FlexiAvatar: Unified 3D Gaussian Human Avatars Under Arbitrary Body Visibility

Yihalem Yimolal Tiruneh[1], Muhammad Salman Ali[1], Uyoung Jeong[1], Muneeb A. Khan[1], MD Khalequzzaman Chowdhury Sayem[1], Allanur Bayramgeldiyev[1], Binod Bhattarai[2,3,4], and Seungryul Baek[1]

[1]UNIST, South Korea [2]University of Aberdeen, UK [3]University College London, UK [4]Fogsphere, UK

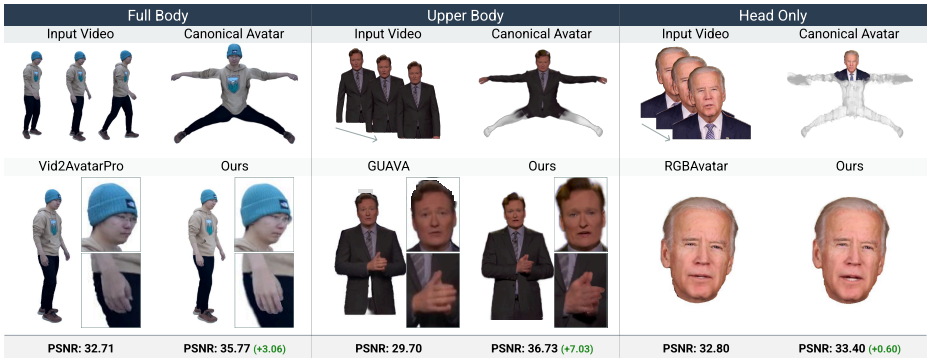

Fig. 1: From monocular video under arbitrary body visibility, FlexiAvatar reconstructs animatable 3D Gaussian avatars within a single pipeline, outperforming state-of-the-art full-body, upper-body, and head-only methods.

Abstract. Reconstructing animatable 3D human avatars from monocular video is a fundamental problem in computer vision with broad applications in AR/VR and digital content creation. Existing approaches typically couple parametric body models with neural rendering or 3D Gaussian splatting and optimize all body regions jointly from short videos, which often degrades fidelity in the visible areas. To overcome this limitation, we introduce **FlexiAvatar**, a unified framework that explicitly optimizes only the visible body regions, effectively eliminating artifacts arising from unobserved limbs. Our method integrates occlusion-robust SMPL-X tracking with part-specific residual refinement to capture high-frequency geometric and appearance details. To complete entirely unseen regions (e.g., back views), we leverage a diffusion-based approach to generate texture consistent with the observed appearance. Experiments on full-body (NeuMan, ZJU-MoCap, WildAvatar), upper/half-body (talk-show clips), and head-only (INSTA) inputs show that FlexiAvatar delivers consistently higher reconstruction quality, outperforming state-of-the-art methods by an average PSNR improvement of approximately 3% across datasets. Finally, by restricting optimization to observed regions, our method reduces the effective number of Gaussians that must be optimized and rendered, leading to reduced runtime and memory overhead in partial-visibility scenarios. The project page can be found here.

Keywords: 3D Gaussian Splatting · Human Avatar Reconstruction · Visibility-Aware Optimization · Partial Body Visibility · SMPL-X

1 Introduction

Realistic, animatable human avatars are essential for applications such as AR/VR [33, 46,65], film production, remote collaboration [10,40], and interactive entertainment [20]. However, most real-world and in-the-wild video data (video calls, talk-show clips, and social media footage) spans a wide visibility spectrum, containing severe cropping, partial-body visibility, and occlusions, rather than clean, full-body studio recordings. Reconstructing a complete, high-fidelity, animatable avatar from such monocular partial-body input remains an open and largely unsolved challenge.

Classical approaches such as SMPL [31], FLAME [30], and SMPL-X [38] employ parametric mesh models that provide strong priors for articulated body, face, and hand reconstruction. However, these models are inherently constrained by their fixed topology and often struggle to reproduce fine-grained geometric and textural fidelity [2, 62]. Neural rendering methods such as NeRF [34] improve visual realism over parametric approaches, but typically require lengthy per-subject optimization and lack real-time rendering efficiency [25].

3D Gaussian Splatting (3DGS) [22] has enabled fast, high-quality reconstruction of dynamic scenes [1]. Building on this representation, avatar methods such as GART [27], GaussianAvatar [14], ExAvatar [35] and Vid2AvatarPro [12] achieve high visual fidelity and strong animation quality. Despite these advances, all existing 3DGS avatar methods share a common assumption: the entire body is visible throughout training. In portrait recordings, video calls, and talk-show clips where only the face, head, or upper body ever appears, this assumption is systematically violated. Continuing to supervise Gaussians for non-visible limbs introduces hallucinated geometry and texture drift that corrupts even the visible regions. No existing method [12, 35] directly targets this visibility spectrum problem within the optimization loop itself.

Portrait-level approaches such as InsTaG [28], HRAvatar [57], RGBAvatar [29], and FATE [59] achieve high-fidelity reconstruction of the head region. In contrast, upper-body methods such as GUAVA [58] extend the reconstruction scope beyond the face; however, they remain constrained to a fixed upper-body input setting and often exhibit degraded quality in occluded or unseen regions. Neither class of methods generalizes across the entire visibility spectrum: full-body methods [12, 27, 35, 52] optimize all body regions regardless of visibility, inevitably hallucinating geometry and appearance in unobserved areas, while localized methods, whether head-only [28,41,59] or upper-body [58] are architecturally constrained to a specific body coverage input setting. This fragmentation forces practitioners to maintain separate, purpose-built pipelines for each input visibility setting, a significant practical barrier to real-world deployment [6].

To address these limitations, we propose FlexiAvatar, a visibility-aware Gaussian avatar reconstruction framework that explicitly restricts optimization to observed body regions, updating geometry and appearance only where visual evidence exists. This design eliminates ghost geometry and texture drift, and enables high-quality animation from full-body, upper-body, and head-only inputs within a single unified pipeline, as illustrated in Fig. 1. Our method combines structured priors from SMPL-X with region-level part-specific residual refinement and a visibility-guided learning objective. We employ occlusion-robust SMPL-X tracking to maintain accurate

articulation across varying visibility conditions, and introduce part-specific residual modules to recover high-frequency appearance details, such as facial expressions and finger creases, that exceed the capacity of global Gaussian representations. For unobserved regions such as the back, we leverage a diffusion-based approach to generate subject-consistent auxiliary views, providing appearance supervision for unseen areas without introducing hallucinations or propagating errors to the observed regions. By integrating (i) visibility-aware optimization, (ii) occlusion-robust SMPL-X tracking, (iii) part-specific residual refinement, and (iv) generative texture completion, FlexiAvatar achieves high-quality reconstruction across varying visibility (full-body, upper-body, and head-only) inputs within a unified pipeline.

Our main contributions are:

- We propose FlexiAvatar, a visibility-aware Gaussian avatar reconstruction framework that restricts optimization to visually observed body regions, updating geometry and appearance only where image evidence is present. This formulation eliminates hallucinated artifacts and texture drift, and enables high-quality animation from full-body, upper-body, or head-only inputs (Sec. 3).
- Our method incorporates an occlusion-robust SMPL-X tracking pipeline for accurate body registration, residual modules to recover high-frequency facial and hand details, and a diffusion-based generative texture completion module that synthesizes novel views to fill unseen regions (Fig. 2).
- Extensive quantitative and qualitative experiments across full-body [16, 19, 39], upper-body [54], and head-only [65] datasets demonstrate that FlexiAvatar consistently outperforms state-of-the-art methods (Sec. 4).

2 Related Work

3D-based Avatar Reconstruction. Template-based parametric models like FLAME [30], SMPL [31], and SMPL-X [38] provide controllable body, face, and hand representations and remain standard backbones in monocular human-capture pipelines. Recent advances in neural rendering have shifted from implicit NeRF-style fields to explicit representations to improve fidelity and enable differentiable primitives. 3DGS [22] further enables real-time, high-fidelity rendering through an explicit representation, making it a strong foundation for animatable avatars.

Person-specific Gaussian avatars from monocular video achieve high fidelity and motion controllability, yet uniformly assume full-body visibility during training. Early methods like Vid2Avatar [11] introduced self-supervised scene decomposition to reconstruct avatars from unmasked in-the-wild videos. 3DGS-Avatar [42] leverages a non-rigid deformation network with 3D Gaussian Splatting to achieve fast training and real-time rendering of clothed avatars from monocular video. GART [27] employs template-driven Gaussians with learnable skinning for articulation, while GaussianAvatar [14] introduces pose-conditioned deformation for dynamic appearance. ExAvatar [35] integrates Gaussian splats to the SMPL-X topology for improved facial and hand expressivity. GauHuman [15] employs body-prior-guided Gaussian splatting for efficient high-fidelity avatar reconstruction from monocular video, while ToMiE [56] decomposes texture and geometry into a modular implicit representation, and Vid2AvatarPro [12] incorporates universal generative priors for in-the-wild tex-

ture completion. Under partial views, all these methods hallucinate unseen regions, introduce artifacts, or propagate errors to visible areas.

For portrait-level avatar reconstruction, GaussianAvatars [41] employs rigged 3DGS; InsTaG [28] learns personalized talking heads from short monocular sequences; HRAvatar [57] yields relightable head avatars; and FATE [59] supports 360° full-head reconstruction with texture editing. RGBAvatar [29] compresses Gaussian blendshapes for online modeling, GEM [64] uses lightweight eigen-bases, and MeGA [48] adopts component-wise modeling. While these approaches achieve high fidelity in head avatars, they remain limited to localized reconstruction and do not generalize to full or upper-body avatars.

Recently, generalizable upper or half-body Gaussian avatars from *minimal* input have gained attention. GUAVA [58] reconstructs an animatable upper-body avatar from a *single* image via inverse texture mapping and neural refinement, while AniGS [43] uses inconsistent-Gaussian reconstruction to improve generalization. However, single-image pipelines lack multi-view constraints and often struggle with unseen-region completion and visual fidelity. Our approach bridges this gap by preserving the practicality of sparse inputs while introducing visibility-aware optimization that focuses on observed geometry, mitigating hallucination of unobserved regions and enabling stable optimization and reconstruction under partial views.

2D-based Human Animation. 2D motion transfer and diffusion-based generation methods provide strong appearance priors but lack a persistent 3D identity representation. DreamPose [21] synthesizes fashion videos from a single image and pose, while MagicPose [3] enables identity-aware pose and expression retargeting through latent diffusion. Champ [63] introduces SMPL-conditioned depth and normal priors for controllable animation, and MimicMotion [61] extends this direction with confidence-aware pose guidance for smooth long-term motion. Unlike these image-space methods, we leverage a diffusion-based approach to generate texture-consistent auxiliary views for unobserved regions.

3 Methodology

Our proposed method, FlexiAvatar, reconstructs a high-fidelity, animatable 3D avatar from a monocular video, even under partial-body visibility (e.g., upper-body or head-only). Our method consists of six components (as illustrated in Fig. 2): (i) performing accurate, visibility-aware SMPL-X registration to obtain a personalized template (Sec. 3.1); (ii) augmenting the input with diffusion-generated views to complete unseen textures (Sec. 3.2); (iii) learning a triplane-conditioned hybrid mesh–Gaussian representation for geometry and appearance (Sec. 3.3); (iv) enforcing visibility-aware optimization to prevent hallucination (Sec. 3.4); (v) refining high-frequency facial and hand appearance with a part-specific residual module (Sec. 3.5); and (vi) animating the avatar with SMPL-X poses and rendering it using 3D Gaussian Splatting (Sec. 3.6).

3.1 Accurate SMPL-X Registration

Given an input sequence, we extract frame-wise SMPL-X [38] parameters and 2D whole-body keypoints using off-the-shelf estimators [53, 55]. These provide an initialization of body pose $\theta \in \mathbb{R}^{55\times3}$, shape $\beta \in \mathbb{R}^{100}$, and facial expression $\psi \in \mathbb{R}^{50}$,

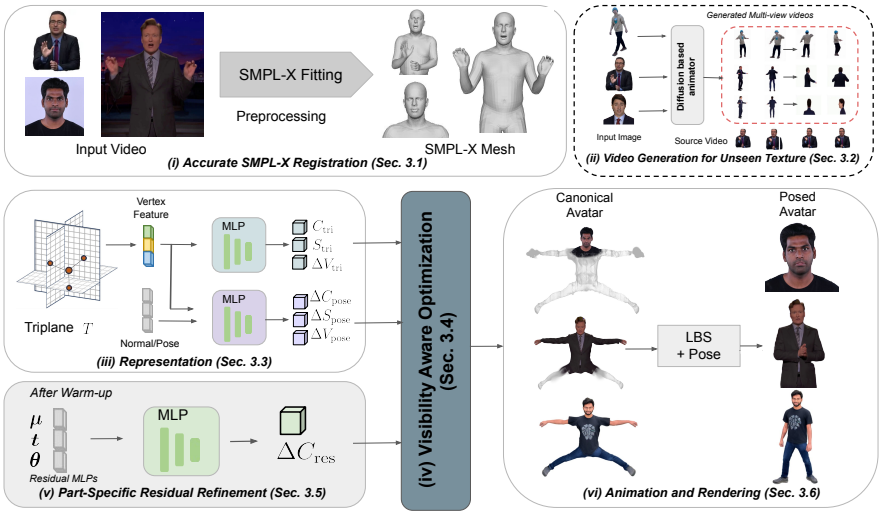

Fig. 2: Overview of our *FlexiAvatar* pipeline. From a monocular video, we perform visibility-aware SMPL-X registration (Sec. 3.1) and generate auxiliary multi-view sequences to supervise unseen regions (Sec. 3.2). Canonical vertices are projected onto a learnable triplane whose MLPs regress static and pose-dependent Gaussian parameters (Sec. 3.3). After warm-up, visibility-aware optimization prunes low-evidence Gaussians (Sec. 3.4) and residual MLPs enhance detail (Sec. 3.5). Gaussians are animated by SMPL-X linear blend skinning and rendered with 3DGS (Sec. 3.6).

along with keypoints $\mathbf{K} \in \mathbb{R}^{J\times 2}$ and confidence scores $\mathbf{s} \in \mathbb{R}^{J}$. Directly using all detected keypoints leads to artifacts when large portions of the body are outside the camera's field of view. To mitigate this, we introduce a binary visibility mask that excludes low-confidence or occluded joints from the reprojection objective. This ensures the stable fitting of the observed body parts without distorting unobserved regions. To capture identity-specific geometric nuances, we further optimize joint offsets $\Delta\mathbf{J}$ and facial offsets $\Delta\mathbf{V}_{\text{face}}$, producing a personalized canonical template. The final objective is:

$$\mathcal{L}_{\text{regi}} = \mathcal{L}_{\text{kpt}} + \lambda_{\text{init}}\mathcal{L}_{\text{init}} + \lambda_{\text{sface}}\mathcal{L}_{\text{sface}} + \lambda_{\text{sreg}}\mathcal{L}_{\text{sreg}}, \tag{1}$$

where $\mathcal{L}_{\text{regi}}$ denotes the overall SMPL-X registration objective, $\mathcal{L}_{\text{kpt}}$ is the 2D keypoint reprojection loss, $\mathcal{L}_{\text{init}}$ penalizes deviations from the off-the-shelf initialization, $\mathcal{L}_{\text{sface}}$ aligns the SMPL-X face geometry to a pre-fitted FLAME template to preserve identity-specific facial details and $\mathcal{L}_{\text{sreg}}$ regularizes both the shape symmetry and the learned joint-offset parameters. To improve robustness under occlusion, we apply a masking strategy to the 2D reprojection loss $\mathcal{L}_{\text{kpt}}$, defined as:

$$\mathcal{L}_{\text{kpt}} = \sum_{j} \mathbf{m}_j \left\| \Pi\left(V_j(\theta,\beta,\Delta J)\right) - \mathbf{K}_j \right\|_1, \tag{2}$$

where Π is the camera projection function, $\mathbf{V}_j$ is the j-th 3D joint location derived from the personalized SMPL-X model (which includes parameters θ, β, and the joint offset $\Delta\mathbf{J}$), $\mathbf{K}_j$ is the corresponding target 2D keypoint, and $\mathbf{m}_j \in \{0,1\}$ is a per-joint binary mask derived from detector confidence scores. Here $\mathbf{m} \in \mathbb{R}^{J}$ zeros

out low-confidence or occluded joints that fall below the threshold τ, restricting supervision to well-localized, visible keypoints only and preventing unreliable 2D evidence from corrupting the pose update. Based on empirical evidence, we set $\tau = 0.4$. $\mathcal{L}_{\text{init}}$ is defined as:

$$\mathcal{L}_{\text{init}} = \|\theta - \theta^*\|_1 + \|\beta - \beta^*\|_1 + \|\psi - \psi^*\|_1, \tag{3}$$

$\mathcal{L}_{\text{init}}$ term acts as a strong prior, leveraging the robust output of the initial regressor and preventing the optimization from drifting into implausible states. $\mathcal{L}_{\text{sface}}$ is a composite loss that aligns the SMPL-X face geometry (including the personalized offset ΔV_{face}) to the pre-fitted FLAME model [9, 30]:

$$\mathcal{L}_{\text{sface}} = \mathcal{L}_{\text{vertex}} + \mathcal{L}_{\text{lap}} + \mathcal{L}_{\text{edge}}, \tag{4}$$

where $\mathcal{L}_{\text{vertex}} = \|(\mathbf{V}_{\text{face}} + \Delta\mathbf{V}_{\text{face}}) - \mathbf{V}_{\text{FLAME}}\|_1$ is an L_1 loss on vertex positions, directly minimizing the distance between the SMPL-X face vertices and the target FLAME vertices to capture fine-grained identity details. $\mathcal{L}_{\text{lap}}$ is an L_2 loss on the mesh Laplacians, $\|\Delta(\mathbf{V}_{\text{face}} + \Delta\mathbf{V}_{\text{face}}) - \Delta\mathbf{V}_{\text{FLAME}}\|_2$, which ensures the facial mesh curvature and local smoothness properties are consistent. $\mathcal{L}_{\text{edge}}$ is an L_1 loss on the edge lengths of the two meshes, preserving the local topology and preventing unnatural stretching. Finally, $\mathcal{L}_{\text{sreg}}$ is a regularization term to ensure plausible and stable optimization:

$$\mathcal{L}_{\text{sreg}} = \mathcal{L}_{\text{shape}} + \mathcal{L}_{\text{jo}} + \mathcal{L}_{\text{sym}}, \tag{5}$$

where $\mathcal{L}_{\text{shape}} = \|\beta\|_2^2$ is a squared L_2 norm on the shape parameters β, keeping the body shape within the plausible distribution of the SMPL-X model. $\mathcal{L}_{\text{jo}} = \|\Delta\mathbf{J}\|_2^2$ is a squared L_2 norm on the joint offsets, preventing extreme and non-human-like deviations in the skeleton. $\mathcal{L}_{\text{sym}}$ enforces bilateral symmetry on the personalized offsets $\Delta\mathbf{J}$ and $\Delta\mathbf{V}_{\text{face}}$ by penalizing differences between corresponding left and right-side offsets. This formulation yields a geometry-consistent and identity-preserving reconstruction that remains robust across varying levels of body visibility. We employ SMPL-X [38] as the unified body representation across the entire visibility spectrum, including head-only inputs, instead of switching to a dedicated facial model such as FLAME [30]. This unified formulation enables a single reconstruction pipeline for all input scenarios. To preserve facial fidelity, we align the SMPL-X facial region with a DECA-initialized [9] FLAME mesh using the facial alignment loss $\mathcal{L}_{\text{sface}}$ (Eq. 4) and transfer facial expressions through the shared expression basis.

3.2 Video Generation for Unseen Texture

In unconstrained videos with severe cropping or occlusion, existing avatar methods often produce artifacts in regions that are never observed. While some recent works [26, 47] regularize these unseen areas using diffusion priors, they assume input videos contain full-body coverage, which limits their applicability to partial-body scenarios. To overcome this limitation, we introduce a part-specific motion video generation strategy. Using MimicMotion [61], we synthesize auxiliary videos conditioned on the input, and jointly use these with the original footage to construct the final avatar.

This augmentation effectively recovers missing texture information for consistently unobserved regions. Specifically, we generate a novel video of the subject performing a full 360° rotation. This synthetic motion reveals unobserved views, such as the back, providing the necessary appearance cues for texture completion.

3.3 Representation

We construct our model by integrating the parametric SMPL-X [38] prior with explicit 3D Gaussian splatting [22]. In line with [35], we use a hybrid mesh–Gaussian representation. From the co-registered SMPL-X model (shape β, offsets $\Delta\mathbf{J}$,$\Delta\mathbf{V}_{\text{face}}$), we generate a canonical mesh $\bar{\mathbf{V}}$ that is up-sampled to N vertices while preserving the underlying triangular connectivity of the SMPL-X template. We associate with each of those N vertices a learnable 3D Gaussian asset (mean position, scale, color), thus forming a surface-conditioned Gaussian representation. To encode identity and enable animation, we use a learnable triplane field $\mathbf{T} \in \mathbb{R}^{3\times C\times H\times W}$. Canonical positions $\bar{\mathbf{P}} \in \mathbb{R}^{N\times 3}$ are orthogonally projected onto the three planes, and bilinear interpolation yields a per-vertex feature $\mathbf{f}_i$ (Fig. 2). Two MLP groups decode these features. The first predicts pose-independent Gaussian parameters (offset $\Delta\mathbf{V}_{\text{tri},i}$, scale $\mathbf{S}_{\text{tri},i}$, color logit $\mathbf{C}_{\text{tri},i}$) that capture static shape and appearance. The second is pose-dependent: one MLP maps $\mathbf{f}_i$ and body pose (excluding root) to $\Delta\mathbf{V}_{\text{pose},i}$,$\Delta\mathbf{S}_{\text{pose},i}$, and a color MLP maps pose and normal $\mathbf{n}_i$ to $\Delta\mathbf{C}_{\text{pose},i}$. This disentangles static identity from dynamic articulation, supporting robust learning under limited pose diversity while preserving high-frequency detail.

3.4 Visibility-Aware Optimization

In upper-body or head-only videos, large portions of the body (e.g., legs and feet) are never observed. Methods that assume a full one-to-one correspondence with the SMPL-X surface inevitably hallucinate these unseen regions, leading to unrealistic geometry and texture artifacts in the visible regions. To mitigate this, we introduce a visibility-aware optimization that gathers *per-Gaussian* visibility cues and restricts supervision to only the observed regions. Visibility information, indicating which Gaussians contribute to each rendered pixel, is obtained directly from the rasterizer and used to guide optimization.

Per-Gaussian Evidence. After each refinement stage, we compute a confidence signal for every Gaussian i, defined by its visibility rate:

$$v_i = \frac{1}{F}\sum_{f=1}^{F} v_i^f \in [0,1], \tag{6}$$

where v_i^f denotes the binary visibility flag for Gaussian i at frame f, and F denotes the total number of processed frames. This score reflects how confidently a Gaussian is supported by visual evidence. We first optimize all Gaussians without filtering for the initial 2,000 iterations. This allows coarse geometry, pose, and color to stabilize, building reliable statistics for v_i and preventing early false negatives from aggressively removing still-emerging regions. After warm-up, we determine the visibility threshold automatically via Otsu's method [36]: the distribution of $\{v_i\}$ across all Gaussians

is treated as a bimodal histogram, and the threshold τ^* that minimizes intra-class variance between the visible and invisible Gaussians is selected. This avoids the need for a manually tuned threshold and adapts to the observed visibility distribution of each input sequence. Gaussians whose visibility score falls below the threshold τ^* are excluded from the forward rasterization pass, reducing computational overhead in proportion to the number of invisible Gaussians. This visibility-based filtering is similarly applied to the regularization terms, specifically the Laplacian constraints on Gaussian color, scale, and position: these are masked for invisible Gaussians, preventing them from contributing to the loss and ensuring that regularization does not introduce artifacts into the visible regions.

Unlike conventional visibility masks, v_i is used to guide optimization rather than as a rendering-time filter. Since all Gaussians are decoded from the shared triplane representation T, gradients originating from observed pixels can propagate to Gaussians in unobserved regions and, through the shared representation, indirectly influence neighboring visible Gaussians. By masking invisible Gaussians in both the reconstruction and regularization losses (Sec. 3.4), we suppress these spurious gradients at their source, ensuring that the optimization is driven only by regions supported by visual evidence. As illustrated in Fig. 3, on the *Oliver* sequence, where the lower limbs are never observed, conventional optimization propagates gradients into the invisible lower body, which subsequently degrades the reconstructed torso. In contrast, our visibility-aware optimization confines gradient propagation to the observed upper body, resulting in more stable and accurate reconstruction.

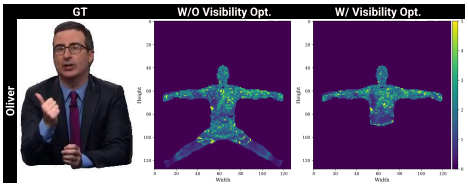

Fig. 3: Gradient magnitude on the shared triplane for the upper-body *Oliver* sequence.

3.5 Part-Specific Residual Refinement

Capturing fine-grained facial expressions and accurate hand details is crucial for conveying emotion and maintaining realism [17, 44]. To enhance these high-frequency regions, we introduce an implicit function–based residual appearance refinement module that enhances color predictions for the face and hands [13]. Given a Gaussian center $\boldsymbol{\mu} \in \mathbb{R}^3$, time t, and pose θ, we construct the input by concatenating their positional encodings. A lightweight MLP then predicts a color residual conditioned on this spatiotemporal encoding, allowing precise recovery of dynamic appearance details beyond the base Gaussian representation.

$$\Delta\mathbf{C}_{\text{res}}(\boldsymbol{\mu},t,\theta) = F_\theta\big(\text{concat}\big(\gamma(\boldsymbol{\mu}),\gamma(t)\big),\theta\big) \tag{7}$$

where $\Delta\mathbf{C}_{\text{res}}$ represents the predicted color residuals, F_θ denotes the learnable MLP, and $\gamma(\cdot)$ is the positional encoding function. In practice, we implement F_θ using three separate, lightweight MLPs specialized for the face, left hand, and right hand, respectively, each of which is only activated when the corresponding body part is visible according to our visibility-aware optimization (Sec. 3.4). We first train the coarse avatar without these residual modules for 2,000 iterations to obtain a stable geometry and base color representation. Once convergence is reached, the three

residual networks are activated, and their predicted color offsets are progressively added to enhance fine-grained appearance. The final color is then updated as follows:

$$\mathbf{C}_{\text{final}} = \mathbf{C}_{\text{tri}} + \Delta\mathbf{C}_{\text{pose}} + \Delta\mathbf{C}_{\text{res}}. \tag{8}$$

3.6 Animation and Rendering

The 3D human avatar is constructed in a canonical space and animated using SMPL-X poses θ and facial expression code ψ. Identity-specific offsets $\Delta\mathbf{V}_{\text{tri}}$ are combined with pose-dependent offsets $\Delta\mathbf{V}_{\text{pose}}$, while face and hand regions directly adopt SMPL-X deformations for accuracy. Facial expression offsets $\Delta\mathbf{V}_{\text{expr}}$ from ψ are added to face vertices, eliminating the need to learn the expression space. The posed geometry is then skinned via linear blend skinning (LBS) using SMPL-X weights. Rendering is performed using 3D Gaussian Splatting (3DGS) [22].

Training Objective. We jointly optimize all learnable components, including the 3D Gaussian parameters (positions, scales, colors, and opacities), triplane feature fields, residual RGB MLPs, alongside the sequence-specific SMPL-X parameters. We supervise the framework using a composite image reconstruction objective comprising L_1, SSIM, and perceptual LPIPS terms between rendered outputs and ground-truth images. We further incorporate a dedicated facial consistency loss and apply Laplacian-style regularization following [35] to encourage spatial smoothness in geometry and appearance. During mixed-source training, captured and diffusion-generated frames are sampled with equal probability; to account for potential artifacts in synthetic views, we down-weight their reconstruction losses following [47]. Formally, the overall training objective $\mathcal{L}$ is a weighted sum defined as:

$$\mathcal{L} = \lambda_{L1}\mathcal{L}_1 + \lambda_{\text{ssim}}\mathcal{L}_{\text{ssim}} + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}} + \lambda_{\text{face}}\mathcal{L}_{\text{face}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} \tag{9}$$

where $\mathcal{L}_1$, $\mathcal{L}_{\text{ssim}}$, and $\mathcal{L}_{\text{lpips}}$ are computed on a cropped human region for efficiency. The L_1 term penalizes per-pixel deviations between the rendered image I_{pred} and the ground-truth I_{gt}, $\mathcal{L}_{\text{ssim}}$ encourages local structural and contrast consistency, and $\mathcal{L}_{\text{lpips}}$ compares deep feature activations to promote sharp, perceptually realistic textures. The facial loss $\mathcal{L}_{\text{face}}$ enforces that facial Gaussians remain aligned with a stable, identity-preserving facial template. Following [35], we render a facial image I mesh using a differentiable mesh renderer and a fixed UV texture obtained by averaging registration textures, and supervise it with $\mathcal{L}_{\text{face}} = \left\| I_{\text{mesh}} - I_{\text{gt}} \right\|_1$ over the facial region. This constrains the facial Gaussians to maintain a consistent appearance and reduces drift under novel expressions.

The regularization term $\mathcal{L}_{\text{reg}}$ aggregates our geometry and appearance priors. Its main components are Laplacian smoothness penalties, which we apply in a visibility-aware manner. Using the binary visibility mask $\mathbf{m}_i \in \{0,1\}$ (derived from the visibility rate v_i in Sec. 3.4), we ensure that smoothing is applied only to well-observed Gaussians, preventing over-regularization of unseen regions.

Table 1: Quantitative results on the NeuMan [19] dataset for 3D human avatar reconstruction.

Method	PSNR↑	SSIM↑	LPIPS↓
HumanNeRF [50]	27.06	0.967	1.90
InstantAvatar [18]	28.47	0.972	2.80
NeuMan [19]	29.32	0.972	1.40
Vid2Avatar [11]	30.70	0.980	1.40
GaussianAvatar [14]	29.94	0.980	1.20
3DGS-Avatar [42]	28.99	0.974	1.60
ExAvatar [35]	34.80	0.984	0.90
Vid2AvatarPro [12]	32.71	0.983	1.19
Ours	**35.77**	**0.987**	**0.83**

Table 2: ZJU-MoCap dataset results (subjects 377, 386, 387, 392, 393, 394).

Method	377	386	387	392	393	394
PSNR↑						
GauHuman [15]	32.63	36.47	32.99	32.47	32.31	33.81
ToMiE [56]	33.63	37.29	34.42	34.46	32.73	34.26
Ours	**34.98**	**39.50**	**36.11**	**36.15**	**34.66**	**36.63**
SSIM↑						
GauHuman [15]	0.974	0.973	0.970	0.967	0.965	0.967
ToMiE [56]	0.977	0.976	0.973	0.973	0.966	0.968
Ours	**0.991**	**0.990**	**0.986**	**0.991**	**0.985**	**0.986**
LPIPS↓						
GauHuman [15]	2.07	2.99	2.75	3.35	3.14	2.57
ToMiE [56]	1.80	2.91	2.50	2.88	3.02	2.67
Ours	**0.66**	**0.85**	**0.99**	**0.80**	**1.15**	**1.01**

4 Experiments

We present empirical results on standard benchmarks. We first cover implementation details, datasets, and metrics (Sec. 4.1), then quantitative and qualitative comparisons (Sec. 4.2, Sec. 4.3), and finally an ablation analysis (Sec. 4.4).

4.1 Implementation Details

Our framework is implemented in PyTorch [37] and optimized using the Adam optimizer [23]. All experiments are conducted on a single NVIDIA RTX A6000 GPU for 30k iterations with a batch size of 1. The base learning rate for all network parameters is set to 1×10^{-3}. Body masks are extracted using SAM [24]. To reduce sensitivity to imperfect foreground masks, we use the same masking strategy as [35]. Detailed hyperparameter settings, including the λ weights used in Eq. (1) and Eq. (9), are provided in the Supplementary Material.

Datasets. We evaluate our method across varying visibility inputs, (i) full-body, (ii) upper-body, and (iii) head-only, to comprehensively evaluate the performance of FlexiAvatar. For full-body evaluation, we use the NeuMan dataset [19], selecting the `bike`, `citron`, `jogging`, and `seattle` sequences, which provide broad body coverage with minimal motion blur. We adopt the official train/test split following the established protocol [14]. To further validate generalization on a standard full-body benchmark, we additionally evaluate on ZJU-MoCap [39] using six subjects (377, 386, 387, 392, 393, 394). For each subject, we reserve the last 100 frames for testing and utilize the remaining frames for training. To further assess robustness in unconstrained real-world scenarios, we evaluate our method on WildAvatar [16], a large-scale human-centric dataset collected from YouTube. We randomly select seven subjects, reserving the last 100 frames of each sequence for testing and using the remaining frames for training. For upper-body evaluation, we use three subjects (`Oliver`, `Conan`, and `Chemistry`) from the TalkShow dataset [54]. Each 10-second clip (∼300 frames) is split by reserving the last 30 frames for testing and using the remaining frames for training. Finally, for head-only reconstruction, we evaluate on seven videos (`bala`, `malte_1`, `biden`, `justin`, `marcel`, `nf_01`, `nf_03`) from the INSTA dataset [65]. Following prior work, we reserve the final 350 frames for testing [29].

Evaluation Protocol. For quantitative evaluation, we report PSNR, SSIM [49], and LPIPS (×100) [60]. In all result tables, the best score is highlighted in **bold**. For Tab. 1,

Table 3: Quantitative comparisons of 3D human avatars on the TalkShow (upper-body) [54] dataset.

Subject	Method	PSNR↑	SSIM↑	LPIPS↓
Oliver	ExAvatar [35]	29.13	0.938	2.29
	GART [27]	24.14	0.927	7.76
	GUAVA [58]	26.76	0.890	12.17
	Ours	**29.80**	**0.952**	**1.81**
Conan	ExAvatar [35]	35.66	0.980	3.61
	GART [27]	26.89	0.974	5.56
	GUAVA [58]	29.70	0.928	7.69
	Ours	**36.73**	**0.984**	**2.82**
Chemistry	ExAvatar [35]	25.67	0.922	10.29
	GART [27]	22.28	0.914	11.89
	GUAVA [58]	26.85	0.920	**7.15**
	Ours	**27.80**	**0.935**	9.67

Table 4: Quantitative comparisons of 3D human avatars on the INSTA (Head-only) [65] dataset.

Method	PSNR↑	SSIM↑	LPIPS↓
SplattingAvatar [45]	29.03	0.932	10.35
GaussianAvatars [41]	29.10	0.945	8.61
FlashAvatar [51]	27.44	0.912	11.05
MonoGaussianAvatar [5]	28.91	0.945	7.43
GaussianBlendShapes [32]	30.01	0.947	9.17
FATE [59]	27.85	0.942	5.68
RGBAvatar [29]	32.72	**0.953**	6.04
Ours	**33.04**	**0.953**	**5.63**

Table 5: Quantitative comparisons of 3D human avatars on the WildAvatar [16] dataset.

Method	PSNR↑	SSIM↑	LPIPS↓
GauHuman [15]	28.31	0.957	3.83
ToMiE [56]	28.65	0.957	3.81
Ours	**30.98**	**0.973**	**2.93**

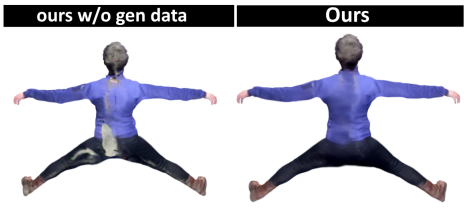

Fig. 4: Avatar in canonical space with and without generated data in our method. Generated data fills missing texture without noticeable artifacts.

all baseline numbers are taken directly from the corresponding papers [12, 14, 35]. Consistent with prior work [4, 14, 18, 42], evaluation on NeuMan is performed by optimizing SMPL-X parameters for test frames using only the image loss while keeping all other model parameters fixed. For quantitative evaluation, we follow the background masking strategy of ExAvatar [35]. For Tab. 3, we report metrics computed using the official implementations released by the respective methods. For the ZJU-MoCap benchmark, quantitative results for GauHuman [15] and ToMiE [56] were obtained by executing their officially released source code on our exact 100-frame test split to ensure a fair comparison. For Tab. 4, baseline numbers are adopted from [29, 59] , with RGBAvatar [29] computed from its official code.

4.2 Quantitative Comparison

We conduct a comprehensive evaluation against SOTA methods across three settings: full-body (NeuMan, ZJU-MoCap), upper-body (TalkShow), and head-only (INSTA). **Full-Body Avatars on NeuMan.** We evaluate our method against state-of-the-art full-body avatar approaches on the NeuMan dataset [19]. As reported in Tab. 1, our model achieves the highest performance across all evaluation metrics. To further assess the fidelity of fine-grained reconstruction, we conduct focused evaluations on the facial and hand regions that exhibit complex, high-frequency motion and appearance variations. As shown in the Supplementary Material (Sec. E.6, Tab. A7), the proposed part-specific residual refinement (Sec. 3.5) significantly improves the reconstruction of fine-grained facial and hand details, consistently outperforming ExAvatar on cropped face and hand regions with notably higher PSNR. These results demonstrate the

Fig. 5: Qualitative comparison on the NeuMan full-body dataset. Our method recovers sharper clothing texture and more accurate hand detail.

Fig. 6: Qualitative comparison on the TalkShow upper-body dataset. Our method preserves facial sharpness and recovers hand detail.

effectiveness of our framework in recovering detailed and expressive geometry and appearance.

Full-Body Avatars on ZJU-MoCap. We further validate our method on the ZJU-MoCap dataset [39], a standard full-body benchmark, comparing against GauHuman [15] and ToMiE [56] across six subjects (377, 386, 387, 392, 393, 394). As reported in Tab. 2, our method achieves the best performance across all subjects and all three metrics, outperforming both baselines especially in LPIPS. This confirms that our visibility-aware optimization preserves full-body reconstruction quality while generalizing across constrained studio captures and partial-visibility settings within one pipeline.

Full-Body Avatars on WildAvatar. To evaluate generalization beyond curated benchmarks, we compare FlexiAvatar against GauHuman [15] and ToMiE [56] on seven in-the-wild subjects from the WildAvatar dataset [16]. As reported in Tab. 5, FlexiAvatar consistently outperforms both baselines across all three metrics, achieving higher PSNR and lower LPIPS. These results demonstrate that our visibility-aware optimization generalizes effectively to unconstrained, in-the-wild videos.

Upper-Body Avatars on TalkShow. We also evaluate our method in the upper-body setting using three subjects from the TalkShow dataset. As shown in Tab. 3, our approach consistently outperforms all baselines (ExAvatar [35], GART [27], GUAVA [58]) across all subjects. This confirms the robustness and high fidelity of our method for upper-body reconstruction, particularly in challenging regions with limited visibility. The effectiveness of our framework in recovering detailed and expressive geometry and appearance further highlights its ability to generalize across varying input coverage, maintaining high-frequency details even when only partial observations are available.

Head-Only Avatars on INSTA. Finally, we evaluate our model on head-only reconstruction using seven subjects from the INSTA dataset [65]. The results are reported in Tab. 4. As evident, our method achieves the best PSNR and LPIPS scores, highlighting its strong generalization ability for dynamic head reconstruction. These results further demonstrate that our visibility-aware design and high-frequency residual modeling remain effective even under extremely limited input coverage.

4.3 Qualitative Comparison

For qualitative evaluation, we compare FlexiAvatar with a wide range of state-of-the-art methods under three visibility settings: full-body, upper-body, and head-only inputs. For full-body reconstruction, we compare against Vid2Avatar, Vid2AvatarPro, and ExAvatar on NeuMan (Fig. 5), and against GauHuman and ToMiE on ZJU-MoCap. Our method consistently reconstructs sharper clothing textures and more accurate hand structures on NeuMan, while producing clearer skin textures and more consistent limb geometry on ZJU-MoCap (see Fig. A3 in the Supplementary Material). For upper-body reconstruction (Fig. 6), we compare with GUAVA, ExAvatar, and GART on the TalkShow dataset. Methods designed under the assumption of full-body visibility degrade under partial observations: GART produces severe artifacts and distorted textures, while ExAvatar yields blurred appearance and fails to capture high-frequency facial and hand details. In contrast, FlexiAvatar preserves facial sharpness and reconstructs fine hand structures, outperforming both general and upper-body-specific baselines such as GUAVA. For head-only reconstruction (Fig. 1), we additionally compare against RGBAvatar, where FlexiAvatar maintains high visual fidelity and produces clean avatar reconstructions, demonstrating strong generalization even under extremely limited input coverage.

4.4 Ablation Studies

To assess the contribution of each core component, we conduct ablation studies on both the NeuMan and INSTA datasets. The quantitative results are reported in Tab. 6 and Tab. 7. Additional experiments and visualizations are provided in the Supplementary Material, including implementation details, network architectures, and the full SMPL-X registration procedure. We also present extended qualitative comparisons, canonical-space visualizations, and additional novel-pose animation results.

w/o visibility-aware optimization. Disabling our visibility-aware optimization on the INSTA dataset leads to a consistent performance drop across all metrics (Tab. 7). The reductions in PSNR and LPIPS confirm that this component is essential for achieving accurate, high-fidelity head reconstruction in monocular, head-only scenarios where visible evidence is limited. Beyond preserving visual fidelity, this visibility-aware pruning inherently restricts the total number of Gaussians that must be optimized and rendered. As detailed in Tab. 8, this structural efficiency yields a nearly 50% reduction in memory footprint for head-only avatars, resulting in notably faster animation and rendering speeds compared to full-body baselines.

w/o SMPL-X optimization. Removing confidence-weighted keypoint masking leads to the largest single-component drop as shown in Tab. 7. Under head-only input, off-the-shelf estimators produce unreliable pose estimates for occluded joints;

Table 6: Impact of refinement heads and generated views on reconstruction quality on NeuMan [19] dataset.

Method	PSNR↑	SSIM↑	LPIPS↓
Ours (full)	**35.77**	**0.987**	**0.83**
w/o refinement	34.90	0.985	0.86
w/o gen. data	35.29	0.986	0.92

Table 7: Impact of visibility-aware optimization and SMPL-X optimization on reconstruction quality for the INSTA [65] dataset.

Method	PSNR↑	SSIM↑	LPIPS↓
Ours (full)	**33.04**	**0.953**	**5.63**
w/o visibility-aware opt.	32.15	0.949	5.80
w/o SMPL-X opt.	31.07	0.934	7.60

Table 8: Detailed runtime and memory statistics averaged for Head-only (INSTA) and Upper-body (TalkShow) settings. #G denotes the number of human Gaussians. Asset (MB) is the summed tensor footprint of the Gaussian asset dictionary (mean, opacity, scale, rotation, rgb). Animation (Anim) and rendering (Rend) speeds are reported in milliseconds.

Setting	Method	Resolution	#G	Asset (MB)	Anim (ms)	Rend (ms)	Total (ms)	FPS ↑
Head	ExAvatar	512^2	167,390	8.94	32.175	2.098	34.272	29.18
	Ours	512^2	**84,095** (-49.7%)	**4.49** (-49.8%)	**24.135** (-25.0%)	**1.410** (-32.8%)	**25.545** (-25.5%)	**39.15** (+34.2%)
Upper body	ExAvatar	512^2	167,390	8.94	30.239	1.966	32.206	31.06
	Ours	512^2	**144,350** (-13.8%)	**7.71** (-13.8%)	**28.127** (-7.0%)	**1.408** (-28.4%)	**29.536** (-8.2%)	**33.87** (+9.1%)

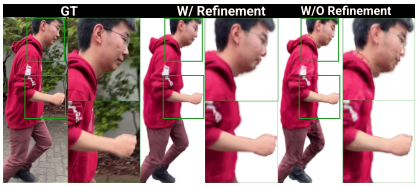

Fig. 7: Ablation of the part-specific residual refinement module on the NeuMan sequence. FlexiAvatar recovers fine-grained facial and hand details.

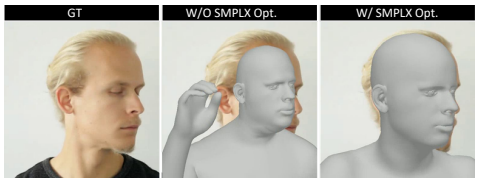

Fig. 8: Effect of occlusion-robust SMPL-X optimization. Our confidence-weighted joint masking recovers a well-aligned identity-specific template.

without visibility-masked refinement these errors corrupt the canonical template and misaligned Gaussian attachment points, degrading even fully observed facial regions. The qualitative impact is shown in Fig. 8.

w/o refinement. Removing the refinement heads substantially degrades reconstruction quality, particularly in high-frequency regions. As shown in Tab. 6, PSNR drops markedly, and the declines in SSIM and LPIPS further indicates the model's reduced ability to synthesize sharp textures and fine geometric detail Fig. 7.

w/o generated data. Excluding generated training views leads to artifacts when we render it from novel views, as shown in Fig. 4. This suggests that the generated data plays an important role in enhancing realism, improving texture completion for invisible region. We down-weight the diffusion-generated views' reconstruction loss to keep synthetic artifacts from overriding observed evidence (Supp. Sec. E.1, Tab. A2). To confirm these views preserve identity, we measure Face Consistency (FC), the ArcFace [8] cosine similarity between rendered and ground-truth faces (Supp. Sec. E.2, Tab.A3), where our method achieves the highest FC among all baselines.

5 Conclusion

We present FlexiAvatar, a visibility-aware framework for reconstructing generalizable 3D Gaussian human avatars from monocular videos. In contrast to previous methods

that implicitly assume full-body visibility, our approach restricts optimization to only the visible regions, effectively eliminating ghost artifacts and enhancing fidelity under partial-view inputs. Through the integration of SMPL-X-based body tracking, part-specific high-frequency refinement, visibility-aware optimization, and a diffusion-based generative completion module, FlexiAvatar provides a unified solution capable of high-quality reconstruction across full-body, upper-body, and head-only settings.

Limitations. Although FlexiAvatar delivers high-fidelity reconstructions across varying visibility conditions, it still shares a major limitation with existing 3DGS-based avatar methods [7, 11, 35, 42]: modeling dynamically deforming clothing remains challenging. As an avenue for future work, our visibility-aware formulation could be extended to better identify and adaptively refine regions undergoing nonrigid deformations, enabling more accurate reconstruction of complex clothing dynamics.

Acknowledgements

This work is supported by NRF grants (No. RS-2025-00521013 60%, No. RS-2025-02216916 10%) and IITP grants (No. RS2020-II201336 Artificial intelligence graduate school program(UNIST) 10%; No. RS-2025-25442149 LG AI STAR Talent Development Program for Leading Large-Scale Generative AI Models in the Physical AI Domain 10%), funded by the Korean government (MSIT). This work is also supported by the InnoCORE program of the Ministry of Science and ICT(26-InnoCORE-01) 10%.

References

1. Ali, M.S., Zhang, C., Cagnazzo, M., Valenzise, G., Tartaglione, E., Bae, S.H.: Compression in 3d gaussian splatting: A survey of methods, trends, and future directions. IEEE Transactions on Circuits and Systems for Video Technology (2026)
2. Alldieck, T., Magnor, M., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single rgb camera. In: CVPR (2019)
3. Chang, D., Shi, Y., Gao, Q., Xu, H., Fu, J., Song, G., Yan, Q., Zhu, Y., Yang, X., Soleymani, M.: Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In: ICML (2024)
4. Chen, J., Zhang, Y., Kang, D., Zhe, X., Bao, L., Jia, X., Lu, H.: Animatable neural radiance fields from monocular rgb videos. arXiv preprint arXiv:2106.13629 (2021)
5. Chen, Y., Wang, L., Li, Q., Xiao, H., Zhang, S., Yao, H., Liu, Y.: Monogaussianavatar: Monocular gaussian point-based head avatar. In: SIGGRAPH (2024)
6. Cho, I., Shin, E., Ali, M.S., Bae, S.H.: Dynamic structured pruning with novel filter importance and leaky masking based on convolution and batch normalization parameters. IEEE Access **9**, 165005–165013 (2021)
7. Choi, H., Moon, G., Armando, M., Leroy, V., Lee, K.M., Rogez, G.: Mononhr: Monocular neural human renderer. In: 3DV (2022)
8. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019)
9. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. ToG (2021)

10. Fidalgo, C.G., Sousa, M., Mendes, D., Dos Anjos, R.K., Medeiros, D., Singh, K., Jorge, J.: Magic: Manipulating avatars and gestures to improve remote collaboration. In: 2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR) (2023)
11. Guo, C., Jiang, T., Chen, X., Song, J., Hilliges, O.: Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In: CVPR (2023)
12. Guo, C., Li, J., Kant, Y., Sheikh, Y., Saito, S., Cao, C.: Vid2avatar-pro: Authentic avatar from videos in the wild via universal prior. In: CVPR (2025)
13. Haider, A., Ali, M.S., Qamar, M., Khalil, T., Kim, S.Y., Oh, J., Tartaglione, E., Bae, S.H.: I-inr: iterative implicit neural representations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 4520–4528 (2026)
14. Hu, L., Zhang, H., Zhang, Y., Zhou, B., Liu, B., Zhang, S., Nie, L.: Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In: CVPR (2024)
15. Hu, S., Hu, T., Liu, Z.: Gauhuman: Articulated gaussian splatting from monocular human videos. In: CVPR (2024)
16. Huang, Z., Hu, S., Wang, G., Liu, T., Zang, Y., Cao, Z., Li, W., Liu, Z.: Wildavatar: Learning in-the-wild 3d avatars from the web. In: CVPR (2025)
17. Jeong, U., Tiruneh, Y.Y., Chang, H.J., Baek, S., Kim, K.I.: Thom: Generating physically plausible hand-object meshes from text. In: IEEE Conf. Comput. Vis. Pattern Recog. Findings. pp. 3653–3664 (June 2026)
18. Jiang, T., Chen, X., Song, J., Hilliges, O.: Instantavatar: Learning avatars from monocular video in 60 seconds. In: CVPR (2023)
19. Jiang, W., Yi, K.M., Samei, G., Tuzel, O., Ranjan, A.: Neuman: Neural human radiance field from a single video. In: ECCV (2022)
20. Kang, J.H., Ali, M.S., Jeong, H.W., Choi, C.K., Kim, Y., Jeong, S.Y., Bae, S.H., Kim, H.Y.: A super-resolution-based feature map compression for machine-oriented video coding. IEEE Access **11**, 34198–34209 (2023)
21. Karras, J., Holynski, A., Wang, T.C., Kemelmacher-Shlizerman, I.: Dreampose: Fashion image-to-video synthesis via stable diffusion. In: ICCV (2023)
22. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ToG (2023)
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
24. 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: ICCV (2023)
25. Kopanas, G., Philip, J., Leimkühler, T., Drettakis, G.: Point-based neural rendering with per-view optimization. Computer Graphics Forum (Proceedings of the Eurographics Symposium on Rendering) (2021)
26. Lee, I., Kim, B., Joo, H.: Guess the unseen: Dynamic 3d scene reconstruction from partial 2d glimpses. In: CVPR (2024)
27. Lei, J., Wang, Y., Pavlakos, G., Liu, L., Daniilidis, K.: Gart: Gaussian articulated template models. In: CVPR (2024)
28. Li, J., Zhang, J., Bai, X., Zheng, J., Zhou, J., Gu, L.: Instag: Learning personalized 3d talking head from few-second video. In: CVPR (2025)
29. Li, L., Li, Y., Weng, Y., Zheng, Y., Zhou, K.: Rgbavatar: Reduced gaussian blendshapes for online modeling of head avatars. In: CVPR (2025)
30. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ToG (2017)
31. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ToG (2015)
32. Ma, S., Weng, Y., Shao, T., Zhou, K.: 3d gaussian blendshapes for head avatar animation. In: SIGGRAPH (2024)

33. Ma, S., Simon, T., Saragih, J., Wang, D., Li, Y., De La Torre, F., Sheikh, Y.: Pixel codec avatars. In: CVPR (2021)
34. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
35. Moon, G., Shiratori, T., Saito, S.: Expressive whole-body 3d gaussian avatar. In: ECCV (2024)
36. Otsu, N., et al.: A threshold selection method from gray-level histograms. Automatica (1979)
37. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019)
38. 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: CVPR (2019)
39. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: CVPR (2021)
40. Piumsomboon, T., Lee, G.A., Hart, J.D., Ens, B., Lindeman, R.W., Thomas, B.H., Billinghurst, M.: Mini-me: An adaptive avatar for mixed reality remote collaboration. In: Proceedings of the 2018 CHI conference on human factors in computing systems (2018)
41. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In: CVPR (2024)
42. Qian, Z., Wang, S., Mihajlovic, M., Geiger, A., Tang, S.: 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In: CVPR (2024)
43. Qiu, L., Zhu, S., Zuo, Q., Gu, X., Dong, Y., Zhang, J., Xu, C., Li, Z., Yuan, W., Bo, L., et al.: Anigs: Animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. In: CVPR (2025)
44. Sayem, M., Chowdhury, M.T., Tiruneh, Y.Y., Khan, M.A., Ali, M.S., Bhattarai, B., Baek, S.: Handvqa: Diagnosing and improving fine-grained spatial reasoning about hands in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2515–2525 (2026)
45. Shao, Z., Wang, Z., Li, Z., Wang, D., Lin, X., Zhang, Y., Fan, M., Wang, Z.: SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In: CVPR (2024)
46. Shen, K., Guo, C., Kaufmann, M., Zarate, J.J., Valentin, J., Song, J., Hilliges, O.: X-avatar: Expressive human avatars. In: CVPR (2023)
47. Sim, G., Moon, G.: Persona: personalized whole-body 3d avatar with pose-driven deformations from a single image. In: ICCV (2025)
48. Wang, C., Kang, D., Sun, H., Qian, S., Wang, Z., Bao, L., Zhang, S.H.: Mega: Hybrid mesh-gaussian head avatar for high-fidelity rendering and head editing. In: CVPR (2025)
49. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The thrity-seventh asilomar conference on signals, systems & computers, 2003 (2003)
50. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: Humannerf: Free-viewpoint rendering of moving people from monocular video. In: CVPR (2022)
51. Xiang, J., Gao, X., Guo, Y., Zhang, J.: Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In: CVPR (2024)
52. Xiang, T., Sun, A., Delp, S., Kozuka, K., Fei-Fei, L., Adeli, E.: Wild2avatar: Rendering humans behind occlusions. arXiv preprint arXiv:2401.00431 (2023)
53. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: ICCV (2023)

54. Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., Black, M.J.: Generating holistic 3d human motion from speech. In: CVPR (2023)
55. Yin, W., Cai, Z., Wang, R., Zeng, A., Wei, C., Sun, Q., Mei, H., Wang, Y., Pang, H.E., Zhang, M., et al.: Smplest-x: Ultimate scaling for expressive human pose and shape estimation. TPAMI (2025)
56. Zhan, Y., Zhu, Q., Niu, M., Ma, M., Zhao, J., Zhong, Z., Sun, X., Qiao, Y., Zheng, Y.: Towards explicit exoskeleton for the reconstruction of complicated 3d human avatars. In: ICCV (2025)
57. Zhang, D., Liu, Y., Lin, L., Zhu, Y., Chen, K., Qin, M., Li, Y., Wang, H.: Hravatar: High-quality and relightable gaussian head avatar. In: CVPR (2025)
58. Zhang, D., Liu, Y., Lin, L., Zhu, Y., Li, Y., Qin, M., Li, Y., Wang, H.: Guava: Generalizable upper body 3d gaussian avatar. In: ICCV (2025)
59. Zhang, J., Wu, Z., Liang, Z., Gong, Y., Hu, D., Yao, Y., Cao, X., Zhu, H.: Fate: Full-head gaussian avatar with textural editing from monocular video. In: CVPR (2025)
60. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
61. Zhang, Y., Gu, J., Wang, L.W., Wang, H., Cheng, J., Zhu, Y., Zou, F.: Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. In: ICML (2025)
62. Zheng, W., Zhao, J., Liu, X., Pan, Y., Gan, Z., Han, H., Liu, N.: Flame-based multi-view 3d face reconstruction. In: Computer Graphics International Conference (2023)
63. Zhu, S., Chen, J.L., Dai, Z., Dong, Z., Xu, Y., Cao, X., Yao, Y., Zhu, H., Zhu, S.: Champ: Controllable and consistent human image animation with 3d parametric guidance. In: ECCV (2024)
64. Zielonka, W., Bolkart, T., Beeler, T., Thies, J.: Gaussian eigen models for human heads. In: CVPR (2025)
65. Zielonka, W., Bolkart, T., Thies, J.: Instant volumetric head avatars. In: CVPR (2023)