This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Multi-view Consistent 3D Gaussian Head Avatars 'without' Multi-view Generation

Aviral Chharia, Fernando De la Torre
Carnegie Mellon University
{achharia, ftorre}@cs.cmu.edu

Figure 1. ***MVCHead*** achieves state-of-the-art for unconditional generation of high fidelity, multi-view consistent 3D Gaussian head avatars in "minimal resource setting", without requiring intermediate views, or even 3D data. The generated Gaussian heads capture complex textures and fine facial micro-structure, including wrinkles, hair wisps, ear rims, lip contours, skin blemishes, eyes, and accessories.

Abstract

High-fidelity 3D Gaussian head avatar generation is critical for applications such as AR/VR, telepresence, and digital humans. Existing methods depend on multi-view datasets, 3D captures, or intermediate 2D view synthesis. In contrast, we learn both conditional and unconditional 3D head models from randomly sampled 2D images alone, without using multi-view data, 3D supervision, or intermediate view generation. We introduce MVCHead, a single-shot state space model that enforces multi-view consistency (MVC) directly in the 3D representation while regressing 3D Gaussians under these constraints. At its core, we propose a Hierarchical State Space (HiSS) block that progressively refines Gaussians from coarse to fine, while capturing long-range dependencies. Within each HiSS block, we modify Mamba's standard unidirectional scan with the proposed Hierarchical Bi-directional State Scan (HiBiSS) that aligns recurrence with the axes along which multi-view inconsistencies are strongest. Finally, we design an SE(3) Multi-view Critic that judges whether a set of self-renders arises from a single underlying 3D configuration, rewarding cross-view pixel alignment without observing real multi-view pairs. MVCHead achieves state-of-the-art perceptual quality, surpasses prior methods in both texture and geometric consistency, and maintains comparable shape consistency. To demonstrate scalability, we release FaceGS-10K, the first large-scale dataset of ready-to-use 3D Gaussian head assets for training and evaluation of 3D head models. Project Page and code: https://humansensinglab.github.io/MVCHead/

1. Introduction

High-fidelity 3D Gaussian head avatars have become central to AR/VR, telepresence, digital characters, and large-scale content creation in film and games [2, 36, 45, 53, 55,

75, 79, 83]. These applications demand vast numbers of realistic yet non-identifiable 3D head avatars that are consistent across views but correspond to no real individual–avoiding privacy concerns and enabling rapid content creation. Generating such assets in a minimal-resource setting (e.g., from 2D images alone) is practically important, especially for studios that cannot afford dense multi-view capture rigs or high-end 3D scanning. Moreover, multi-view diffusion pipelines that first synthesize intermediate views are computationally heavy and often require additional training data. Motivated by these constraints, we explore this 'minimal-resource setting'.

Recent work on 3D Gaussian head avatar generation falls into three broad categories that differ primarily in supervision, data requirements, and scalability (see Fig. 2). *First*, multi-view optimization-based methods [3, 10, 20, 26, 61, 70, 71] reconstruct a full 3D head from high-resolution studio-captured sequences with dense multi-view coverage. These pipelines, using datasets such as NeRSemble [42] or RenderMe-360 [60] (with $\sim 10^4$ frames per subject), achieve impressive photorealism and strong MVC (see Fig. 2(a)). However, reliance on costly capture setups and heavy per-subject optimization limits scalability.

A *second* class of methods [18, 19, 24, 31, 32, 51, 54, 69, 78, 84] encompasses multi-view diffusion approaches that start from a single image and first synthesize intermediate views, typically including side views of the subject via off-the-shelf image or video diffusion models (see Fig. 2(b)). A separate reconstructor then lifts these images into a 3DGS representation [41]. While fidelity is high, MVC becomes tightly coupled to intermediate view quality: pixel-aligned cross-view losses are not optimized since there is no end-to-end differentiability, and identity drift persists: tiny per-view deviations (e.g., subtle shifts in hair, ear contours, or jawline shading) may not correspond to any consistent 3D explanation. Moreover, generating dense intermediate views per asset is computationally prohibitive at scale.

A *third* line of works [5, 6, 38, 44] includes feed-forward 3D generators that directly produce 3D Gaussian head avatars in an end-to-end differentiable manner. These methods aim for unconditional generation of 3D Gaussian heads from a learned prior, enabling the creation of diverse, non-existent identities while avoiding per-subject optimization. GGHead [44], GS-GAN [38], and CGS-GAN [6] improve stability, yet enforcing MVC without explicit multi-view supervision remains open, particularly in minimal-resource settings when the model never observes real multi-view pairs. In this work, we tackle this highly challenging minimal-resource setting (see Fig. 2(c)): achieving large-scale, real-time synthesis of multi-view consistent 3D Gaussian head avatars via a single-shot, end-to-end differentiable model that operates *(i) without generating intermediate views and (ii) without relying on 3D ground truth*.

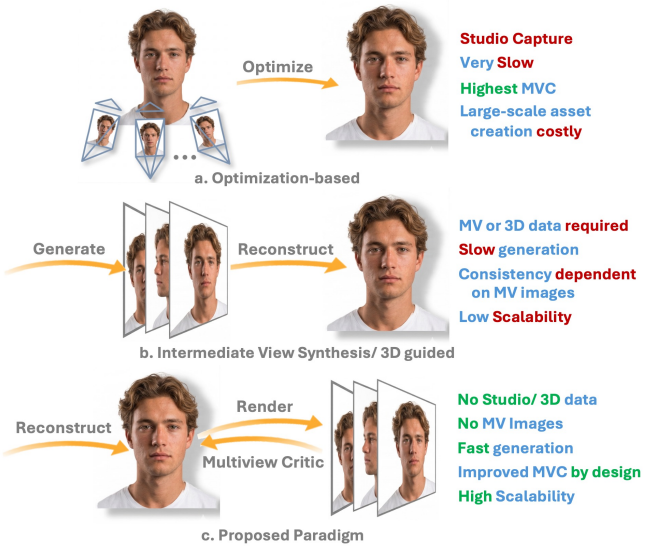


Figure 2. **Motivation.** Paradigms for 3D Gaussian head avatar generation. (a) Requires expensive studio captures; (b) Synthesizes intermediate views before reconstruction; (c) Learns an unconditional 3D Gaussian head directly from 2D images w/o intermediate generation or even 3D data.

To address this, we introduce *MVCHead*, a novel state space model tailored to this setting. To the best of our knowledge, MVCHead is the *first* to leverage state space modeling for 3D Gaussian head generation. It takes a latent code and produces a complete set of 3D Gaussians in a single forward pass. MVCHead consists of a series of Hierarchical State Space (HiSS) blocks that organize Gaussians in a hierarchy and guide finer levels through offsets anchored to coarser parent Gaussians. Within each HiSS block, we apply the proposed Hierarchical Bi-directional State Scan (HiBiSS), which enforces grid-aligned coherence to reconcile typical view-to-view drift. Finally, we propose an SE(3) Multi-view Critic that rewards cross-view pixel alignment, inducing multi-view consistency by design. Taken together, MVCHead combines architectural improvements with a learned consistency critic to generate 3D Gaussian head avatars of high visual quality and strong multi-view consistency (see Fig. 1). Our main contributions include:

- We highlight the challenge of MVC and analyze how it can be induced by design, arguing that intermediate view generation is counterproductive for scalability. We propose an SE(3) Multi-view Critic that rewards cross-view pixel alignment without real multi-view pairs.
- We introduce *MVCHead*, the first to leverage visual Mamba for 3D Gaussian head generation: a fast, single-shot state space model that directly predicts Gaussians and improves MVC in unconditional 3D head synthesis.
- We modify Mamba's traditional unidirectional scan into a Hierarchical Bi-directional State Scan (HiBiSS) that aligns recurrence with principal axes of multi-view drift.

- MVCHead surpasses the state-of-the-art in perceptual quality and along all three MVC axes, achieving superior texture and geometric consistency while maintaining comparable shape consistency.
- We release FaceGS-10K, a large-scale dataset of ready-to-use 3D Gaussian heads for large-scale training, benchmarking, and evaluation of 3D-aware head models.

2. Related Works

2.1. 3D Gaussian Head Avatars

Multi-view optimization-based methods. A large body of work [3, 10, 20, 26, 61, 70, 71] reconstructs detailed 3D heads by optimizing Gaussians against dense, high-resolution studio-captured multi-view video sequences such as RenderMe-360 [60] and NeRSemble [42], which largely guarantee MVC. GaussianAvatars [61] rigs Gaussians to FLAME [50]; SplattingAvatar [63] leverages monocular video; GaussianHeadAvatars [74] and MonoGaussianAvatar [12] exploit multi-view data but from relatively sparse or monocular views. These set an upper bound on quality but offer low scalability due to expensive capture and per-subject optimization.

Multi-view diffusion methods. These models [18, 19, 24, 54, 68, 69, 78, 84] generate 3D head avatars from a single input image by first synthesizing several intermediate views [31, 32, 51] via off-the-shelf image or video diffusion [54] and subsequently reconstructing the avatar. Zero-1-to-A [84], FaceLift [54], Cap4D [69], FaceCraft4D [78], SpinMeRound [24], and Portrait4D [18, 19] have achieved impressive fidelity in this two-stage setup. Cap4D [69] and FaceCraft4D [78] target 4D controllability; Portrait4D [18, 19] variants improve identity stability across expression and view changes; FaceLift [54] couples multi-view diffusion with Gaussian reconstruction. While fidelity is high, MVC hinges on the intermediate view generator; pixel-aligned cross-view losses are not optimized end-to-end, and identity drift across synthesized views persists. Moreover, dense multi-view generation for each asset is computationally prohibitive at scale. Other works leverage monocular or multi-view videos [22, 25, 43, 47, 67, 80].

Feed-forward and other methods. These methods [6, 14, 15, 35, 38, 44, 48, 59] generate avatars directly in 3D through a feed-forward mapping from latent codes to Gaussians. Recent large Gaussian reconstruction models such as LAM [35], GAGAvatar [14], PanoLAM [48], PercHead [59], and GPAvatar [15] reintroduce end-to-end differentiability but rely on large-scale video datasets [72], multi-view data from Cafca [7], or studio-collected 3D data [42, 60] to impose MVC. GGHead [44] uses a 2D CNN model to predict Gaussian attributes in a UV-template head and regularizes geometry via a total-variation loss. Hyun *et al.* [38] introduce hierarchical Gaussians to stabilize training; Barthel *et al.* [6] address the challenge of view conditioning. Despite this progress, enforcing MVC without paired multi-view supervision remains a key bottleneck.

2.2. State Space Models

State Space Models (SSMs) originate from classical linear dynamical systems and Kalman filtering [39]. Gu *et al.* introduced the modern Structured State Space Sequence (S4) family, which demonstrated strong long-range dependency modeling [28, 29]. Mamba [27] extends S4 by replacing its fixed hidden-space projection matrices with an input-dependent selective projection mechanism. Recent variants [49, 52, 85] adapt SSM scanning to 2D and higher-dimensional inputs. Hybrid Mamba-Transformer architectures [34] have achieved SOTA performance on ImageNet-1K [16] classification and multiple vision tasks [13, 21]. Despite this, the use of SSMs in 3D generative modeling remains largely unexplored. Gamba [64] combines Mamba with 3DGS for single-view reconstruction but shows limited texture quality; MVGamba [77] targets simple objects for content creation rather than human heads. MVCHead is the first to leverage SSMs for 3D head avatar generation. We use SSMs to align recurrence with the axes along which multi-view inconsistencies manifest, making state space propagation instrumental in improving MVC.

3. MVCHead

We aim to learn a generative mapping from a latent code z to a 3D head, represented as a set of anisotropic Gaussians [41]. Unlike methods that rely on expensive studio captures or additional view synthesis, we operate in a minimal-resource setting, supervising solely on 2D images.

Notation and Preliminaries. For a latent code $z \sim \mathcal{N}(0, I)$, we generate a set of anisotropic Gaussians [41], $\mathcal{S}_\theta(z) = \{g_i\}_{i=1}^N$. Each individual Gaussian g_i is defined by the tuple $g_i = (\mu_i, s_i, q_i, \alpha_i, c_i)$. Here, $\mu_i \in \mathbb{R}^3$ denotes the 3D center, $s_i \in \mathbb{R}^3_+$ encodes positive axis-aligned scales, $q_i \in \mathbb{H}$ is a unit quaternion defining a rotation matrix $R(q_i) \in SO(3)$, $\alpha_i \in (0, 1)$ is an opacity value, and $c_i \in [0, 1]^3$ is an RGB color. We fix the Gaussian budget at N= 240K, which is sufficient for high-fidelity modeling of facial features. A differentiable splatting renderer $\mathcal{R}$ maps $\mathcal{S}_\theta(z)$ and a camera pose $T \in SE(3)$ to an image: $\mathbf{I} = \mathcal{R}(\mathcal{S}_\theta(z), T) \in \mathbb{R}^{H \times W \times 3}$. Crucially, the only supervision comes from 2D images sampled from large face corpora; these images provide a texture and appearance distribution but no ground truth cross-view correspondences.

Overview. The proposed model architecture is illustrated in Fig. 3. We build on the transformer-based GSGAN [38], making three key departures: a novel Dual-Mixer architecture leveraging the state space blocks; the proposed HiBiSS scan; and the SE(3) Multi-view Critic as an explicit MVC

reward. The resulting architecture, *MVCHead*, is an end-to-end differentiable pipeline that enforces MVC through structural design and a learned geometric reward, rather than relying on explicit 3D supervision. **(1)** MVCHead comprises a stack of HiSS blocks that progressively refine the Gaussian representation from coarse to fine. These blocks employ the proposed HiBiSS to propagate geometric and appearance cues across a token grid, ensuring local and global consistency when regressing the 3D Gaussian head. **(2)** The resulting set of 3D Gaussians is processed by a 3DGS rasterizer [41]. This allows us to render the avatar from arbitrary camera poses. **(3)** During training, these renders are evaluated by two distinct critics: an adversarial texture discriminator that ensures high-frequency realism and stylistic alignment with the training distribution; and an SE(3) Multi-view Critic that enforces MVC by rewarding pixel-aligned cross-view agreement.

3.1. Hierarchical State Space (HiSS) Blocks

We represent the head as a composition of Gaussians that are progressively refined across a hierarchy of L HiSS blocks. Unlike conventional 3DGS [41], here Gaussians serve a dual role: they provide a partial, coarse approximation of the 3D head and simultaneously guide the regression of subsequent finer-level Gaussians.

Anchor-based Refinement. Fine-level Gaussians are parametrized explicitly as offsets from coarser-level anchors [38]. This architectural bias ensures that new primitives lie near established structure, forcing details to refine existing geometry rather than drifting arbitrarily. As synthesis progresses, the Gaussian count grows by an upsampling ratio r per block, enabling progressively detailed synthesis of facial features. Specifically, each subsequent block upsamples its input points [33, 57, 81] and attaches new Gaussians to existing ones. The final avatar is rendered jointly in a single splatting pass using the aggregated set of $\sum_{l=0}^{L-1} Nr^l$ primitives.

Conditioning and Disentanglement. The initial HiSS block ($l = 0$) takes as input a scaffold of randomly initialized learnable tokens of size 512×3 [66]. To increase the representational capacity, these tokens are lifted to a higher-dimensional feature grid via multi-frequency positional encoding, yielding a dense $H \times W$ grid. To ensure identity-consistent synthesis, we apply disentangled appearance conditioning via AdaIN layers [37]. Tokens are modulated by a learned scale and bias predicted from a mapped latent $w \in W$, which empirically helps decouple appearance from geometry throughout the hierarchy. The same conditioning is applied to all HiSS blocks, ensuring appearance coherence while geometry is refined. Notably, following CGSGAN [6], we explicitly omit camera conditioning within these HiSS blocks. By introducing camera poses only during rendering and the SE(3) Multi-view Critic, we prevent the model from collapsing into view-specific 2D heuristics and ensure the MVC signal remains anchored to the 3D geometry.

Dual-Mixer Architecture. Within each HiSS block, tokens pass through two complementary mixers: a self-attention that aggregates global semantics and captures long-range dependencies not strongly tied to spatial axes (such as overall facial identity or global cues), and a state space block that enforces local grid-aligned coherence along horizontal and vertical directions via scanning mechanisms described below. The output tokens are then fed to per-attribute MLP heads that directly regress the Gaussian parameters. HiSS blocks operate on a fixed-resolution token grid at all levels, preserving spatial coherence.

Insight 1: Axis-aligned Multi-view Drift

Multi-view inconsistencies in 3D Gaussian heads are strongly aligned with the image axes. *Yaw* changes mainly cause horizontal shifts: silhouettes move left-right, ear visibility changes, jawlines shift laterally, and fine structures such as hair appear to translate sideways. In contrast, *Pitch* changes mainly produce vertical drift: shading on the nose, cheekbones, eyebrows, and chin shift up-down, and the apparent heights of facial features change. Since most training data normalize heads to be upright and centered, the dominant components of multi-view inconsistency lie along the horizontal and vertical image directions.

3.2. Hierarchical Bi-directional State Space Scanning (HiBiSS)

SSMs offer a natural mechanism for imposing architectural constraints along the specific axes where multi-view inconsistencies typically manifest. However, standard unidirectional scans (i.e., left-to-right) [27] are insufficient for 3D head generation as they lack vertical propagation and introduce causal biases that prevent global context integration. We therefore introduce HiBiSS, which applies four complementary 2D scans: row-wise left-to-right ($\rightarrow$), row-wise right-to-left ($\leftarrow$), column-wise top-to-bottom ($\downarrow$), and column-wise bottom-to-top ($\uparrow$), creating bidirectional recurrent paths that connect any two tokens along both axes. We implement it by adapting SS2D [52] to the hierarchical Gaussian prediction setting: tokens are linearly projected, reshaped into an $H \times W$ grid, processed by four symmetric scan trajectories, fused, and re-projected back to the original token space, preserving one-to-one correspondence between spatial positions and token identities.

Motivation. Consider a camera with intrinsics $\mathbf{K} = \mathrm{diag}(f_x, f_y, 1)$ and a canonical pose ($R = I,\ t = \mathbf{0}$). A 3D point $X = (X, Y, Z)^\top$ on the head surface projects to

pixel coordinates: $\mathbf{u} = (x, y)^\top = \left(f_x \frac{X}{Z},\ f_y \frac{Y}{Z}\right)^\top$. Small yaw and pitch rotations about the vertical and horizontal axes, with angles $\delta\theta_y$ and $\delta\theta_x$ respectively, induce a first-order displacement: $\delta\mathbf{u} \approx J_x(X)\,\delta\theta_x + J_y(X)\,\delta\theta_y$, where $J_x(X) = \frac{\partial \mathbf{u}}{\partial \theta_x}$ and $J_y(X) = \frac{\partial \mathbf{u}}{\partial \theta_y}$ are the pitch and yaw Jacobians at X. For upright, centered heads, where depth Z varies smoothly and the face is approximately centered on the optical axis, we typically observe: $\left|\frac{\partial x}{\partial \theta_y}\right| \gg \left|\frac{\partial y}{\partial \theta_y}\right|, \left|\frac{\partial y}{\partial \theta_x}\right| \gg \left|\frac{\partial x}{\partial \theta_x}\right|$, i.e., yaw mainly produces horizontal displacement, while pitch produces vertical displacement. This motivates encoding cross-view corrections with state-space recurrences aligned to rows and columns.

HiBiSS Architecture. Based on this motivation, we introduce HiBiSS to encode cross-view corrections using state space recurrences aligned to the rows and columns. Let $F \in \mathbb{R}^{H \times W \times d}$ denote the 2D token grid, with row index $i \in \{1, \ldots, H\}$, column index $j \in \{1, \ldots, W\}$, and channel dimension d. The horizontal forward scan along row i is defined by the recurrence:

$$h^{\rightarrow}_{i,j+1} = A_h\, h^{\rightarrow}_{i,j} + B_h\, F_{i,j}, \quad \tilde{F}^{\text{hor}}_{i,j} = C_h\, h^{\rightarrow}_{i,j} + D_h\, F_{i,j},$$

where $h^{\rightarrow}_{i,j} \in \mathbb{R}^d$ is the hidden state at position (i, j) and $A_h, B_h, C_h, D_h \in \mathbb{R}^{d \times d}$ are structured state space matrices following the parameterization of [52]. The vertical forward scan is defined analogously. HiBiSS runs all four directional scans hierarchically and fuses the resulting features into an updated grid $\tilde{F}$. Thus, state-space propagation is explicitly aligned with the directions where $\|\partial \mathbf{u}/\partial \theta\|$ is largest, implementing an anisotropic, pose-aware smoothing that targets the principal axes of inconsistency drift. HiBiSS is applied *before* per-level upsampling and attribute regression. Applying it after upsampling would increase compute and dilute the recurrence over near-duplicate tokens, while applying it during per-attribute prediction would deprive the model of a shared, geometry-aware context. Since the `Attn+MLP` mixer operates on the same appearance-conditioned features, and passing them through HiBiSS beforehand enables coherent propagation of both appearance and geometric cues, improving multi-view agreement across the full set of Gaussian attributes.

Insight 2: Consistency by Construction

While intermediate view generation may produce inconsistent views, self-renders from any fixed 3D model are, *by construction*, geometrically consistent (see Fig. 4). This allows us to train a critic that distinguishes plausible 3D configurations from inconsistent ones without requiring explicit multi-view data.

3.3. SE(3) Multi-view Critic

The Critic is an extrinsic-aware encoder E_ψ that maps a set of images and corresponding camera poses to a scalar consistency score $s = E_\psi(\{\hat{I}_k\}, \{T_k\}) \in \mathbb{R}$. For a given latent z, we render K views $\{\hat{I}_k\}_{k=1}^K$ of the generated avatar under a set of canonicalized camera poses $\{T_k\}_{k=1}^K$. The Critic jointly processes both images and poses to produce a score that is higher when the set is mutually consistent. During training, the model maximizes this score, so that improving multi-view agreement directly improves the objective:

$$\mathscr{L}_{mvc} = -\mathbb{E}_{z,\{T_k\}}\left[E_\psi\left(\{\mathcal{R}(\mathcal{S}_\theta(z), T_k)\}_{k=1}^K,\ \{T_k\}_{k=1}^K\right)\right]$$

Training Strategy. To ensure that E_ψ provides a meaningful MVC signal, we train it as a binary set classifier. The positive set $\mathcal{S}^+ = \{(\mathcal{R}(\mathcal{S}_\theta(z), T_k), T_k)\}_{k=1}^K$ consists of K views rendered from the *same* avatar under different poses T_k. The negative set $\mathcal{S}^- = \{(\mathcal{R}(\mathcal{S}_\theta(z_k), T_k), T_k)\}_{k=1}^K$ comprises views each rendered from a different latent but sharing the same T_k's. The Critic is optimized with a binary cross-entropy loss on its logits, encouraging it to assign higher scores to positive sets than to negative ones. Although the negative sets exhibit obvious identity variation, the Critic must additionally learn subtle geometric and textural cues of consistency such as silhouette coherence and shading continuity. Once trained, E_ψ serves as a differentiable reward term: the HiSS blocks are updated to maximize $E_\psi(\mathcal{S}^+)$, pushing the model to produce avatars whose self-renders exhibit stronger cross-view consistency.

Geometric Transform Attention. The Critic's consistency score should depend only on the relative view arrangement, not absolute camera placement or intrinsics. While standard cross-attention lacks this invariance, Geometric Transform Attention (GTA) [56] addresses it by embedding SE(3) structure directly into the attention, ensuring equivariance to global rigid transforms and invariance to intrinsics. Architecturally, E_ψ follows a ViT-style design augmented with GTA [56]. Each image is patchified into tokens. We inject extrinsics by anchoring all poses relative to the first view $\tilde{T}_k = T_k T_1^{-1}$, and align tokens across views by pre-transforming the attention queries and keys with lightweight, block-diagonal linear maps derived from these relative extrinsics. Since GTA aligns tokens using SE(3) relations, i.e., without camera intrinsics, the score s is invariant to intrinsics and cropping, and stable across rig changes, yielding pose-only invariance. Moreover, since the scene and all cameras undergo the same transform, the set of relative transforms is unchanged, s is preserved, providing global-rigid equivariance.

Training Objective. The total loss for MVCHead is a multi-task objective that combines geometric consistency, textural realism, and structural regularization. Given only 2D images, the joint model optimizes the parameters of the Gaussian decoder (HiSS blocks), the SE(3) Multi-view Critic (E_ψ), and the adversarial texture discriminator (D_ϕ). This joint training pushes the model to produce 3D config-

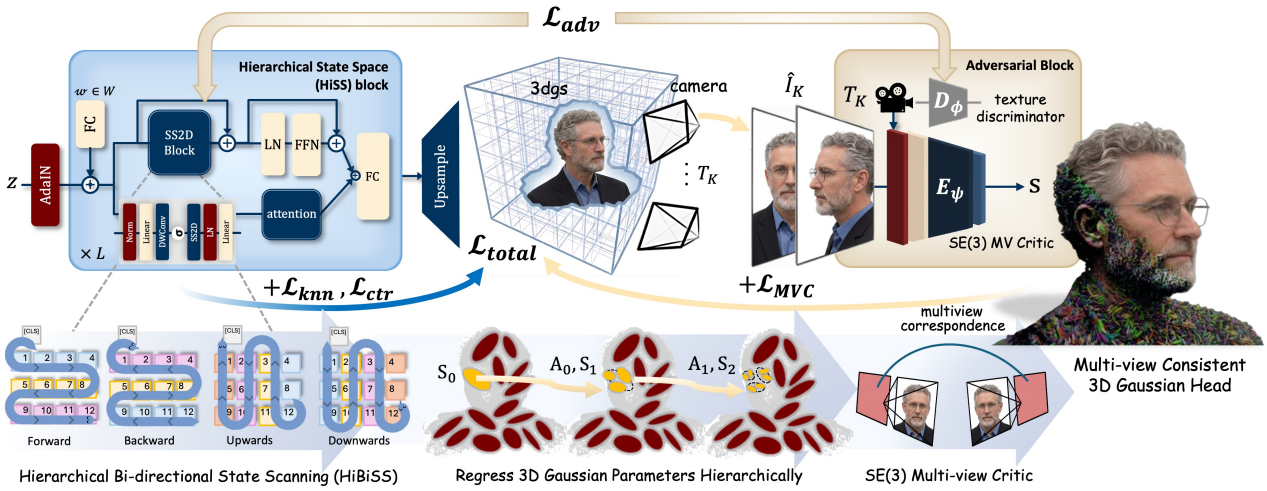


Figure 3. **Model Architecture.** MVCHead along with its key proposed components, including HiSS blocks which hierarchically regress the 3D Gaussian parameters (Gaussian S_0 becomes the anchor A_0 for computing the next Gaussian S_1, and so on), and perform Hierarchical Bi-directional State Scan (HiBiSS) in all directions, and the SE(3) Multi-view Critic, which enforces MVC.

urations that are both multi-view consistent and statistically indistinguishable from real images.

The total loss combines: **(1) SE(3) Multi-view Critic ($\mathscr{L}_{mvc}$):** Encourages cross-view geometric consistency. This constitutes a key departure from prior works such as GS-GAN [38] and CGSGAN [6], which rely primarily on adversarial and conditional losses. **(2) Adversarial texture term ($\mathscr{L}_{adv}$):** A standard camera-conditioned adversarial loss with an R1 gradient penalty. It ensures that the projected textures of generated avatars match the distribution of real training images across K sampled views [6]. **(3) Spatial regularization ($\mathscr{L}_{knn}$ and $\mathscr{L}_{ctr}$):** These constrain the Gaussian point cloud [38]. $\mathscr{L}_{knn}$ penalizes excessive spacing between neighboring Gaussians to maintain surface density, while $\mathscr{L}_{ctr}$ penalizes Gaussian centers' drift from their hierarchical anchors to ensure structural stability. The combined loss is defined as follows:

$$\begin{aligned}\mathscr{L}_{\text{total}} = & \underbrace{\lambda_{mvc}\Big(-\mathbb{E}_{z,\{T_k\}}\big[E_\psi(\{\hat{I}_k\},\{T_k\})\big]\Big)}_{\text{SE(3) Multi-view Critic}} \\ & + \underbrace{\mathbb{E}_z \frac{1}{K}\sum_{k=1}^{K} \text{softplus}\big(-D_\phi(\mathcal{R}(\mathcal{S}_\theta(z),T_k),T_k)\big)}_{\text{Camera-conditioned Adv., K-view AVG. for Texture Consistency}} \\ & + \underbrace{\lambda_{knn}\mathscr{L}_{knn} + \lambda_{ctr}\mathscr{L}_{ctr}}_{\text{Gaussian Regularizer (Local spacing, Center drift)}}\end{aligned}$$

4. Experiments and Results

Datasets. For a fair comparison, we train MVCHead under the established experimental protocol on the FFHQ [40] and FFHQ-C [6] datasets, and benchmark against SOTA generative 3D head models trained on the same datasets.

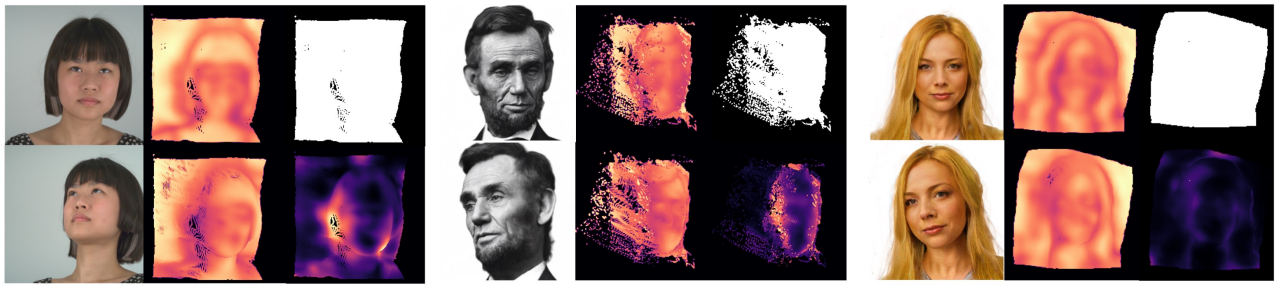
Evaluation Metrics. Following prior works, we report Fréchet Inception Distance (FID) and $\text{FID}_{3\text{D}}$ to measure the perceptual realism of generated avatars. However, these metrics fail to capture MVC. Quantitative evaluation of MVC remains an open challenge in 3D head synthesis, with no universally accepted metrics currently. To address this gap, we adapt scores from two SOTA frameworks: MVG-Bench [73] and MEt3R [4], providing the first comprehensive quantitative assessment of MVC for 3D head avatars.

Training. MVCHead was trained for 10M steps on FFHQ and FFHQ-C using Adam optimizer on 4 NVIDIA H100 GPUs over 3 days. Additional training details, including hyperparameters, are presented in the supplementary.

4.1. Main Results

We evaluate MVCHead by training independently on FFHQ [40] and FFHQ-C [6] to ensure a fair comparison. The synthesized avatars demonstrate SOTA visual quality, capturing fine facial features such as wrinkles, hair wisps, and skin blemishes (see Fig. 1).

Realism (FID↓, $\text{FID}_{3\text{D}}$ ↓). We use FID to assess the perceptual realism of rendered views from generated 3D Gaussian head avatars. FID measures the distributional similarity in the Inception-V3 feature space over 50K renders. Since standard FID evaluates only near-frontal views, we also report $\text{FID}_{3\text{D}}$ [6], in which camera poses are randomly sampled across a wider range of viewpoints to probe realism under arbitrary viewing angles. The results are summarized in Table 1 and 2. Under this minimal-resource setting, MVCHead achieves SOTA scores, producing visually coherent renders that remain plausible across diverse synthetic identities and viewpoints.



(a) NeRSemble (SIGGRAPH 23)
MEt3R↓ = 0.207 (High MVC)

(b) CAP4D (CVPR 25)
MEt3R↓ = 0.312 (Low MVC)

(c) MVCHead (Ours)
MEt3R↓ = 0.231 (Near-Natural MVC)

Figure 4. **Self-Renders provide strong MVC prior.** We evaluate MVC between view pairs from (a) studio-captured data [42], (b) intermediate view synthesis [69], and (c) self-renders from 3D. Using MASt3R [46] for estimating epipolar-consistent correspondence and FeatUp-DINO [8, 23] for measuring feature agreement with a view-invariant encoder, we compute a per-pixel consistency score map over the overlapping region. For each case, we visualize: *Left:* inputs; *Middle:* reprojected views A→B and B→A; *Right:* overlap mask and consistency map (dark = consistent, bright = inconsistent). MEt3R [4] is the spatial average of the error.

Table 1. **Perceptual Realism.** Comparison of FID scores. 512×512 resolution was used for the experiments. †Uses super-resolution network. *We report the results from the original paper.

Method	Venue	FID↓	
		FFHQ [40]	FFHQ-C [6]
StyleSDF [58]	CVPR 2022	11.2†	-
EpiGRAF [65]	NeurIPS 2022	9.92	-
VoxGRAF [62]	NeurIPS 2022	9.0	-
GMPI [82]	ECCV 2022	8.29	-
StyleNeRF [30]	ICLR 2022	7.80†	-
EG3D [9]	CVPR 2022	4.70†	-
Mimic3D [11]	ICCV 2023	5.37	-
GSGAN [38]	NeurIPS 2024	5.60	5.17
GSM [1]	CVPR 2024	28.19	-
GGHead [44]	SIGGRAPH Asia 24	5.15*	5.37
CGSGAN [6]	NeurIPS 2025	4.94	4.53
MVCHead	**Ours**	**4.39**	**3.94**

Table 2. **Perceptual Realism at extremes.** Comparison of $\text{FID}_{3\text{D}}$ scores. 512×512 resolution was used for the experiments.

Method	Venue	$\text{FID}_{3\text{D}}$ ↓	
		FFHQ [40]	FFHQ-C [6]
GSGAN [38]	NeurIPS 2024	10.50	7.68
GGHead [44]	SIGGRAPH Asia 24	7.90	7.78
CGSGAN [6]	NeurIPS 2025	4.94	4.53
MVCHead	**Ours**	**4.39**	**3.94**

Shape Consistency (CD↓, depth↓). We assess shape consistency using Chamfer Distance and depth error. For each generated identity, we construct two independent 3DGS representations, G_1 and G_2, by optimizing from two disjoint subsets of multi-view renders produced by the same avatar. Each 3DGS is then downsampled to a fixed-budget point cloud of 60K points, yielding P_1 and P_2, and we compute $e_{\text{cd}}(G_1, G_2) = d_{\text{CD}}(P_1, P_2)$. In addition, we render K depth maps $\pi_i^d(G)$ per 3DGS and measure a masked depth error e_d over the overlapping foreground regions across views. Intuitively, e_{cd} captures global shape discrepancies, while e_d is sensitive to local errors along silhouettes and fine structures. As reported in Table 3, MVCHead achieves lower CD, indicating improved global shape consistency. The depth error is comparable between the two methods, suggesting that local depth accuracy is similar.

Table 3. **Multi-view Consistency.** Consistency scores of the synthesized 3D Gaussian heads averaged over 100 avatars.

Method	**Shape** [73]		**Texture** [73]			**Geometric** [4]
	CD↓	depth↓	cPSNR↑	cSSIM↑	cLPIPS↓	MEt3R↓
CGSGAN [6]	0.6724	**6.6624**	21.852	0.7434	0.0622	0.2814
MVCHead	**0.6654**	6.6649	**22.082**	**0.7636**	**0.0528**	**0.2620**

Texture Consistency (cPSNR↑, cSSIM↑, cLPIPS↓). To evaluate cross-view texture stability, for each avatar we fit *two* independent 3DGS representations, G_1 and G_2, from *disjoint* multi-view subsets rendered from the same underlying avatar. We then render each 3DGS into K RGB images $\pi_i(G_1)$ and $\pi_i(G_2)$ under a fixed camera rig and compute MVC metrics between corresponding views: $e_m(G_1, G_2) = \frac{1}{K}\sum_{i=1}^{K} d_m\big(\pi_i(G_1), \pi_i(G_2)\big)$, where $m \in \{\text{cPSNR}, \text{cSSIM}, \text{cLPIPS}\}$ and d_m denotes the corresponding image-space metric. Since G_1 and G_2 are reconstructed from non-overlapping view subsets, they coincide only when textures are self-consistent across viewpoints; discrepancies reveal cross-view texture drift. These metrics quantify how well fine texture patterns, such as eyebrows, lip color, skin blemishes, and hair edges, remain stable under pose changes. As reported in Table 3, MVCHead exhibits strong texture consistency.

Geometric Consistency (MEt3R↓). To evaluate geometric consistency under larger camera changes, we adopt MEt3R [4], which measures MVC directly between image pairs without requiring known camera poses or 3D ground truth. Given a pair of self-renders (I_1, I_2) of a single avatar, we first use MASt3R [46] to obtain dense, pose-

Table 4. **Ablation Study.** Performed on the FFHQ-C [6] dataset with 512×512 resolution to verify the proposed components.

	Method	FID↓	MEt3R↓ [4]
	Full Model	**3.94**	**0.2620**
Loss	w/o $\mathscr{L}_{adv}$	collapse	
Loss	w/o $\mathscr{L}_{mvc}$	5.41	0.3144
Model Arch.	w/o HiSS block	5.28	0.2948
Model Arch.	w/o HiBiSS	4.78	0.2873

free stereo reconstructions $X_1, X_2 \in \mathbb{R}^{H\times W\times 3}$ in the coordinate frame of I_1. We then extract semantic features with DINO [8] and upsample them with FeatUp [23] to obtain high-resolution feature maps F_1 and F_2. Using the MASt3R point maps, these features are unprojected into 3D and reprojected into the frame of I_1, yielding aligned feature maps $\hat{F}_1$ and $\hat{F}_2$. A masked, pixel-wise cosine similarity between $\hat{F}_1$ and $\hat{F}_2$ over the overlapping region defines a directional consistency score $S(I_1, I_2)$. The final MEt3R(I_1, I_2) score is computed as $1 - 0.5 \cdot \big(S(I_1, I_2) + S(I_2, I_1)\big)$. We adapt this pipeline to head avatars by sampling camera pairs uniformly along yaw and pitch around a canonical rig and computing MEt3R over many random view pairs per identity. As reported in Table 3, MVCHead achieves a lower MEt3R score than SOTA, indicating stronger geometric consistency under large pose changes.

4.2. Ablation Study

To verify the effectiveness of each component, we perform an ablation study (see Table 4). Removing the adversarial loss $\mathscr{L}_{adv}$ leads to training collapse, confirming its necessity for maintaining image realism. Dropping the MVC loss $\mathscr{L}_{mvc}$ degrades both FID and MEt3R, underscoring the importance of the SE(3) Multi-view Critic for enforcing cross-view consistency. Removing `SS2D+LN+FFN` (i.e., the state space component) from each HiSS block results in a noticeable decline in MVC, confirming that the SSM contributes meaningfully beyond what attention alone provides. Finally, replacing HiBiSS with a standard unidirectional scan degrades performance, validating that axis-aligned, bidirectional recurrence is critical for reconciling multi-view drift.

4.3. Extensions

FaceGS-10K Dataset. To demonstrate a direct application of unconditional 3D head generation at scale, we construct *FaceGS-10K*—to our knowledge, the first large-scale dataset of ready-to-use 3D Gaussian head assets that is independent of any parametric 3D head model. Each asset contains 240K anisotropic Gaussians, along with 24 renderings over the frontal hemisphere at a resolution of 512×512. We generate the dataset by sampling diverse latent codes from the trained MVCHead model, retaining only identities that meet both a cross-view consistency threshold and a frontal realism filter. In contrast to purely 2D datasets [40], multi-view image collections without underlying 3D representations [42], or FLAME-registered meshes [50, 76], FaceGS-10K stores raw Gaussian attributes that can be directly rendered using off-the-shelf 3DGS rasterizers [41]. *FaceGS-10K* can support a range of downstream applications, including providing 3D supervision for reconstruction methods, and enabling privacy-preserving synthetic identity generation for AR/VR and content creation. More details in the supplementary material.

Conditional Generation. We adapt MVCHead for personalized avatar creation from a single image using optimization-based inversion. Given an input face image, we recover a latent code and minimize an ArcFace-based identity preservation loss [17]. Because the architecture remains unchanged, multi-view consistency (MVC) is naturally preserved in the personalized results. We emphasize that conditional generation is not the primary focus of this work; rather, it demonstrates that MVCHead is sufficiently structured to support inversion while preserving MVC. More details in the supplementary material.

5. Conclusion and Future Work

We present MVCHead, the *first* state space model for 3D Gaussian heads designed to address multi-view consistency (MVC) in the minimal-resource setting. MVCHead generates high-fidelity, multi-view consistent 3D head avatars in a single forward pass, achieving SOTA performance on five of six metrics. At its core are the HiSS block, which aligns SSM recurrence with the principal axes of drift via HiBiSS scanning, and the SE(3) Multi-view Critic, which enhances MVC by *design* without studio data or intermediate view synthesis. To our knowledge, this is the first comprehensive analysis of multi-view consistency in 3D head avatars.

Limitations. Despite its strengths, MVCHead has several limitations. First, it is only trained on front and side views, and cannot generate full 360° avatars; future work could add back-of-head coverage. Second, its geometric priors are learned entirely from 2D supervision. More explicit structural constraints (e.g., bilateral symmetry) could further reduce the search space. Additionally, harder negatives for the critic, e.g., geometrically perturbed views of the same identity, can further strengthen the consistency signal.

Acknowledgments. The computational resources were supported by PSC Bridges-2 through the Advanced Cyberinfrastructure Coordination Ecosystem: Services and Support (ACCESS) program allocation CIS250961. The authors thank Francisco Vicente Carrasco, Saswat Subhajyoti Mallick, Jianjin Xu, and José Pedro Gomes for their suggestions and feedback that improved the work.

References

[1] Rameen Abdal, Wang Yifan, Zifan Shi, Yinghao Xu, Ryan Po, Zhengfei Kuang, Qifeng Chen, Dit-Yan Yeung, and Gordon Wetzstein. Gaussian shell maps for efficient 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9441–9451, 2024. 7

[2] Shivangi Aneja, Artem Sevastopolsky, Tobias Kirschstein, Justus Thies, Angela Dai, and Matthias Nießner. Gaussianspeech: Audio-driven personalized 3d gaussian avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13065–13075, 2025. 1

[3] Shivangi Aneja, Sebastian Weiss, Irene Baeza, Prashanth Chandran, Gaspard Zoss, Matthias Niessner, and Derek Bradley. Scaffoldavatar: High-fidelity gaussian avatars with patch expressions. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025. 2, 3

[4] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6034–6044, 2025. 6, 7, 8

[5] Florian Barthel, Arian Beckmann, Wieland Morgenstern, Anna Hilsmann, and Peter Eisert. Gaussian splatting decoder for 3d-aware generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7963–7972, 2024. 2

[6] Florian Barthel, Wieland Morgenstern, Paul Hinzer, Anna Hilsmann, and Peter Eisert. Cgs-gan: 3d consistent gaussian splatting gans for high resolution human head synthesis. *arXiv preprint arXiv:2505.17590*, 2025. 2, 3, 4, 6, 7, 8

[7] Marcel C Buehler, Gengyan Li, Erroll Wood, Leonhard Helminger, Xu Chen, Tanmay Shah, Daoye Wang, Stephan Garbin, Sergio Orts-Escolano, Otmar Hilliges, et al. Cafca: High-quality novel view synthesis of expressive faces from casual few-shot captures. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 3

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 7, 8

[9] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini de Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022. 7

[10] Peng Chen, Xiaobao Wei, Qingpo Wuwu, Xinyi Wang, Xingyu Xiao, and Ming Lu. Mixedgaussianavatar: Realistically and geometrically accurate head avatar via mixed 2d-3d gaussian splatting. *arXiv preprint arXiv:2412.04955*, 2024. 2, 3

[11] Xingyu Chen, Yu Deng, and Baoyuan Wang. Mimic3d: Thriving 3d-aware gans via 3d-to-2d imitation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2338–2348. IEEE Computer Society, 2023. 7

[12] Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. Monogaussianavatar: Monocular gaussian point-based head avatar. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–9, 2024. 3

[13] Aviral Chharia, Wenbo Gou, and Haoye Dong. Mv-ssm: multi-view state space modeling for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11590–11599, 2025. 3

[14] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. *Advances in Neural Information Processing Systems*, 37:57642–57670, 2024. 3

[15] Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. Gpavatar: Generalizable and precise head avatar from image (s). *arXiv preprint arXiv:2401.10215*, 2024. 3

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[17] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 8

[18] Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7119–7130, 2024. 2, 3

[19] Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. In *European Conference on Computer Vision*, pages 316–333. Springer, 2024. 2, 3

[20] Helisa Dhamo, Yinyu Nie, Arthur Moreau, Jifei Song, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Headgas: Real-time animatable head avatars via 3d gaussian splatting. In *European Conference on Computer Vision*, pages 459–476. Springer, 2024. 2, 3

[21] Haoye Dong, Aviral Chharia, Wenbo Gou, Francisco Vicente Carrasco, and Fernando De la Torre. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. *arXiv preprint arXiv:2407.09646*, 2024. 3

[22] Wei-Qi Feng, Dong Han, Ze-Kang Zhou, Shunkai Li, Xiaoqiang Liu, Pengfei Wan, Di Zhang, and Miao Wang. Gpavatar: High-fidelity head avatars by learning efficient gaussian projections. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 250–259, 2025. 3

[23] Stephanie Fu, Mark Hamilton, Laura E. Brandt, Axel Feldmann, Zhoutong Zhang, and William T. Freeman. Featup: A model-agnostic framework for features at any resolution. In *The Twelfth International Conference on Learning Representations*, 2024. 7, 8

[24] Stathis Galanakis, Alexandros Lattas, Stylianos Moschoglou, Bernhard Kainz, and Stefanos Zafeiriou. Spinmeround: Consistent multi-view identity generation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14346–14356, 2025. 2, 3

[25] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Mononphm: Dynamic head reconstruction from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10747–10758, 2024. 3

[26] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Npga: Neural parametric gaussian avatars. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2, 3

[27] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024. 3, 4

[28] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 3

[29] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021. 3

[30] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *International Conference on Learning Representations*, 2022. 7

[31] Yuming Gu, Hongyi Xu, You Xie, Guoxian Song, Yichun Shi, Di Chang, Jing Yang, and Linjie Luo. Diffportrait3d: Controllable diffusion for zero-shot portrait view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10456–10465, 2024. 2, 3

[32] Yuming Gu, Phong Tran, Yujian Zheng, Hongyi Xu, Heyuan Li, Adilbek Karmanov, and Hao Li. Diffportrait360: Consistent portrait diffusion for 360 view synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26263–26273, 2025. 2, 3

[33] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational visual media*, 7(2):187–199, 2021. 4

[34] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25261–25270, 2025. 3

[35] Yisheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Weihao Yuan, Zilong Dong, and Liefeng Bo. Lam: Large avatar model for one-shot animatable gaussian head. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–13, 2025. 3

[36] Ding-Jiun Huang, Yuanhao Wang, Shao-Ji Yuan, Albert Mosella-Montoro, Francisco Vicente Carrasco, Cheng Zhang, and Fernando De la Torre. From blurry to believable: Enhancing low-quality talking heads with 3d generative priors. *arXiv preprint arXiv:2602.06122*, 2026. 1

[37] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 4

[38] Sangeek Hyun and Jae-Pil Heo. Gsgan: Adversarial learning for hierarchical generation of 3d gaussian splats. *Advances in Neural Information Processing Systems*, 37:67987–68012, 2024. 2, 3, 4, 6, 7

[39] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960. 3

[40] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 6, 7, 8

[41] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 3, 4, 8

[42] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 2, 3, 7, 8

[43] Tobias Kirschstein, Simon Giebenhain, and Matthias Nießner. Diffusionavatars: Deferred diffusion for high-fidelity 3d head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5481–5492, 2024. 3

[44] Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. Gghead: Fast and generalizable 3d gaussian heads. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2, 3, 7

[45] Tobias Kirschstein, Javier Romero, Artem Sevastopolsky, Matthias Nießner, and Shunsuke Saito. Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars. *arXiv preprint arXiv:2502.20220*, 2025. 1

[46] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 7

[47] Linzhou Li, Yumeng Li, Yanlin Weng, Youyi Zheng, and Kun Zhou. Rgbavatar: Reduced gaussian blendshapes for online modeling of head avatars. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10747–10757, 2025. 3

[48] Peng Li, Yisheng He, Yingdong Hu, Yuan Dong, Weihao Yuan, Yuan Liu, Siyu Zhu, Gang Cheng, Zilong Dong, and Yike Guo. Panolam: Large avatar model for gaussian full-head synthesis from one-shot unposed image. *arXiv preprint arXiv:2509.07552*, 2025. 3

[49] Shufan Li, Harkanwar Singh, and Aditya Grover. Mamba-nd: Selective state space modeling for multi-dimensional data. In *European Conference on Computer Vision*, pages 75–92. Springer, 2024. 3

[50] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 3, 8

[51] Tingting Liao, Yujian Zheng, Yuliang Xiu, Adilbek Karmanov, Liwen Hu, Leyang Jin, and Hao Li. Soap: Style-omniscient animatable portraits. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025. 2, 3

[52] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024. 3, 4, 5

[53] Zhibin Liu, Haoye Dong, Aviral Chharia, and Hefeng Wu. Human-vdm: Learning single-image 3d human gaussian splatting from video diffusion models. *arXiv preprint arXiv:2409.02851*, 2024. 1

[54] Weijie Lyu, Yi Zhou, Ming-Hsuan Yang, and Zhixin Shu. Facelift: Learning generalizable single image 3d face reconstruction from synthetic heads. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12691–12701, 2025. 2, 3

[55] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec avatar studio: Paired human captures for complete, driveable, and generalizable avatars. In *Advances in Neural Information Processing Systems*, pages 83008–83023. Curran Associates, Inc., 2024. 1

[56] Takeru Miyato, Bernhard Jaeger, Max Welling, and Andreas Geiger. Gta: A geometry-aware attention mechanism for multi-view transformers. In *International Conference on Learning Representations (ICLR)*, 2024. 5

[57] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 4

[58] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13503–13513, 2022. 7

[59] Antonio Oroz, Matthias Nießner, and Tobias Kirschstein. Perchead: Perceptual head model for single-image 3d head reconstruction & editing. *arXiv preprint arXiv:2511.02777*, 2025. 3

[60] Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, and Kwan-Yee Lin. Renderme-360: A large digital asset library and benchmarks towards high-fidelity head avatars. *Advances in Neural Information Processing Systems*, 36:7993–8005, 2023. 2, 3

[61] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 2, 3

[62] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. *Advances in Neural Information Processing Systems*, 35:33999–34011, 2022. 7

[63] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1606–1616, 2024. 3

[64] Qiuhong Shen, Zike Wu, Xuanyu Yi, Pan Zhou, Hanwang Zhang, Shuicheng Yan, and Xinchao Wang. Gamba: Marry gaussian splatting with mamba for single-view 3d reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3

[65] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. *Advances in Neural Information Processing Systems*, 35:24487–24501, 2022. 7

[66] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 4

[67] Jiapeng Tang, Davide Davoli, Tobias Kirschstein, Liam Schoneveld, and Matthias Niessner. Gaf: Gaussian avatar reconstruction from monocular videos via multi-view diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5546–5558, 2025. 3

[68] Felix Taubner, Ruihang Zhang, Mathieu Tuli, Sherwin Bahmani, and David B Lindell. Mvp4d: Multi-view portrait video diffusion for animatable 4d avatars. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–11, 2025. 3

[69] Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B Lindell. Cap4d: Creating animatable 4d portrait avatars with morphable multi-view diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5318–5330. IEEE Computer Society, 2025. 2, 3, 7

[70] Kartik Teotia, Hyeongwoo Kim, Pablo Garrido, Marc Habermann, Mohamed Elgharib, and Christian Theobalt. Gaussianheads: End-to-end learning of drivable gaussian head avatars from coarse-to-fine representations. *ACM Transactions on Graphics (TOG)*, 43(6):1–12, 2024. 2, 3

[71] Yating Wang, Xuan Wang, Ran Yi, Yanbo Fan, Jichen Hu, Jingcheng Zhu, and Lizhuang Ma. 3d gaussian head avatars with expressive dynamic appearances by compact tensorial representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21117–21126, 2025. 2, 3

[72] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022. 3

[73] Xianghui Xie, Chuhang Zou, Meher Gitika Karumuri, Jan Eric Lenssen, and Gerard Pons-Moll. Mvgbench: Comprehensive benchmark for multi-view generation models. *arXiv preprint arXiv:2507.00006*, 2025. 6, 7

[74] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2024. 3

[75] Peizhi Yan, Rabab Ward, Qiang Tang, and Shan Du. Gaussian déjà-vu: Creating controllable 3d gaussian head-avatars with enhanced generalization and personalization abilities. *arXiv preprint arXiv:2409.16147*, 2024. 2

[76] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 601–610, 2020. 8

[77] Xuanyu Yi, Zike Wu, Qiuhong Shen, Qingshan Xu, Pan Zhou, Joo-Hwee Lim, Shuicheng Yan, Xinchao Wang, and Hanwang Zhang. Mvgamba: Unify 3d content generation as state space sequence modeling. *Advances in Neural Information Processing Systems*, 37:7580–7607, 2024. 3

[78] Fei Yin, Chun-Han Yao, Rafal K Mantiuk, Varun Jampani, et al. Facecraft4d: Animated 3d facial avatar generation from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11612–11621, 2025. 2, 3

[79] Dongbin Zhang, Yunfei Liu, Lijian Lin, Ye Zhu, Kangjie Chen, Minghan Qin, Yu Li, and Haoqian Wang. Hravatar: High-quality and relightable gaussian head avatar. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26285–26296, 2025. 2

[80] Jiawei Zhang, Zijian Wu, Zhiyang Liang, Yicheng Gong, Dongfang Hu, Yao Yao, Xun Cao, and Hao Zhu. Fate: Full-head gaussian avatar with textural editing from monocular video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5535–5545, 2025. 3

[81] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 4

[82] Xiaoming Zhao, Fangchang Ma, David Güera, Zhile Ren, Alexander G Schwing, and Alex Colburn. Generative multiplane images: Making a 2d gan 3d-aware. In *European conference on computer vision*, pages 18–35. Springer, 2022. 7

[83] Xiaozheng Zheng, Chao Wen, Zhaohu Li, Weiyi Zhang, Zhuo Su, Xu Chang, Yang Zhao, Zheng Lv, Xiaoyuan Zhang, Yongjie Zhang, et al. Headgap: Few-shot 3d head avatar via generalizable gaussian priors. In *2025 International Conference on 3D Vision (3DV)*, pages 946–957. IEEE, 2025. 2

[84] Zhenglin Zhou, Fan Ma, Hehe Fan, and Tat-Seng Chua. Zero-1-to-a: Zero-shot one image to animatable head avatars using video diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15941–15952, 2025. 2, 3

[85] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 3