Michael J. Black, Otmar Hilliges, and Andreas Geiger. AG3D: Learning to generate 3D avatars from 2D image collections. In *ICCV*, 2023. 3

[18] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 3

[19] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen-Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *ECCV*, pages 729–747, 2022. 3, 6

[20] Xuan Gao, Haiyao Xiao, Chenglai Zhong, Shimin Hu, Yudong Guo, and Juyong Zhang. Portrait video editing empowered by multimodal generative priors. In *SIGGRAPH Asia*, 2024. 3

[21] Xuan Gao, Jingtao Zhou, Dongyu Liu, Yuqi Zhou, and Juyong Zhang. Controlling avatar diffusion with learnable gaussian embedding. In *SIGGRAPH Asia*, 2025. 3, 5

[22] Jiazhi Guan, Quanwei Yang, Kaisiyuan Wang, Hang Zhou, Shengyi He, Zhiliang Xu, Haocheng Feng, Errui Ding, Jingdong Wang, Hongtao Xie, Youjian Zhao, and Ziwei Liu. Talk-act: Enhance textural-awareness for 2d speaking avatar reenactment with diffusion model. In *SIGGRAPH Asia*, 2024. 3

[23] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *CVPR*, 2023. 3

[24] Chen Guo, Zhuo Su, Jian Wang, Shuang Li, Xu Chang, Zhaohu Li, Yang Zhao, Guidong Wang, and Ruqi Huang. Sega: Drivable 3d gaussian head avatar from a single image, 2025. 3

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

[26] Yisheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Weihao Yuan, Zilong Dong, and Liefeng Bo. Lam: Large avatar model for one-shot animatable gaussian head. In *SIGGRAPH*, 2025. 1, 3

[27] Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafiirah Hosenie, Thomas J Cashman, Julien Valentin, Darren Cosker, and Tadas Baltrušaitis. Look ma, no markers: holistic performance capture without the hassle. *ACM TOG*, 43(6), 2024. 3, 5

[28] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections. *ICLR*, 2022. 3

[29] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *CVPR*, 2022. 3

[30] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. In *ICLR*, 2024. 1, 3, 4

[31] Yangyi Huang, Ye Yuan, Xueting Li, Jan Kautz, and Umar Iqbal. Adahuman: Animatable detailed 3d human generation with compositional multiview diffusion. In *ICCV*, pages 13533–13543, 2025. 3

[32] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM TOG*, 42(4):1–12, 2023. 2, 5

[33] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *CVPR*, pages 12753–12762, 2021. 3

[34] Yudong Jin, Sida Peng, Xuan Wang, Tao Xie, Zhen Xu, Yifan Yang, Yujun Shen, Hujun Bao, and Xiaowei Zhou. Diffuman4d: 4d consistent human view synthesis from sparse-view videos with spatio-temporal diffusion models. In *ICCV*, 2025. 3

[35] Yash Kant, Ethan Weber, Jin Kyu Kim, Rawal Khirodkar, Su Zhaoen, Julieta Martinez, Igor Gilitschenski, Shunsuke Saito, and Timur Bagautdinov. Pippo: High-resolution multi-view humans from a single image. In *CVPR*, 2025. 3, 5

[36] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE TPAMI*, 43(12):4217–4228, 2021. 3

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

[38] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *ECCV*, 2022. 3

[39] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Zhaoen Su, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *ECCV*, 2024. 4, 5

[40] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 6

[41] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM TOG*, 2023. 2, 6

[42] Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. GGHead: Fast and Generalizable 3D Gaussian Heads. In *SIGGRAPH Asia*, 2024. 3

[43] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *NeurIPS*, 36:10516–10529, 2023. 3

[44] Jason Lawrence, Danb Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G. Desloge, Tommy Fortes, Eric M. Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, Claude Knaus, Brian Kuschak, Ricardo Martin-Brualla, Harris Nover, Andrew Ian Russell, Steven M. Seitz, and Kevin Tong. Project starline: a high-fidelity telepresence system. *ACM TOG*, 40(6), 2021. 3

[45] Heyuan Li, Ce Chen, Tianhao Shi, Yuda Qiu, Sizhe An, Guanying Chen, and Xiaoguang Han. Spherehead: Stable 3d full-head synthesis with spherical tri-plane representation. In *ECCV*, 2024. 3

[46] Heyuan Li, Kenkun Liu, Lingteng Qiu, Qi Zuo, Keru Zheng, Zilong Dong, and Xiaoguang Han. Hyplanehead: Rethinking tri-plane-like representations in full-head image synthesis. In *NeurIPS*, 2025. Poster. 3

[47] Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, et al. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In *CVPR*, 2025. 3, 6

[48] Junxuan Li, Chen Cao, Gabriel Schwartz, Rawal Khirodkar, Christian Richardt, Tomas Simon, Yaser Sheikh, and Shunsuke Saito. Uravatar: Universal relightable gaussian codec avatars. In *SIGGRAPH*, 2024. 3

[49] Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Mengfei Li, Xiaowei Chi, Siyu Xia, Wei Xue, et al. Pshuman: Photorealistic single-view human reconstruction using cross-scale diffusion. In *CVPR*, 2025. 3

[50] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM TOG*, 36(6):194:1–194:17, 2017. 1

[51] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *CVPR*, pages 19711–19722, 2024. 3

[52] Tingting Liao, Yujian Zheng, Adilbek Karmanov, Liwen Hu, Leyang Jin, Yuliang Xiu, and Hao Li. Soap: Style-omniscient animatable portraits. In *SIGGRAPH*, 2025. 3

[53] Gaojie Lin, Jianwen Jiang, Chao Liang, Tianyun Zhong, Jiaqi Yang, Zerong Zheng, and Yanbo Zheng. Cyberhost: A one-stage diffusion framework for audio-driven talking body generation. In *ICLR*, 2025. 3

[54] Haiyang Liu, Xingchao Yang, Tomoya Akiyama, Yuantian Huang, Qiaoge Li, Shigeru Kuriyama, and Takafumi Taketomi. Tango: Co-speech gesture video reenactment with hierarchical audio motion embedding and diffusion interpolation. In *ICLR*, 2025. 3

[55] Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. Humangaussian: Text-driven 3d human generation with gaussian splatting. In *CVPR*, 2024. 3

[56] Yixing Lu, Junting Dong, Youngjoong Kwon, Qin Zhao, Bo Dai, and Fernando De la Torre. Gas: Generative avatar synthesis from a single image. In *ICCV*, 2025. 3

[57] Julieta Martinez, Emily Kim, Javier Romero, et al. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. *NeurIPS*, 2024. 2, 5

[58] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3

[59] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3D gaussian avatar. In *ECCV*, 2024. 3

[60] OpenAI. Introducing gpt-5, 2025. Blog post. 5

[61] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 6

[62] Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, and Kwan-Yee Lin. Renderme-360: Large digital asset library and benchmark towards high-fidelity head avatars. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 6

[63] Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. Humansplat: Generalizable single-image human gaussian splatting with structure priors. In *NeurIPS*, 2024. 3

[64] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 1

[65] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 3

[66] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 3

[67] Malte Prinzler, Egor Zakharov, Vanessa Sklyarova, Berna Kabadayi, and Justus Thies. Joker: Conditional 3D head synthesis with extreme facial expressions. In *3DV*, 2025. 3

[68] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *CVPR*, 2023. 3

[69] Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, and Liefeng Bo. Lhm: Large animatable human reconstruction model from a single image in seconds. In *ICCV*, 2025. 1, 3, 4, 5

[70] Lingteng Qiu, Shenhao Zhu, Qi Zuo, Xiaodong Gu, Yuan Dong, Junfei Zhang, Chao Xu, Zhe Li, Weihao Yuan, Liefeng Bo, et al. Anigs: Animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. In *CVPR*, 2025. 3

[71] Lingteng Qiu, Peihao Li, Heyuan Li, Qi Zuo, Xiaodong Gu, Yuan Dong, Weihao Yuan, Rui Peng, Siyu Zhu, Xiaoguang Han, Guanying Chen, and Zilong Dong. Lhm++: An efficient large human reconstruction model for pose-free images to 3d, 2026. 6

[72] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3

[73] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, pages 2304–2313, 2019. 3

[74] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *CVPR*, 2024. 3

[75] Yuxiang Shi, Zhe Li, Yanwen Wang, Hao Zhu, Xun Cao, and Ligang Liu. Dex-portrait: Disentangled and expressive portrait animation via explicit and latent motion representations. *arXiv preprint arXiv:2512.15524*, 2025. 3

[76] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *ICLR*, 2024. 3

[77] Qwen Team. Qwen2.5 technical report, 2025. 5

[78] Qwen-Image Team. Qwen-image technical report, 2025. 5

[79] Wan Team. Wan: Open and advanced large-scale video generative models, 2025. 5

[80] Timo Teufel, Xilong Zhou, Umar Iqbal, Pramod Rao, Pulkit Gera, Jan Kautz, Vladislav Golyanik, and Christian Theobalt. Humanolat: A large-scale dataset for full-body human relighting and novel-view synthesis. In *ICCV*, 2025. 2

[81] Phong Tran, Egor Zakharov, Long-Nhat Ho, Liwen Hu, Adilbek Karmanov, Aviral Agarwal, McLean Goldwhite, Ariana Bermudez Venegas, Anh Tuan Tran, and Hao Li. Voodoo xp: Expressive one-shot head reenactment for vr telepresence. *ACM TOG*, 2024. 3

[82] Hanzhang Tu, Ruizhi Shao, Xue Dong, Shunyuan Zheng, Hao Zhang, Lili Chen, Meili Wang, Wenyu Li, Siyan Ma, Shengping Zhang, Boyao Zhou, and Yebin Liu. Tele-aloha: A telepresence system with low-budget and high-authenticity using sparse rgb cameras. In *SIGGRAPH*, 2024. 3

[83] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025. 6

[84] Lizhen Wang, Zhiyua Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *CVPR*, 2022. 2

[85] Lizhen Wang, Xiaochen Zhao, Jingxiang Sun, Yuxiang Zhang, Hongwen Zhang, Tao Yu, and Yebin Liu. Styleavatar: Real-time photo-realistic portrait avatar from a single video. In *SIGGRAPH*, 2023. 6

[86] Li Wang, Yiyu Zhuang, Yanwen Wang, Xun Cao, Chuan Guo, Xinxin Zuo, and Hao Zhu. Sketch2posenet: Efficient and generalized sketch to 3d human pose prediction. In *SIGGRAPH Asia*, 2025. 3

[87] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. In *ICLR*, 2024. 3, 4

[88] Shaofei Wang, Tomas Simon, Igor Santesteban, Timur Bagautdinov, Junxuan Li, Vasu Agrawal, Fabian Prada, Shoou-I Yu, Pace Nalbone, Matt Gramlich, Roman Lubachersky, Chenglei Wu, Javier Romero, Jason Saragih, Michael Zollhoefer, Andreas Geiger, Siyu Tang, and Shunsuke Saito. Relightable full-body gaussian codec avatars. In *SIGGRAPH*, 2025. 3

[89] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *CVPR*, pages 4563–4573, 2023. 3

[90] Yanwen Wang, Yiyu Zhuang, Jiawei Zhang, Li Wang, Yifei Zeng, Xun Cao, Xinxin Zuo, and Hao Zhu. Tera: Rethinking text-guided realistic 3d avatar generation. In *ICCV*, pages 10686–10697, 2025. 3

[91] Yue Wu, Sicheng Xu, Jianfeng Xiang, Fangyun Wei, Qifeng Chen, Jiaolong Yang, and Xin Tong. Aniportraitgan: Animatable 3d portrait generation from 2d image collections. In *SIGGRAPH Asia*, 2023. 3

[92] Yiqian Wu, Hao Xu, Xiangjun Tang, Xien Chen, Siyu Tang, Zhebin Zhang, Chen Li, and Xiaogang Jin. Portrait3d: Text-guided high-quality 3d portrait generation using pyramid representation and gans prior. *ACM TOG*, 2024.

[93] Yiqian Wu, Malte Prinzler, Xiaogang Jin, and Siyu Tang. Text-based animatable 3d avatars with morphable model alignment. In *SIGGRAPH*, 2025.

[94] Yiqian Wu, Hao Xu, Xiangjun Tang, Yue Shangguan, Hongbo Fu, and Xiaogang Jin. 3dportraitgan: Learning one-quarter headshot 3d gans from a single-view portrait dataset with diverse body poses. *IEEE TCSVT*, pages 1–1, 2025. 3

[95] Zijian Wu, Boyao Zhou, Liangxiao Hu, Hongyu Liu, Yuan Sun, Xuan Wang, Xun Cao, Yujun Shen, and Hao Zhu. Uika: Fast universal head avatar from pose-free images. *arXiv preprint arXiv:2601.07603*, 2026. 3

[96] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *CVPR*, 2024. 3

[97] Jun Xiang, Yudong Guo, Leipeng Hu, Boyang Guo, Yancheng Yuan, and Juyong Zhang. Expressive talking human from single-image with imperfect priors. In *ICCV*, 2025. 3

[98] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *CVPRW*, 2022. 3, 6

[99] Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, et al. Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In *CVPR*, pages 19801–19811, 2024. 2

[100] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *CVPR*, pages 13296–13306, 2022. 3

[101] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *CVPR*, 2023. 3

[102] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *CVPR*, pages 1931–1941, 2024. 3

[103] Zhen Xu, Sida Peng, Chen Geng, Linzhan Mou, Zihan Yan, Jiaming Sun, Hujun Bao, and Xiaowei Zhou. Relightable and animatable neural avatar from sparse-view video. In *CVPR*, 2024. 3

[104] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *CVPR*, 2020. 2

[105] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, Wayne Wu, Chen Qian, Dahua Lin, Ziwei Liu, and Lei Yang. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In *ICCV*, pages 20282–20292, 2023. 3

[106] Houteng Yu, Hao Zhu, and Xun Cao. Realityavatar: Comprehensive head avatar generation with 360° rendering. In *ICME*, pages 1–6, 2025. 3

[107] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *CVPR*, 2021. 2

[108] Zhengming Yu, Tianye Li, Jingxiang Sun, Omer Shapira, Seonwook Park, Michael Stengel, Matthew Chan, Xin Li, Wenping Wang, Koki Nagano, and Shalini De Mello. GAIA: Generative animatable interactive avatars with expression-conditioned gaussians. In *SIGGRAPH Asia*, 2025. 3

[109] Bowen Zhang, Yiji Cheng, Chunyu Wang, Ting Zhang, Jiaolong Yang, Yansong Tang, Feng Zhao, Dong Chen, and Baining Guo. Rodinhd: High-fidelity 3d avatar generation with diffusion models. In *ECCV*, 2024. 3

[110] Dongbin Zhang, Yunfei Liu, Lijian Lin, Ye Zhu, Yang Li, Minghan Qin, Yu Li, and Haoqian Wang. Guava: Generalizable upper body 3d gaussian avatar. In *ICCV*, 2025. 3

[111] Jiawei Zhang, Zijian Wu, Zhiyang Liang, Yicheng Gong, Dongfang Hu, Yao Yao, Xun Cao, and Hao Zhu. Fate: Full-head gaussian avatar with textural editing from monocular video. In *CVPR*, pages 5535–5545, 2025. 3

[112] Xuanmeng Zhang, Jianfeng Zhang, Chacko Rohan, Hongyi Xu, Guoxian Song, Yi Yang, and Jiashi Feng. Getavatar: Generative textured meshes for animatable human avatars. In *ICCV*, 2023. 3

[113] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *CVPR*, pages 9936–9947, 2024. 3

[114] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M Avatar: Implicit morphable head avatars from videos. In *CVPR*, 2022. 3

[115] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *CVPR*, 2023. 3

[116] Zhenglin Zhou, Fan Ma, Hehe Fan, Zongxin Yang, and Yi Yang. Headstudio: Text to animatable head avatars with 3d gaussian splatting. In *ECCV*, 2024. 3

[117] Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3d human creation from a single image. In *CVPR*, pages 26308–26319, 2025. 3, 4, 5