

EAvatar: Expression-Aware Head Avatar Reconstruction with Generative Geometry Priors

SHIKUN ZHANG, Department of Data Science and AI, Monash University, Australia

CUNJIAN CHEN, Department of Data Science and AI, Monash University, Australia

YIQUN WANG, College of Computer Science, Chongqing University, China

QIUHONG KE, Department of Data Science and AI, Monash University, Australia

YONG LI, College of Computer Science, Chongqing University, China

High-fidelity head avatar reconstruction plays a crucial role in AR/VR, gaming, and multimedia content creation. Recent advances in 3D Gaussian Splatting (3DGS) have demonstrated effectiveness in modeling complex geometry with real-time rendering capability and are now widely used in high-fidelity head avatar reconstruction tasks. However, existing 3DGS-based methods still face significant challenges in capturing fine-grained facial expressions and preserving local texture continuity, especially in highly deformable regions. To mitigate these limitations, we propose a novel 3DGS-based framework termed EAvatar for head reconstruction that is both expression-aware and deformation-aware. Our method introduces a sparse expression control mechanism, where a small number of key Gaussians are used to influence the deformation of their neighboring Gaussians, enabling accurate modeling of local deformations and fine-scale texture transitions. Furthermore, we leverage high-quality 3D priors from pretrained generative models to provide a more reliable facial geometry, offering structural guidance that improves convergence stability and shape accuracy during training. Experimental results demonstrate that our method produces more accurate and visually coherent head reconstructions with improved expression controllability and detail fidelity. Project: <https://kkun12345.github.io/EAvatar>.

CCS Concepts: • **Computing methodologies** → **Shape modeling; Rendering; Animation.**

Additional Key Words and Phrases: Head Avatar, Gaussian Splatting, Expression Modeling, High-fidelity Rendering

ACM Reference Format:

Shikun Zhang, Cunjian Chen, Yiqun Wang, Qihong Ke, and Yong Li. 2025. EAvatar: Expression-Aware Head Avatar Reconstruction with Generative Geometry Priors. 1, 1 (August 2025), 20 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 Introduction

With the rapid advancement of VR/AR, visual effects, and game character generation technologies, high-quality and animatable 3D head avatar modeling has become a critical research topic in computer graphics and 3D vision fields [7, 38]. In practical scenarios such as human-computer interaction [21], real-time expression-driven animation [19], and personalized digital asset generation [14], systems often require accurate modeling of head geometry and facial dynamics, along with fine-grained local control and real-time rendering capabilities. These demands pose significant challenges to

Authors' Contact Information: Shikun Zhang, Department of Data Science and AI, Monash University, Melbourne, Victoria, Australia; Cunjian Chen, Department of Data Science and AI, Monash University, Melbourne, Victoria, Australia; Yiqun Wang, College of Computer Science, Chongqing University, Chongqing, China; Qihong Ke, Department of Data Science and AI, Monash University, Melbourne, Victoria, Australia; Yong Li, College of Computer Science, Chongqing University, Chongqing, China.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

existing 3D reconstruction methods, particularly in structural representation, detail fidelity, and deformation flexibility. Current approaches to 3D face modeling can be broadly classified into three categories. The first category encompasses traditional mesh-based 3D Morphable Models (3DMMs), exemplified by the pioneering model developed by Blanz et al. [2] and the more expressive FLAME model introduced by Li et al. [17], which simultaneously models facial shape, expression, and pose. The second category comprises neural implicit function-based approaches, such as NeRF, NeuS [29] and its derivatives tailored for dynamic face reconstruction (e.g., HyperNeRF [22] and AvatarMe [14]). While these methods offer certain levels of personalization and expression control, they still suffer from several limitations. For instance, traditional 3DMMs can efficiently reconstruct facial geometry but often rely on global linear blend weights for expression control [3], making it difficult to achieve fine-grained local edits [17]. Meanwhile, NeRF and NeuS-based methods are capable of modeling continuous implicit fields, but they often struggle with capturing high-frequency geometric details and maintaining local coherence in regions with significant deformations. These limitations become more pronounced in high-resolution settings or under extreme facial expressions, leading to texture blurring and geometric drifting artifacts [11].

Recently, 3D Gaussian Splatting (3DGS) [15] has emerged as a robust explicit representation for real-time rendering, modeling scenes through a collection of discrete 3D Gaussians while facilitating high-quality rasterization. As the third type of approach in head avatar reconstruction, this Gaussian-based strategy has been adopted by several recent works to achieve promising outcomes in head avatar reconstruction. For instance, HeadGaS [6] and PointAvatar [36] both exploit dynamic Gaussians to build animatable head avatars with high rendering efficiency and realism. HeadGaS models expression changes using a global linear blending of parameters, which provides fast animation performance. However, it lacks fine-grained control over local areas. In contrast, PointAvatar relies on preset facial models, resulting in poor generalization when handling extreme expressions. To address coarse expression modeling, GaussianAvatars [23] combines Gaussians with a parametric face model to improve global expression controllability. Nevertheless, its reliance on mesh-based priors limits the representation of out-of-distribution local geometry. Another recent approach, SplattingAvatar [26], improves local detail rendering by embedding Gaussians into a deformable mesh structure, but its expressiveness is still constrained by the mesh’s low-frequency motion, especially under extreme expressions.

In summary, while these methods utilize 3DGS to improve facial avatar reconstruction, they frequently encounter difficulties in accurately modeling fine-grained expression in regions of high deformation, thereby constraining their capacity to capture intricate and subtle facial dynamics. To address the challenges, we propose a novel 3D head avatar modeling approach that integrates both expression-aware and deformation-aware control mechanisms. Specifically, we introduce a controllable Gaussian mechanism that identifies key Gaussians exhibiting substantial expression-induced deformation by applying an experimentally determined threshold to the predicted displacement. Subsequently, a spatial propagation strategy is employed to adjust neighboring Gaussians, facilitating more precise and localized control in expressive regions, such as the mouth and eyebrows. To further improve the geometric fidelity in highly deformable areas, we implement a Gaussian splitting strategy that adaptively duplicates Gaussians upon detecting large offsets. This approach enhances the representation of complex structures while preserving computational efficiency. Additionally, to mitigate the geometry instability observed during the early training stages of existing methods, we introduce a generative prior derived from a pretrained large model. By designing a mesh alignment and structural supervision mechanism, this prior continuously guides the optimization of the geometry throughout training, significantly improving the reconstruction quality of head contours and occluded regions. Through the above components, our framework achieves high-fidelity expression reenactment across identities, as illustrated in Fig. 1 and Fig. 4.

The main contributions of this work are summarized below.

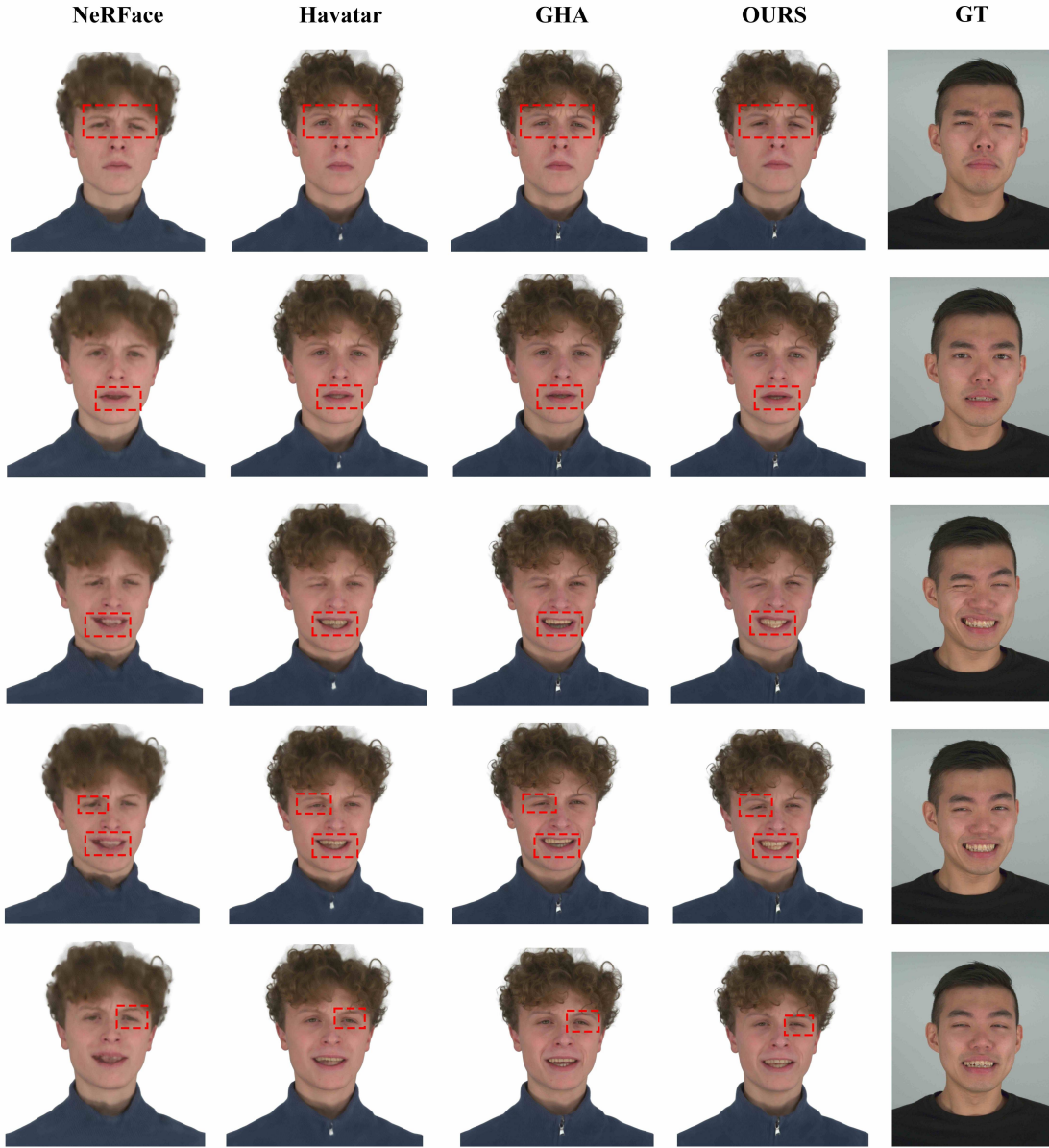


Fig. 1. Qualitative comparisons of different methods on the cross-identity reenactment task. Each column represents a different method, and each row shows results at consecutive time steps in the reenactment sequence. From left to right: NeRFace, HAvatar, GHA, and Ours. Our method more faithfully transfers expressions and maintains identity across frames, producing photo-realistic results.

- We propose a novel expression-aware 3D head avatar reconstruction framework with a controllable Gaussian mechanism that enables expression-driven animation and accurate reproduction of fine-grained expressive details.
- We design a Gaussian splitting strategy to enhance the geometric expressiveness in high-deformation regions.
- We introduce a structure-aware geometry modeling module guided by generative priors from a large-scale generative model, which improves early-stage training stability and ensures globally consistent geometry.
- Our method is evaluated on multiple expression-driven benchmarks. The results demonstrate superior performance in terms of expression reconstruction accuracy, detail preservation, and identity consistency, showing strong generalization and practical value.

2 Related Work

Explicit 3D Morphable Models (3DMM). Traditional 3D avatar modeling approaches are commonly built upon 3D Morphable Models, which enable controllable editing of shape and texture via low-dimensional parameters. Extensions such as FLAME [17] incorporate anatomical priors, RingNet [25] improves fitting accuracy through deep learning, and DECA [9] enables fine-grained expression control. However, parameterized representations remain deficient in modeling complex geometry, especially in highly deformable regions with fine structure or texture details.

Implicit Neural Representations. With the rise of neural rendering, implicit field-based methods like NeRF [20] and NeuS [29] have enabled continuous, differentiable 3D modeling via volume rendering and signed distance functions. Several efforts extend this paradigm to avatar modeling: FaceNeRF [10] introduces expression-conditioned volumetric fields, but suffers from slow inference and limited controllability over expressions. NeRFace [11] fits an expression-conditioned dynamic NeRF using a deep MLP, enabling controllable face rendering via 3DMM-driven deformation, but the expression control relies on low-dimensional parameter regression, making it difficult to accurately model complex local variations such as eye and mouth details.

Gaussian Splatting-based Approaches. Recently, 3D Gaussian Splatting has emerged as a promising alternative, balancing rendering efficiency and representational power. GaussianHead [28] models dynamic heads using deformable 3D Gaussians and a compact tri-plane with learnable derivations, achieving high-fidelity reconstruction from monocular videos. D3GA [37] reconstructs multi-layered, drivable full-body avatars from multi-view videos by embedding 3D Gaussian primitives into tetrahedral cages, with separate cages for the body, garments, and face, where the facial component can model dynamic head motions. GHA [31] uses implicit initialization and expression-aware decomposition for high-fidelity dynamic modeling, while HumanGaussian [33] demonstrates monocular Gaussian reconstruction from a single image. Despite notable improvements in overall quality, existing methods face limitations in fine-grained deformation modeling and local consistency.

Neural Methods with Structural Priors. To bridge quality and control, recent efforts have integrated 3DMM priors into neural frameworks [5, 8, 12, 24]. IMAvatar [35] blends shape bases and skinning fields for expression and pose deformation, but is dependent on tracking and iterative ray marching. i3DMM [32] decouples shape into a reference geometry and deformation field, enabling dense correspondence but introducing noise in hair regions. Neural Head Avatars [14] separates geometry and texture via color-aware energy terms to improve local consistency under large expression or pose changes. However, these methods often degrade in under-constrained regions such as teeth or ears when faced with out-of-distribution poses.

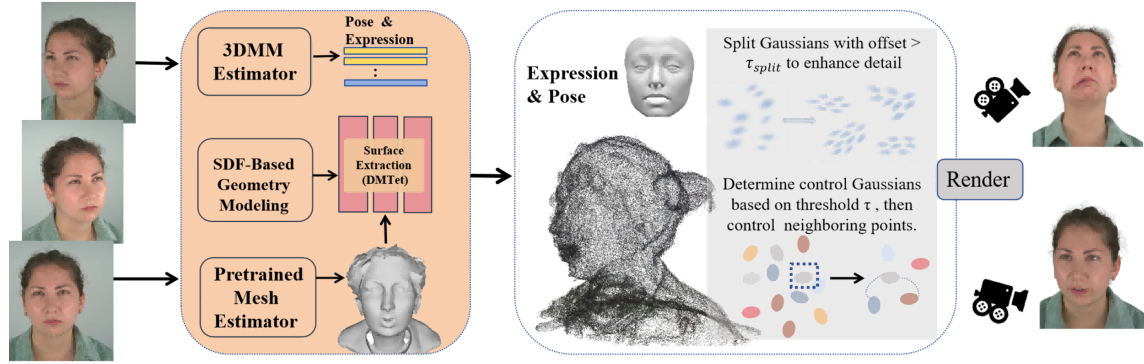


Fig. 2. The overview of the EAvatar rendering and reconstruction. We first learn an implicit SDF-based geometry and extract the surface via DMTet. A high-quality mesh from a large-scale pretrained model is used as a generative prior to stabilize initialization and guide accurate shape construction. In the second stage, we build upon the predicted mesh to further refine the dynamic Gaussian representation. A controllable Gaussian mechanism and a splitting strategy are introduced to improve expression-driven deformation and local detail.

In summary, while existing methods have made notable progress in improving reconstruction quality and control flexibility, there is still room for improvement in handling large deformations, recovering high-frequency details, and ensuring stable training.

3 Proposed Method

We aim to reconstruct a high-quality and expression-controllable 3D head avatar under large deformations. Our method is built upon a dynamic 3D Gaussian representation and consists of several innovative components. In Section 3.1, we introduce the basic modeling pipeline, which builds a neutral Gaussian representation and predicts dynamic changes based on expression and pose parameters. Section 3.2 presents the proposed controllable Gaussian mechanism, which selects Gaussians with large expression-induced deformation and adjusts their neighbors for better local modeling. We also introduce a simple but effective Gaussian splitting strategy that improves geometric details in high-deformation areas. Section 3.3 describes our structure-aware modeling strategy, forming the first stage of training that leverages a generative mesh to stabilize optimization and enhance geometry in occluded regions. Finally, Section 3.4 outlines our training process, which includes two stages. The first stage models geometry by optimizing a signed distance function guided by a prior mesh from a generative model, enabling structured and identity-aware Gaussian initialization. The second stage then jointly refines appearance and expression control. The overall flowchart of the proposed method is illustrated in Fig 2.

3.1 Avatar Representation

Building on the theoretical principles of 3D Gaussian Splatting [15], we have designed a dynamic avatar representation based on 3D Gaussian primitives capable of modelling geometric shapes as well as appearance changes driven by expressions and poses. Our approach uses a canonical Gaussian set as the base shape and applies learnable expression-aware transformations to generate personalized avatar representations.

We begin by fitting a 3D Morphable Model (3DMM) to each frame of a multi-view sequence, extracting two types of control parameters: expression coefficients $\theta \in \mathbb{R}^{d_{\text{exp}}}$ and head pose parameters $\beta \in \mathbb{R}^6$. Based on the canonical head pose, we construct a neutral Gaussian set:

$$\mathcal{G}_0 = \{\mathbf{X}_0, \mathbf{F}_0, \mathbf{Q}_0, \mathbf{S}_0, \mathbf{A}_0\}, \quad (1)$$

where $\mathbf{X}_0 \in \mathbb{R}^{N \times 3}$ are the positions of N Gaussians, $\mathbf{F}_0$ denotes feature vectors, $\mathbf{Q}_0$ the rotations, $\mathbf{S}_0$ the scales, and $\mathbf{A}_0$ the opacities. Dynamic colors are predicted from $\mathbf{F}_0$ through learnable networks, to the extent that neutral colors do not have to be defined.

To model the deformation caused by θ and β , we introduce a dynamic generator Φ consisting of multiple MLPs. Each MLP predicts residuals to update attributes from the neutral template. Specifically, the spatial position is defined as:

$$\mathbf{X}(\theta, \beta) = \mathbf{X}_0 + f_{\text{exp}}^{\text{def}}(\mathbf{X}_0, \theta) + f_{\text{pose}}^{\text{def}}(\mathbf{X}_0, \beta). \quad (2)$$

On the other hand, color attributes are dynamically predicted by:

$$\mathbf{C}(\theta, \beta) = f_{\text{exp}}^{\text{color}}(\mathbf{F}_0, \theta) + f_{\text{pose}}^{\text{color}}(\mathbf{F}_0, \beta). \quad (3)$$

Other properties, including rotation $\mathbf{Q}$, scale $\mathbf{S}$, and opacity $\mathbf{A}$, follow the same residual modeling strategy:

$$\mathbf{Q}(\theta, \beta) = \mathbf{Q}_0 + f_{\text{exp}}^{\text{rot}}(\mathbf{Q}_0, \theta) + f_{\text{pose}}^{\text{rot}}(\mathbf{Q}_0, \beta), \quad (4)$$

$$\mathbf{S}(\theta, \beta) = \mathbf{S}_0 + f_{\text{exp}}^{\text{scale}}(\mathbf{S}_0, \theta) + f_{\text{pose}}^{\text{scale}}(\mathbf{S}_0, \beta), \quad (5)$$

$$\mathbf{A}(\theta, \beta) = \mathbf{A}_0 + f_{\text{exp}}^{\text{alpha}}(\mathbf{A}_0, \theta) + f_{\text{pose}}^{\text{alpha}}(\mathbf{A}_0, \beta). \quad (6)$$

To transform all spatial attributes into the world coordinate system, we apply a rigid transform $T(\cdot)$ to the positions and rotations:

$$\{\mathbf{X}^{\text{world}}, \mathbf{Q}^{\text{world}}\} = T(\{\mathbf{X}(\theta, \beta), \mathbf{Q}(\theta, \beta)\}). \quad (7)$$

The remaining attributes $\mathbf{C}, \mathbf{S}, \mathbf{A}$ are preserved in the local coordinate space. The final expression-aware Gaussian set is defined as:

$$\mathcal{G}(\theta, \beta) = \{\mathbf{X}^{\text{world}}, \mathbf{C}, \mathbf{Q}^{\text{world}}, \mathbf{S}, \mathbf{A}\}. \quad (8)$$

This set $\mathcal{G}(\theta, \beta)$ (Eq. 8) is passed to the differentiable renderer for image synthesis, enabling expression-driven reconstruction and high-fidelity avatar rendering. Unlike prior works that treat all Gaussians uniformly, our full pipeline introduces a control-based selection mechanism and Gaussian splitting strategy (see Sec. 3.2) to enhance local expressiveness and handle high-deformation regions more effectively.

3.2 Controllable Gaussian Mechanism

In real facial expressions, particularly during intense motions such as laughing or frowning, key expressive areas (e.g., the mouth corners or brows) undergo significant deformations, while surrounding areas exhibit subtle changes in texture and geometry. Relying solely on global MLP-based prediction often fails to capture these local variations, resulting in blurry reconstructions or geometric discontinuities.

To address this, we propose a controllable Gaussian mechanism that enhances the expressiveness of highly deformable areas by explicitly modeling local geometric adjustments. Specifically, we automatically identify *Control Gaussians*

based on the magnitude of predicted displacements and propagate their influence to neighboring Gaussians via distance-weighted interpolation. This allows our model to accurately capture localized deformations while preserving consistency in high-expression regions.

Specifically, we start by computing the expression-driven displacement magnitude of each Gaussian using the MLP $f_{\text{exp}}^{\text{def}}$ from the base model:

$$\Delta \mathbf{x}_i = \left\| f_{\text{exp}}^{\text{def}}(\mathbf{x}_i, \theta) \right\|_2, \quad (9)$$

where $\mathbf{x}_i \in \mathbb{R}^3$ is the canonical position of the i -th Gaussian, and $f_{\text{exp}}^{\text{def}}(\mathbf{x}_i, \theta)$ outputs its predicted expression offset. We then define a threshold τ to detect significantly displaced Gaussians. If $\Delta \mathbf{x}_i > \tau$, we classify the i -th Gaussian as a *Control Gaussian*. The set of Control Gaussians is denoted as:

$$C = \{i \mid \Delta \mathbf{x}_i > \tau, i = 1, \dots, N\}. \quad (10)$$

To enhance local consistency, we propagate the influence of each Control Gaussian to its nearby neighbors. For each $\mathbf{x}_i \in C$, we use nearest neighbor search to find a local neighborhood:

$$\mathcal{N}(i) = \{j \mid \|\mathbf{x}_j - \mathbf{x}_i\|_2 < r, j \neq i\}, \quad (11)$$

where r is a predefined radius controlling the influence range. Notably, a neighboring point $\mathbf{x}_j$ may be influenced by multiple control points. To model this multi-source influence, we define a set of control Gaussians C_j that affect each $\mathbf{x}_j$, and update its final position as:

$$\mathbf{x}'_j = \mathbf{x}_j + \sum_{i \in C_j} w_{ij}(\mathbf{x}'_i - \mathbf{x}_i), \quad (12)$$

where the influence weight w_{ij} is computed based on the spatial distance between $\mathbf{x}_j$ and each control point $\mathbf{x}_i$ using a Gaussian kernel, normalized within C_j :

$$w_{ij} = \frac{\exp\left(-\frac{\|\mathbf{x}_j - \mathbf{x}_i\|_2^2}{\sigma^2}\right)}{\sum_{k \in C_j} \exp\left(-\frac{\|\mathbf{x}_j - \mathbf{x}_k\|_2^2}{\sigma^2}\right)}. \quad (13)$$

Here, σ controls the decay of influence with respect to spatial distance. This neighborhood-aware adjustment—realized through Gaussian-weighted aggregation from multiple control points—promotes smoother local deformations while preserving sharp details in highly expressive regions, as reflected in Eq. 12 and Eq. 13. Through this mechanism, our method enhances the representation of fine-grained dynamics and improves realism in highly deformable areas, especially under extreme expressions.

Gaussian Splitting Strategy. In expressive regions with extremely large deformations, a single Gaussian may struggle to accurately model complex local geometry. To address this issue, we introduce a targeted Gaussian splitting strategy. When the predicted displacement of a Gaussian exceeds a higher threshold τ_{split} , we dynamically split it into two split instances. The split instances are initialized close to the original location and inherit its appearance and structural attributes. This process increases the local density of Gaussians in highly deformable regions, enabling a finer representation of geometric variations. Our splitting strategy is deformation-aware and only applies to regions with large displacements. This ensures modeling efficiency while effectively improving local detail preservation and geometric expressiveness under extreme facial motions.

3.3 Structure-Aware Geometry Modeling with Generative Prior Constraints

To provide a stable and structure-aware foundation for downstream surface representation, we introduce a geometry modeling stage guided by a high-quality prior mesh generated by a large-scale generative model. This first-stage module optimizes an implicit signed distance function (SDF) via a neural network, from which a differentiable mesh surface is extracted using DMtet. The resulting mesh serves as reliable and identity-aware guidance for Gaussian initialization in the subsequent stage.

We first represent the geometry using an implicit signed distance function (SDF) modeled by a multi-layer perceptron (MLP) f_{sdf} , which maps a 3D point $\mathbf{x} \in \mathbb{R}^3$ to a scalar SDF value s and a feature vector η :

$$(s, \eta) = f_{\text{sdf}}(\mathbf{x}). \quad (14)$$

Here, $s \in \mathbb{R}$ denotes the signed distance to the surface (positive for outside, negative for inside), and $\eta \in \mathbb{R}^d$ encodes features used for downstream appearance prediction. We then extract the initial implicit surface using Deep Marching Tetrahedra (DMtet) [27], as it supports differentiable surface extraction for end-to-end training:

$$\hat{\mathbf{X}} = \text{DMtet}(f_{\text{sdf}}), \quad (15)$$

where $\hat{\mathbf{X}} \in \mathbb{R}^{M \times 3}$ is the resulting set of M mesh vertices. To improve the robustness and structural accuracy of the extracted mesh in the training stage, we incorporate a high-quality prior mesh as additional geometric guidance. We denote the prior mesh as:

$$\mathbf{X}^{\text{mesh}} = \{\mathbf{x}_m^{\text{mesh}}\}_{m=1}^M, \quad \mathbf{x}_m^{\text{mesh}} \in \mathbb{R}^3 \quad (16)$$

where each $\mathbf{x}_m^{\text{mesh}}$ represents the 3D position of the m -th vertex. We generate this 3D mesh using a large-scale pretrained generative model [30], which adopts a structured latent representation and a sparse-aware transformer architecture, enabling more accurate and identity-specific initialization shapes.

Subsequently, we apply Iterative Closest Point (ICP) [1] to align the prior mesh to our predicted mesh, ensuring coordinate consistency. After alignment, we treat the prior mesh as a structural constraint to guide the optimization toward more accurate target shapes. Specifically, we introduce a global alignment loss that enforces consistency in the overall shape and scale of the predicted mesh with respect to the prior. Rather than relying on point-wise vertex supervision—which tends to be noisy and overly restrictive at this stage—we design a lightweight yet effective constraint based on two stable geometric features: the mesh center and scale. We compute each mesh’s global center as the mean of its vertices, and define the scale as the mean distance from the vertices to the center. Given the prior mesh (pre-aligned via ICP) and the predicted mesh, we denote their centers as $\mathbf{c}^{\text{mesh}}$ and $\hat{\mathbf{c}}$, and their scales as s^{mesh} and $\hat{s}$, respectively. The alignment loss is defined as:

$$\mathcal{L}_{\text{mesh}} = \|\mathbf{c}^{\text{mesh}} - \hat{\mathbf{c}}\|_2^2 + (s^{\text{mesh}} - \hat{s})^2. \quad (17)$$

This prior-guided constraint ensures global geometric consistency while avoiding overly strict local supervision, thereby enhancing the stability and structural integrity of the predicted mesh. Unlike a fixed facial template, our prior mesh is generated from the input image using a powerful generative model, enabling identity- and shape-specific adaptation. This adaptation leads to more accurate geometry modeling and stronger structural alignment during optimization. As shown in Fig. 7, our method yields more plausible and consistent geometry when trained for the same number of epochs in the first stage, demonstrating improved convergence behavior. Our ablation studies (see Fig. 7 and Table 4) confirm that removing this structure-aware modeling strategy leads to inferior surface reconstruction and unstable optimization, demonstrating its critical importance to the overall framework.

In summary, the proposed structure-aware geometry modeling module, guided by a generative prior and optimized through global alignment, significantly enhances geometric accuracy and consistency. It establishes a stable structural foundation that facilitates high-fidelity expression reconstruction in subsequent stages.

3.4 Training

We employ a two-stage training strategy to ensure stable convergence and high-quality avatar reconstruction.

Stage I: Structural Geometry Modeling. In the first stage, we jointly optimize the expression- and pose-driven deformation networks along with the neutral Gaussian set and the implicit surface. The objective integrates several commonly used supervision signals, including RGB reconstruction, silhouette consistency, landmark proximity, Laplacian regularization, and deformation offset regularization. Importantly, we introduce a novel global alignment loss based on the global center and scale of the predicted and prior meshes. As presented in Sec. 3.3, this large-scale prior mesh alignment loss constrains the predicted geometry using a high-quality mesh generated by a large-scale pretrained model. This global constraint is robust to local noise and significantly improves early-stage geometric stability. The total loss for this stage is defined as:

$$\begin{aligned} \mathcal{L}_{\text{init}} = & \lambda_{\text{rgb}} \mathcal{L}_{\text{rgb}} + \lambda_{\text{sil}} \mathcal{L}_{\text{sil}} + \lambda_{\text{offset}} \mathcal{L}_{\text{offset}} \\ & + \lambda_{\text{lmk}} \mathcal{L}_{\text{lmk}} + \lambda_{\text{lap}} \mathcal{L}_{\text{lap}} + \lambda_{\text{mesh}} \mathcal{L}_{\text{mesh}}. \end{aligned} \quad (18)$$

To balance the contributions of different supervision signals during training, we empirically set the loss weights as follows: $\lambda_{\text{rgb}} = 1.0$, $\lambda_{\text{sil}} = 0.1$, $\lambda_{\text{offset}} = 0.01$, $\lambda_{\text{lmk}} = 0.1$, $\lambda_{\text{lap}} = 100$, and $\lambda_{\text{mesh}} = 1.0$. These values are kept fixed throughout all experiments to ensure consistency and fair comparison across different model variants.

Stage II: Dynamic Optimization. In the second stage, we continue to jointly optimize the expression, pose, and attribute networks along with all other components in an end-to-end manner. This stage further refines the dynamic Gaussian representation in terms of geometry, appearance, and controllability. We adopt a combination of full-image RGB reconstruction loss and LPIPS perceptual loss on randomly cropped local patches:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rgb}} \mathcal{L}_{\text{rgb}} + \lambda_{\text{lpi}} \mathcal{L}_{\text{lpi}}. \quad (19)$$

We empirically set $\lambda_{\text{rgb}} = 1.0$ and $\lambda_{\text{lpi}} = 0.1$. This two-stage pipeline enables a smooth transition from stable geometry initialization to expressive, high-fidelity avatar reconstruction.

4 Experiments

In this section, we present the details of our experimental setup, datasets, ablation studies, and comparisons with existing methods. We begin by describing the implementation details, including training configurations and hardware specifications. Next, we introduce the datasets used for training and evaluation. Our method is evaluated on two tasks: self-reenactment and cross-identity reenactment. We report both qualitative and quantitative results, and compare our approach against state-of-the-art methods to demonstrate its effectiveness. We also provide a more visual video comparison of the results in the project. Additionally, we conduct a series of ablation studies to investigate the impact of each key module in our framework. By incrementally adding the proposed modules, we analyze their individual contributions to the final performance through both visual and numerical comparisons.

4.1 Datasets.

We conduct our experiments on multi-view video data from the NeRSemble dataset [16], which contains 16 camera view sequences for each subject. Each camera captures multiple expression sequences, such as EMO-1-shout+laugh. For each identity, we use sequences labeled as “FREE” for evaluation and the remaining sequences for training. Following the preprocessing pipeline of GHA [31], we first remove the background from each frame using a segmentation model [18], and extract 68 facial 2D landmarks using a standard landmark detector [4]. We then fit the multi-view Basel Face Model (BFM) [13] to estimate 3D landmarks, expression coefficients, and head pose corresponding to the detected 2D landmarks.

4.2 Evaluation Metrics.

For quantitative evaluation, we adopt three widely used metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). These metrics assess the reconstruction quality from the perspectives of pixel-level fidelity, structural consistency, and perceptual similarity, respectively.

4.3 Training Details.

We use the Adam optimizer for both training stages. In the structural geometry modeling stage, the learning rate is set to 1×10^{-3} for all networks. We train this stage for 10,000 iterations with a batch size of 4. In the second stage, we jointly optimize all components of the Gaussian avatar model. The learning rates are set to 1×10^{-4} for the color, deformation, and attribute MLPs; 1×10^{-5} for the neutral positions $\mathbf{X}_0$ and feature vectors $\mathbf{F}_0$; 1×10^{-5} for rotation $\mathbf{Q}_0$; 3×10^{-5} for scale $\mathbf{S}_0$. This stage is trained for 500,000 iterations with a batch size of 1. It is worth noting that using the generative prior from a large-scale pretrained model to guide structural geometry modeling reduces the overall training time by 11% compared to not using the prior. We empirically set the control and splitting thresholds to 0.3 and 0.2, respectively, based on the typical size and movement range of expressive facial regions. The higher control threshold focuses on clearly deformed areas, while the lower splitting threshold enables earlier refinement in moderately changing regions. We will further discuss the reasonableness of these threshold choices in Sec. 4.5.4

4.4 Results and Comparisons

Self-Reenactment Evaluation. We qualitatively and quantitatively compare our method with several representative approaches on the self-reenactment task. Specifically, NeRFace [11] employs a deep MLP to model an expression-conditioned dynamic NeRF, where facial expressions and head poses derived from a 3DMM are used as conditioning inputs to enable controllable face geometry and appearance modeling in a canonical space. HAvatar [34] utilizes 3DMM mesh as a conditioning input and applies tri-plane based neural radiance fields for high-fidelity reconstruction. For fair comparison, we replace the adversarial GAN loss with the VGG perceptual loss following the practice in Gaussian Head Avatar [31]. Gaussian Head Avatar (GHA) builds an explicit head representation based on a set of dynamic 3D Gaussians, enabling ultra high-fidelity image synthesis. To ensure fairness, we use the same input sequences and render all methods through a unified rendering pipeline.

Qualitative comparisons are shown in Fig. 3, where our method achieves sharper reconstruction and more faithful appearance in high-frequency regions such as teeth and hair. Furthermore, our model demonstrates better expression transferability than previous methods, for example in the accuracy of eye closure. More detailed video results can be found in our project.

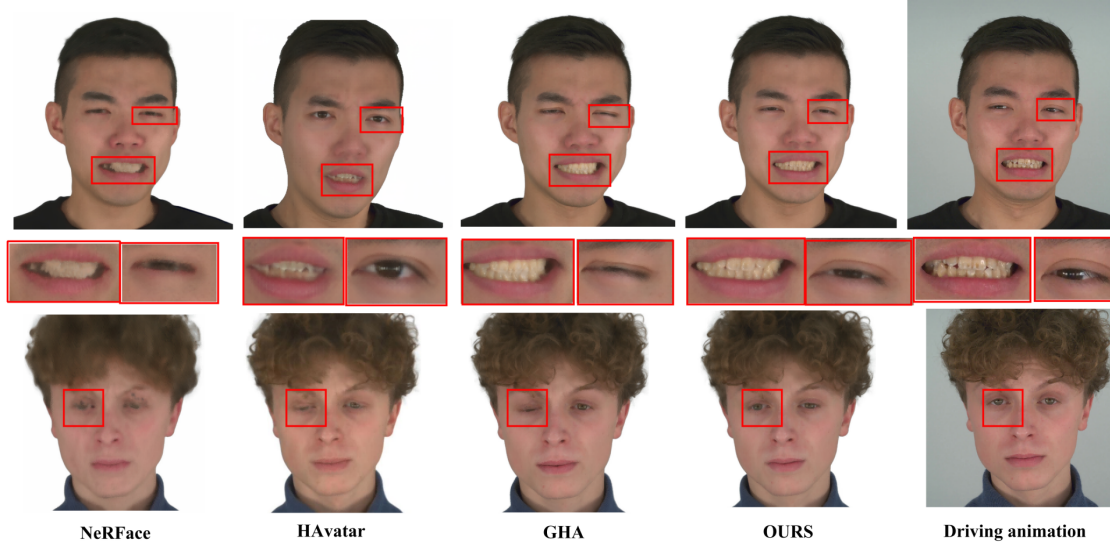


Fig. 3. Qualitative comparisons of different methods on self-reenactment task. From left to right: NeRFace, HAvatar, GHA and Ours. Our method can reconstruct details like eyes, teeth, etc. with high quality.

Table 1. Quantitative evaluation results of NeRFace, HAvatar, GHA and our full method on the self-reenactment task. $\downarrow$ indicates lower is better, $\uparrow$ indicates higher is better.

Method	Case 1			Case 2			Case 3		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
NeRFace	21.38	0.746	0.283	21.08	0.847	0.279	20.19	0.817	0.223
HAvatar	22.62	0.802	0.247	21.63	0.861	0.264	22.16	0.878	0.184
GHA	24.02	0.814	0.203	25.47	0.879	0.147	24.56	0.902	0.144
Ours	24.24	0.814	0.203	26.72	0.884	0.141	24.82	0.903	0.141

Method	Case 4			Case 5			Average		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
NeRFace	20.58	0.825	0.219	21.04	0.793	0.239	20.85	0.805	0.248
HAvatar	21.70	0.839	0.215	21.54	0.801	0.231	21.93	0.836	0.228
GHA	25.49	0.859	0.164	23.28	0.812	0.217	24.56	0.853	0.175
Ours	25.92	0.863	0.161	23.66	0.822	0.196	25.07	0.857	0.168

We also perform a quantitative comparison across four identities using three standard metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). Unlike the previous method Gaussian Head Avatar [31], which selects six random camera views for evaluation, we compute the average metric over all 16 camera views to comprehensively assess multi-view performance. In addition, we retain the full head region—including the neck and shoulders—without using facial parsing to remove body areas, since these regions are also important for identity modeling. The evaluation results are summarized in Table 1. Our method shows



Fig. 4. Another cross-identity example evaluated under the same setting as Fig. 1. Columns correspond to methods and rows to consecutive frames. Our approach continues to exhibit consistent identity preservation and expression fidelity over time.

slight improvements in LPIPS and SSIM, while achieving a significant boost in PSNR, indicating better preservation of high-frequency details in reconstruction.

Cross-Identity Reenactment. We also compare our method with previous leading methods on the cross-identity reenactment task. Cross-identity reenactment is a generative task without ground-truth targets under novel expressions,

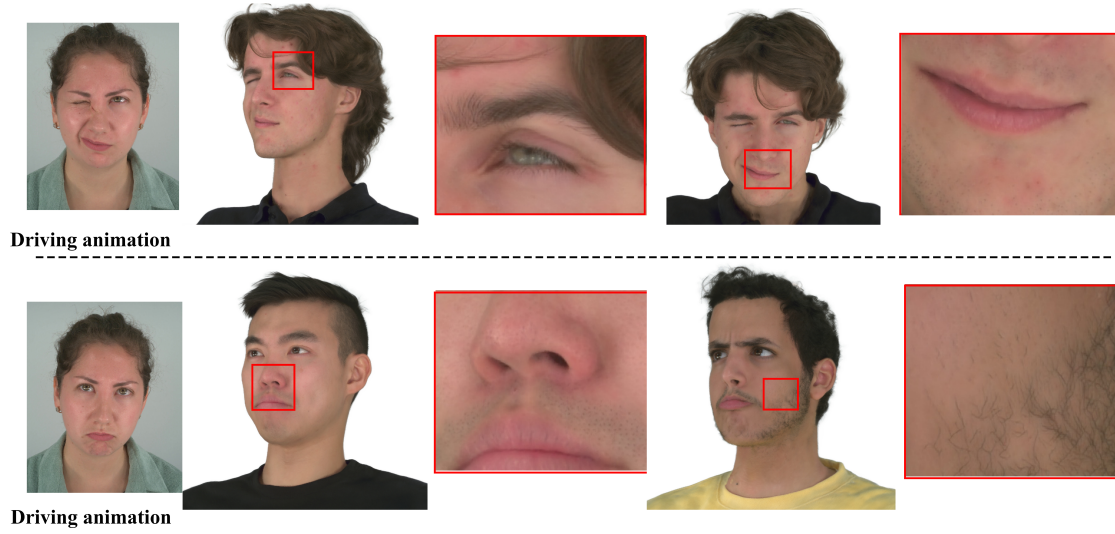


Fig. 5. Results of our expression-controllable head avatar generation. As demonstrated, the generated results accurately capture facial expressions while preserving fine-grained local details.

so metrics cannot be computed. Many prior works rely on visual comparisons to evaluate expression transfer quality. To this end, we provide a qualitative comparison across identities. As shown in Fig. 1 and Fig. 4, our approach generates clearer and more realistic results, with more accurate expression transfer. And as shown in Fig. 5, our method effectively captures facial expressions while preserving fine-grained local details.

Table 2. Quantitative evaluation of our method and other SOTA methods on 3D consistency for novel view synthesis.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
NeRFace	20.43	0.842	0.215
HAvatar	21.16	0.878	0.184
GHA	22.89	0.880	0.167
Ours	23.22	0.882	0.165

Novel View Synthesis. To evaluate the model’s ability to generate consistent images from new viewpoints, we conduct experiments on novel view synthesis. This task assesses the 3D consistency of the learned geometry and the model’s generalization across unseen camera angles. We train the model using data from only 8 camera views and test it on the remaining 8 unseen views. Fig 6 shows qualitative results under novel viewpoints. In addition, we perform a quantitative comparison with previous methods by computing PSNR, SSIM, and LPIPS on the test views. The average assessment results for the 8 test camera views are summarised in Table 2. These results demonstrate that our method generalizes better to unseen views and maintains more consistent geometry across different camera angles.

4.5 Ablation Study

4.5.1 Structure-aware modeling strategy. To evaluate the effectiveness of our structure-aware modeling strategy, we conduct a comparison with a baseline that does not use any generative prior. In the baseline setup, the initial mesh

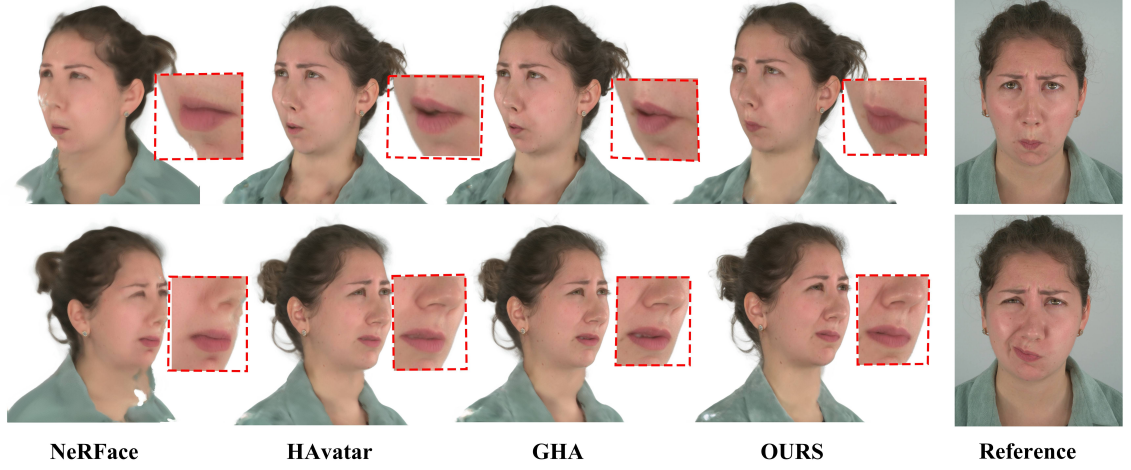


Fig. 6. Qualitative comparison with prior methods on the novel view synthesis task. We use 8-view synchronized videos for training the avatar and the remaining 8 new views were used to test.

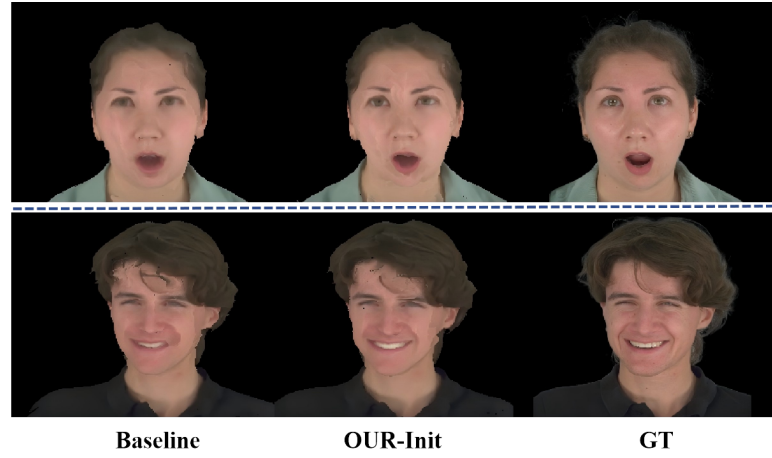


Fig. 7. Ablation study of the structure-aware modeling strategy: by introducing an explicit mesh as a shape constraint, our strategy ensures that better contour and pentagonal shapes are obtained in the first stage.

is obtained by fitting an implicit SDF and color field without any geometric constraints. The resulting mesh vertices are directly used as Gaussian positions, and other attributes are optimized using multi-view image supervision. In contrast, our method adds a mesh-based constraint generated by a large-scale generative model. As shown in Fig 7, with the same epoch of training our method can achieve more stable and coherent modelling in regions such as contours and facial features. Moreover, by comparing the fourth and fifth rows in Table 4, we observe that incorporating our structure-aware modeling strategy (Ours, fifth row) leads to significant improvements across all reconstruction metrics. These results confirm that the proposed strategy improves the overall mesh quality and provides a stronger starting point for subsequent optimization.

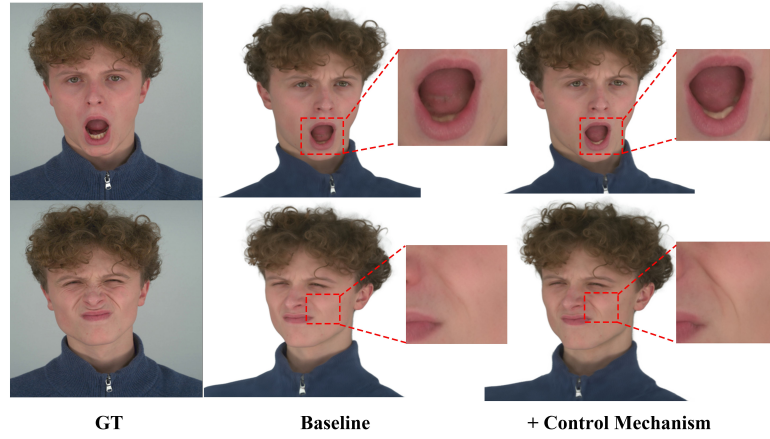


Fig. 8. Ablation study on controllable Gaussian mechanism. The figure shows the results of self-reenactment task.

4.5.2 Controllable Gaussian Mechanism. To verify the effectiveness of our proposed controllable Gaussian mechanism for modeling expressive details, we compare it with a baseline that does not use control point propagation. In the baseline setting, all Gaussian attributes are directly predicted by the MLPs, without any local structural constraints or guidance. In contrast, our method explicitly selects Gaussians with large predicted displacements as control points and applies a neighborhood propagation strategy to adjust nearby Gaussians. This mechanism improves the local continuity and geometric consistency, especially in regions undergoing strong facial deformation.

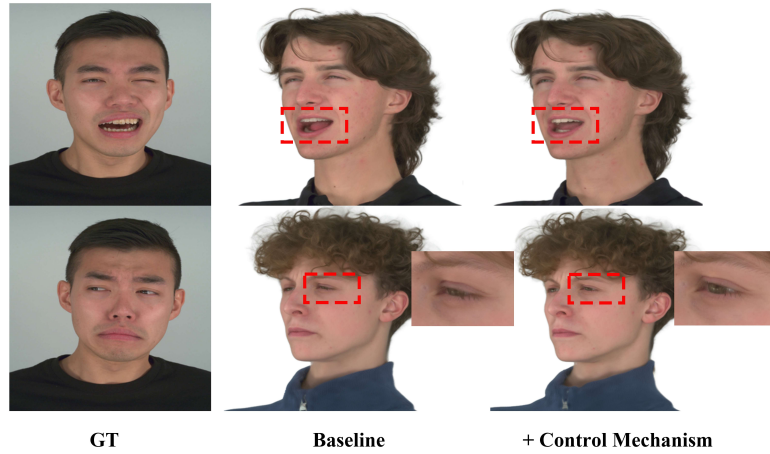


Fig. 9. Ablation study of the controllable Gaussian mechanism on the cross-identity reenactment task.

We show the results of the visualisation comparison on the self-reenactment and cross-identity Reenactment tasks in Fig 8 and Fig 9, respectively. As shown in these two figures, under expressions such as mouth opening or eyebrow raising, the baseline model tends to produce geometric distortions and blurry artifacts around key regions like the

mouth corners and eyes. In comparison, our method generates clearer boundaries and more consistent local details by explicitly guiding deformation through the proposed control point propagation. We also performed a comparison of the reconstruction accuracy on the self-reenactment task for validation, refer to Table 4.

4.5.3 Splitting Strategy. Compared to opacity-based splitting strategies, our approach determines Gaussian splitting based on the degree of attribute displacement. Specifically, when a Gaussian undergoes significant shifts in its properties, we trigger a split. This strategy enables our method to more faithfully capture fine-grained local variations, especially those that occur during dynamic expression changes.

Fig 10 presents a comparison between our splitting strategy and the baseline method, which relies solely on opacity to guide the splitting process. As shown in the figure, by tracking significant changes in Gaussian attributes, our method effectively adapts to local deformations and produces more accurate and natural reconstructions under large expression changes. For example, the appearance of the inner mouth and teeth is rendered more realistically and naturally.

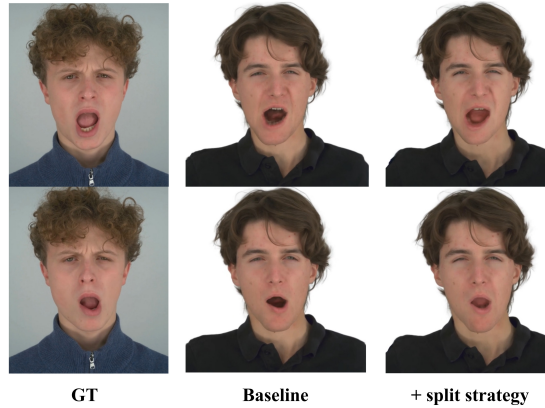


Fig. 10. Ablation study of the split strategy on the cross-identity reenactment task.

4.5.4 Threshold Selection and Analysis. Although facial geometries differ across individuals, we observe that the degree and pattern of expression-induced deformation, relative to the neutral face, remain largely consistent, particularly in active facial regions such as the mouth, eyes, and brows. This indicates that the displacement magnitude and variation trend of expressive areas are generally similar across identities, supporting the feasibility of a unified control strategy.

We further analyzed the typical amplitude of motion in expressive regions and their spatial coverage across identities, and found them to consistently fall within moderate ranges. This observation informed our choice of a control threshold of 0.3, which effectively captures significant expression-driven displacements while avoiding over-propagation. Similarly, a splitting threshold of 0.2 was found to provide early refinement in moderately active regions without introducing unnecessary growth. These values align with early-stage experiments and were fixed throughout all evaluations.

To further validate the choice of the control threshold, we conduct quantitative comparisons across different values on representative cases. As shown in Table 3, a threshold of 0.3 achieves the highest PSNR and SSIM, while lower values such as 0.15 tend to over-propagate motion, and higher values such as 0.4 fail to capture subtle deformations.

Unlike the control threshold, the splitting threshold primarily governs the number of regions selected for refinement at each iteration, thereby controlling the number of newly introduced Gaussians with the aim of capturing fine-grained

Table 3. Impact of different control thresholds on two representative cases.

Threshold	Case 1		Case 2	
	PSNR $\uparrow$	SSIM $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$
0.15	23.86	0.795	24.72	0.858
0.30	24.24	0.814	24.82	0.903
0.40	24.01	0.796	24.48	0.901

local details. A lower threshold (e.g., 0.1) triggers more frequent splitting operations, resulting in nearly twice as many new Gaussians per iteration compared to a threshold of 0.2. However, as shown in Fig 11, it brings almost no perceptual improvement over the 0.2 threshold. In contrast, a higher threshold (e.g., 0.3) delays region updates and fails to timely capture changes in visual details, especially around subtle facial regions such as lips and teeth.

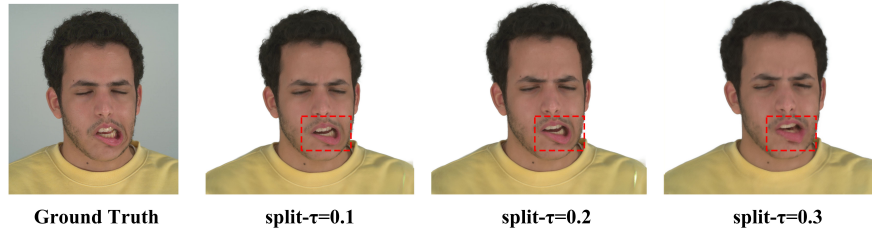


Fig. 11. Ablation study on splitting threshold. The figure shows the results of self-reenactment task.

4.5.5 Quantitative Analysis. To comprehensively evaluate the contribution of each proposed module, we conduct an ablation study starting from a baseline that excludes expression control and structure-aware modeling strategy. Table 4 compares different combinations by progressively adding the expression control mechanism, the Gaussian splitting strategy, and the structure-aware modeling strategy. Overall, we observe consistent improvements across all metrics—PSNR, SSIM, and LPIPS—as modules are gradually introduced. Specifically, incorporating the expression control mechanism leads to a notable gain in PSNR, indicating better reconstruction of overall image structure. While SSIM and LPIPS show smaller or slightly fluctuating changes in some cases, the model remains stable in performance. Nevertheless, these metrics may overlook certain perceptual aspects. As shown in Fig 8 and Fig 9, our method yields consistent qualitative improvements—especially in mouth and eye movement as well as teeth visibility—revealing subtle yet meaningful changes that go beyond what standard perceptual metrics can capture. Finally, with the addition of the structure-aware geometry modeling module, our method achieves the best results across all three metrics, demonstrating the strong synergy among the proposed components.

4.5.6 Efficiency Analysis. Regarding efficiency, we clarify that the Signed Distance Function (SDF) is used as a geometry prior to stabilize the early phase of Gaussian optimization. This process is executed once per subject. The full training pipeline, including SDF-based initialization, Gaussian attribute learning, and expression control optimization, takes approximately 2.5 days per subject on a single NVIDIA RTX 3090 GPU. This training time is shorter than GHA (3 days), and significantly shorter than HAvatar, which requires around 7 days. As shown in Table 5.

Table 4. Impact of the different modules proposed in our approach on the self-reenactment task.

Method	Case 1			Case 2		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Baseline	23.28	0.812	0.217	25.47	0.879	0.147
+control	23.36	0.816	0.214	25.55	0.879	0.147
+control & split	23.47	0.818	0.199	25.74	0.880	0.145
+control & split & structure: (Ours)	23.66	0.822	0.196	26.72	0.884	0.141

Table 5. Comparison of training time and rendering speed across methods.

Method	Training Time $\downarrow$	Rendering FPS $\uparrow$
NeRFace	4 days	~ 0.06
HAvatar	7 days	~ 7
GHA	3 days	~ 32
Ours	2.5 days	~ 32

At inference time, our method uses rasterization-based rendering of 3D Gaussians and achieves real-time performance: approximately 32 FPS on an RTX 3090. This enables smooth and responsive rendering suitable for interactive or near real-time applications.

5 Conclusion

In this paper, we propose an expression-aware and deformation-aware 3D avatar reconstruction framework leveraging dynamic 3D Gaussian representations. By introducing a controllable Gaussian mechanism and a deformation-aware splitting strategy, our method improves geometric expressiveness and local consistency in regions with complex expression changes. Additionally, we introduce a structure-aware geometry modeling module guided by a pretrained large-scale avatar prior, which significantly improves geometric stability and structural consistency during training. Experimental results show that our approach outperforms existing methods across multiple metrics, enabling high-quality synthesis and more realistic expression-driven avatar reconstruction.

References

- [1] Paul J Besl and Neil D McKay. 1992. Method for registration of 3-D shapes. In *Sensor fusion IV: control paradigms and data structures*, Vol. 1611. Spie, 586–606.
- [2] Volker Blanz and Thomas Vetter. 2023. A morphable model for the synthesis of 3D faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 157–164.
- [3] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. 2018. Large scale 3D morphable models. *International Journal of Computer Vision* 126, 2 (2018), 233–254.
- [4] Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*. 1021–1030.
- [5] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 0–0.
- [6] Helisa Dhamo, Yinyu Nie, Arthur Moreau, Jifei Song, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. 2024. Headgas: Real-time animatable head avatars via 3d gaussian splatting. In *European Conference on Computer Vision*. Springer, 459–476.
- [7] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhler, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 2020. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (ToG)* 39, 5 (2020), 1–38.

- [8] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. 2021. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (ToG)* 40, 4 (2021), 1–13.
- [9] Yao Feng, Fanzi Wu, Xiaohua Shao, Yiyi Wang, and Xing Zhou. 2021. DECA: Detailed Expression Capture and Animation from a Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2886–2897.
- [10] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2021. FaceNeRF: A Geometry-Aware 3D Facial Appearance Model via Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [11] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2021. NerFace: Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [12] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. 2019. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1155–1164.
- [13] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. 2018. Morphable face models-an open framework. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 75–82.
- [14] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18653–18664.
- [15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* 42, 4 (2023), 139–1.
- [16] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. 2023. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–14.
- [17] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 6 (2017), 194–1.
- [18] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2021. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8762–8771.
- [19] Stephen Lombardi, Jason Saraghi, Tomas Simon, and Yaser Sheikh. 2018. Deep appearance models for face rendering. *ACM Transactions on Graphics (ToG)* 37, 4 (2018), 1–13.
- [20] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- [21] Maja Pantic and Leon JM Rothkrantz. 2003. Toward an affect-sensitive multimodal human-computer interaction. *Proc. IEEE* 91, 9 (2003), 1370–1390.
- [22] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. 2021. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228* (2021).
- [23] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. 2024. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20299–20309.
- [24] Sami Romdhani and Thomas Vetter. 2005. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 2. IEEE, 986–993.
- [25] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. 2019. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7763–7772.
- [26] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. 2024. Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1606–1616.
- [27] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. 2021. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 6087–6101.
- [28] Jie Wang, Jiu-Cheng Xie, Xianyan Li, Feng Xu, Chi-Man Pun, and Hao Gao. 2025. Gaussianhead: High-fidelity head avatars with learnable gaussian derivation. *IEEE Transactions on Visualization and Computer Graphics* (2025).
- [29] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689* (2021).
- [30] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. 2024. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506* (2024).
- [31] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. 2024. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1931–1941.
- [32] Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. 2021. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12803–12813.
- [33] Wei Zhang, Y. Liu, et al. 2024. HumanGaussian: 3D Human Reconstruction from a Single Image using Gaussian Splatting. *arXiv preprint arXiv:2402.00842* (2024).
- [34] Xiaochen Zhao, Lizhen Wang, Jingxiang Sun, Hongwen Zhang, Jinli Suo, and Yebin Liu. 2023. Havatar: High-fidelity head avatar via facial model conditioned neural radiance field. *ACM Transactions on Graphics* 43, 1 (2023), 1–16.

- [35] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. 2022. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13545–13555.
- [36] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. 2023. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 21057–21067.
- [37] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. [n. d.]. Drivable 3D Gaussian Avatars. In *International Conference on 3D Vision 2025*.
- [38] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. 2018. State of the art on monocular 3D face reconstruction, tracking, and applications. In *Computer graphics forum*, Vol. 37. Wiley Online Library, 523–550.