Identity-Preserving Human Reconstruction from a Single Image via 3D Token Inference

Yanqi Bao[1,2], Jiaxiang Shang[3], Yang Gao[1], Yingchun Liu[3], Jing Huo[1⋆], and Jing Liao[2⋆]

[1] Key Laboratory of New Software Technology, Nanjing University, Nanjing, China
[2] City University of Hong Kong, China
[3] Central Media Technology Institute, Huawei, China
yanqibao1997@gmail.com, huojing@nju.edu.cn

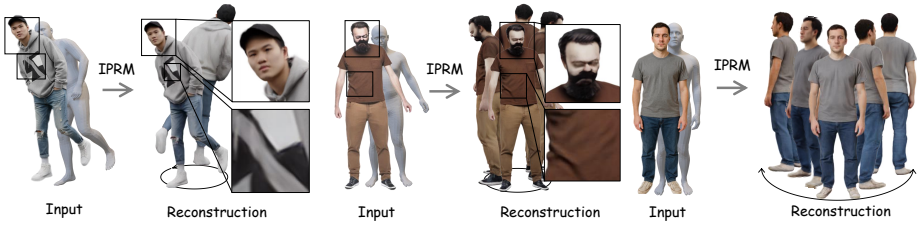

Fig. 1: Identity-Preserving Large Human Reconstruction Model (IPRM), a feed-forward framework that reconstructs photorealistic, clothed 3D humans from a single in-the-wild image while preserving **3D consistency and identity consistency**.

Abstract. We present the Identity-Preserving Large Human Reconstruction Model (IPRM), a feed-forward framework that reconstructs photorealistic, clothed 3D humans from a single in-the-wild image while preserving 3D identity. Recent works predominantly infer 3D structure from 2D features, making it challenging to achieve 3D consistency and preserve human identity in 3D space. To alleviate these challenges, IPRM anchors the single-view 3D human reconstruction by constructing a human-based 3D feature space and explicitly preserves the human 3D identity features during inference. Specifically, we introduce an efficient and robust SMPL-based sparse voxel representation to transform 2D image identity features into 3D space, categorizing them into 3D visible identity tokens and 3D invisible tokens to be inferred. Using these 3D tokens, an identity-aware 3D token inference module is proposed to propagate projected 3D identity features from visible to invisible tokens, ensuring that only unobserved regions are predicted while the observed identity remains intact. We further design an encoder–decoder architecture that decodes SMPL-based 3D features into either a 3D Gaussian Splatting or mesh representation, and augment it with a 3D ID Adapter for identity preservation. Instead of conventional conditioning on image tokens (2D ID Adapter), this adapter utilizes 3D identity tokens extracted from a parallel identity branch as guidance to inject token-wise identity information. Comprehensive experiments on existing benchmarks and in-the-wild data show IPRM surpasses state-of-the-art methods in reconstruction performance, efficiency, 3D consistency and identity consistency.

⋆ Corresponding author.

Keywords: Large Human Reconstruction Model · Feed-forward Framework · Identity-aware

1 Introduction

The pursuit of photorealistic 3D clothed human models represents a rapidly evolving field with substantial commercial impact across gaming, film production, fashion design, and AR/VR applications. Although dense-view capture systems with scene-specific reconstruction models demonstrate success in achieving this goal [16], their hardware requirements and substantial computational cost impede widespread deployment. Therefore, a user-friendly system is needed that can reconstruct 3D humans from any in-the-wild image in a feed-forward manner.

This task is challenging, where the core target lies in how to **infer the complete 3D representation from monocular input while preserving human identity** [22]. Recent approaches employ multiview or video-based generation to enhance 2D view completeness [12,22,26,33], but struggle to maintain 3D consistency, leading to artifacts such as severed fingers. Subsequent work [28] addresses this by directly using 2D image feature tokens as condition to update 3D geometric tokens sampled on parametric body models (SMPL) [24] (referred to as 2D-condition Token Inference). However, this fails to ensure explicit alignment between inferred 3D identity features and 2D condition features, as Fig. 2.

Therefore, we propose the Identity-Preserving Large Human Reconstruction Model (IPRM), a novel framework that anchors the single-view human reconstruction process in a human-based identity-aligned 3D SMPL feature space. In contrast to related work [45, 51], this space is specifically designed to explicitly preserve human 3D identity in a global, transformer-based 3D token inference paradigm. However, implementing this framework presents several significant technical challenges: 1) Sampling point-tokens from the coarse SMPL model [26, 28] struggles to accurately project 2D identity features onto these visible 3D points, while the large number of points compromises efficiency. 2) Methods that directly update all 3D token features conditioned on 2D image tokens [28] struggle to preserve the projected 3D identity features. 3) The decoding process from the SMPL-based feature space to 3D representations is often underexplored, despite its potential to cause identity drift and compromise the fidelity of 3D identity features in SMPL-based feature space.

To address these challenges, 1) We introduce a **sparse voxel representation in SMPL space** for 3D human modeling, which is more efficient compared to SMPL-point sampling and exhibits greater robustness for SMPL-induced feature projection errors. 2) We design an **identity-aware 3D token inference module**, which includes multiple visibility-based cross-attention blocks to infer invisible tokens from visible identity tokens only, and an adaptive 3D Human Feature to align entire 3D tokens to the human-specific knowledge. 3) IPRM introduces a **3D ID Adapter** based on a parallel identity branch to maintain the consistency of identity-related tokens at the token level during the repre-

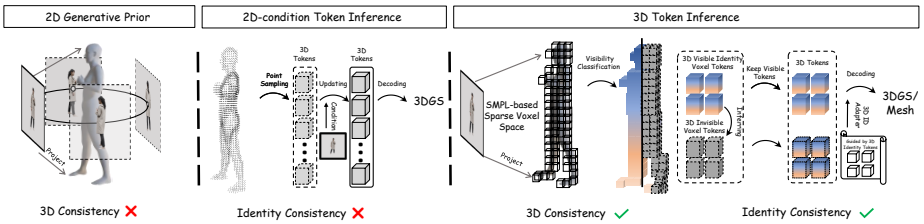

Fig. 2: Existing methods based on 2D Generative Priors or 2D-condition Token Inference often over-rely on 2D feature space and fail to maintain 3D consistency and explicitly align 3D human identity with the 2D input. To address these, **IPRM** directly achieves 3D Token Inference from 3D visible tokens to 3D invisible tokens and explicitly preserves consistency on 3D identity-related tokens during the entire process.

sentation decoding process. Specifically, IPRM consists of two stages. In the first stage, we leverage the SMPL as 3D spatial prior to sample visible (identity) and invisible sparse voxels, and project image features onto these voxels to obtain the corresponding tokens. Based on these tokens, the identity-aware 3D token inference module reasons invisible tokens using visibility-based cross-attention blocks, which can explicitly preserve 3D identity tokens unchanged in this process. Subsequently, they are further refined by the adaptive 3D Human Feature, which includes a visible Human Feature derived from the input image and an invisible Human Feature predicted from the inferred 3D invisible tokens to represent the entire 3D human structure. In the second stage, we train an encoder-decoder architecture to decode these SMPL-based token-features, supporting both Gaussian Splatting (3DGS) [3, 18] and mesh representations. To prevent identity drift during this process, we design a parallel identity branch to predict the 3D identity condition tokens from the given image, which provides token-wise guidance for each self-attention in the encoder and decoder.

In summary, IPRM is a novel 3D human reconstruction framework capable of achieving fast and high-fidelity human reconstruction from single-image features within **one second**, while utilizing minimal GPU memory. Quantitative and qualitative experiments on extensive benchmarks and in-the-wild data demonstrate the superiority of IPRM, particularly in preserving identity features. Furthermore, IPRM can be directly extended to multi-view input settings and animated by SMPL. Our main contributions are as follows:

- IPRM establishes a novel framework for reconstructing humans via 3D token inference based on SMPL-based 3D sparse voxel representation while preserving 3D identity features explicitly.
- We design an identity-aware 3D token inference module, which includes visibility-based cross-attention blocks to maintain identity features consistency during the 3D inference process, and an adaptive 3D Human Feature for further refinement with human-specific knowledge.
- We design a 3D ID Adapter as a critical 3D guidance module to mitigate identity drift throughout the encoder-decoder process, which, compared to conventional 2D ID Adapters, significantly enhances identity consistency at the 3D token level.

– IPRM achieves efficient inference of 3D human representations from image features in approximately 0.6 seconds. Qualitative and quantitative evaluations validate the framework's superiority over existing methods and demonstrate the effectiveness of its components for identity preservation and 3D consistency.

2 Related Work

2.1 Diffusion-based Human Reconstruction

Reconstructing 3D humans from a single view is inherently an ill-posed problem. Therefore, a straightforward idea is to introduce 2D diffusion models for providing additional priors [38,40]. Early work [2,37,41,49] attempts to leverage Score Distillation Sampling (SDS) [15,36] as auxiliary supervision to optimize 3D structures from pre-trained 2D diffusion models, but these approaches often require time-consuming optimization. Later, leveraging advances in diffusion models, SiTH [12] addresses missing 3D information by dividing the task into generative hallucination and reconstruction, using diffusion models to hallucinate unseen back-view appearances from the input image. Building on this idea, subsequent work adopts a similar generation-reconstruction paradigm [5, 6, 33, 48]. Among these work, [1] focuses on preserving the shape and structural details of the underlying 3D representation, while PSHuman [22] introduces an explicit iterative human carving approach to achieve similar goals. In addition to multi-view generation models, HumanSplat [26] utilizes pre-trained video generation models to complete continuous 3D information, while DiffHuman [33] extends image generation to different domains such as normal and depth maps. Although these 2D diffusion model-based methods have achieved great success in recent years, 2D-space generation struggles to align with 3D structures, inevitably leading to 3D inconsistencies. Additionally, the multi-step nature of diffusion models inevitably sacrifices efficiency. To address these issues, IPRM introduces a novel framework that directly infers invisible features in a SMPL-based 3D feature space without diffusion processes and maintains token-wise identity consistency in 3D space by an additional 3D ID Adapter.

2.2 Large Human Reconstruction Model

Single-image 3D human reconstruction is traditionally tackled by directly regressing the parameters of a parametric body model [17, 21, 32] or by learning implicit pixel-aligned surfaces that capture clothing details [30, 31, 44, 45]. However, such methods suffer from poor geometric quality and lack explicit representations. The development of Large Reconstruction Models (LRMs) [4, 13, 35] has revitalized this field by enabling generalizable feed-forward 3D target reconstruction from a single image or a small number of images. In human reconstruction, several work adopts this paradigm to replace traditional reconstruction methods [52]. An early work, Human-LRM [42] uses triplane features for coarse reconstruction and improves representation with diffusion models. Human-Splat [26] utilizes generated multi-view image features to implicitly update SMPL-based human geometric point-token features and decodes them into

3DGS representations. Similarly, [5] also adopts this paradigm and introduces additional optimization for generating back-view images. In contrast, LHM [28] and a derivative work [29] abandon the multi-view generation process, propose a video-based training paradigm, and introduce a multimodal body-head transformer to further enhance the reconstruction quality of head tokens.

Although these methods achieve better 3D consistency compared to 2D-diffusion-based approaches [23,39,47], they suffer from two key limitations. First, direct update on all 3D token features based on the 2D image condition does not establish explicit correspondence between input image and 3D tokens, leading to identity drift in 3D space, Fig. 5. Second, existing methods are based on sampling numerous feature points from SMPL [28], which not only compromises computational efficiency but also struggles to handle cases involving loose clothing that deviates from the SMPL topology. To address these challenges, IPRM adopts an SMPL-based sparse voxel representation as the foundational 3D structure. This representation effectively captures the 3D neighborhood features of SMPL. Combined with the dedicatedly trained encoder-decoder, this sparse representation enables transformation to realistic 3DGS or mesh representations, improving robustness on incorrect SMPL estimation. Based on this representation, unlike methods that rely on 2D-condition token inference, we develop a novel identity-aware 3D token inference paradigm to directly infer invisible 3D voxel tokens from visible identity tokens while explicitly preserving 3D identity features.

3 Method

3.1 Human Parametric Model

SMPL [24] and its variant SMPL-X [27] serve as essential parametric body model priors in human reconstruction tasks, establishing a coarse 3D geometry that facilitates subsequent reconstruction processes. In this work, we utilize SMPL-X to establish SMPL-based sparse voxel representations. The SMPL-X model employs shape parameters and pose parameters to represent human body configurations, which can be directly converted into mesh to estimate the occupied voxels and the visibility mask in IPRM.

3.2 IPRM Framework

Given an in-the-wild human image $\mathbf{I}$ and the corresponding SMPL, our goal is to reconstruct a 3D-consistent and photorealistic human representation in one second while explicitly preserving identity with the input image. To achieve this, we propose IPRM, a feed-forward framework that directly infers 3DGS and mesh representations aligned with the input image with low computational overhead. The entire pipeline is illustrated in Fig. 3.

Previous 2D-condition Token Inference methods [26, 28] rely on sampling points directly on the SMPL surface. This approach involves extensively sampling points and relies heavily on the coarse SMPL prior, leading to inefficiencies

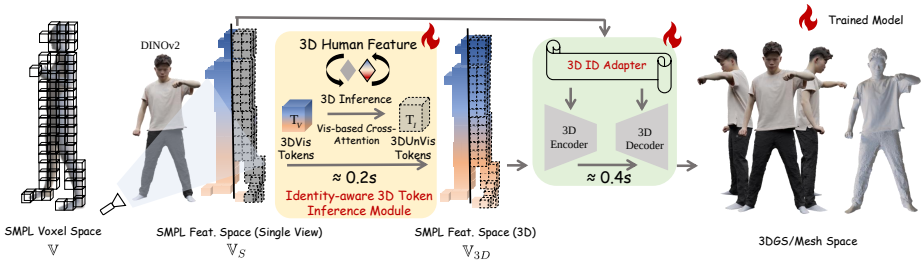

Fig. 3: Overview of IPRM's Inference Process: IPRM consists of two stages. In the first stage, after constructing an SMPL-based sparse voxel space $\mathbb{V}$ and projecting image features into it $\mathbb{V}_S$, IPRM classifies all voxels as visible ($\mathbf{T}_V$) or invisible ($\mathbf{T}_I$). Then IPRM leverages **Identity-aware 3D Token Inference Module** to infer invisible tokens based on visible voxel tokens, obtaining complete 3D features $\mathbb{V}_{3D}$. In the second stage, these SMPL-based 3D features are processed through a **3D ID Adapter** and encoder-decoder architecture to achieve identity-preserving 3DGS/mesh.

and challenges in achieving accurate pixel-aligned feature projection due to inaccuracies in SMPL geometry. Inspired by Trellis [43], IPRM adopts SMPL-based sparse voxels as the foundational representation. By operating on active voxels rather than individual surface points, this method significantly improves efficiency and mitigates the impact of 2D-3D projection errors. Combined with the specifically trained encoder-decoder, it enables conversion from the SMPL feature space to 3DGS and mesh representations, enhancing robustness for SMPL.

Based on this representation, IPRM consists of two stages. In the first stage, we propose an identity-aware 3D token inference module. It uses projected visible voxel tokens (hereafter called identity tokens) to infer invisible 3D tokens while preserving identity consistency in the 3D space, resulting in the complete 3D SMPL-based human feature space. In the second stage, based on such 3D human feature space, IPRM introduces a 3D ID Adapter in the representation encoder-decoding process, which utilizes 3D identity features from a dedicated identity branch as guidance to mitigate identity drift in a token-by-token manner.

3.3 SMPL-based Sparse Voxel Tokens

Sparse voxels are an efficient 3D representation designed to coarsely capture the occupied regions of a 3D space surrounding an object [9, 25]. Unlike Trellis [43], which directly defines in real geometric space, our sparse voxels are adapted to the SMPL space by identifying active voxels based on whether they are occupied by the SMPL geometry. Instead of sampling exact points on the SMPL surface, this voxel-based representation captures the 3D space surrounding SMPL and outlines the human's coarse geometry on the 3D grid [43], while being less sensitive to feature projection inaccuracies on the SMPL surface, especially for loose clothing. Correspondingly, to enable high-quality conversion from the SMPL space to 3DGS and mesh, a dedicated encoder-decoder structure is trained, further enhancing robustness against projection error. The effectiveness of this representation is discussed in Sec. 4.3 and Fig. 6(a).

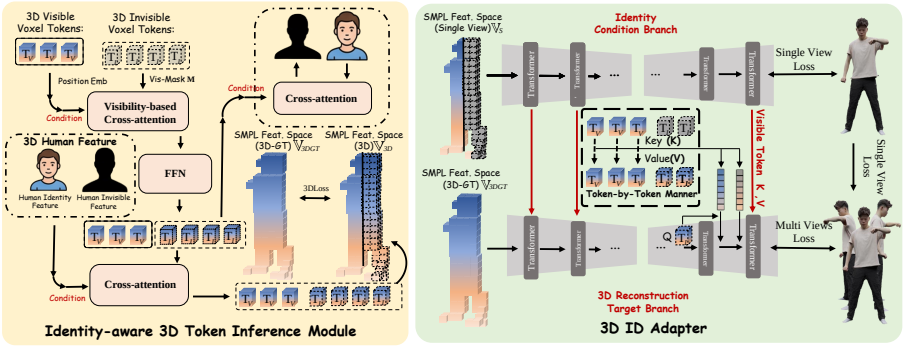

Fig. 4: The training process of the Identity-aware 3D Token Inference Module and 3D ID Adapter. Given voxel tokens, the **identity-aware 3D token inference module** uses multiple visibility-based cross-attention blocks and an adaptive 3D Human Feature to update invisible voxel tokens and refine all 3D tokens with human-specific knowledge, respectively. When decoding these features into 3DGS/mesh, the **3D ID Adapter** introduces an additional identity branch to predict 3D identity features, which serve as guidance to preserve the identity of 3D representations for the target branch at the token level.

Based on the SMPL-based sparse voxel $\mathbb{V}$ and input image features $\mathbf{F} = \text{DINOv2}(\mathbf{I})$, we project $\mathbf{F}$ into the 3D voxel space and serialize these projections as tokens $\mathbf{T} = (z_i, f_i)_{i=0}^{N-1}$, where $z_i \in \{0, 1, \ldots, n-1\}^3$ is the positional index of the active voxels and $f_i \in \mathbb{R}^C$ represents the projected feature from $\mathbf{F}$. We refer to it as the SMPL feature space (single view), denoted as $\mathbb{V}_S$. Subsequently, IPRM determines token visibility through a two-step process. First, it performs back-face culling by computing the angular relationship between the SMPL-based voxel normals (*Norm*) and the ray from the camera origin to the voxel center. Second, to mitigate occlusion issues, it optionally traverses all intermediate voxels along each ray path; if any occupied voxel is encountered before reaching the target, the target voxel is marked as invisible. Based on these two criteria, tokens are classified into visible tokens $\mathbf{T}_V$ (i.e., identity tokens) and invisible tokens $\mathbf{T}_I$ (i.e., tokens to be inferred), and a visibility mask $\mathbf{M}$ is obtained according to the token positional indices z_i.

3.4 Identity-aware 3D Token Inference Module

Given $\mathbf{T}_V$, IPRM aims to infer $\mathbf{T}_I$ while preserving the integrity and consistency of the known identity features $\mathbf{T}_V$. To address this objective, we introduce the identity-aware 3D token inference module, which incorporates a visibility-based cross-attention mechanism with adaptive 3D Human Feature guidance. It enables effective transformation from $\mathbb{V}_S$ into the SMPL feature space (3D), called $\mathbb{V}_{3D}$.

To pre-process $\mathbf{T}_I$ and $\mathbf{T}_V$, IPRM first incorporates sinusoidal positional encodings based on active voxel positions and concatenates them token-wise. The resulting tokens are subsequently passed through cross-transformer blocks. Specifically, given the visibility mask $\mathbf{M}$, visibility-based cross-attention blocks are used to infer $\mathbf{T}_I$ conditioned on $\mathbf{T}_V$. Compared to conventional self-attention [28],

this effectively maintains consistency in visible tokens. This process enables reasoning from known 3D regions to unknown regions, but this approach struggles to align with human-specific knowledge and maintain 3D continuity, particularly at the visible-invisible tokens boundaries. Therefore, we introduce an additional human-specific guidance, termed 3D Human Feature, to further refine them.

Prior approaches [28] typically depend on human-specific 2D image encoders [19], leveraging a 2D single-view human image feature as condition, referred to as visible Human Feature $\mathcal{P}_V$, to accomplish this goal. However, relying exclusively on this 2D feature to guide the updating of all tokens $\mathbf{T}$ proves insufficient for 3D inference, as 2D features from a single viewpoint inherently fail to align with the complete 3D spatial representation. Therefore, we propose an adaptive 3D Human Feature, which, based on $\mathcal{P}_V$, introduces an additional invisible Human Feature $\mathcal{P}_I$ representing the invisible 3D regions that are not captured from the given viewpoint. Subsequently, the IPRM leverages the concatenation ($||$) of visible $\mathcal{P}_V$ and invisible Human Feature $\mathcal{P}_I$ (as 3D Human Feature) as a condition to collectively guide the updating of all 3D tokens. To obtain $\mathcal{P}_I$, we use inferred $\mathbf{T}_I$ (after visibility-based cross-attention blocks) as condition to update original $\mathcal{P}_V$, which is achieved through the cross-attention blocks, enabling the Human Feature to transition from visible to invisible regions, as shown in Fig. 4.

$$\begin{aligned} \mathcal{P}_I &= \text{Cross-Attention}(\mathcal{P}_V; \mathbf{T}_I), \\ \mathbf{T} &= \text{Cross-Attention}(\mathbf{T}_V||\mathbf{T}_I; \mathcal{P}_V||\mathcal{P}_I). \end{aligned} \tag{1}$$

3.5 3D ID Adapter

After inferring invisible 3D tokens from SMPL-based single-view features, IPRM decodes all 3D voxel features $\mathbb{V}_{3D}$ into 3DGS or mesh representations. Existing work [28, 43] has addressed this process by an encoder-decoder architecture, but often overlooked potential identity drift. This drift stems from the ambiguity in directly decoding 3D tokens defined in SMPL space into realistic 3D representation space. To mitigate this, a naive approach incorporates the 2D image condition to guide 3D token (ID Adapter [7]) decoding by cross-attention layers.

We explore the integration of human features [19] and face features [8] as 2D conditioning inputs in the cross-attention layers to guide the encoder-decoder [28]. However, experiments show that this approach yields only modest improvements in Tab. 4 and Fig. 6 (b), likely due to the difficulty of establishing explicit correspondence between 2D image tokens and 3D tokens, as discussed above. To address this, IPRM introduces 3D identity condition tokens, which directly guide 3D token updating in a token-by-token manner and achieve better alignment within the same feature domain. To obtain these 3D identity condition tokens, we design an additional identity branch that operates in parallel with the target branch. Unlike the target (reconstruction) branch, which takes $\mathbb{V}_{3D}$ ($\mathbb{V}_{3DGT}$) as input and focuses on reconstructing 3D representations from SMPL-based features, the identity branch processes $\mathbb{V}_S$ as input to reconstruct only the input identity image using visible Gaussian primitives.

By optimizing the identity branch using only an identity image, this branch acquires the capability to infer 3D identity condition tokens from a 2D input image, which serve as guidance for the target branch. Motivated by the design of ReferenceNet [14], we replace the keys and values in each self-attention layer of the target branch with the corresponding keys and values from the identity branch, while preserving the query from the target branch. Additionally, since the identity branch primarily focuses on visible feature extraction, this 3D token-wise guidance is applied exclusively to the visible identity tokens, while the inferred invisible tokens retain the original keys and values from the target branch.

3.6 Training Strategy

IPRM adopts a two-stage training paradigm to separately train the identity-aware 3D token inference module and the encoder-decoder structure with the 3D ID Adapter, as shown in Fig. 4.

In the first stage, we take a single image and SMPL as inputs to predict SMPL-based 3D voxel features $\mathbb{V}_{3D}$ from $\mathbb{V}_S$. We supervise this process using 3D MSE loss against 3D-GT voxel features. To obtain the 3D-GT voxel features $\mathbb{V}_{3DGT}$ for each human, we project and aggregate the DINO features $\mathbf{F}$ from all viewpoint images into the 3D SMPL voxel space. In the second stage, we fine-tune the pre-trained encoder-decoder [43] to decode $\mathbb{V}_{3DGT}$ directly into 3DGS or mesh representations, without variational sampling. We also introduce a parallel identity branch with the 3D ID Adapter to preserve identity features. We separately train the 3DGS and mesh decoders. For 3DGS, the decoder maps 3D voxel features into Gaussian parameters (position offsets, colors, scales, opacities, and rotations), optimized with L_1, D-SSIM, and LPIPS losses. For meshes, 3D features are transformed into signed distance values at voxel vertices, from which outputs are extracted via isosurfaces and optimized using L_1 loss on depth and normal maps. The identity branch is supervised using only the input view. To ensure consistency, IPRM introduces an additional single-view loss on the identity view between the two branches. All losses are summed for optimizing.

4 Experiments

4.1 Experimental Details

Dataset and Evaluation Metrics: We employ Human4Dit [34], THuman2.1 [46], CustomHumans [11], 2K2K [10], and in-the-wild data for training and evaluating IPRM. To ensure fairness, we follow the evaluation benchmarks of LHM [28] and PSHuman [22] by using the same test cases and viewpoints for assessment while ensuring these test cases are excluded from training. To quantitatively evaluate the performance of 3DGS rendering, we use traditional metrics such as PSNR, SSIM, and LPIPS for 3D human reconstruction quality. Additionally, we introduce Input-view PSNR, which computes PSNR exclusively on the input identity viewpoint, referred to as PSNR (I), and Face Consistency [28] to evaluate IPRM's identity preservation. For mesh reconstruction quality, we employ three metrics: one-directional point-to-surface (P2S), L1 Chamfer Distance (CD), and Normal Consistency (NC) [22].

Table 1: Comparisons of methods on two benchmarks [22, 28], Synthetic Data and THuman2.1 for 3DGS, with inference time and memory usage.

Methods	Synthetic Data					THuman2.1			Inference	
	PSNR↑	SSIM↑	LPIPS↓	PSNR(I)↑	FC↓	PSNR↑	SSIM↑	LPIPS↓	Time	Memory
GTA (NeurIPS 2023)	17.03	0.919	0.09	17.66	0.051	19.61	0.834	0.10	0.68s	≈8GB
SIFu (CVPR 2024)	16.68	0.917	0.09	19.22	0.060	19.44	0.831	0.10	0.65s	≈9GB
DreamGaussian (ICLR 2024)	18.54	0.917	0.08	19.47	0.056	-	-	-	2 min	≈8GB
SITH (CVPR 2024)	-	-	-	-	-	18.46	0.820	0.10	45.12s	≈21GB
PSHuman (CVPR 2025)	17.56	0.921	0.08	21.44	0.037	20.85	0.864	0.08	1 min	≈40GB
Trellis* (CVPR 2025)	21.67	0.921	0.05	21.99	0.058	21.33	0.886	0.06	4.20s	≈11GB
LHM-0.5B (ICCV 2025)	25.18	0.951	**0.03**	26.64	0.035	-	-	-	2.01s	≈18GB
Ours	**27.04**	**0.954**	**0.03**	**28.90**	**0.033**	**26.74**	**0.948**	**0.05**	**0.63s**	≈9GB

Table 2: Quantitative comparison for mesh reconstruction [22].

Method	Opt	THuman2.1		
		CD↓	P2S↓	NC↑
ICON	×	0.6146	0.5934	0.8493
ECON	×	0.6725	0.6331	0.8362
GTA	×	0.5791	0.5587	0.8491
SIFU	×	0.5754	0.5576	0.8500
HiLo	×	0.5977	0.5892	0.8405
SITH	×	0.6474	0.5810	0.8264
PSHuman	✓	**0.4399**	**0.4077**	0.8504
Ours	×	0.4458	0.4331	**0.8504**

Table 3: Comparison on CustomHumans Dataset [28].

Method	CustomHumans			
	PSNR↑	SSIM↑	LPIPS↓	FC↓
SiTH	27.07	0.938	0.07	-
LHM-0.5B	28.31	0.953	0.05	0.039
Ours	**29.33**	**0.971**	**0.04**	**0.035**

Table 4: Ablation study of the core modules.

Method	Synthetic Data		
	PSNR(I)	SSIM	LPIPS
SMPL-based Sparse Voxel			
w/o Feature Project	17.01	0.911	0.10
Voxel n=64 (128 by default)	27.36	0.930	0.05
Identity-aware 3D Token Inference Module			
w/o Visibility-based Cross-attention [Self-Att]	27.87	0.952	0.04
w/o Adaptive 3D Human Feature	28.54	0.953	0.03
3D ID Adapter			
w/o Condition	25.80	0.950	0.04
w 2D Image Condition	26.42	0.952	0.04
Ours	**28.90**	**0.954**	**0.03**

Table 5: Application on more inputs.

Method	THuman2.1		
	PSNR↑	SSIM↑	LPIPS↓
Single View (Front image)	26.74	0.948	0.05
Two Views (+Back Image)	27.31	0.952	0.04
All Views	**27.69**	**0.958**	**0.03**

Preprocessing Pipeline: Given an in-the-wild image, we first utilize Segment Anything [20] to segment the foreground human. Subsequently, we generate a square and centralized image and estimate SMPL-X and camera pose following SiTH [12] for reconstruction.

Training and Inference Details: We reproject the multi-view rendered image features back to the 3D SMPL voxel space and average them within each SMPL-based voxel [43] to obtain the 3D-GT SMPL feature space $\mathbb{V}_{3DGT}$ for training the identity-aware 3D token inference module. Subsequently, along with the rendered multi-view images, we train the 3DGS-based encoder-decoder architecture and 3D ID Adapter, and utilize the GT human mesh along with its corresponding normals to train the mesh-related decoder. Additionally, in comparative experiments, we use the same method to process human data for fine-tuning Trellis [43], including its diffusion model and encoder-decoder architecture. During inference, **we do not use** $\mathbb{V}_{3DGT}$. The IPRM can directly construct $\mathbb{V}_S$ from a single in-the-wild image and its corresponding SMPL, which is used to infer invisible tokens by the identity-aware 3D token inference module, obtaining $\mathbb{V}_{3D}$. Finally, $\mathbb{V}_{3D}$ can be decoded into 3DGS or mesh representations.

Training Configuration: IPRM adopts a two-stage training paradigm and sets the voxel size n to **128** in all experiments. Specifically, the identity-aware 3D token inference module is trained from scratch with an initial learning rate of 2×10^{-3} over 200,000 iterations. In the second stage, we fine-tune the pre-trained

Fig. 5: Qualitative comparison under 3DGS representation on Internet data. We input a single **in-the-wild image** (the input after SAM and centralization) and compare with recent SOTA work. Our method demonstrates better *identity consistency* in facial and hand regions in rows 1-4, as well as better *3D consistency* across different viewpoints (side view) in rows 5-6. We zoom in on some details in the red&black boxes.

encoder-decoder architecture [43], where we introduce an identity branch with the same pre-trained initialization. We optimize them with an initial learning rate of 5×10^{-4} over 100,000 iterations.

4.2 Comparative Experiments

We compare against three types of recent methods: 1) methods leveraging 2D generative priors [22], 2) 2D-condition token inference methods [28], and 3) general 3D generation models fine-tuned on human data [43]. For fair comparison, all methods are evaluated on same test cases using consistent views.

Quantitative Evaluation: As evidenced by the experimental results in Tab. 1 and Tab. 3, our proposed IPRM achieves substantial improvements across all evaluation metrics compared to existing state-of-the-art methods. These performance gains can be primarily attributed to the framework's ability to perform 3D token inference directly in 3D space, enabling superior 3D consistency, particularly when compared to approaches based on 2D generative prior. The PSNR(I) and FC metrics show that IPRM better preserves human identity compared

to LHM [28]. This improvement stems from our SMPL-based sparse voxel and identity-aware 3D token inference module, which achieves alignment between 3D tokens and 2D input features while explicitly preserving identity-related tokens invariant. Compared to Trellis [43] fine-tuned on human datasets, IPRM achieves greater consistency with the original image and excels in capturing fine-grained details, such as facial features, owing to the integration of SMPL priors and the 3D ID Adapter. Furthermore, as shown in Tab. 2, IPRM significantly outperforms most existing methods in mesh reconstruction, delivering results comparable to PSHuman [22]. This achievement is particularly notable, as PSHuman relies on iterative optimization for each individual case, whereas IPRM operates within a feed-forward paradigm, offering both efficiency and effectiveness.

Efficiency and Memory Usage Compared to existing algorithms, IPRM achieves superior computational efficiency, with faster inference speed and reduced memory usage, as shown in Tab. 1. The first stage of IPRM takes approximately 0.2s, which includes the identity-aware 3D token inference module, while the second stage takes approximately 0.4s and includes the encoder-decoder and 3D ID Adapter, as shown in Fig. 3. This efficiency is primarily attributed to avoiding the iterative denoising process in diffusion models and adopting sparse voxel representation.

Qualitative Evaluation Consistent with the quantitative results, IPRM also demonstrates remarkable performance in the qualitative evaluations on *wild data* presented in Fig. 5. We design the qualitative analysis from two perspectives: Identity Consistency and 3D Consistency. For **Identity Consistency**, IPRM achieves identity preservation with the input image in both geometry and appearance. Compared to existing methods based on 2D generative priors, such as PSHuman [22], IPRM demonstrates better 3D geometric consistency, particularly in the hand and facial regions, as shown in the hat regions in row 2 and the head and hand regions in rows 3-4. Similarly, 2D-condition token inference approaches, such as LHM [28], exhibit noticeable inconsistencies in preserving input pose and identity, as shown in the facial regions in rows 1-4 and the hand regions in row 3. In contrast, IPRM effectively addresses both challenges within a unified framework, maintaining identity fidelity through the proposed identity-aware 3D token inference module and 3D ID Adapter. Moreover, compared to fine-tuned 3D generative models [43], IPRM achieves superior input consistency, particularly in high-detail regions such as facial features and clothing, while Trellis [43] suffers from inconsistency due to its stochastic sampling nature.

For **3D Consistency**, IPRM also achieves better 3D consistency, especially for side views. Specifically, existing methods such as PSHuman [22] and LHM [28] exhibit severe artifacts, such as discontinuous appearances and distorted geometry, as shown in the side regions of the arms and face in rows 5-6. In contrast, IPRM maintains multi-view consistency and significantly reduces artifacts while preserving identity consistency, which benefits from our proposed 3D token inference paradigm. More cases are in the *supplementary materials and videos*.

As illustrated in Fig. 9, we provide a qualitative comparison of IPRM for mesh reconstruction. While PSHuman [22] relies on iterative optimization for each

Fig. 6: Ablation on Loose Clothing. Compared to SMPL-point-based feature sampling [28], Sparse Voxel demonstrates better robustness. Compared to 2D attention (ID Adapter [7]), the 3D ID Adapter exhibits better identity consistency.

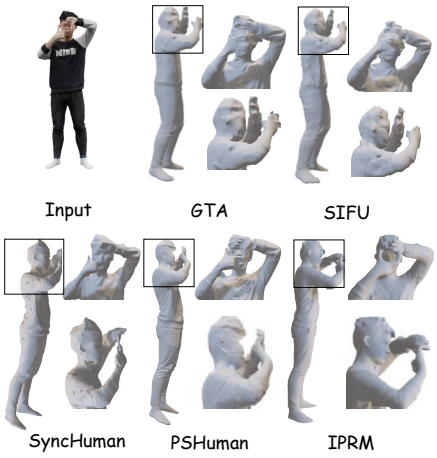

Fig. 7: Qualitative comparison on the test set. We compare with recent SOTA work on unseen cases. The results further validate that IPRM better preserves identity consistency, such as in the hand and facial regions highlighted by the boxes.

individual case, IPRM achieves more complete reconstruction results (e.g., facial and hand regions) through a more efficient feed-forward paradigm, which benefits from the 3D consistency of IPRM. Furthermore, compared to GTA [50] and SyncHuman [6], IPRM produces smoother surfaces and more precise geometry, demonstrating its robustness in maintaining 3D consistency.

4.3 Ablation Study

We analyze the functional mechanisms of the SMPL-based sparse voxel, identity-aware 3D token inference module, and 3D ID Adapter in Tab. 4.

SMPL-based Sparse Voxel: Beyond its contribution to higher efficiency, we further validate the impact of the SMPL-based sparse voxel representation on overall performance enhancement, especially for **loose clothing** cases. Unlike existing approaches that directly sample point features on SMPL surfaces [28], the sparse voxel representation defines features in the 3D space surrounding the SMPL geometry. This design enhances robustness against inaccurate SMPL fitting, leading to more reliable reconstruction for loose clothing, as shown in Fig. 6. Moreover, projecting features into 3D space introduces significant improvements to IPRM in Tab. 4 ("w/o Feature Project" initializes $\mathbb{V}_S$ tokens only with positional encoding as LHM [28]), enabling identity-consistent reconstructions. We also compare the impact of different voxel sizes n on model performance. Tab. 4 shows that as the voxel resolution increases, IPRM's performance improves, but inference time increases accordingly, approximately 0.48s v.s. 0.63s.

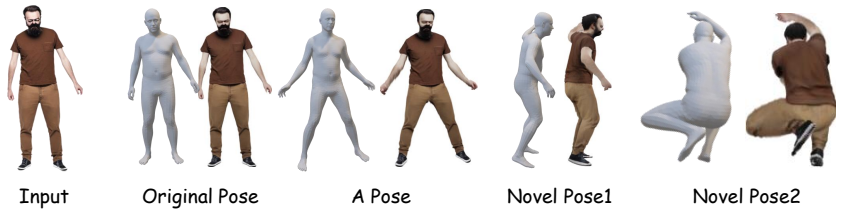

Fig. 8: Rendering under novel pose and view. IPRM can achieve rendering under novel poses by leveraging SMPL and linear blend skinning.

Identity-aware 3D Token Inference Module: In Tab. 4, the visibility-based cross-attention block proves to be critical for identity preservation compared to traditional self-attention layers [28], as it explicitly retains identity-relevant tokens when inferring invisible tokens. Furthermore, the adaptive 3D Human Feature leverages human-specific priors to globally refine 3D tokens, also yielding performance gains and proving indispensable.

3D ID Adapter: Experimental results demonstrate that utilizing 2D image as condition (conventional ID Adapter [7]) provides some benefit in maintaining identity features during the encoder-decoder process in Tab. 4; however, the impact remains limited. This highlights the challenge of establishing explicit correspondences between 2D image tokens and 3D inferred tokens. In contrast, our proposed 3D ID Adapter exhibits a significantly greater impact on performance, which is also demonstrated in Fig. 6(b). This improvement can be attributed to the identity branch effectively preserving 3D identity features and enhancing token-to-token correspondence through 3D identity tokens, which better align with the 3D token inference process.

Robustness to SMPL Fitting: Although utilizing SMPL-based framework enables reconstruction under novel poses, SMPL fitting is often inaccurate, especially in cases with loose clothing, which leads to performance degradation in SMPL-based methods [22, 28]. Compared to sampling points on SMPL [26, 28], the SMPL-based sparse voxel representation and corresponding 3D feature encoder-decoder architecture proposed by IPRM improve its robustness to SMPL fitting errors, as shown in the skirt and pants regions in Fig. 6(a).

4.4 Applications

More Inputs: IPRM inherently supports any number of input images and can even directly process inputs from all views. To further explore its potential, we extend IPRM to multi-input scenarios in Tab. 5. Experimental results demonstrate that as the number of visible tokens increases, the prediction performance improves significantly, especially when back-view images are introduced.

Novel Poses Reconstruction: Benefiting from the SMPL-based sparse voxel representation, IPRM enables novel pose reconstruction. As shown in Fig. 8, given an input image and corresponding SMPL, IPRM reconstructs the 3DGS representation, which can be deformed to novel poses by linear blend skinning.

Fig. 9: Failure cases in challenging scenarios, including difficult poses on a white background, and exaggerated clothing or complex props on a black background.

5 Conclusion

This paper presented IPRM, a feed-forward 3D human reconstruction framework from single image. By explicitly inferring invisible 3D regions from visible identity tokens within the SMPL-based sparse voxel space, IPRM achieves 3D-consistent and identity-preserving reconstruction. Specifically, we propose three key technical contributions—the SMPL-based sparse voxel representation, identity-aware 3D token inference module, and 3D ID Adapter—which are integrated to achieve this goal. Extensive experiments demonstrate that IPRM achieves state-of-the-art performance across multiple benchmarks, substantially surpassing existing methods in 3D consistency and identity consistency, while being significantly faster than prior work and requiring minimal memory.
Limitation: Although IPRM achieves state-of-the-art results, its reconstruction performance remains limited in challenging cases, such as difficult poses, exaggerated clothing, and complex props. This limitation may be attributed to the limited amount of training data and the constraints of human-specific pretrained models [19], as well as the fact that some clothing or props can severely exceed the SMPL body space. As shown in Fig. 9, even when the SMPL predictions are accurate, IPRM may still produce noticeable reconstruction errors in these out-of-body regions. Future work could alleviate these limitations by combining the SMPL pose prior with pretrained 3D generative priors to better model loose clothing and props. In addition, benefiting from IPRM's efficient inference, its reconstruction results can serve as initialization for per-case optimization.

Acknowledgements

We thank all anonymous reviewers and area chairs for their valuable comments. This work is supported in part by New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0122904), National Natural Science Foundation of China (62192783, 62276128, No. CRS_HKUST605/25), Jiangsu Natural Science Foundation (BK20243051), Jiangsu Science and Technology Major Project (BG2024031, BG2025035), the Fundamental Research Funds for the Central Universities (14380128, KG202514), the "111 Center"(B26023), a GRF grant from the Research Grants Council (RGC) of the Hong Kong Special Administrative Region, China (Project No. CityU 11208123), the NSFC/RGC Collaborative Research Scheme jointly sponsored by the Research Grants Council of the Hong Kong Special Administrative Region, China, the Central Media Technology Institute, Huawei (Project No. TC20240927042) and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

1. AlBahar, B., Saito, S., Tseng, H.Y., Kim, C., Kopf, J., Huang, J.B.: Single-image 3d human digitization with shape-guided diffusion. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–11 (2023)
2. Babiloni, F., Lattas, A., Deng, J., Zafeiriou, S.: Id-to-3d: Expressive id-guided 3d heads via score distillation sampling. Advances in Neural Information Processing Systems **37**, 97167–97183 (2024)
3. Bao, Y., Ding, T., Huo, J., Liu, Y., Li, Y., Li, W., Gao, Y., Luo, J.: 3d gaussian splatting: Survey, technologies, challenges, and opportunities. IEEE Transactions on Circuits and Systems for Video Technology **35**(7), 6832–6852 (2025)
4. Bao, Y., Liao, J., Huo, J., Gao, Y.: Distractor-free generalizable 3d gaussian splatting. arXiv preprint arXiv:2411.17605 (2024)
5. Chen, J., Li, C., Zhang, J., Zhu, L., Huang, B., Chen, H., Lee, G.H.: Generalizable human gaussians from single-view image. arXiv preprint arXiv:2406.06050 (2024)
6. Chen, W., Li, P., Zheng, W., Zhao, C., Li, M., Zhu, Y., Dou, Z., Wang, R., Liu, Y.: Synchuman: Synchronizing 2d and 3d generative models for single-view human reconstruction. arXiv preprint arXiv:2510.07723 (2025)
7. Cui, S., Guo, J., An, X., Deng, J., Zhao, Y., Wei, X., Feng, Z.: Idadapter: Learning mixed features for tuning-free personalization of text-to-image models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 950–959 (2024)
8. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
9. Gao, Q., Xu, Q., Su, H., Neumann, U., Xu, Z.: Strivec: Sparse tri-vector radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17569–17579 (2023)
10. Han, S.H., Park, M.G., Yoon, J.H., Kang, J.M., Park, Y.J., Jeon, H.G.: High-fidelity 3d human digitization from single 2k resolution images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12869–12879 (2023)
11. Ho, H.I., Xue, L., Song, J., Hilliges, O.: Learning locally editable virtual humans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21024–21035 (2023)
12. Ho, I., Song, J., Hilliges, O., et al.: Sith: Single-view textured human reconstruction with image-conditioned diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 538–549 (2024)
13. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. arXiv preprint arXiv:2311.04400 (2023)
14. Hu, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8153–8163 (2024)
15. Huang, Y., Yi, H., Xiu, Y., Liao, T., Tang, J., Cai, D., Thies, J.: Tech: Text-guided reconstruction of lifelike clothed humans. In: 2024 International Conference on 3D Vision (3DV). pp. 1531–1542. IEEE (2024)
16. Işık, M., Rünz, M., Georgopoulos, M., Khakhulin, T., Starck, J., Agapito, L., Nießner, M.: Humanrf: High-fidelity neural radiance fields for humans in motion. ACM Transactions on Graphics (TOG) **42**(4), 1–12 (2023)

17. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7122–7131 (2018)
18. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
19. Khirodkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models. In: European Conference on Computer Vision. pp. 206–228. Springer (2024)
20. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
21. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2252–2261 (2019)
22. Li, P., Zheng, W., Liu, Y., Yu, T., Li, Y., Qi, X., Chi, X., Xia, S., Cao, Y.P., Xue, W., et al.: Pshuman: Photorealistic single-image 3d human reconstruction using cross-scale multiview diffusion and explicit remeshing. arXiv preprint arXiv:2409.10141 (2024)
23. Li, P., Sun, Y., Cheng, H.: Pointdico: Contrastive 3d representation learning guided by diffusion models. In: 2025 International Joint Conference on Neural Networks (IJCNN). pp. 1–9. IEEE (2025)
24. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
25. Lu, T., Yu, M., Xu, L., Xiangli, Y., Wang, L., Lin, D., Dai, B.: Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20654–20664 (2024)
26. Pan, P., Su, Z., Lin, C., Fan, Z., Zhang, Y., Li, Z., Shen, T., Mu, Y., Liu, Y.: Humansplat: Generalizable single-image human gaussian splatting with structure priors. Advances in Neural Information Processing Systems **37**, 74383–74410 (2024)
27. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10975–10985 (2019)
28. Qiu, L., Gu, X., Li, P., Zuo, Q., Shen, W., Zhang, J., Qiu, K., Yuan, W., Chen, G., Dong, Z., et al.: Lhm: Large animatable human reconstruction model from a single image in seconds. arXiv preprint arXiv:2503.10625 (2025)
29. Qiu, L., Li, P., Zuo, Q., Gu, X., Dong, Y., Yuan, W., Zhu, S., Han, X., Chen, G., Dong, Z.: Pf-lhm: 3d animatable avatar reconstruction from pose-free articulated human images. arXiv preprint arXiv:2506.13766 (2025)
30. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2304–2314 (2019)
31. Saito, S., Simon, T., Saragih, J., Joo, H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 84–93 (2020)
32. Saleem, M.U., Pinyoanuntapong, E., Wang, P., Xue, H., Das, S., Chen, C.: Genhmr: Generative human mesh recovery. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6749–6757 (2025)

33. Sengupta, A., Alldieck, T., Kolotouros, N., Corona, E., Zanfir, A., Sminchisescu, C.: Diffhuman: Probabilistic photorealistic 3d reconstruction of humans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1439–1449 (2024)
34. Shao, R., Pang, Y., Zheng, Z., Sun, J., Liu, Y.: Human4dit: Free-view human video generation with 4d diffusion transformer. arXiv e-prints pp. arXiv–2405 (2024)
35. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In: European Conference on Computer Vision. pp. 1–18. Springer (2024)
36. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023)
37. Tang, Z., Yao, Y., Cui, M., Bo, L., Yang, H.: Gaussianip: Identity-preserving realistic 3d human generation via human-centric diffusion prior. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 348–358 (2025)
38. Wang, J., Lin, C., Nie, L., He, J., Li, H., Liao, K., Zhao, Y., et al.: Jasmine: Harnessing diffusion prior for self-supervised depth estimation. Advances in Neural Information Processing Systems **38**, 79901–79930 (2026)
39. Wang, J., Lin, C., Sun, L., Cao, Z., Yin, Y., Nie, L., Yuan, Z., Chu, X., Wei, Y., Liao, K., et al.: Geometry-guided reinforcement learning for multi-view consistent 3d scene editing. arXiv preprint arXiv:2603.03143 (2026)
40. Wang, J., Lin, C., Sun, L., Liu, R., Nie, L., Li, M., Liao, K., Chu, X.: From editor to dense geometry estimator. arXiv preprint arXiv:2509.04338 (2025)
41. Wang, W., Ye, H., Hong, F., Yang, X., Zhang, J., Wang, Y., Liu, Z., Pan, L.: Geneman: Generalizable single-image 3d human reconstruction from multi-source human data. arXiv preprint arXiv:2411.18624 (2024)
42. Weng, Z., Liu, J., Tan, H., Xu, Z., Zhou, Y., Yeung-Levy, S., Yang, J.: Template-free single-view 3d human digitalization with diffusion-guided lrm. arXiv preprint arXiv:2401.12175 (2024)
43. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21469–21480 (2025)
44. Xiu, Y., Yang, J., Cao, X., Tzionas, D., Black, M.J.: Econ: Explicit clothed humans optimized via normal integration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 512–523 (2023)
45. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: Icon: Implicit clothed humans obtained from normals. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13286–13296. IEEE (2022)
46. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021) (June 2021)
47. Zhang, D., Sun, Y., Li, P., Liu, Y., Lin, H., Xu, H., Mu, X., Lin, L., Yan, W., Yang, N., et al.: Pointcot: A multi-modal benchmark for explicit 3d geometric reasoning. arXiv preprint arXiv:2602.23945 (2026)
48. Zhang, G., Yao, N., Zhang, S., Zhao, H., Pang, G., Shu, J., Wang, H.: Multigo: Towards multi-level geometry learning for monocular 3d textured human reconstruction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 338–347 (2025)
49. Zhang, J., Li, X., Zhang, Q., Cao, Y., Shan, Y., Liao, J.: Humanref: Single image to 3d human generation via reference-guided diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1844–1854 (2024)

50. Zhang, Z., Sun, L., Yang, Z., Chen, L., Yang, Y.: Global-correlated 3d-decoupling transformer for clothed avatar reconstruction. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
51. Zheng, Z., Yu, T., Liu, Y., Dai, Q.: Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. IEEE transactions on pattern analysis and machine intelligence **44**(6), 3170–3184 (2021)
52. Zhuang, Y., Lv, J., Wen, H., Shuai, Q., Zeng, A., Zhu, H., Chen, S., Yang, Y., Cao, X., Liu, W.: Idol: Instant photorealistic 3d human creation from a single image. arXiv preprint arXiv:2412.14963 (2024), https://arxiv.org/abs/2412.14963