FFAvatar: Feed-Forward 4D Head Avatar Reconstruction from Sparse Portrait Images

Jianjiang Yao, Ke Xian*, Renxiang Dai, and Robert Caiming Qiu

School of Electronic Information and Communications,
Huazhong University of Science and Technology, Wuhan 430074, China
{jjyao, kxian, rxdai, caiming}@hust.edu.cn

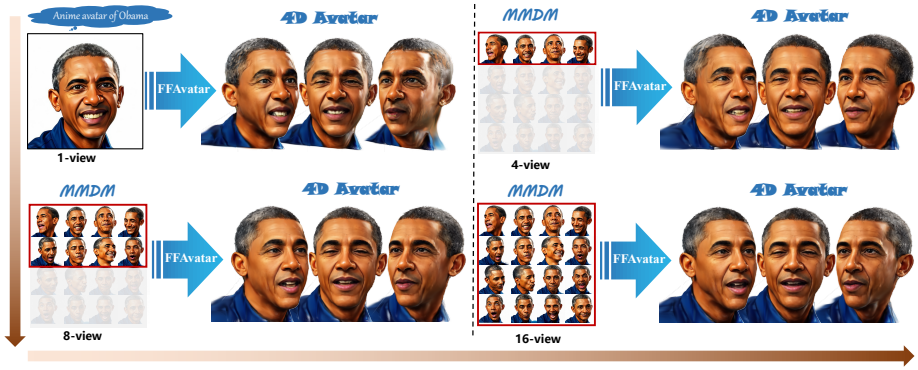

Fig. 1: FFAvatar. We present FFAvatar, a feed-forward framework that reconstructs high-fidelity 4D portrait avatars from one or more reference portrait images. The reference images can be virtual characters synthesized by large-scale generative models from text prompts. Furthermore, by leveraging diffusion-based MMDM [56] to synthesize novel-view and novel-expression 2D images, our approach further enhances reconstruction quality and view-consistent appearance.

Abstract. We present FFAvatar, a Transformer-based 3D Gaussian framework for fast construction of high-quality and animatable 4D head avatars from one or more reference portrait images. Unlike existing feed-forward approaches that require a fixed number of input views, FFAvatar supports *incremental reconstruction*, progressively refining the avatar representation as additional reference images become available. At the core of our method is an alternating attention mechanism that disentangles identity appearance from expression and viewpoint variations, enabling the reconstruction of a canonical 3D appearance that remains consistent across poses and facial expressions. To balance visual fidelity and computational efficiency, we introduce a sparse-to-dense learning paradigm. Coarse appearance features are first learned using sparse primitives anchored to the FLAME vertex level and are subsequently densified in the UV domain to capture fine-grained geometric and texture details.

* Corresponding author.

We further propose a plug-and-play motion refinement module that enables subject-specific dynamic personalization by modeling residual motion beyond parametric deformation. Extensive experiments demonstrate that FFAvatar efficiently produces high-fidelity and controllable 4D head avatars, achieving superior flexibility, driving efficiency, and identity-consistent rendering across diverse expressions and viewpoints. **Project Page:** https://jj-yao.github.io/ffavatar/

1 Introduction

High-quality 4D avatar head reconstruction plays a critical role in a wide range of applications, including virtual reality, digital humans, telepresence, and immersive content creation. An ideal 4D avatar head should faithfully preserve a person's identity while supporting realistic and temporally stable facial dynamics across diverse expressions, head poses, and viewpoints.

Despite the remarkable progress in neural rendering [47, 48] and avatar modeling [3, 19, 24, 30, 49], existing approaches still suffer from several key limitations. ❶ **Limited generalization under sparse observations.** Many existing methods rely on densely captured multi-view observations and perform object-specific optimization [1, 6, 22, 36, 37, 52, 63, 65, 66]. While these approaches can achieve high-quality reconstruction, they generalize poorly under single- or few-shot settings. Feed-forward 3D reconstruction provides a promising alternative for sparse-input scenarios. However, most existing feed-forward 4D avatar methods [8, 12, 13, 23, 70] mainly focus on the single-reference setting, where limited viewpoint cues often lead to incomplete geometry and ambiguous appearance. Some recent methods [28, 35, 62] extend feed-forward reconstruction to few-shot settings by constructing multiple view-dependent canonical subspaces from the input images. However, such a multi-canonical design tends to introduce redundant Gaussian representations and higher computational overhead as more reference views are used, limiting its scalability in sparse-view reconstruction. ❷ **Entanglement between identity appearance and motion.** Effectively disentangling identity-related appearance from facial expressions and head pose remains a challenging problem. In many existing approaches, identity features are intertwined with motion-related variations, which often leads to visual artifacts, identity drift, and rendering inconsistencies when synthesizing novel expressions or viewpoints [8, 23]. ❸ **Trade-off between representation fidelity and efficiency.** High-fidelity avatar rendering often relies on dense UV representations or a large number of 3D Gaussians [8, 23, 29, 33, 63]. Although these dense representations significantly improve rendering quality, they also introduce substantial memory consumption and computational overhead.

To address these challenges, we propose FFAvatar, a novel feed-forward framework for incremental 4D portrait avatar reconstruction from one or more reference portrait images. Our method is built upon three key strategies: ❶ **Flexible multi-view feature aggregation and unified canonical field construction.** To address these limitations, we introduce an alternating attention mechanism, inspired by VGGT [58], to aggregate identity-aware appearance informa-

tion from a variable number of reference images while disentangling expression and viewpoint variations. Instead of maintaining multiple view-dependent canonical subspaces, we leverage the FLAME prior [38] to construct a unified global canonical Gaussian field, where the aggregated global appearance representation is injected through cross-modal alignment. This unified design avoids redundant view-dependent representations and enables high-quality 4D head avatar reconstruction in a single forward pass, leading to improved flexibility, efficiency, and scalability under sparse observations. ❷ **Decoupled modeling of identity and motion.** We explicitly model identity appearance and motion variations separately. For identity modeling, the global identity representation produced by the alternating attention module is decoded into an expression- and view-independent 3D appearance. For motion modeling, we introduce a plug-and-play motion refinement module that captures dynamic motion patterns beyond the expressiveness of the FLAME template, enabling subject-specific dynamic personalization. ❸ **Sparse-to-dense hierarchical representation learning.** To balance representation fidelity and computational efficiency, we propose a sparse-to-dense learning paradigm. The model first learns coarse appearance features using sparse primitives anchored to the FLAME vertex level, and then progressively densifies the representation in the UV domain to capture high-frequency geometric and texture details. Compared with directly optimizing uniformly dense Gaussian representations, this hierarchical design significantly reduces computational cost while maintaining high rendering quality.

As illustrated in Fig. 1, our framework reconstructs high-fidelity 4D avatars from one or more reference portrait images, including AI-generated virtual portraits. Unlike existing feed-forward avatar head reconstruction methods [8, 12, 13,23,35,70], **FFAvatar supports incremental reconstruction**, allowing the avatar representation to be progressively refined as additional images become available. This property enables broader applicability in practical scenarios. For instance, by leveraging the generative capability of diffusion models [56, 61], our method can synthesize high-quality 4D avatars even under a single-image setting, which is difficult to achieve with previous feed-forward single-image reconstruction approaches [8, 12, 13, 23, 70].

Overall, we make the following contributions:

- We present FFAvatar, an incremental feed-forward framework that enables rapid reconstruction of high-fidelity 4D head avatars from one or more reference portrait images.
- We propose a sparse-to-dense hierarchical feature learning paradigm, learning coarse yet semantically stable features at the sparse FLAME vertex level and densifying them in UV space to capture high-frequency geometric and texture details, achieving a better trade-off between efficiency and fidelity.
- We introduce a motion refinement module that enables subject-specific dynamic personalization. Built upon parametric deformation, this module further enhances motion realism by refining identity-dependent dynamic details.

2 Related Work

2.1 Mesh-based Facial Reenactment

Mesh-based facial reenactment has been widely studied [7, 10, 14, 15, 17, 18, 32, 57, 68, 69, 71, 75] for explicit face animation using parametric 3D models. Existing methods can be broadly categorized into deformation-based and graphics-based approaches. Deformation-based methods warp source images or meshes by estimating explicit motion fields [14, 68, 69], based on mesh correspondences, geometry-guided dense flows, or sparse 3D landmark motions. However, they often degrade under large head rotations or significant geometric changes, where explicit deformation modeling becomes less robust. Graphics-based approaches [10, 17, 18, 75] instead reconstruct animatable head meshes and synthesize images through differentiable or hybrid rendering pipelines. Early works rely on parametric 3D morphable models to estimate geometry and appearance from monocular inputs, while recent learning-based methods recover detailed geometry, expression-dependent deformations, and identity-specific shapes from in-the-wild data. However, despite improved fidelity and expression realism, mesh-based pipelines [10, 17, 18, 26, 57, 69, 75] still depend on costly differentiable or hybrid rendering systems and often struggle to capture fine-grained facial details.

2.2 Feed-Forward 3D Reconstruction Models

Feed-forward 3D reconstruction methods aim to recover 3D geometry and appearance directly through a single forward pass of a neural network, avoiding expensive per-instance optimization at inference time [25, 41, 51, 53, 54, 58, 67]. Early works primarily focused on reconstructing coarse 3D shapes from single images using volumetric [4,20,76], point-cloud [74], or mesh-based [2,21,27,45,73] representations, demonstrating the feasibility of fast inference but often suffering from limited geometric detail and poor generalization to complex poses or expressions. With the advent of neural implicit representations [47, 48], feed-forward models have been extended to learn continuous geometry and appearance fields from images [25]. Approaches based on signed distance functions or neural radiance fields have significantly improved reconstruction quality, yet many of them [5, 44, 52, 59, 63, 74] still require per-subject fine-tuning or rely on multi-view inputs [35] during inference, limiting their practical efficiency. More recently, feed-forward neural rendering frameworks have explored hybrid representations that combine explicit geometry with learned appearance features [54, 58], enabling faster inference while maintaining high visual fidelity.

2.3 3D Avatar Head Reconstruction

Reconstructing high-quality 3D avatar heads with realistic appearance and expressive dynamics has long been a fundamental problem in computer vision and graphics [56]. Recent advances in 3D Gaussian Splatting [31] have significantly improved the realism of head reconstruction. By representing scenes with 3D

Gaussian ellipsoids, these methods can effectively model fine geometric details and view-dependent appearance. However, most existing approaches [1, 5, 36, 50, 63, 74] require per-subject optimization and dense multi-view data acquisition, resulting in high computational cost and limited scalability. In addition, identity appearance is often tightly coupled with facial expressions and head poses, which can lead to artifacts or identity drift when synthesizing novel expressions or viewpoints. To address these challenges, recent studies have explored hybrid representations that combine explicit geometric priors with learned appearance models. Examples include mesh-based neural textures and more recent Gaussian-based avatar representations, which enable efficient differentiable rendering and improved visual fidelity. Although these representations reduce rendering overhead, current methods still frequently rely on iterative optimization [5, 16, 36, 44, 52, 59, 63, 66, 74] or constrained input conditions [8, 9, 12, 13, 23, 39, 40, 42, 43, 70], limiting their ability to perform rapid reconstruction from unconstrained data. Some feed-forward methods [28,35,62] attempt to address this issue by constructing and fusing multiple view-dependent canonical subspaces from the input images. However, under multi-view input settings, their number of 3D Gaussians, creation time, and GPU memory consumption increase substantially with the number of input views.

In contrast, FFAvatar focuses on fast feed-forward 4D head avatar reconstruction by leveraging the FLAME prior to build a unified global canonical space. As a result, the number of Gaussians and the animation speed are independent of the number of input images. Our method can efficiently reconstruct a 4D head avatar in a feed-forward manner from hundreds of input images on a single A800 GPU. By explicitly disentangling identity appearance from expression and pose, and by introducing a Transformer-based global appearance aggregation mechanism within a unified canonical space, FFAvatar achieves efficient reconstruction while maintaining strong identity consistency and generalization to unseen expressions and viewpoints.

3 Method

As illustrated in Fig. 2, our FFAvatar framework consists of two main stages. (1) **Static Appearance Canonical Field Generation**: we extract a global appearance representation from multi-view and multi-expression images, perform sparse-to-dense cross-modal alignment, and decode an expression- and viewpoint-invariant static 3D appearance. (2) **3D Head Avatar Animation**: the static appearance is animated using the FLAME model, followed by motion refinement through a Motion-Aware Refinement Module to capture fine-grained dynamic details.

3.1 Static Appearance Canonical Field Generation

Given a set of head images $\mathcal{I} = \{I_i\}_{i=1}^{N}$ captured under different viewpoints and expressions, our goal is to reconstruct a viewpoint- and expression-invariant

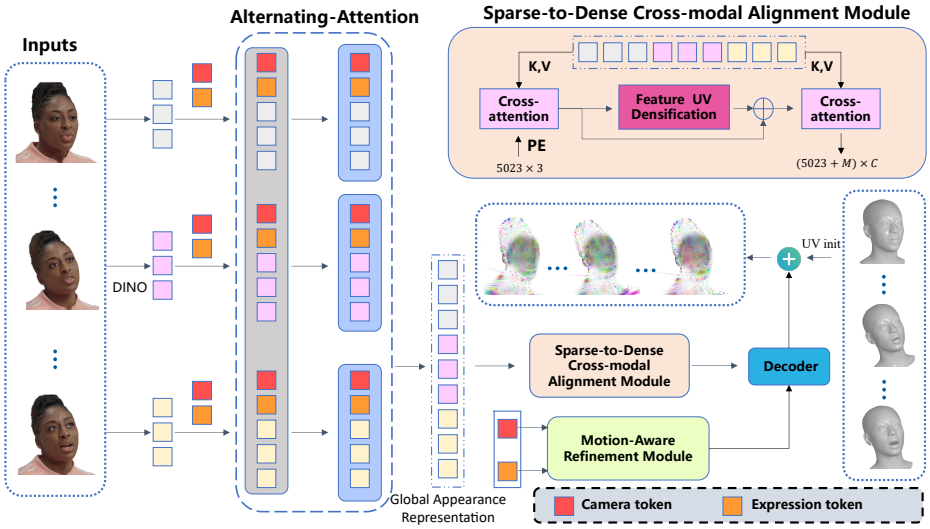

Fig. 2: Overview of FFAvatar. **Canonical field modeling.** Visual features extracted by DINOv3 are concatenated with camera pose and expression encodings. An alternating attention mechanism performs intra- and inter-image matching to infer a consistent global appearance representation across all inputs. This representation is then aligned with the FLAME template through the proposed Sparse-to-Dense Cross-modal Alignment Module, producing an expression- and viewpoint-invariant static 3D appearance. **Deformable field modeling.** Facial motions are driven by the FLAME model using standard linear blend skinning (LBS) and corrective blendshapes, and further refined by the proposed Motion-Aware Refinement Module to capture fine-grained, identity-dependent dynamic details.

canonical 3D head appearance field represented by 3D Gaussian primitives, denoted as G.

Feature Extraction and Global Appearance Aggregation. For each image I_i, we extract dense visual features using a pretrained encoder [55]:

$$I_i^{\text{feat}} = \textsc{DinoV3}(I_i) \in \mathbb{R}^{H \times W \times C} \tag{1}$$

Each feature map is augmented with camera and expression embeddings, denoted as $\mathbf{Z}^{\text{cam}}$ and $\mathbf{Z}^{\text{exp}}$, respectively, and processed by an alternating attention architecture that interleaves intra-image and inter-image attention. This design aggregates identity-consistent appearance cues across images while suppressing viewpoint- and expression-specific variations. The final global appearance representation is obtained by collecting tokens from all images:

$$\mathbf{Z}^{\text{app}} \leftarrow \textsc{AlterAtt}([\{I_i^{\text{feat}}\}_{i=1}^{N}, \mathbf{Z}^{\text{cam}}, \mathbf{Z}^{\text{exp}}]) \in \mathbb{R}^{NHW \times C} \tag{2}$$

Sparse-to-Dense Cross-modal Alignment. To balance computational efficiency and representation fidelity, we adopt a sparse-to-dense alignment strategy. We first initialize sparse Gaussian centers using FLAME template vertices

$\mathbf{V}_0 \in \mathbb{R}^{5023\times 3}$ and align them with the global appearance representation via cross-attention:

$$\mathbf{T}_s \leftarrow \textsc{CrossAtt}(\mathrm{PE}(\mathbf{V}_0), \mathbf{Z}^{\mathrm{app}}) \tag{3}$$

To recover high-frequency appearance details, we further densify features in the UV domain. Given a UV resolution uv_size, the FLAME mesh is rasterized into a planar UV grid of size uv_size × uv_size. Each valid UV location corresponds to a triangle on the mesh and is associated with barycentric weights defined over its three vertices. Dense UV features are obtained by interpolating the aligned sparse features:

$$\mathbf{T}_{\mathrm{uv}}(u,v) = \alpha\,\mathbf{T}_s(i) + \beta\,\mathbf{T}_s(j) + \gamma\,\mathbf{T}_s(k) \tag{4}$$

where (i, j, k) denote the triangle vertex indices and (α, β, γ) are the corresponding barycentric coefficients. Flattening all valid UV samples yields dense features $\mathbf{T}_{\mathrm{uv}} \in \mathbb{R}^{M\times C}$, where $M < \mathrm{uv\_size}^2$ denotes the number of valid UV locations.

We then combine sparse vertex-aligned features and dense UV features to form a unified multi-resolution representation. While UV densification introduces fine-grained spatial detail, it does not explicitly enforce global appearance consistency across multiple observations. Therefore, we perform a second-stage cross-modal alignment to further refine the fused representation using the global appearance tokens:

$$\mathbf{T}_d \leftarrow \textsc{CrossAtt}\left([\mathbf{T}_{\mathrm{uv}}, \mathbf{T}_s], \mathbf{Z}^{\mathrm{app}}\right) \tag{5}$$

This hierarchical alignment progressively transfers identity-consistent appearance information from global image observations to both sparse structural anchors and dense surface samples. As a result, the refined feature set $\mathbf{T}_d$ captures semantically stable coarse geometry together with high-frequency surface detail, yielding an efficient and expressive representation for canonical 3D Gaussian field decoding.

Canonical 3D Gaussian Field Decoding. The refined feature set $\mathbf{T}_d$ is decoded into a canonical 3D Gaussian head representation using a feed-forward decoder $\mathcal{D}$static. For each Gaussian primitive, the decoder predicts both geometric and appearance attributes:

$$G = \{f_n, o_n, c_n, s_n, r_n\}_{n=1}^{M+5023} = \mathcal{D}_{\mathrm{static}}(\mathbf{T}_d) \tag{6}$$

here, $\mathbf{f}_n \in \mathbb{R}^3$ denotes the positional offset relative to the underlying FLAME head template, $o_n \in \mathbb{R}$ represents opacity, $\mathbf{c}_n \in \mathbb{R}^3$ encodes color, $\mathbf{s}_n \in \mathbb{R}^3$ parameterizes anisotropic scaling, and $\mathbf{r}_n \in \mathbb{R}^4$ denotes rotation represented as a quaternion. The resulting set of M+5023 Gaussian primitives defines a viewpoint- and expression-invariant canonical appearance field, which serves as the static structural basis for subsequent dynamic animation.

3.2 3D Head Avatar Animation

Given the canonical static 3D Gaussian head representation G obtained in Section 3.1, we animate the avatar under arbitrary facial expressions and head poses

by combining the FLAME parametric head model [38] with a Motion-Aware Refinement Module (MARM).

FLAME-based Mesh Deformation. FLAME parameterizes head geometry using pose and expression coefficients. For each animation frame, the deformation parameters are defined as

$$\boldsymbol{\theta} = \{\boldsymbol{\theta}^{\text{pose}}, \boldsymbol{\theta}^{\text{exp}}\}, \tag{7}$$

where $\boldsymbol{\theta}^{\text{pose}}$ models rigid head motion and jaw articulation, and $\boldsymbol{\theta}^{\text{exp}}$ controls non-rigid facial expressions. Given the canonical FLAME template mesh $\mathbf{V}_0$, the posed mesh is obtained via linear blend skinning (LBS) with corrective blendshapes:

$$\mathbf{V}(\boldsymbol{\theta}) = \text{LBS}(\mathbf{V}_0, \boldsymbol{\theta}) \tag{8}$$

Following FLAME deformation, we propagate the vertex-level motion to a dense set of surface points using the same UV rasterization strategy as in the canonical construction. Specifically, the deformed mesh $\mathbf{V}(\boldsymbol{\theta})$ is unwrapped into the UV domain, yielding a set of valid UV pixels Ω_{uv} with cardinality M. Each UV pixel $(u, v) \in \Omega_{\text{uv}}$ is associated with barycentric weights $\boldsymbol{\alpha}(u, v) = (\alpha, \beta, \gamma)$ and vertex indices (i, j, k) on the deformed mesh. The corresponding dense surface position is computed as:

$$\tilde{\mathbf{x}}_n = \alpha\, \mathbf{V}_i(\boldsymbol{\theta}) + \beta\, \mathbf{V}_j(\boldsymbol{\theta}) + \gamma\, \mathbf{V}_k(\boldsymbol{\theta}), \quad n = 1, \ldots, M. \tag{9}$$

In addition to the densified surface points, we also retain the original deformed FLAME vertices $\{\mathbf{V}_m(\boldsymbol{\theta})\}_{m=1}^{5023}$ to ensure geometric consistency. By concatenating both sets, we obtain a pose-dependent coarse surface representation with M+5023 points:

$$\tilde{\mathbf{X}}(\boldsymbol{\theta}) = \left[\{\tilde{\mathbf{x}}_n\}_{n=1}^{M}, \{\mathbf{V}_m(\boldsymbol{\theta})\}_{m=1}^{5023}\right] \tag{10}$$

Gaussian Position Update. Each Gaussian primitive is anchored to a corresponding point in $\tilde{\mathbf{X}}(\boldsymbol{\theta})$. During animation, FLAME provides coarse, physically consistent surface motion, while the learned canonical Gaussian offsets are preserved to retain identity-specific geometry. Consequently, the Gaussians follow pose- and expression-driven deformation while maintaining personalized shape and fine-scale details.

Motion-Aware Refinement. While FLAME provides physically plausible coarse deformation, it cannot fully capture identity-specific nonlinear motion patterns, such as subtle muscle dynamics or personalized expression styles. To compensate for these limitations, we introduce a MARM that predicts residual Gaussian updates conditioned on the current animation state. Specifically, we learn a motion refinement network $\mathcal{R}$motion that takes as input the positional encoding of canonical FLAME vertices, the expression and pose parameters $\boldsymbol{\theta}$, and the camera parameters $\boldsymbol{\theta}^{\text{cam}}$. The module predicts residual Gaussian attribute offsets:

$$\Delta\mathbf{g} \leftarrow \mathcal{R}\text{motion}(\text{PE}(\mathbf{V}_0), \boldsymbol{\theta}, \boldsymbol{\theta}^{\text{cam}}) \tag{11}$$

Here, $\Delta\mathbf{g}$ models motion-dependent corrections beyond template-driven deformation, allowing the system to refine geometry and appearance in a view-aware manner. This design avoids directly regressing full Gaussian parameters, instead focusing on compact residual updates that preserve canonical identity while adapting to dynamic conditions. The final animated Gaussian field is obtained by applying the predicted residuals to the FLAME-deformed canonical representation:

$$G(\boldsymbol{\theta}) = G' \oplus \Delta\mathbf{g} \tag{12}$$

where G' denotes the FLAME-driven coarse Gaussian field and $\oplus$ represents attribute-wise updates.

4 Experiments

4.1 Experimental Settings

Implementation Details. Our framework is implemented in PyTorch and optimized using Adam with a learning rate of 4.0×10^{-5} for 300,000 iterations. The DINOv3 [55] feature backbone is frozen during training, while all other components are optimized end-to-end. For each batch, we randomly sample 1∼8 frames from a monocular video to form the input set. The sparse-to-dense cross-modal alignment module and the motion-aware refinement module are implemented using Transformer architectures. We use the GAGAvatar tracker [8] to extract camera-pose and facial-expression conditions, which are used to control the viewpoint and expression of the reconstructed avatar during animation. More implementation details are provided in the supplementary material.

Baselines. In the single-reference setting, we compare with Portrait4D-v2 [13], Real3D-Portrait [70], GAGAvatar [8], and LAM [23]. In the few-reference setting, we further compare with the optimization-based methods FlashAvatar [63], GHA [66], and GaussianAvatars [52], as well as the feed-forward approaches GAPAvatar [9] and FastAvatar [62]. We do not include Avat3r [35] and Fast-GHA [28] in the main comparison because their official implementations are not publicly available.

Datasets and Evaluation. We train the model on the VFHQ dataset [64], which contains 15,204 monocular video clips (about 3M frames). For each extracted frame, we detect the facial region, enlarge the bounding box to include sufficient contextual information, and crop the region of interest. To enhance data usability, we perform camera pose estimation and FLAME parameter tracking for each frame, following a pipeline similar to GAGAvatar [8]. All cropped images are resized to 512×512 pixels for consistency. In addition, background removal is applied to separate the subject from the background. On the VFHQ test set, 8 expressions are randomly selected as inputs, while the remaining frames are used for evaluation. We further evaluate on the NeRSemble dataset [34] under two settings: *novel view synthesis* (8 input views and 8 unseen views) and *novel expression synthesis* (16 input expressions with unseen poses and expressions). Rendering quality is measured using PSNR, SSIM [60], and LPIPS [72], while

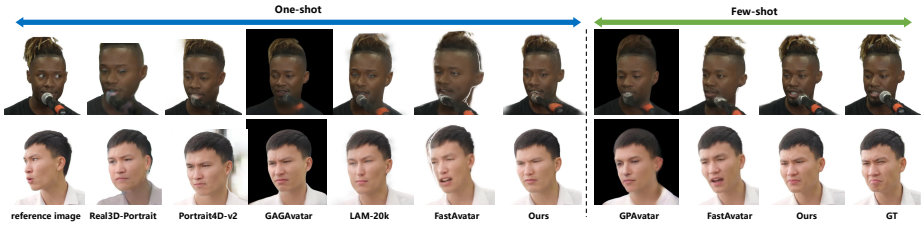

Fig. 3: Novel expression synthesis. Qualitative comparison of novel expression synthesis on the VFHQ monocular test set. We compare our method with Real3D-Portrait [70], Portrait4D-v2 [13], GAGAvatar [8], LAM [23], GPAvatar [9], and FastAvatar [62].

Table 1: Quantitative comparison on the VFHQ dataset under the novel expression setting. ↑ indicates higher is better, ↓ indicates lower is better. We mark the **best** and second-best results.

Method	Setting	Novel Expressions						Efficiency	
		PSNR↑	SSIM↑	LPIPS↓	CSIM↑	AED↓	APD↓	Creation in↓	FPS↑
Real3DPortrait [70]	One-shot	20.88	0.780	0.154	0.750	0.150	0.268	3.5s	15
Portrait4D-v2 [13]		21.34	0.794	0.144	0.717	0.117	0.187	2.9s	11
GAGAvatar [8]		21.83	0.818	0.128	0.816	0.111	0.135	1.6s	63
LAM [23]		22.65	0.829	0.109	0.822	0.102	0.134	1.1s	219
FastAvatar [62]		17.85	0.813	0.167	0.679	0.136	0.328	2.6s	339
Ours		21.82	0.843	0.108	0.817	0.109	0.149	1.3s	31
GPAvatar [9]	Few-shot	22.91	0.795	0.154	0.765	0.138	0.189	**0.7**s	5
FastAvatar [62]		18.12	0.819	0.153	0.781	0.116	0.321	12.2s	97
Ours(fast)		23.20	0.862	0.088	0.852	0.084	0.117	2.1s	**468**
Ours		**23.35**	**0.864**	**0.081**	**0.861**	**0.079**	**0.114**	2.1s	31

identity consistency and motion accuracy are evaluated using CSIM [11], AED, and APD. To provide a more complete evaluation against recent feed-forward avatar reconstruction methods, we also conduct qualitative comparisons with Avat3r [35] and FastGHA [28] on the Ava-256 dataset [46] using the same color-calibrated input data.

4.2 Head Avatar Reconstruction

Quantitative and Qualitative Results on VFHQ. Table 1 reports quantitative results on the VFHQ monocular test set under the novel-expression setting. Compared with one-shot feed-forward reconstruction methods, a key advantage of our approach lies in its ability to flexibly incorporate multiple reference images for incremental reconstruction. Benefiting from this capability, under the few-shot setting our method achieves the best performance across all image quality metrics (PSNR, SSIM, and LPIPS) for novel-expression rendering, while also obtaining higher identity consistency (CSIM) and better motion fidelity (AED and APD). Efficiency comparisons further highlight the practicality of our approach.

Fig. 4: Novel View Synthesis. Qualitative comparison of novel view synthesis results on the NeRsemble multi-view subset. For the single-reference setting, we compare our method with Portrait4D-v2 [13], GAGAvatar [8], LAM [23], and FastAvatar [62]. For the multi-reference setting, we further compare against GaussianAvatars [52], FlashAvatar [63], GHA [66], and GPAvatar [9].

Table 2: Quantitative comparison on the NeRSemble dataset under novel-view and novel-expression settings. ↑ indicates higher is better, ↓ indicates lower is better. We mark the **best** and second-best results.

Method	Setting	Novel Views			Novel Expressions		
		LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑
Real3DPortrait [70]	One-shot	0.197	0.785	16.22	0.165	0.821	17.48
Portrait4D-v2 [13]		0.172	0.797	16.81	0.152	0.814	18.24
GAGAvatar [8]		0.129	0.833	**22.52**	0.095	0.857	**25.87**
LAM-20K [23]		0.175	0.819	16.43	0.122	0.834	20.55
FastAvatar [62]		0.232	0.800	14.78	0.185	0.821	19.41
Ours		0.121	0.839	19.18	0.106	0.851	20.23
FlashAvatar [63]	Train From Scratch	0.209	0.785	17.84	0.221	0.764	16.94
GHA [66]		0.269	0.722	13.93	-	-	-
GaussianAvatars [52]		0.164	0.813	17.99	0.178	0.822	17.56
GPAvatar [9]	Few-shot	0.163	0.822	22.26	0.154	0.829	22.58
FastAvatar [62]		0.158	0.824	20.11	0.135	0.845	22.49
Ours		**0.098**	**0.858**	21.95	**0.075**	**0.881**	24.08

The fast variant (without the motion refinement module) achieves real-time rendering at 468 FPS, significantly surpassing all baselines while maintaining competitive visual quality. Although the full model prioritizes rendering fidelity, it remains computationally efficient and does not require per-subject optimization, resulting in only moderate avatar creation time. Qualitative comparisons are shown in Fig. 3.

Quantitative and Qualitative Results on NeRSemble. It is worth noting that our model is trained exclusively on the VFHQ monocular dataset. The NeRSemble multi-view dataset is not used during training and serves solely

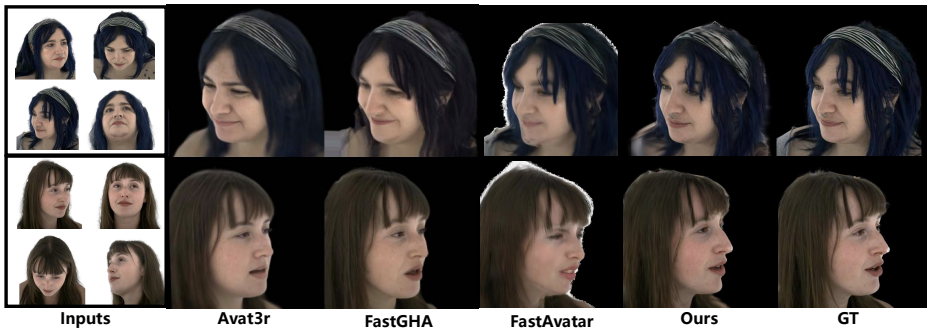

Fig. 5: Qualitative comparisons with FastGHA and Avat3r. Using the same color-calibrated Ava-256 inputs [46], FFAvatar preserves sharper facial details and more stable identity consistency under comparable reference-input settings.

for evaluation. Table 2 presents quantitative comparisons on the NeRSemble multi-view dataset under both novel-view and novel-expression settings. In the one-shot scenario, our method achieves competitive performance compared with existing feed-forward approaches, outperforming most baselines in perceptual quality metrics while maintaining stable rendering quality across unseen viewpoints and expressions. More importantly, when multiple reference images are available, our framework demonstrates clear advantages. Under the few-shot setting, our method achieves the best performance in LPIPS and SSIM for both novel-view and novel-expression synthesis, indicating superior perceptual fidelity and structural consistency. Although GAGAvatar reports slightly higher PSNR in the one-shot setting, our method provides a better overall balance between perceptual quality and geometric consistency, particularly when leveraging multiple input views. Figure 4 further presents qualitative comparisons for novel-view synthesis. Compared with Gaussian-based or train-from-scratch baselines, our method produces sharper facial structures, fewer view-dependent artifacts, and more coherent multi-view appearance, whereas competing approaches often exhibit blur or geometric inconsistencies.

Qualitative Comparisons on Ava-256. As discussed in the main paper, we do not include Avat3r [35] and FastGHA [28] in the main quantitative comparison because their official implementations are not publicly available. For completeness, Fig. 5 provides additional qualitative comparisons on the Ava-256 dataset [46]. Following the meta-review suggestion, we conduct the comparison using the same color-calibrated input data for all methods, ensuring a consistent input setting. These results complement the main evaluation and provide a visual reference for recent feed-forward avatar reconstruction methods. Compared with FastGHA and Avat3r, FFAvatar better preserves fine facial details and maintains more stable identity consistency across the rendered views.

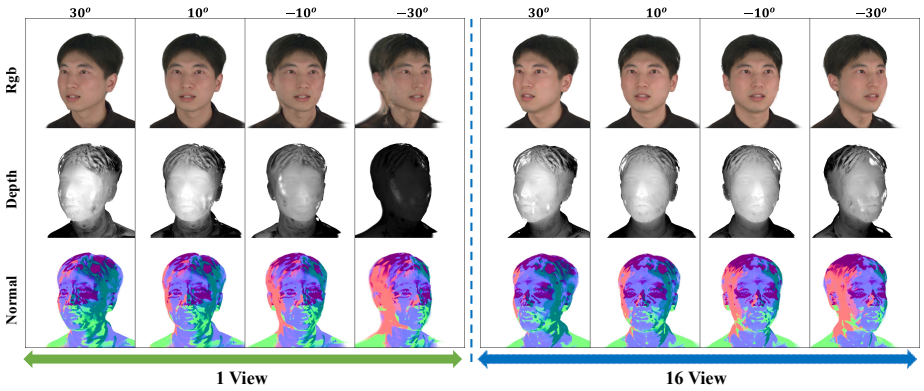

Fig. 6: Multi-view rendering results of our method under different input conditions: single-image input (left) and 16-image input (right). The rendered viewpoints correspond to yaw angles of 30°, 10°, −10°, and −30°. From top to bottom, we show the RGB renderings, depth maps, and surface normal maps produced by 3D Gaussian splatting.

4.3 Ablation Studies

Incremental Reconstruction. FFAvatar supports incremental reconstruction from one or more reference portrait images with diverse expressions and viewpoints. As shown in Table 3, reconstruction quality improves as the number of input frames increases. The most notable gain is observed when increasing the inputs from 1 to 8 frames, suggesting that additional multi-view and multi-expression cues are particularly beneficial for identity stabilization and motion alignment. Further increasing the number of inputs to 16 or 32 frames brings smaller but consistent improvements, indicating that FFAvatar can effectively integrate additional observations. Figure 6 provides qualitative comparisons. With only a single input image, the reconstructed avatar already preserves plausible geometry and appearance, although minor artifacts and view-dependent inconsistencies may appear under large pose changes. When more input images are provided, such as 16 frames, facial structures become sharper, view transitions are smoother, and identity preservation becomes more stable across viewpoints.
Sparse-to-Dense Learning Paradigm. We further evaluate the effectiveness of the proposed sparse-to-dense appearance modeling strategy. Our design first learns coarse yet semantically stable appearance features at the sparse FLAME vertex level, and then densifies the representation in UV space to capture high-frequency geometric and photometric details. Without such hierarchical modeling, directly optimizing sparse primitives leads to insufficient representation capacity for fine structures (e.g., hair strands and subtle facial textures), whereas directly optimizing dense Gaussian primitives significantly increases training cost and memory consumption, and may result in suboptimal deformation due to insufficiently constrained optimization. As shown in Table 4, the sparse-only model

Table 3: Ablation study on the number of input images under the novel expression setting on the VFHQ dataset. We analyze reconstruction quality, motion fidelity, identity consistency, and creation time as the number of input frames increases.

Method	Inputs	Novel Expressions						Efficiency
		PSNR↑	SSIM↑	LPIPS↓	CSIM↑	AED↓	APD↓	Creation in↓
Ours	1	21.82	0.843	0.108	0.817	0.109	0.149	**1.3**s
	4	22.75	0.858	0.091	0.852	0.094	0.119	1.7s
	8	23.35	0.864	0.081	0.861	0.079	0.114	2.1s
	16	23.36	0.866	0.080	0.872	0.078	**0.110**	4.3s
	32	**23.38**	**0.867**	**0.077**	**0.874**	**0.078**	0.111	11.6s

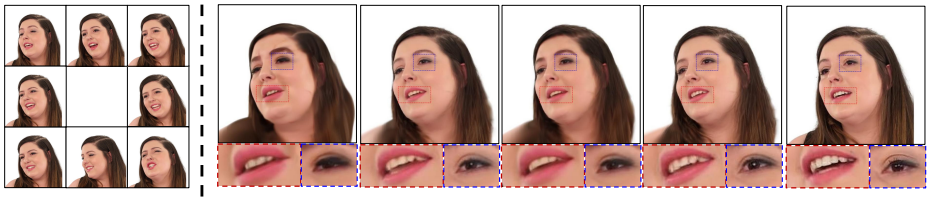

Fig. 7: Ablation study of key design components. (a) Optimization using the original FLAME vertices as Gaussian primitives results in insufficient texture detail in several regions; (b) direct optimization of dense point clouds leads to suboptimal deformation, with some primitives not well optimized; (c) removing the motion refinement module causes inaccurate dynamic motion representation; (d) final results of the full model, producing the most detailed appearance and accurate motion.

(S-5K) exhibits clear performance degradation across all quality and motion metrics, indicating limited representational capability. In contrast, the dense-only optimization (D-20K, D-64K) improves certain reconstruction metrics but incurs substantially higher training time and GPU memory usage, and even becomes infeasible at higher resolutions due to memory overflow. Qualitative comparisons in Fig. 7 further support these findings. Direct sparse optimization produces overly smooth appearance with missing high-frequency details, while direct dense optimization leads to unstable or imperfect deformation in certain regions. In contrast, the proposed sparse-to-dense paradigm enables both detailed appearance modeling and an efficient training scheme, achieving a more favorable trade-off between reconstruction accuracy and computational efficiency.

Motion Refinement Module. We evaluate the effectiveness of the motion refinement module, which enhances dynamic deformation beyond the FLAME template. As shown in Table 1, enabling motion refinement (Ours) consistently improves perceptual quality, identity consistency, and motion accuracy compared to the fast variant without refinement (Ours(fast)), while maintaining

Table 4: Effect of the sparse-to-dense generation strategy. S denotes optimization using the original FLAME vertex set (5,023 points), D denotes direct optimization of dense point clouds, and S2D represents the proposed sparse-to-dense generation process. "–" indicates the number of Gaussian primitives. OOM indicates out-of-memory under a 40 GB GPU memory constraint. Time refers to the wall-clock time required for 1,000 training iterations.

Method	UV Resolution	Novel Expressions						Train	
		PSNR↑	SSIM↑	LPIPS↓	CSIM↑	AED↓	APD↓	Time↓	Memory↓
S-5K	-	19.69	0.837	0.174	0.689	0.131	0.141	40min	22.5GB
S2D-20K	128	23.35	0.864	0.081	0.861	0.079	0.114	47min	25.8GB
D-20K	128	23.41	0.869	0.097	0.858	0.087	0.115	120min	38.2GB
S2D-64K	256	**23.42**	**0.871**	**0.077**	**0.866**	**0.078**	**0.111**	72min	37.1GB
D-64K	256	-	-	-	-	-	-	-	OOM

similar creation time. Fig. 7(c) further shows that removing this module leads to less accurate and oversimplified motion, whereas the full model produces more natural and detailed dynamic behavior.

5 Conclusion

We present **FFAvatar**, an incremental feed-forward framework for efficient and high-fidelity reconstruction of 4D avatar heads from sparse and heterogeneous image inputs. Through an alternating attention mechanism, our method flexibly disentangles identity information from expression and viewpoint across a variable number of reference images, enabling the learning of a stable canonical 3D appearance and ensuring strong cross-view and cross-expression consistency. We propose a sparse-to-dense learning paradigm, where sparse point representations first capture the coarse geometry of the 3D avatar and are subsequently densified in UV space to recover fine-grained geometric and texture details, achieving a better balance between fidelity and efficiency. In addition, a motion refinement module is incorporated to further enhance the realism of dynamic expressions. Overall, FFAvatar provides a flexible, efficient, and robust solution for reconstructing controllable 4D avatars from one or more portrait images.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62406120, the Hubei Provincial Natural Science Foundation of China under Grant No. 2026AFB533 and the CCF-Zhipu Large Model Innovation Fund (NO.CCF-Zhipu202411).

References

1. Aneja, S., Sevastopolsky, A., Kirschstein, T., Thies, J., Dai, A., Nießner, M.: Gaussianspeech: Audio-driven gaussian avatars. arXiv preprint arXiv:2411.18675 (2024)
2. Athar, S., Saito, S., Yang, Z., Pidhorskyi, S., Cao, C.: Bridging the gap: Studio-like avatar creation from a monocular phone capture. In: European Conference on Computer Vision. pp. 72–88. Springer (2024)
3. Athar, S., Xu, Z., Sunkavalli, K., Shechtman, E., Shu, Z.: Rignerf: Fully controllable neural 3d portraits. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 20364–20373 (2022)
4. Bai, Z., Tan, F., Huang, Z., Sarkar, K., Tang, D., Qiu, D., Meka, A., Du, R., Dou, M., Orts-Escolano, S., et al.: Learning personalized high quality volumetric head avatars from monocular rgb videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16890–16900 (2023)
5. Chen, Y., Wang, L., Li, Q., Xiao, H., Zhang, S., Yao, H., Liu, Y.: Monogaussianavatar: Monocular gaussian point-based head avatar. In: ACM SIGGRAPH 2024 conference papers. pp. 1–9 (2024)
6. Cho, K., Lee, J., Yoon, H., Hong, Y., Ko, J., Ahn, S., Kim, S.: Gaussiantalker: Real-time talking head synthesis with 3d gaussian splatting. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 10985–10994 (2024)
7. Chu, X., Goswami, N., Cui, Z., Wang, H., Harada, T.: Artalk: Speech-driven 3d head animation via autoregressive model. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–9 (2025)
8. Chu, X., Harada, T.: Generalizable and animatable gaussian head avatar. Advances in Neural Information Processing Systems **37**, 57642–57670 (2024)
9. Chu, X., Li, Y., Zeng, A., Yang, T., Lin, L., Liu, Y., Harada, T.: Gpavatar: Generalizable and precise head avatar from image (s). arXiv preprint arXiv:2401.10215 (2024)
10. Daněček, R., Black, M.J., Bolkart, T.: Emoca: Emotion driven monocular face capture and animation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20311–20322 (2022)
11. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
12. Deng, Y., Wang, D., Ren, X., Chen, X., Wang, B.: Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7119–7130 (2024)
13. Deng, Y., Wang, D., Wang, B.: Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. In: European Conference on Computer Vision. pp. 316–333. Springer (2024)
14. Doukas, M.C., Zafeiriou, S., Sharmanska, V.: Headgan: One-shot neural head synthesis and editing. In: Proceedings of the IEEE/CVF International conference on Computer Vision. pp. 14398–14407 (2021)
15. Drobyshev, N., Chelishev, J., Khakhulin, T., Ivakhnenko, A., Lempitsky, V., Zakharov, E.: Megaportraits: One-shot megapixel neural head avatars. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 2663–2671 (2022)
16. Feng, W.Q., Han, D., Zhou, Z.K., Li, S., Liu, X., Wan, P., Zhang, D., Wang, M.: Gpavatar: High-fidelity head avatars by learning efficient gaussian projections. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 250–259 (2025)

17. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. ACM Transactions on Graphics (ToG) **40**(4), 1–13 (2021)
18. Filntisis, P.P., Retsinas, G., Paraperas-Papantoniou, F., Katsamanis, A., Roussos, A., Maragos, P.: Visual speech-aware perceptual 3d facial expression reconstruction from videos. arXiv preprint arXiv:2207.11094 (2022)
19. Gafni, G., Thies, J., Zollhofer, M., Nießner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8649–8658 (2021)
20. Giebenhain, S., Kirschstein, T., Georgopoulos, M., Rünz, M., Agapito, L., Nießner, M.: Mononphm: Dynamic head reconstruction from monocular videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10747–10758 (2024)
21. Grassal, P.W., Prinzler, M., Leistner, T., Rother, C., Nießner, M., Thies, J.: Neural head avatars from monocular rgb videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18653–18664 (2022)
22. Guo, Y., Chen, K., Liang, S., Liu, Y.J., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5784–5794 (2021)
23. He, Y., Gu, X., Ye, X., Xu, C., Zhao, Z., Dong, Y., Yuan, W., Dong, Z., Bo, L.: 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. pp. 1–13 (2025)
24. Hong, Y., Peng, B., Xiao, H., Liu, L., Zhang, J.: Headnerf: A real-time nerf-based parametric head model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20374–20384 (2022)
25. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. arXiv preprint arXiv:2311.04400 (2023)
26. Huang, Z., Shi, M., Liu, C., Xian, K., Cao, Z.: Simhmr: A simple query-based framework for parameterized human mesh reconstruction. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 6918–6927 (2023)
27. Ichim, A.E., Bouaziz, S., Pauly, M.: Dynamic 3d avatar creation from hand-held video input. ACM Transactions on Graphics (ToG) **34**(4), 1–14 (2015)
28. Ji, X., Weiss, S., Kansy, M., Naruniec, J., Cao, X., Solenthaler, B., Bradley, D.: Fastgha: Generalized few-shot 3d gaussian head avatars with real-time animation. In: The Fourteenth International Conference on Learning Representations (2026)
29. Jiang, Y., Liao, Q., Li, X., Ma, L., Zhang, Q., Zhang, C., Lu, Z., Shan, Y.: Uv gaussians: Joint learning of mesh deformation and gaussian textures for human avatar modeling. Knowledge-Based Systems **320**, 113470 (2025)
30. Kania, K., Yi, K.M., Kowalski, M., Trzciński, T., Tagliasacchi, A.: Conerf: Controllable neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18623–18632 (2022)
31. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
32. Khakhulin, T., Sklyarova, V., Lempitsky, V., Zakharov, E.: Realistic one-shot mesh-based head avatars. In: European Conference on Computer Vision. pp. 345–362. Springer (2022)
33. Kirschstein, T., Giebenhain, S., Tang, J., Georgopoulos, M., Nießner, M.: Gghead: Fast and generalizable 3d gaussian heads. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)

34. Kirschstein, T., Qian, S., Giebenhain, S., Walter, T., Nießner, M.: Nersemble: Multi-view radiance field reconstruction of human heads. ACM Transactions on Graphics (TOG) **42**(4), 1–14 (2023)
35. Kirschstein, T., Romero, J., Sevastopolsky, A., Nießner, M., Saito, S.: Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12089–12100 (2025)
36. Li, J., Zhang, J., Bai, X., Zheng, J., Ning, X., Zhou, J., Gu, L.: Talkinggaussian: Structure-persistent 3d talking head synthesis via gaussian splatting. In: European Conference on Computer Vision. pp. 127–145. Springer (2024)
37. Li, J., Zhang, J., Bai, X., Zhou, J., Gu, L.: Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7568–7578 (2023)
38. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph. **36**(6), 194–1 (2017)
39. Li, W., Zhang, L., Wang, D., Zhao, B., Wang, Z., Chen, M., Zhang, B., Wang, Z., Bo, L., Li, X.: One-shot high-fidelity talking-head synthesis with deformable neural radiance field. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17969–17978 (2023)
40. Li, X., De Mello, S., Liu, S., Nagano, K., Iqbal, U., Kautz, J.: Generalizable one-shot 3d neural head avatar. Advances in Neural Information Processing Systems **36**, 47239–47250 (2023)
41. Liang, H., Ren, J., Mirzaei, A., Torralba, A., Liu, Z., Gilitschenski, I., Fidler, S., Oztireli, C., Ling, H., Gojcic, Z., et al.: Feed-forward bullet-time reconstruction of dynamic scenes from monocular videos. arXiv preprint arXiv:2412.03526 (2024)
42. Liang, H., Ge, Z., Majee, S., Tiwari, A., Godaliyadda, G., Veeraraghavan, A., Balakrishnan, G.: Fastavatar: Instant 3d gaussian splatting for faces from single unconstrained poses. arXiv preprint arXiv:2508.18389 (2025)
43. Ma, H., Zhang, T., Sun, S., Yan, X., Han, K., Xie, X.: Cvthead: One-shot controllable head avatar with vertex-feature transformer. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6131–6141 (2024)
44. Ma, S., Weng, Y., Shao, T., Zhou, K.: 3d gaussian blendshapes for head avatar animation. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–10 (2024)
45. Ma, S., Simon, T., Saragih, J., Wang, D., Li, Y., De La Torre, F., Sheikh, Y.: Pixel codec avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 64–73 (2021)
46. Martinez, J., Kim, E., Romero, J., Bagautdinov, T., Saito, S., Yu, S.I., Anderson, S., Zollhöfer, M., Wang, T.L., Bai, S., et al.: Codec avatar studio: Paired human captures for complete, driveable, and generalizable avatars. Advances in Neural Information Processing Systems **37**, 83008–83023 (2024)
47. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
48. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM transactions on graphics (TOG) **41**(4), 1–15 (2022)
49. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5865–5874 (2021)

50. Peng, Z., Hu, W., Shi, Y., Zhu, X., Zhang, X., Zhao, H., He, J., Liu, H., Fan, Z.: Synctalk: The devil is in the synchronization for talking head synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 666–676 (2024)
51. Qi, D., Yang, T., Wang, B., Zhang, X., Zhang, W.: Predicting 3d representations for dynamic scenes. arXiv preprint arXiv:2501.16617 (2025)
52. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20299–20309 (2024)
53. Ren, J., Xie, C., Mirzaei, A., Kreis, K., Liu, Z., Torralba, A., Fidler, S., Kim, S.W., Ling, H., et al.: L4gm: Large 4d gaussian reconstruction model. Advances in Neural Information Processing Systems **37**, 56828–56858 (2024)
54. Shen, Y., Zhang, Z., Qu, Y., Zheng, X., Ji, J., Zhang, S., Cao, L.: Fastvggt: Training-free acceleration of visual geometry transformer. arXiv preprint arXiv:2509.02560 (2025)
55. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
56. Taubner, F., Zhang, R., Tuli, M., Lindell, D.B.: Cap4d: Creating animatable 4d portrait avatars with morphable multi-view diffusion models. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5318–5330. IEEE Computer Society (2025)
57. Tewari, A., Elgharib, M., Bharaj, G., Bernard, F., Seidel, H.P., Pérez, P., Zollhofer, M., Theobalt, C.: Stylerig: Rigging stylegan for 3d control over portrait images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6142–6151 (2020)
58. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
59. Wang, J., Xie, J.C., Li, X., Xu, F., Pun, C.M., Gao, H.: Gaussianhead: High-fidelity head avatars with learnable gaussian derivation. IEEE Transactions on Visualization and Computer Graphics (2025)
60. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
61. Wu, R., Gao, R., Poole, B., Trevithick, A., Zheng, C., Barron, J.T., Holynski, A.: Cat4d: Create anything in 4d with multi-view video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26057–26068 (2025)
62. Wu, Y., Chen, X., Wu, Y., Li, W., Lu, Y., Feng, K.: Fastavatar: Towards unified and fast 3d avatar reconstruction with large gaussian reconstruction transformers. arXiv preprint arXiv:2508.19754 (2025)
63. Xiang, J., Gao, X., Guo, Y., Zhang, J.: Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1802–1812 (2024)
64. Xie, L., Wang, X., Zhang, H., Dong, C., Shan, Y.: 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. pp. 657–666 (2022)

65. Xu, S., Chen, G., Yang, J., Zhang, Y., Deng, Y., Lin, S., Guo, B.: Vasa-3d: Lifelike audio-driven gaussian head avatars from a single image. arXiv preprint arXiv:2512.14677 (2025)
66. Xu, Y., Chen, B., Li, Z., Zhang, H., Wang, L., Zheng, Z., Liu, Y.: Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1931–1941 (2024)
67. Yang, J., Huang, J., Chen, Y., Wang, Y., Li, B., You, Y., Sharma, A., Igl, M., Karkus, P., Xu, D., et al.: Storm: Spatio-temporal reconstruction model for large-scale outdoor scenes. arXiv preprint arXiv:2501.00602 (2024)
68. Yang, K., Chen, K., Guo, D., Zhang, S.H., Guo, Y.C., Zhang, W.: Face2face ρ: Real-time high-resolution one-shot face reenactment. In: European conference on computer vision. pp. 55–71. Springer (2022)
69. Yao, G., Yuan, Y., Shao, T., Zhou, K.: Mesh guided one-shot face reenactment using graph convolutional networks. In: Proceedings of the 28th ACM international conference on multimedia. pp. 1773–1781 (2020)
70. Ye, Z., Zhong, T., Ren, Y., Yang, J., Li, W., Huang, J., Jiang, Z., He, J., Huang, R., Liu, J., et al.: Real3d-portrait: One-shot realistic 3d talking portrait synthesis. arXiv preprint arXiv:2401.08503 (2024)
71. Zeng, B., Liu, B., Li, H., Liu, X., Liu, J., Chen, D., Peng, W., Zhang, B.: Fnevr: Neural volume rendering for face animation. Advances in Neural Information Processing Systems **35**, 22451–22462 (2022)
72. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
73. Zheng, Y., Abrevaya, V.F., Bühler, M.C., Chen, X., Black, M.J., Hilliges, O.: Im avatar: Implicit morphable head avatars from videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13545–13555 (2022)
74. Zheng, Y., Yifan, W., Wetzstein, G., Black, M.J., Hilliges, O.: Pointavatar: Deformable point-based head avatars from videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21057–21067 (2023)
75. Zielonka, W., Bolkart, T., Thies, J.: Towards metrical reconstruction of human faces. In: European conference on computer vision. pp. 250–269. Springer (2022)
76. Zielonka, W., Bolkart, T., Thies, J.: Instant volumetric head avatars. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4574–4584 (2023)