GRMM: Real-Time High-Fidelity Gaussian Morphable Head Model with Learned Residuals

Mohit Mendiratta[1*] Mayur Deshmukh[1] Kartik Teotia[1] Vladislav Golyanik[1] Adam Kortylewski[1,2]
Christian Theobalt[1]
[1] Max Planck Institute for Informatics and Saarland University
[2] University of Freiburg

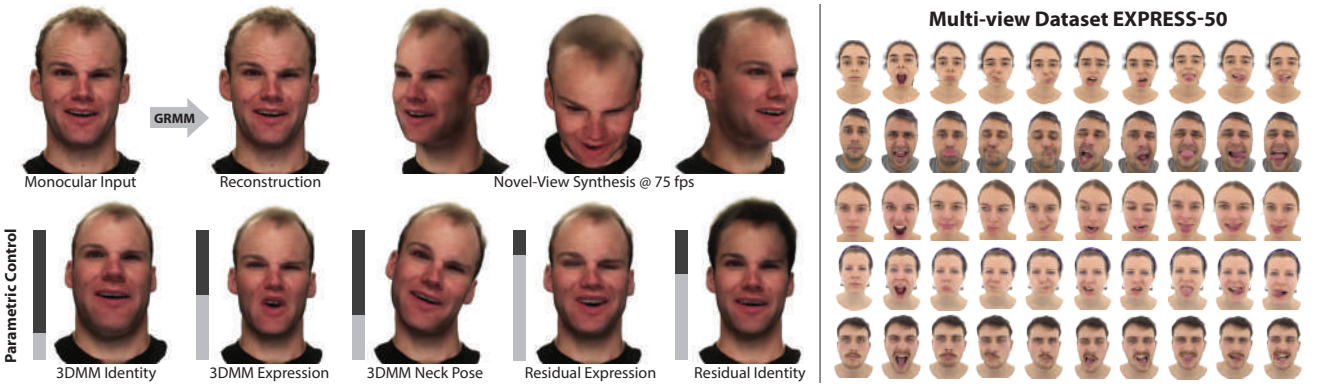

Figure 1. GRMM provides disentangled control over a base 3DMM and learned residuals, fitting unseen identities from input images and enabling novel view synthesis and expression editing while preserving identity. It produces photorealistic 1K-resolution full-head renderings with diverse expressions in real time, achieving up to 75 fps. As part of this work, we also contribute EXPRESS-50, a dataset of 50 identities with 60 distinct expressions aligned across subjects, enabling consistent modeling of expression residuals.

Abstract

3D Morphable Models (3DMMs) enable controllable facial geometry and expression editing for reconstruction, animation, and AR/VR, but traditional PCA-based mesh models are limited in resolution, detail, and photorealism. Neural volumetric methods improve realism but remain too slow for interactive use. Recent Gaussian Splatting (3DGS) based facial models achieve fast, high-quality rendering but still depend solely on a mesh-based 3DMM prior for expression control, limiting their ability to capture fine-grained geometry, expressions, and full-head coverage. We introduce GRMM, the first full-head Gaussian 3D morphable model that augments a base 3DMM with residual geometry and appearance components, additive refinements that recover high-frequency details such as wrinkles, fine skin texture, and hairline variations. GRMM provides disentangled control through low-dimensional, interpretable parameters (e.g., identity shape, facial expressions) while separately modelling residuals that capture subject- and expression-specific detail beyond the base model's capacity. Coarse decoders produce vertex-level mesh deformations, fine decoders represent per-Gaussian appearance, and a lightweight CNN refines rasterised images for enhanced realism, all while maintaining 75 FPS real-time rendering. To learn consistent, high-fidelity residuals, we present EXPRESS-50, the first dataset with 60 aligned expressions across 50 identities, enabling robust disentanglement of identity and expression in Gaussian-based 3DMMs. Across monocular 3D face reconstruction, novel-view synthesis, and expression transfer, GRMM surpasses state-of-the-art methods in fidelity and expression accuracy while delivering interactive real-time performance.

1. Introduction

High-fidelity 3D face modelling is essential for VR, AR, animation, and digital avatar creation. A widely used class of methods, 3D Morphable Models (3DMMs) [2, 10] compactly and controllably represent facial geometry and appearance using low-dimensional parametric spaces. However, achieving photorealism, real-time efficiency, and expressiveness remains challenging, as current methods struggle to capture fine details while maintaining real-time performance. Traditional mesh-based 3DMMs [2, 3, 20, 46] are efficient and interpretable but limited in resolution and unable to represent fine-scale geometry and texture variation. Neural rendering [12, 47] and volumetric meth-

ods [13, 14, 52] improve visual fidelity, but are computationally heavy and struggle with large deformations or extreme expressions. More recently, 3D Gaussian Splatting (3DGS)[15] has enabled high-resolution rendering at interactive rates. However, 3DGS-based facial models[45, 50] still rely solely on coarse mesh-based 3DMM priors for expression control, limiting their *expressivity* and ability to capture subtle, identity- or expression-specific detail. Furthermore, these models are not publicly available.

An additional bottleneck is the scarcity of datasets with both *high expression diversity* and *cross-identity expression alignment*. While recent datasets [19, 24, 25] improve identity coverage, they typically provide limited expression variability and lack the alignment required for disentangled identity–expression learning.

To address these limitations, we propose the Gaussian Residual Morphable Model (GRMM), the first open-source, full-head 3D Gaussian morphable model with learned residuals: additive refinements to both geometry and appearance that recover high-frequency details such as wrinkles, skin microstructure, and hairline variation beyond the capacity of a base 3DMM. GRMM offers independent control over (i) interpretable low-dimensional 3DMM parameters (identity shape, facial expression, head pose) and (ii) residual parameters that encode fine-scale, identity- and expression-specific deviations. This separation enables precise, compositional manipulation of facial attributes without sacrificing realism (Figure 1).

Technically, GRMM is based on a mesh-based base 3DMM [20] to produce a coarse head shape and predicts residual deformations using lightweight MLPs driven by low-dimensional identity and expression codes. UV-anchored Gaussian primitives deform coherently with the mesh, preserving spatial consistency and cross-identity correspondence. Convolutional decoders predict per-Gaussian appearance for detailed geometry and texture, while an image-space CNN refines rasterised images to recover surface details not fully captured by the Gaussians. The result is real-time 1K-resolution rendering at 75 FPS.

To support learning disentangled residuals, we introduce EXPRESS-50, a multi-view dataset containing 50 identities each performing 60 *consistently aligned* expressions. Expression alignment is achieved via over 150 hours of frame-by-frame manual annotation, ensuring that all subjects exhibit semantically matched expressions (including challenging motions such as tongue movement). This alignment enables cross-identity supervision for robust residual disentanglement and improves generalisation to unseen subjects and expressions.

In summary, our contributions are:

- GRMM, the first open-source, full-head residual Gaussian morphable model that achieves high expressivity, fine-grained control, and real-time 1K rendering at 75 FPS, outperforming prior morphable face models in both quality and flexibility.
- A novel architecture that separates base 3DMM control from learned residual geometry and appearance, combined with enhanced mesh topology and UV mapping that explicitly models teeth and inner-mouth regions. Thus, enabling high rendering quality, speed, and expressivity.
- EXPRESS-50, i.e., a new multi-view image dataset with 50 identities and 60 aligned facial expressions, extends the corpus of existing datasets in the literature and serves as an essential ingredient to obtaining the results demonstrated in this paper.

2. Related Work

Mesh-based head models. Parametric 3D face models represent facial geometry, expression, and identity using low-dimensional parameters. The seminal 3D Morphable Model (3DMM) by Blanz and Vetter [2] aligns a fixed-topology mesh to 3D scans through non-rigid registration and learns shape and appearance spaces via PCA [1]. Subsequent works [20, 40] introduced multilinear models with separate control over facial components (e.g., jaw and eyes), which have become standard priors for reconstruction and tracking [17, 35, 36]. FaceScape [46] improved realism with high-resolution geometry and diverse expressions, but mesh-based 3DMMs remain constrained by linear subspaces and limited expressiveness for fine details. To overcome this, non-linear mesh 3DMMs [29, 38, 39] use deep networks to learn complex mappings from latent codes to mesh geometry, improving reconstruction quality and facial variation. However, these models often act as black boxes, sacrificing interpretability and editability, and typically remain limited to facial regions without supporting full-head modelling. Delta models such as DECA [11] and EMOCA [8] enhance detail with UV-space displacements but remain restricted to the facial region. Generative approaches, such as Morphable Diffusion [6], leverage diffusion models conditioned on 3DMMs to synthesise avatars from a single image; however, they lack explicit control over identity and expression, and cannot represent the mouth interior or hair. Our method unifies the controllability of mesh-based 3DMMs with learned full-head per-vertex deltas and 3D Gaussian refinement, thereby retaining interpretability while capturing high-frequency details that exceed the limits of linear or face-only models.

Implicit parametric head models. Implicit representations have driven significant progress in neural parametric head modelling. SDF-based methods [12, 47] avoid fixed mesh topology and better capture complex structures like hair. NeRF-based approaches [14, 42, 52] achieve photorealistic heads without explicit geometry, while hybrid techniques [5, 13, 23] combine mesh priors with volumetric

fields for improved controllability and realism, often using large-scale capture datasets. However, NeRF-based models suffer from low rendering efficiency, forcing trade-offs between quality and speed. In contrast, our method predicts mesh-based deformations in a delta space and adds fine-scale details using 3D Gaussians, enabling efficient rendering while preserving control and high-frequency detail.

3D Gaussians-based head representations. 3D Gaussian splatting (3DGS) has recently emerged as a powerful approach for photorealistic novel-view rendering with real-time performance [15, 41]. Initially developed for rigid scenes, it has been extended to dynamic domains, including human heads and faces. The 3D Gaussian Parametric Head Model (GPHM) [45] adapts 3DGS for facial geometry by representing the head with a dense set of Gaussians trained on datasets such as NeRSemble [19] and FaceVerse [43], achieving high-quality synthesis. However, GPHM relies on MLP decoding, lacks a clear separation between coarse geometry and fine detail, and relies on 3DMM fitting and keypoints for reconstruction, which limits its expressiveness. HeadGAP [50] builds on FLAME with part-based Gaussian offsets but remains constrained by FLAME's fixed topology and shape space, while also inheriting the MLP overhead. Furthermore, HeadGAP cannot sample new identities or expressions, reducing its generative flexibility. Other re-enactment approaches, such as GAGAvatar [7] and Portrait4D-v2 [9], use captured FLAME parameters but can only replay observed motions. Despite these advances, building a generative, expressive, and efficient head model remains an open challenge. Our approach combines mesh-based 3DMM control with full-head geometric deltas and 3D Gaussian refinement. We utilise a lightweight MLP for vertex geometry and convolutional decoders for per-Gaussian parameters, thereby avoiding the need for heavy per-Gaussian MLPs. This design reduces runtime, separates coarse geometry from fine details, and allows for sampling of identities and expressions.

Multiview head datasets. Several multiview head datasets have been developed to advance 3D head modelling. FaceScape [46] captures 938 subjects with 20 expressions using high-resolution multiview images, primarily featuring East Asian identities and excluding hair. NeRSemble [19] comprises 300 identities in controlled setups, offering good subject diversity but limited expression coverage [27]. RenderMe-360 [25] captures 500 subjects with full 360-degree views, including complex hairstyles and accessories, but offers only 12 expressions per subject. AVA-256 [24] extends diversity by supporting 256 identities under consistent illumination and broad expression coverage, but it suffers from background matting and unnatural colour distribution, which complicates 3D reconstruction. Although these datasets improve diversity and fidelity, they lack expression alignment across identities, which is critical for learning morphable models with precise identity-expression control.

We complement these datasets with our EXPRESS-50 dataset, which provides expression alignment across 50 diverse identities. EXPRESS-50 captures a broader range of expressions than existing datasets, ensuring consistent correspondence across subjects. This alignment is essential for learning expressive, identity-disentangled morphable models.

3. Method

We present GRMM, a real-time, high-fidelity full-head morphable model that augments a mesh-based 3D Morphable Model (3DMM) with learned geometry and appearance residuals using 3D Gaussian splatting. Section 3.1 introduces our new EXPRESS-50 dataset and the associated preprocessing pipeline, which together form a key contribution enabling consistent expression alignment across identities. Section 3.2 outlines the Gaussian attributes, image model, and camera model used to define the 3D representation and projection process. Section 3.3 describes the model structure, and Section 3.4 details the training methodology. Finally, Section 3.5 presents the inference process, including refinement steps for full-head reconstruction.

3.1. Expression-Aligned Datasets

We use two datasets to train our model: **EXPRESS-50**, a novel dataset we collected for complex facial expression modelling, and **RenderMe-360** [25], a publicly available 4D human head dataset. EXPRESS-50 was created to provide a rich and diverse set of high-intensity expressions, complementing RenderMe-360's broader identity coverage. EXPRESS-50 contains 50 identities performing 60 distinct expressions, recorded at 24 fps using a 24-camera FaceRig at 3840×2160 resolution. RenderMe-360 features 500 subjects performing 12 expressions, captured with 60 synchronised cameras at a resolution of 2448×2048.

Preprocessing. We preprocess both datasets through expression alignment, tracking, image preprocessing, and depth generation. A key contribution is the manual alignment of expressions across identities, enabling better disentanglement of identity and expression residuals in our morphable model. Age and gender statistics for EXPRESS-50 are in the supplemental (Section 5), and we will publicly release the dataset with annotations, expression labels, and preprocessing outputs.

Expression Alignment. We manually annotate peak expressions in the EXPRESS-50 and RenderMe-360 datasets

to ensure consistent expression alignment across identities. In EXPRESS-50, each subject follows a scripted sequence of 60 expressions demonstrated via a reference video. For a chosen reference identity, we manually select the frame where each expression is most prominent to serve as the canonical example. For the remaining 49 identities, we select the peak expression frame that best matches each canonical frame, ensuring visual and semantic alignment across subjects. In RenderMe-360, each of the 12 expressions is provided as a short video sequence per identity. We observe that the final frame in each sequence typically captures the peak of the intended expression, so we annotate the last frame of each video as the aligned expression frame. We use 280 RenderMe-360 identities for alignment, excluding those with heavy makeup or large accessories that interfere with facial appearance. Together, these datasets comprise 330 unique identities, offering broad coverage of facial shapes and expressions. Examples of consistent expression alignment are included in the supplementary material.

Tracking. To recover coarse facial geometry, we estimate FLAME [20] parameters: neck pose $\boldsymbol{\theta}_{\text{neck}}$, jaw pose $\boldsymbol{\theta}_{\text{jaw}}$, expression $\boldsymbol{\alpha}_{\text{exp}}$, and identity $\boldsymbol{\alpha}_{\text{id}}$. We adopt landmark- and photometry-based tracking using VHAP [28] to fit FLAME to annotated frames, obtaining a tracked mesh $\mathbf{v}_{\text{rec}}$, global pose $(\mathbf{R}, \mathbf{t})$ per frame. For RenderMe-360 and EXPRESS-50, 68 facial landmarks from [4] guide the tracking.

Ground-Truth Depth Generation. To obtain high-quality ground truth depth images $\mathbf{I}^{gt}_{depth}$ for supervision, we adopt ProbeSDF [37], a state-of-the-art surface reconstruction method. We apply it to each time step in our dataset to raycast depth from the optimised 3D surface.

Image Preprocessing. Foreground masks are extracted with RMBG-2.0 [49] for RenderMe-360 and BackgroundMattingV2 [21] for EXPRESS-50, while Sapiens [16] provides additional facial masks to remove the torso and focus on the face.

3.2. Preliminaries

Image Generation Model. We build upon 3D Gaussian Splatting (3DGS) [15], where each primitive is parameterized by position $\mathbf{p}$, rotation $\mathbf{r}$, scale $\mathbf{s}$, opacity $\mathbf{o}$, and color $\mathbf{c}$. Following RaDe-GS [48], we render depth from Gaussian attributes, and further attach a learnable feature vector $\mathbf{f}_i \in \mathbb{R}^{32}$ to each primitive for richer mid-level appearance. The complete attribute set is

$$\mathcal{B} = \{\mathbf{p}, \mathbf{r}, \mathbf{s}, \mathbf{o}, \mathbf{c}, \mathbf{f}\}, \quad \mathbf{I}_{rgb}, \mathbf{I}_{depth}, \mathbf{I}_{feature} = \mathcal{R}(\mathcal{B}, \pi_{\mathbf{K},\mathbf{E}}), \tag{1}$$

where $\mathcal{R}$ is a differentiable rasterizer with camera intrinsics $\mathbf{K}$ and extrinsics $\mathbf{E}$. To ensure a consistent reference frame, we transform cameras into the canonical FLAME space:

$$\pi'_{\mathbf{R},\mathbf{t}} = \mathbf{K} \cdot \mathbf{E} \cdot \begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix}, \tag{2}$$

where $\mathbf{R}, \mathbf{t}$ are the tracked FLAME mesh orientation and translation obtained from the preprocessing step Section 3.1.

GaussianHeads. Our GRMM method adapts the GaussianHeads (GH) [34] architecture, which maps primitives to UV space of a template mesh. Unlike GH, designed for subject-specific reconstruction, our method generalizes across identities and expressions with a different formulation and training strategy. Given expression code $\mathbf{z}_{exp}$ and view direction $\mathbf{d}$, GH predicts primitive attributes as

$$\{\mathbf{v}_\delta, \delta_r, \delta_p, \delta_s, \mathbf{o}, \mathbf{c}\} = \mathcal{D}_{GH}(\mathbf{z}_{exp}, \mathbf{d}), \tag{3}$$

where $\mathbf{v}_\delta$ are mesh vertex deformations, $\{\delta_r, \delta_p, \delta_s\}$ are rotation, translation and scale offsets relative to the template, and $\mathbf{o}, \mathbf{c}$ denote opacity and color.

3.3. Gaussian Residual Morphable Model

Our goal is to generate a high-fidelity head model for unseen identity and expression. Our method comprises a set of decoders (Figure 2), $\mathcal{D}_{\text{GRMM}} := \{\Phi_{mesh}, \Phi_T, \Phi_\alpha, \Phi_{app}\}$ and a refinement network Ψ_{ref}. The mesh decoder Φ_{mesh} predicts vertex deformations for $\mathbf{v}_{\text{rec}}$, the transform decoder Φ_T outputs Gaussian primitive transformations, the opacity decoder Φ_α estimates primitive opacities, and the color decoder Φ_{app} produces view-dependent appearance. All modules take as input the residual identity code $\mathbf{z}_{id}$, residual expression code $\mathbf{z}_{exp}$, neck pose $\boldsymbol{\theta}_{\text{neck}}$, jaw pose $\boldsymbol{\theta}_{\text{jaw}}$, expression coefficients $\boldsymbol{\alpha}_{\text{exp}}$, and view direction $\mathbf{d}$, generating a complete head model with photorealistic rendering.

$$\{\mathbf{v}_\delta, \delta_r, \delta_p, \delta_s, \mathbf{o}, \mathbf{c}, \mathbf{f}\} = \mathcal{D}_{\text{GRMM}}\big(\mathbf{z}_{id}, \mathbf{z}_{exp}, \boldsymbol{\alpha}_{\text{exp}}, \boldsymbol{\theta}_{\text{neck}}, \boldsymbol{\theta}_{\text{jaw}}, \mathbf{d}\big). \tag{4}$$

The refinement network Ψ_{ref} refines the rendered image $\mathbf{I}_{\mathbf{rgb}}$, which is obtained from Equation 1. Each component is introduced in detail in the following sections.

Leveraging Expression Alignment. We represent each subject with a learnable residual identity latent code $\mathbf{z}_{id} \in \mathbb{R}^{512}$ and each facial expression with a learnable global expression residual latent code $\mathbf{z}_{exp} \in \mathbb{R}^{256}$. Each expression is associated with a single $\mathbf{z}_{exp}$ that is shared across all identities. This design promotes a clear separation between identity and expression residuals.

Mesh Decoder The mesh decoder Φ_{mesh} predicts per-vertex identity and expression displacements. We use a shared MLP before the decoder to fuse identity and pose

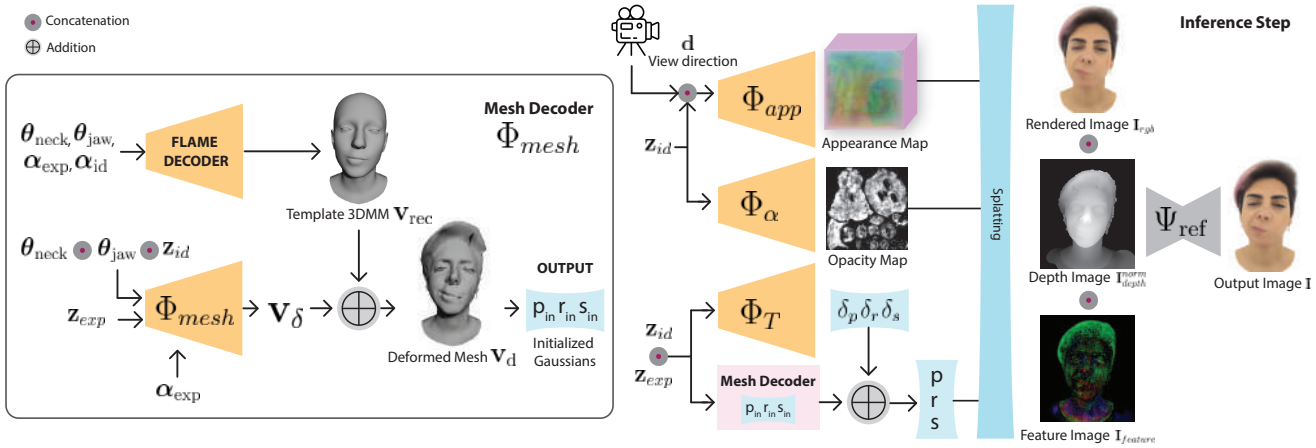

Figure 2. **Method pipeline.** Identity and expression latents $\mathbf{z}_{id} \in \mathbb{R}^{512}$ and $\mathbf{z}_{exp} \in \mathbb{R}^{256}$, together with FLAME pose/expression parameters $(\theta_{\text{neck}}, \theta_{\text{jaw}}, \alpha_{\text{exp}})$, drive the coarse mesh decoder Φ_{mesh} to predict per-vertex displacements $\mathbf{v}_\delta$. Adding these to the tracked mesh $\mathbf{v}_{\text{rec}}$ yields the deformed mesh $\mathbf{M}_{\text{d}} = (\mathbf{v}_{\text{d}}, \mathcal{F})$. UV-anchored 3D Gaussians with initial $(\mathbf{p}_{\text{in}}, \mathbf{r}_{\text{in}}, \mathbf{s}_{\text{in}})$ are placed on $\mathbf{M}_{\text{d}}$. The transformation decoder $\Phi_{\text{T}}(\mathbf{z}_{id}, \mathbf{z}_{exp})$ outputs UV-aligned maps $\delta_p, \delta_r, \delta_s$ to refine position, rotation, and scale; the opacity decoder $\Phi_\alpha(\mathbf{z}_{id})$ and appearance decoder $\Phi_{\text{app}}(\mathbf{z}_{id}, \mathbf{d})$ produce opacity, RGB, and a 32-D feature map. A differentiable rasterizer renders $\mathbf{I}_{\text{rgb}}, \mathbf{I}_{\text{depth}}, \mathbf{I}_{\text{feature}}$, where $\mathbf{I}_{\text{depth}}$ is normalized to $\mathbf{I}^{norm}_{depth}$ and provided as input to the screen-space CNN Ψ_{ref}, which outputs the final RGB image $\mathbf{I}$.

features, which improves conditioning and enables disentangled geometry prediction. A shared MLP processes the identity and pose inputs:

$$\mathbf{f}_{\text{base}} = \text{MLP}_{\text{shared}}([\mathbf{z}_{id}, \boldsymbol{\theta}_{\text{neck}}, \boldsymbol{\theta}_{\text{jaw}}]), \tag{5}$$

where $\mathbf{z}_{id} \in \mathbb{R}^{512}$ is the residual identity code, and $\boldsymbol{\theta}_{\text{neck}}, \boldsymbol{\theta}_{\text{jaw}} \in \mathbb{R}^3$ are pose parameters. Identity displacements are predicted as:

$$\mathbf{v}_{\delta,\text{id}} = \Phi_{\text{mesh,id}}(\mathbf{f}_{\text{base}}), \tag{6}$$

where $\mathbf{v}_{\delta,\text{id}} \in \mathbb{R}^{N_v \times 3}$ and N_v is the number of mesh vertices. FLAME expression modulation is applied to the concatenated identity–expression feature:

$$\mathbf{f}_{\text{exp}} = [\mathbf{f}_{\text{base}}, \mathbf{z}_{exp}], \quad \mathbf{z}_{exp} \in \mathbb{R}^{256}, \tag{7}$$

$$[\boldsymbol{\gamma}, \boldsymbol{\beta}] = \text{MLP}_{\text{FiLM}}(\boldsymbol{\alpha}_{\text{exp}}), \quad \boldsymbol{\alpha}_{\text{exp}} \in \mathbb{R}^{100}, \tag{8}$$

$$\tilde{\mathbf{f}}_{\text{exp}} = \mathbf{f}_{\text{exp}} + \boldsymbol{\gamma} \odot \mathbf{f}_{\text{exp}} + \boldsymbol{\beta}, \tag{9}$$

where MLP_{FiLM} outputs the scale $\boldsymbol{\gamma} \in \mathbb{R}^d$ and shift $\boldsymbol{\beta} \in \mathbb{R}^d$ parameters for feature-wise linear modulation (FiLM). Expression displacements are then computed as:

$$\mathbf{v}_{\delta,\text{exp}} = \Phi_{\text{mesh,exp}}(\tilde{\mathbf{f}}_{\text{exp}}). \tag{10}$$

The final deformed mesh is:

$$\mathbf{v}_{\text{d}} = \mathbf{v}_{\text{rec}} + \mathbf{v}_{\delta,\text{id}} + \mathbf{M}_{\text{face}}\, \mathbf{v}_{\delta,\text{exp}}, \tag{11}$$

where $\mathbf{M}_{\text{face}}$ masks teeth vertices to prevent expression offsets. We add teeth vertices following VHAP and extend FLAME with inner-mouth faces, forming $\mathbf{M}_{\text{d}} = (\mathbf{v}_{\text{d}}, \mathcal{F})$, where $\mathcal{F}$ is the face connectivity. FLAME enhancements are provided and ablated in the supplemental material (Section 7). Additionally, we ablate the importance of Φ_{mesh} in Section 10.

Gaussian Primitive Initialisation. We initialize $\text{N}_{\text{prim}} = \text{N}_{\text{g}}^2$ 3D Gaussians by uniformly sampling the UV space of the 3DMM mesh at a resolution of $\text{N}_{\text{g}} \times \text{N}_{\text{g}}$, with $\text{N}_{\text{g}} = 512$. Each Gaussian is assigned a position $\mathbf{p}_{\text{in}} \in \mathbb{R}^3$ via barycentric interpolation on the deformed mesh $\mathbf{M}_{\text{d}}$ with vertices $\mathbf{v}_{\text{d}}$, and both rotation $\mathbf{r}_{\text{in}} \in \mathbb{R}^3$ and scale $\mathbf{s}_{\text{in}} \in \mathbb{R}^3$ are initialized to zero. To model the mouth interior, we enhance the FLAME UV map by defining separate UV regions for the teeth and mouth interior. More details are provided in the supplemental material (Section 7).

Decoding the Gaussian Primitive Attributes. To enable high-resolution real-time rendering, we decode the properties of each Gaussian using efficient CNN decoders. Following [26, 32], these decoders map identity and expression codes $(\mathbf{z}_{id}, \mathbf{z}_{exp})$ to geometric and appearance attributes, capturing fine transformations and view-dependent colour.

Transformation decoder Φ_T maps $\mathbf{z}_{id}$ and $\mathbf{z}_{exp}$ to an offset map of size $\text{N}_{\text{g}} \times \text{N}_{\text{g}} \times 10$, corresponding to the offsets of position (δ_p), rotation (δ_r) and scale (δ_s) for the initial values $\mathbf{p}_{\text{in}}$, $\mathbf{r}_{\text{in}}$, and $\mathbf{s}_{\text{in}}$. The updated Gaussian parameters are:

$$\mathbf{p} = \mathbf{p}_{\text{in}} + \delta_p, \quad \mathbf{r} = \delta_r, \quad \mathbf{s} = \delta_s. \tag{12}$$

Position offsets δ_p capture fine-scale surface variation, including facial hair and inner-mouth geometry.

Opacity decoder Φ_α predicts a map of size $\mathrm{N_g} \times \mathrm{N_g} \times 1$, where each value represents the opacity o_i of a Gaussian and is only conditioned on $\mathbf{z}_{id}$.

Appearance decoder Φ_{app} predicts a map of size $\mathrm{N_g} \times \mathrm{N_g} \times 35$, where each entry contains RGB colour $c_i \in \mathbb{R}^{3\times1}$ and a learned feature vector $\mathbf{f}_i \in \mathbb{R}^{32\times1}$; this decoder is conditioned on $\mathbf{z}_{id}$ and the view direction $\mathbf{d}$.

Refinement Network. We use a CNN in the screen space, Ψ_{ref}, to refine the rendered results. The image resolution remains unchanged (1K) before and after refinement. Please refer to our video for a clearer illustration. We also ablate the importance of the refinement network in the supplementary material (Section 10). This refinement enhances appearance priors that are difficult to capture for our 3DGS-based model without altering the underlying 3D representation, similar to the approaches in [44, 51].

$$[\mathbf{I}_{rgb}, \mathbf{I}_{feature}, \mathbf{I}_{depth}] = \mathcal{R}(\mathcal{B}, \pi'_{\mathbf{R},\mathbf{t}}), \tag{13}$$

$$\mathbf{I} = \Psi_{\text{ref}}([\mathbf{I}_{rgb}, \mathbf{I}_{feature}, \mathbf{I}^{norm}_{depth}]), \tag{14}$$

The rendered depth image is standardised for Ψ_{ref} by applying min–max normalisation, resulting in $\mathbf{I}^{norm}_{depth}$.

3.4. Training and Losses

Given the GRMM representation, our proposed model is learned end-to-end using multi-view image supervision to train the decoders and refinement network. For this, we optimise the following objective function:

$$\begin{aligned}\mathbf{L} \;&= \mathbf{L}_{\text{rec}}(\mathbf{I}, \mathbf{I}^*) \;+ \mathbf{L}_{\text{rec}}(\mathbf{I}_{rgb}, \mathbf{I}^*) + \\ &\lambda_{depth} \cdot \mathbf{L}_{\text{depth}}(\mathbf{I}_{depth}, \mathbf{I}^{gt}_{depth}) \;+ \mathbf{L}_{\text{reg}}\,.\end{aligned} \tag{15}$$

Here, $\mathbf{I}^*$ denotes the ground-truth RGB image. $\mathbf{L}_{\text{rec}}$ is the reconstruction loss computed between both the image-space prediction $\mathbf{I}$ from refinement network and the rendered image $\mathbf{I}_{rgb}$ against $\mathbf{I}^*$. $\mathbf{L}_{\text{depth}}$ is the L2 loss between the predicted and ground-truth depth images, scaled by the weight λ_{depth}. Finally, $\mathbf{L}_{\text{reg}}$ represents additional regularization terms applied during training. Specifically, image reconstruction loss:

$$\mathbf{L}_{\text{rec}} = \lambda_{l1}\mathbf{L}_{\text{l1}} + \lambda_{ssim}\mathbf{L}_{\text{ssim}} + \lambda_{perc}\mathbf{L}_{\text{perc}} \tag{16}$$

consists of L1 loss $\mathbf{L}_{\text{l1}}$, SSIM loss $\mathbf{L}_{\text{ssim}}$, and perceptual loss $\mathbf{L}_{\text{perc}}$ with the VGG [33] as the backbone. Meanwhile, the training regularization loss:

$$\mathbf{L}_{reg} = \lambda_s\mathbf{L}_{\text{s}} + \lambda_z\mathbf{L}_{\text{z}} + \lambda_{lap}\mathbf{L}_{\text{lap}}, \tag{17}$$

Here, $\mathbf{L}_{\text{s}}$ is a regularisation term on the scale parameters, which encourages the scale of Gaussian primitives $\mathbf{s}$ to stay within a constrained range as follows:

$$\mathbf{L}_{\text{s}} = \text{mean}(l_{\text{s}}),\; l_{\text{s}} = \begin{cases} 1/\max(s_{i,d}, 10^{-7}) & \text{if } s_{i,d} < 0.1 \\ (s_{i,d} - 10.0)^2 & \text{if } s_{i,d} > 10.0 \\ 0 & \text{otherwise,} \end{cases} \tag{18}$$

where $s_{i,d}, d \in \{x, y, z\}$ denotes the scale value of each Gaussian primitive $\mathbf{i}$ along each axis, and $\text{mean}(\cdot)$ is the average operation across all dimensions, similar to [31]. $\mathbf{L}_{\text{lap}}$ represents a smoothness regularization term for the deformed mesh $\mathbf{M}_d$, and $\mathbf{L}_{\text{z}}$ is the $\mathbf{L}_2$-norm of $\mathbf{z}_{id}$ and $\mathbf{z}_{exp}$ to improve the disentanglement.

In our experiments, we set $\lambda_{l1} = 0.8$, $\lambda_{ssim} = 0.2$, $\lambda_{perc} = 0.04$, $\lambda_z = 0.001$, $\lambda_{lap} = 0.01$, $\lambda_s = 0.1$ and $\lambda_{depth} = 0.5$.

3.5. Fitting via Inverse Rendering

Given a single- or multi-view RGB portrait, we obtain 3D face tracking with VHAP and align inputs (Sec. 3.1). We use a two-stage optimisation.

Stage 1 (latent inversion). With decoders fixed, we optimise the latent codes $\mathbf{z}_{id}$ and $\mathbf{z}_{exp}$ by minimising

$$\mathbf{L}^{(1)}_{\text{fit}} = \mathbf{L}_{\text{rec}}(\mathbf{I}, \mathbf{I}^*) + \mathbf{L}_{\text{rec}}(\mathbf{I}_{\text{rgb}}, \mathbf{I}^*) + \lambda_z\,\mathbf{L}_z. \tag{19}$$

Stage 2 (prior-preserving refinement). We then fix $\mathbf{z}_{id}$ and $\mathbf{z}_{exp}$ and fine-tune $\mathcal{D}_{\text{GRMM}}$ for by minimising

$$\mathbf{L}^{(2)}_{\text{fit}} = \mathbf{L}_{\text{rec}}(\mathbf{I}, \mathbf{I}^*) + \mathbf{L}_{\text{rec}}(\mathbf{I}_{\text{rgb}}, \mathbf{I}^*) + \lambda_{loc}\,\mathbf{L}_{\text{loc}}. \tag{20}$$

$\mathbf{L}_{\text{loc}}$ is a PTI [30]-inspired locality regulariser that constrains updates to a small neighbourhood of the pretrained solution, preserving the prior. We set $\lambda_{loc} = 0.1$. We define and ablate $\mathbf{L}_{\text{loc}}$ in detail in the supplementary material (Section 9).

4. Experiments

We evaluate GRMM on the RAVDESS [22] dataset for monocular 3D face reconstruction, using 10 randomly selected identities. To assess novel-view synthesis across diverse expressions, we further sample ten identities from NeRSemble [19], five from RenderMe, and three from EXPRESS-50. RAVDESS provides monocular RGB videos with acted emotions, while NeRSemble contributes multi-view recordings that capture complex expressions and head motion. We report ablation studies in Section 4.1, and present results on downstream applications including monocular fitting, novel-view synthesis, and expression transfer in Section 4.2. Additionally, examples of disentangled parametric control for the inverted identities are included in the supplementary material (Section 12).

4.1. Ablation

No-residuals (direct 3DMM conditioning). In this variant, we remove residual parameterisation and condition the network directly on the FLAME parameters.

$$\{\mathbf{v}_\delta, \delta_r, \delta_p, \delta_s, \mathbf{o}, \mathbf{c}, \mathbf{f}\} = \mathcal{D}_{\text{GRMM}}\big(\boldsymbol{\alpha}_{\text{id}}, \boldsymbol{\alpha}_{\text{exp}}, \boldsymbol{\theta}_{\text{neck}}, \boldsymbol{\theta}_{\text{jaw}}, \mathbf{d}\big). \tag{21}$$

Learning residuals $\mathbf{z}_{id}$ and $\mathbf{z}_{exp}$ over FLAME parameters enhances identity and expression representations, yielding finer hair and appearance details and improved mouth articulation (see Figure 3).

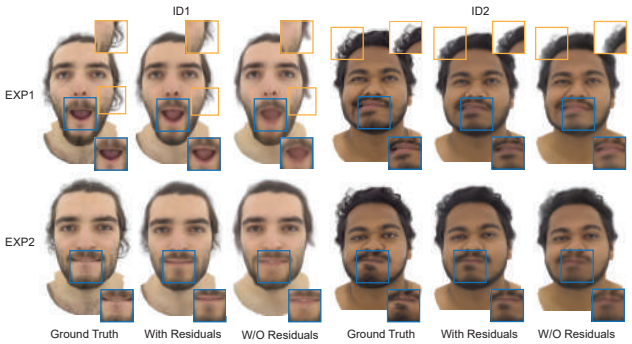

Figure 3. **Residual parameterization improves fidelity.** Qualitative ablation comparing *W/O residuals* vs. *With residuals*. Residuals yield finer hair detail and better mouth articulation (e.g., for ID1, EXP2 the mouth cannot roll in without residuals), with higher PSNR (dB): *W/O residuals* 28.91 vs. *With residuals* 30.54 (+1.63). *Please zoom in for details.*

Combining Datasets. We conduct an ablation study to assess the impact of combining EXPRESS-50 and RenderMe. Using camera views as input, we fit our model to target identities from NeRSemble. Figure 4 and Table 1 compare models trained without EXPRESS-50, without RenderMe, and with both datasets. Joint training clearly improves identity and expression fidelity, highlighting the complementary strengths of RenderMe-360 for identity generalization and EXPRESS-50 for expression generalization.

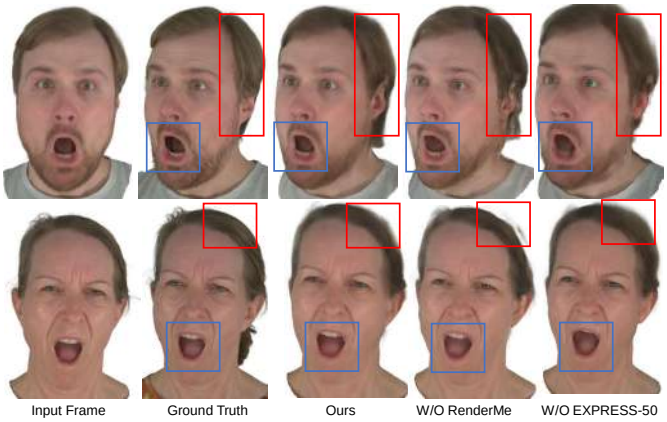

Figure 4. **Combining Datasets.** *(left to right)* Ground Truth, Without EXPRESS-50, without RenderMe and Ours. Our model trained with the combined datasets leads to overall better identity and expression fidelity.

Table 1. **Combining Datasets.** Joint training clearly enhances both identity and expression fidelity, showcasing the complementary strengths of RenderMe-360 for identity generalization and EXPRESS-50 for expression generalization.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
W/O EXPRESS-50	24.56	0.87	0.126
W/O RenderMe	25.27	0.90	0.115
Ours - Full Model	**27.40**	**0.92**	**0.091**

Table 2. **Novel-view synthesis.** Quantitative comparison on held-out views. GRMM achieves the best performance, substantially improving over MoFaNeRF and HeadNeRF.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
MoFaNeRF	13.89	0.55	0.372
HeadNeRF	17.42	0.85	0.178
Ours	**30.85**	**0.97**	**0.072**

Table 3. **Monocular reconstruction.** Quantitative comparison using RMSE and FID. GRMM outperforms MoFaNeRF and HeadNeRF, indicating improved pixel accuracy and perceptual fidelity.

Method	RMSE ↓	FID ↓
MoFaNeRF	0.193	290.786
HeadNeRF	0.067	116.34
Ours	**0.022**	**74.34**

4.2. Comparisons and Application

In this section, we demonstrate applications of GRMM in monocular image fitting, novel-view synthesis, and expression transfer. These applications showcase the generalisation capacity of our model to unseen identities, expressions and views.

Compared Methods. We compare GRMM with publicly available parametric head models, including HeadNeRF [14], MoFaNeRF [52], and the recent Morphable Diffusion [6]. We observe that Morphable Diffusion is not a volume-rendering approach and achieves viewpoint control via a conditioning camera. Consequently, it does not generalise well to camera distributions from unseen datasets; when rendering novel viewpoints using evaluation dataset cameras, the results appear misaligned and slightly distorted. Therefore, we do not report quantitative metrics for Morphable Diffusion, but instead assess its performance through a user study and qualitative analysis. The user study evaluates novel view, expression and identity consistency, with full details and results provided in the supplementary material (Section 8).

Novel-view Synthesis. We evaluate GRMM for the task of novel-view synthesis for different identities in our evaluation set. We fit our model to a single viewpoint as de-

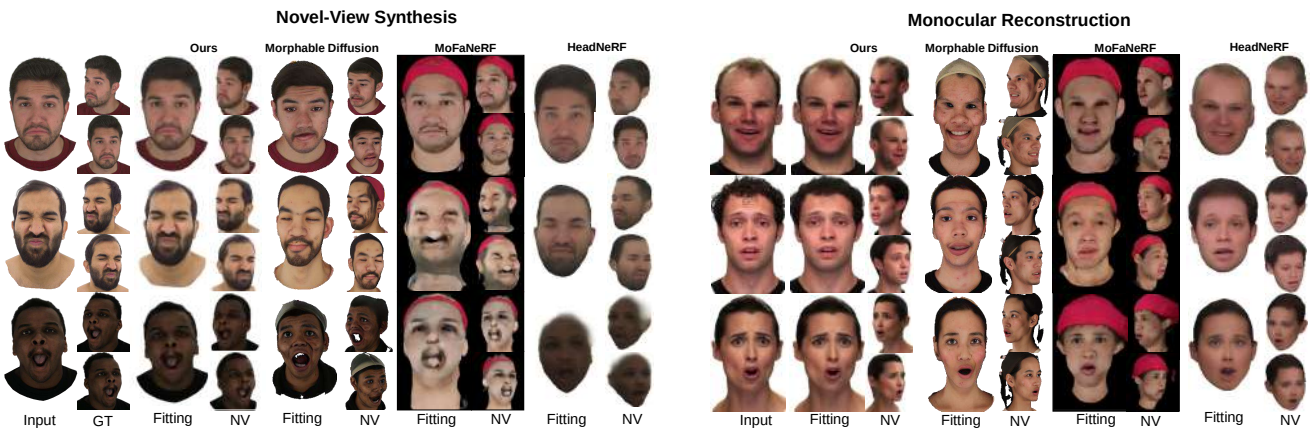

Figure 5. **Novel view synthesis (left) and monocular reconstruction (right).** Left: given one or few posed views, GRMM synthesizes unseen viewpoints while preserving identity and expression. Right: from a single RGB frame, GRMM reconstructs the subject and renders both the input and novel views. We compare against Morphable Diffusion [6], MoFaNeRF [52], and HeadNeRF [14] in both settings. FID is reported only in the monocular inversion setting, where GRMM achieves lower FID than all other baselines. *Please zoom in for details.*

scribed in Section 3.5, and assess our model's performance on two holdout views. Our method shows improved fitting and novel-view synthesis quality, as shown in Figure 5 and Table 2. Note that for the related methods, we use their publicly available inference code without any modifications.

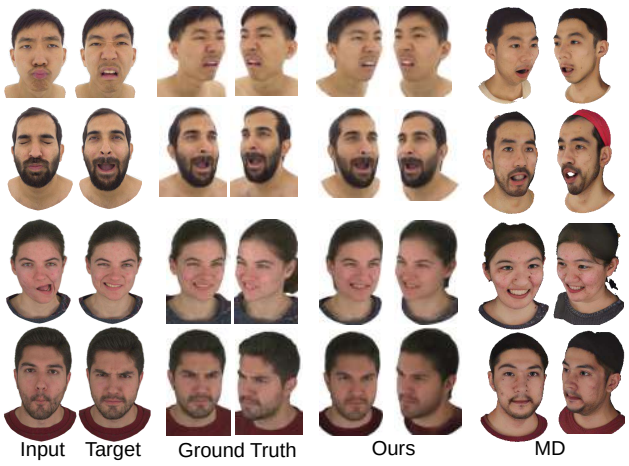

Figure 6. **Expression transfer.** From a single frontal image, we invert (Sec. 3.5), swap expression parameters, and render novel views on EXPRESS-50 and NeRSemble. GRMM preserves identity and subtle expressions with multi-view consistency, whereas Morphable Diffusion (MD) generates inconsistent expressions across views. *Please zoom in for details.*

Expression Transfer. We compare GRMM to Morphable Diffusion for expression transfer by randomly sampling target expressions for selected identities from EXPRESS-50 and NeRSemble. For each identity, we perform inversion (Section 3.5) on a single frontal view, swap the expression parameters, and render the results under novel views. Qualitatively, Morphable Diffusion struggles to capture subtle expressions and produces expression-inconsistent renderings across novel views, whereas GRMM preserves both identity and expressions with multi-view consistency at high resolution and real-time frame rates (see Figure 6 and the supplemental video). A user study (Section 8) corroborates these findings: participants consistently preferred GRMM for both expression accuracy and identity preservation in side-by-side novel-view comparisons.

4.3. Conclusion

We present GRMM, a Gaussian Residual Morphable Model that addresses key limitations of existing 3D morphable head models. By combining the efficiency of 3D Gaussian Splatting with residual augmentation of a mesh-based 3DMM in a coarse-to-fine pipeline, GRMM achieves photorealistic and diverse facial expressions at 1K resolution in real time (75 fps). The model leverages UV-anchored Gaussians and a lightweight refinement stage to enhance rendering quality, speed, and controllable expressivity. We also introduce EXPRESS-50, a multi-view dataset with 50 identities and 60 semantically aligned expressions, and show that joint training with RenderMe-360 improves both identity generalisation and expression fidelity. While advancing the state of the art, GRMM remains challenged by out-of-distribution subjects and lighting variations, suggesting the need for more diverse training data to improve robustness. With its ability to capture exaggerated expressions and render interactively, GRMM is well-suited for applications in computer graphics, virtual and augmented reality, and facial animation.

References

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010. 2

[2] V Blanz and T Vetter. A morphable model for the synthesis of 3d faces. In *26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 1999)*, pages 187–194. ACM Press, 1999. 1, 2

[3] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5543–5552, 2016. 1

[4] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 4

[5] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason Saragih. Authentic volumetric avatars from a phone scan. *ACM Trans. Graph.*, 41(4), 2022. 2

[6] Xiyi Chen, Marko Mihajlovic, Shaofei Wang, Sergey Prokudin, and Siyu Tang. Morphable diffusion: 3d-consistent diffusion for single-image avatar creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10359–10370, 2024. 2, 7, 8

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

[8] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. 2

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

[10] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (ToG)*, 39(5):1–38, 2020. 1

[11] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4): 1–13, 2021. 2

[12] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Learning neural parametric head models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

[13] Yang Haotian, Zheng Mingwu, Ma ChongYang, Lai Yu-Kun, Wan Pengfei, and Huang Haibin. Vrmm: A volumetric relightable morphable head model. In *SIGGRAPH 2024 Conference Proceedings*, 2024. 2

[14] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022. 2, 7, 8

[15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2, 3, 4

[16] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569*, 2024. 4

[17] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Trans. Graph.*, 37(4), 2018. 2

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 4

[19] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), 2023. 2, 3, 6

[20] Tianye Li, Timo Bolkart, Michael J Black, and Javier Romero. Learning a model of facial shape and expression from 4d scans. 2017. 1, 2, 4

[21] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. *arXiv*, pages arXiv–2012, 2020. 4

[22] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5): e0196391, 2018. 6

[23] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Trans. Graph.*, 40(4), 2021. 2

[24] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. *NeurIPS Track on Datasets and Benchmarks*, 2024. 2, 3

[25] Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo

Dai, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, and Kwan-Yee Lin. Renderme-360: A large digital asset library and benchmarks towards high-fidelity head avatars. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[26] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering, 2024. 5

[27] Malte Prinzler, Egor Zakharov, Vanessa Sklyarova, Berna Kabadayi, and Justus Thies. Joker: Conditional 3d head synthesis with extreme facial expressions, 2024. 3

[28] Shenhan Qian. Vhap: Versatile head alignment with adaptive appearance priors, 2024. 4, 1

[29] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3d faces using convolutional mesh autoencoders, 2018. 2

[30] Daniel Roich, Ron Mokady, Amit H. Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images, 2021. 6

[31] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. 2023. 6

[32] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars, 2024. 5

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 6

[34] Kartik Teotia, Hyeongwoo Kim, Pablo Garrido, Marc Habermann, Mohamed Elgharib, and Christian Theobalt. Gaussianheads: End-to-end learning of drivable gaussian head avatars from coarse-to-fine representations. *ACM Trans. Graph.*, 43(6), 2024. 4

[35] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2017. 2

[36] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niessner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395, 2016. 2

[37] Briac Toussaint, Diego Thomas, and Jean-Sébastien Franco. Probesdf: Light field probes for neural surface reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11026–11035, 2025. 4, 1

[38] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model, 2018. 2

[39] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model, 2019. 2

[40] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. *ACM Trans. Graph.*, 24(3):426–433, 2005. 2

[41] Angtian Wang, Peng Wang, Jian Sun, Adam Kortylewski, and Alan Yuille. Voge: A differentiable volume renderer using gaussian ellipsoids for analysis-by-synthesis. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[42] Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo Gotardo. Morf: Morphable radiance fields for multiview neural head modeling. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2

[43] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[44] Yuelang Xu, Hongwen Zhang, Lizhen Wang, Xiaochen Zhao, Huang Han, Qi Guojun, and Yebin Liu. Latentavatar: Learning latent expression code for expressive neural head avatar. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 6

[45] Yuelang Xu, Lizhen Wang, Zerong Zheng, Zhaoqi Su, and Yebin Liu. 3d gaussian parametric head model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2, 3

[46] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 601–610, 2020. 1, 2, 3

[47] T Yenamandra, A Tewari, F Bernard, HP Seidel, M Elgharib, D Cremers, and C Theobalt. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[48] Baowen Zhang, Chuan Fang, Rakesh Shrestha, Yixun Liang, Xiaoxiao Long, and Ping Tan. Rade-gs: Rasterizing depth in gaussian splatting. *arXiv preprint arXiv:2406.01467*, 2024. 4

[49] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, 2024. 4

[50] Xiaozheng Zheng, Chao Wen, Zhaohu Li, Weiyi Zhang, Zhuo Su, Xu Chang, Yang Zhao, Zheng Lv, Xiaoyuan Zhang, Yongjie Zhang, Guidong Wang, and Xu Lan. Headgap: Few-shot 3d head avatar via generalizable gaussian priors. *arXiv preprint arXiv:2408.06019*, 2024. 2, 3

[51] Xiaozheng Zheng, Chao Wen, Zhaohu Li, Weiyi Zhang, Zhuo Su, Xu Chang, Yang Zhao, Zheng Lv, Xiaoyuan Zhang, Yongjie Zhang, Guidong Wang, and Lan Xu. Headgap: Few-shot 3d head avatar via generalizable gaussian priors, 2025. 6

[52] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. In *European conference on computer vision*, pages 268–285. Springer, 2022. 2, 7, 8

GRMM: Real-Time High-Fidelity Gaussian Morphable Head Model with Learned Residuals

Supplementary Material

5. Dataset Details

We utilise two datasets for training: EXPRESS-50 and RenderMe-360, each offering distinct advantages in terms of expression coverage, identity diversity, and multi-view supervision. Their combination enables robust learning of geometry, appearance, and expression disentanglement. EXPRESS-50 is a multi-view dataset containing 50 subjects (29 male, 21 female), each performing 60 aligned expressions. Expressions are matched across all identities, enabling consistent expression conditioning during training. Subjects span ages 23–40 (mean: 28), with the age and gender distribution shown in Figure 7.

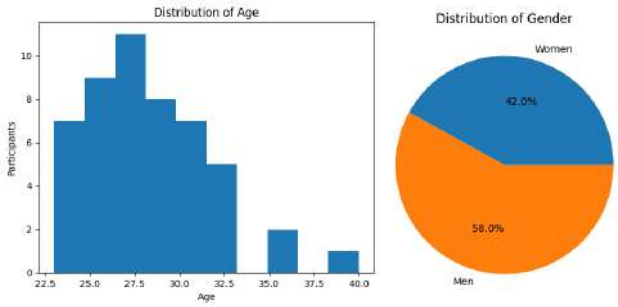

Figure 7. Statistics of the participants in our dataset.

Figure 8 shows different identities under a shared neutral expression, while Figure 9 illustrates variation and consistency across 10 aligned expressions for five sampled subjects.

We select 280 identities from RenderMe-360 after filtering, each performing 12 semantically matched expressions under dense 360° multi-view capture. While it has fewer expressions per subject than EXPRESS-50, its rich identity and view coverage support generalisation across head poses and appearances. Example of aligned expressions from this dataset are shown in Figure 10.

6. Depth Supervision with ProbeSDF

To supervise geometry, we leverage ground-truth depth maps $\mathbf{I}^{gt}_{depth}$ generated with ProbeSDF [37]. These depth images are spatially aligned with the corresponding input RGB views, enabling us to directly measure consistency between the reconstructed geometry and the reference depth. As shown in Figure 11, ProbeSDF provides smooth and consistent depth for learning geometry.

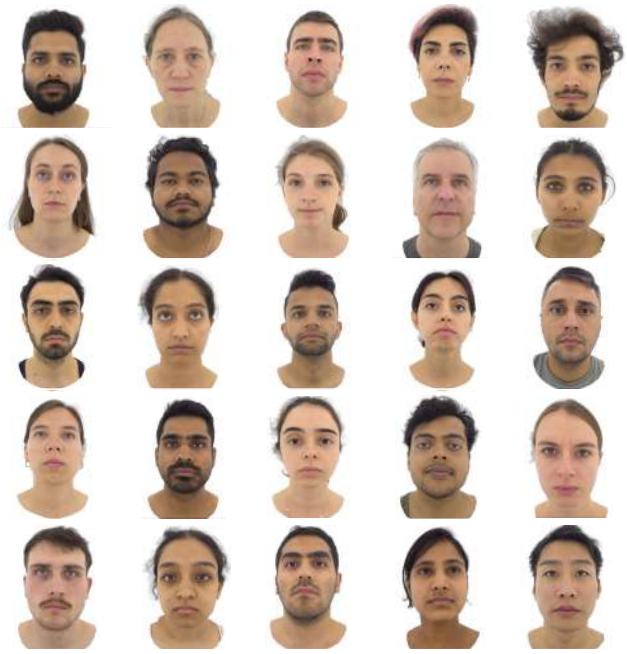

Figure 8. Examples of 25 distinct identities from the EXPRESS-50 dataset, each aligned to the same neutral expression. This snapshot represents a subset of the 50 identities available in the dataset.

7. Mesh and UV Enhancement

Without explicit mouth–interior geometry, the model exploits a shortcut: Mouth interior geometry and appearance is implicitly encoded in the expression residual $\mathbf{z}_{exp}$, entangling expression and intra-oral appearance (see Figure 13). Qualitatively, when we zero out the expression residual code $\mathbf{z}_{exp}$ the mouth interior becomes severely distorted, indicating that tooth and tongue detail is stored in the expression channel rather than in identity residuals $\mathbf{z}_{id}$. To improve geometric expressiveness and enable detailed modeling of the mouth interior, we extend the original FLAME [20] mesh topology by adding vertex sets for the upper and lower teeth, similar to VHAP [28], along with a face for the inner mouth cavity. This modification modestly increases the vertex count while preserving FLAME's semantic structure. To support Gaussian parameter prediction in these new regions, we augment the UV map to cover the extended topology. The resulting UV layout includes the mouth interior, allowing convolutional decoders to assign meaningful color, opacity, and feature values to mouth-interior Gaussians. Figure 12 illustrates the enhanced UV

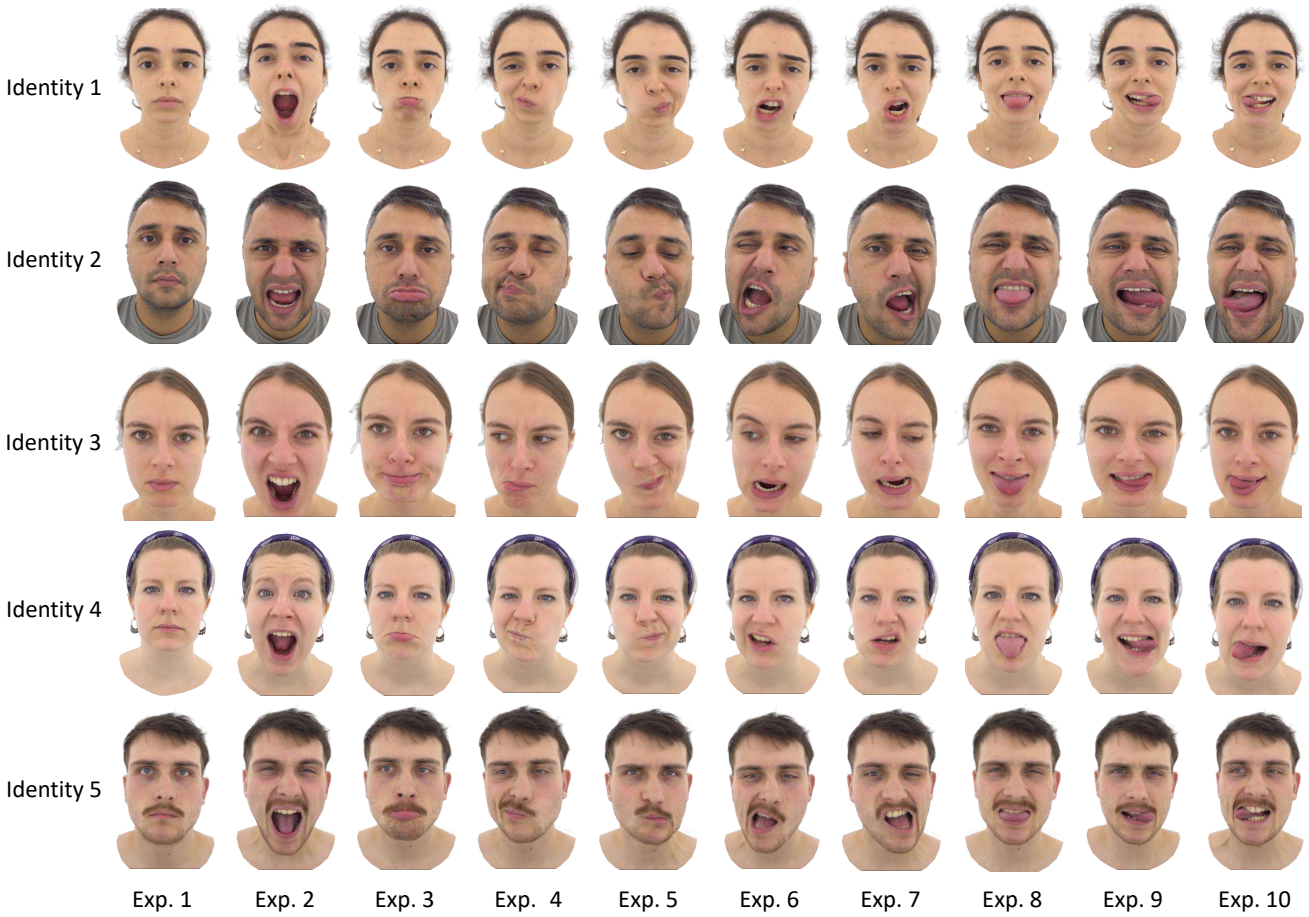

Figure 9. Diverse facial expressions from the EXPRESS-50 dataset are aligned consistently across all identities.

Figure 10. Examples of aligned expressions from the RenderMe360 dataset

layout.

8. User study

We conducted a user study to qualitatively assess how well GRMM and Morphable Diffusion preserve (i) the reference person's facial expression and (ii) identity. Each trial presented three images: a frontal reference (ground truth) on the left and two novel-view renderings from the two methods, labeled **A** and **B**. For every example, participants answered two questions: (1) which variant better preserves the expression, and (2) which variant better preserves the identity. Responses used five options: *Strong preference for A/B*, *Weak preference for A/B*, and *Equally preferred* (tie). The assignment of **A** and **B** to the underlying methods was pre-specified and counterbalanced across examples to avoid label bias. We collected responses from **20** participants, each evaluating **15** examples (two judgments per example: expression and identity), yielding **600** total judgments. As summarized in Table 4, participants showed a clear overall preference for GRMM: **94.2%** of judgments favored GRMM (sum of strong and weak preferences), **4.5%** favored Morphable Diffusion, and **1.3%** were ties.

Figure 11. Examples of reconstructed depth from ProbeSDF.

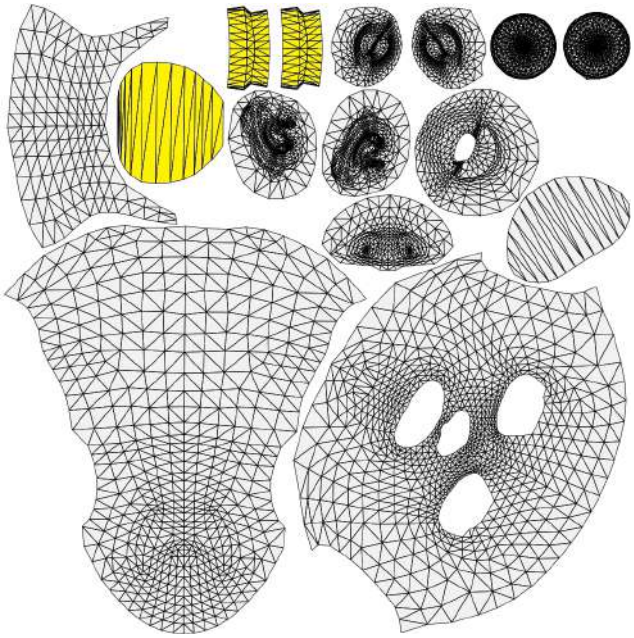

Figure 12. **UV enhancement.** The mouth interior enhancement is highlighted in *yellow*

9. Locality regularisation for inverse rendering

Locality regularization. During the second-stage refinement, we regularize the model to remain close to the pretrained solution for interpolations between the fine-tuned subject and dataset identities. Specifically,

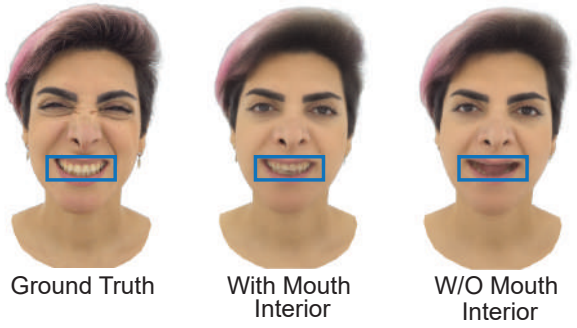

Figure 13. Zeroing the expression residual $\mathbf{z}_{exp}$ exposes a shortcut: intra-oral appearance is entangled with expression, severely distorting the mouth.

Type	Ours		Morphable Diffusion		Tie
	++	+	++	+	
Expression	86.7%	6.0%	3.7%	1.0%	2.7%
Identity	89.3%	6.3%	2.7%	1.7%	0.0%
Overall	88.0%	6.2%	3.2%	1.3%	1.3%

Table 4. User preference distribution comparing **Ours** and **Morphable Diffusion** on expression and identity preservation. Percentages are pooled over all judgments; "++" denotes strong preference and "+" denotes weak preference. Overall: **Ours** 94.2%, **Morphable Diffusion** 4.5%, Tie 1.3%.

we sample a dataset identity r with known parameters $(\mathbf{z}^r_{id}, \mathbf{z}^r_{exp}, \alpha^r_{id}, \alpha^r_{exp}, \theta^r_{jaw}, \theta^r_{neck})$ and consider the fine-tuned subject $(\mathbf{z}^*_{id}, \mathbf{z}^*_{exp}, \alpha^*_{id}, \alpha^*_{exp}, \theta^*_{jaw}, \theta^*_{neck})$. We construct interpolated parameters at a fixed $t = 0.3$:

$$\begin{aligned}
\tilde{\mathbf{z}}_{id} &= (1-t)\,\mathbf{z}^r_{id} + t\,\mathbf{z}^*_{id},\\
\tilde{\mathbf{z}}_{exp} &= (1-t)\,\mathbf{z}^r_{exp} + t\,\mathbf{z}^*_{exp},\\
\tilde{\alpha}_{id} &= (1-t)\,\alpha^r_{id} + t\,\alpha^*_{id},\\
\tilde{\alpha}_{exp} &= (1-t)\,\alpha^r_{exp} + t\,\alpha^*_{exp},\\
\tilde{\theta}_{jaw} &= \mathrm{SLERP}(\theta^r_{jaw}, \theta^*_{jaw}, t),\\
\tilde{\theta}_{neck} &= \mathrm{SLERP}(\theta^r_{neck}, \theta^*_{neck}, t).
\end{aligned}$$

We render two images using the same interpolated parameters: I_{frozen} with the pretrained frozen model, and I_{tuned} with the fine-tuned model. Our *locality loss* is then defined as:

$$\mathbf{L}_{loc} = \mathbf{L}_{rec}(I_{tuned}, I_{frozen}),$$

where $\mathbf{L}_{rec}$ is the image-space reconstruction loss defined in Sec. 3.4. This encourages the refined model to preserve the behaviour of the pretrained prior (Figure 14) along interpolation paths between known dataset identities and the personalized subject, ensuring that refinements remain localized.

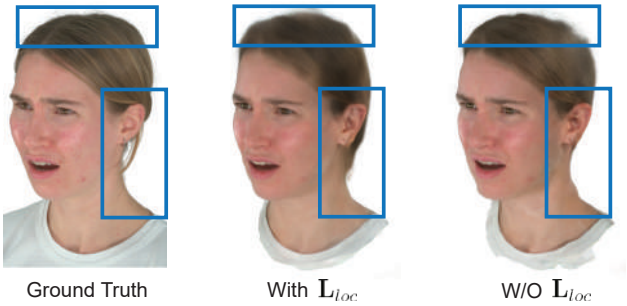

Figure 14. **Locality loss preserves the prior and improves novel views.** Qualitative ablation comparing *With locality loss* vs. *W/O locality loss*. The locality loss preserves the prior and yields sharper detail and fewer artifacts in novel views. PSNR (dB): *With* 28.43 vs. *W/O* 27.75 (+0.68). *Please zoom in for details.*

10. Additional ablations.

No Mesh Decoder. Disabling Φ_{mesh} and learning only fine residual offsets for the Gaussian primitives reduces mouth and cheek articulation, which yields poorer facial expressivity and reduced photorealism, as shown in Figure 15.

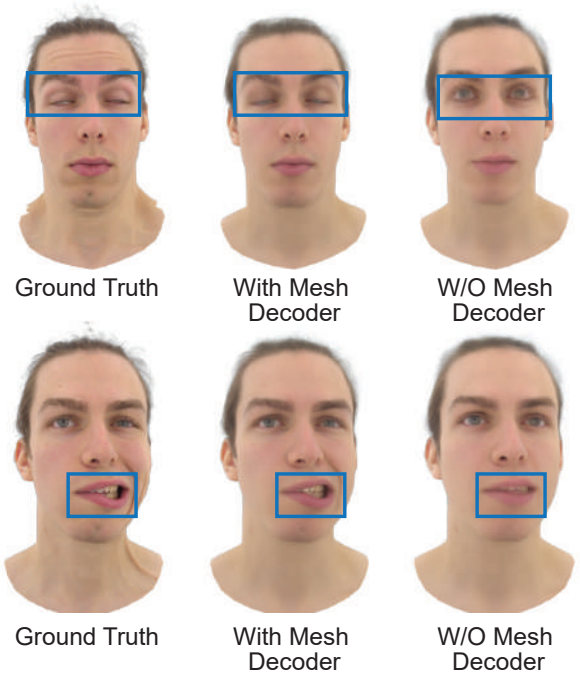

Figure 15. **No mesh decoder harms expressivity and realism.** Qualitative ablation comparing *With mesh decoder* vs. *W/O mesh decoder*. Disabling Φ_{mesh} and learning only residual Gaussian offsets weakens mouth and cheek articulation and degrades photorealism. PSNR (dB): *W/O mesh decoder* 32.34 vs. *With mesh decoder* 34.52 (+2.18). *Please zoom in for details.*

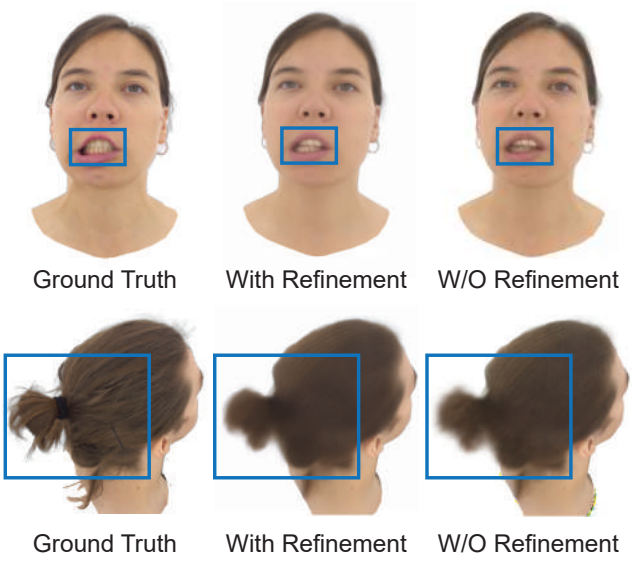

Figure 16. **Screen-space refinement improves fidelity.** Qualitative ablation comparing *With* vs. *W/O* Ψ_{ref}. Eliminating Ψ_{ref} degrades hair texture, removes mouth details, and reduces overall sharpness and fidelity—even after 3D-aware rasterization. PSNR (dB): *W/O* 30.83 vs. *With* 31.60 (+0.77). *Please zoom in for details.*

No Refinement Network. Eliminating Ψ_{ref} leads to degraded hair texture, loss of mouth details, and reduced overall sharpness and fidelity, as shown in Figure 16. This highlights the importance of screen-space correction, even after rasterization.

11. Implementation details.

The model is trained for 250,000 iterations with a batch size of 1 on four NVIDIA A100 GPUs. We use the Adam [18] optimizer with a learning rate of 1×10^{-4} for all learnable parameters.

12. Additional Results.

Disentangled control. We demonstrate disentangled control over 3DMM and residual parameters; see Figure 17.

13. Limitations and Future Work

While significantly advancing the state of the art, our model is not without limitations. One notable limitation of the method lies in its difficulty in handling out-of-distribution subjects, such as individuals with long hair or unconventional facial features, which may deviate significantly from the training data. Additionally, variations in the lighting environment can challenge the model's robustness, potentially leading to artefacts or inaccuracies in rendering. These issues underscore the need to enhance the generalizability of

the approach. A promising direction to address this limitation involves annotating a more diverse set of identities with a wide range of expressions and lighting variations, ensuring that the model can better accommodate subjects with varying appearances and environmental conditions, thereby enhancing the method's ability to generalise to more challenging real-world scenarios.

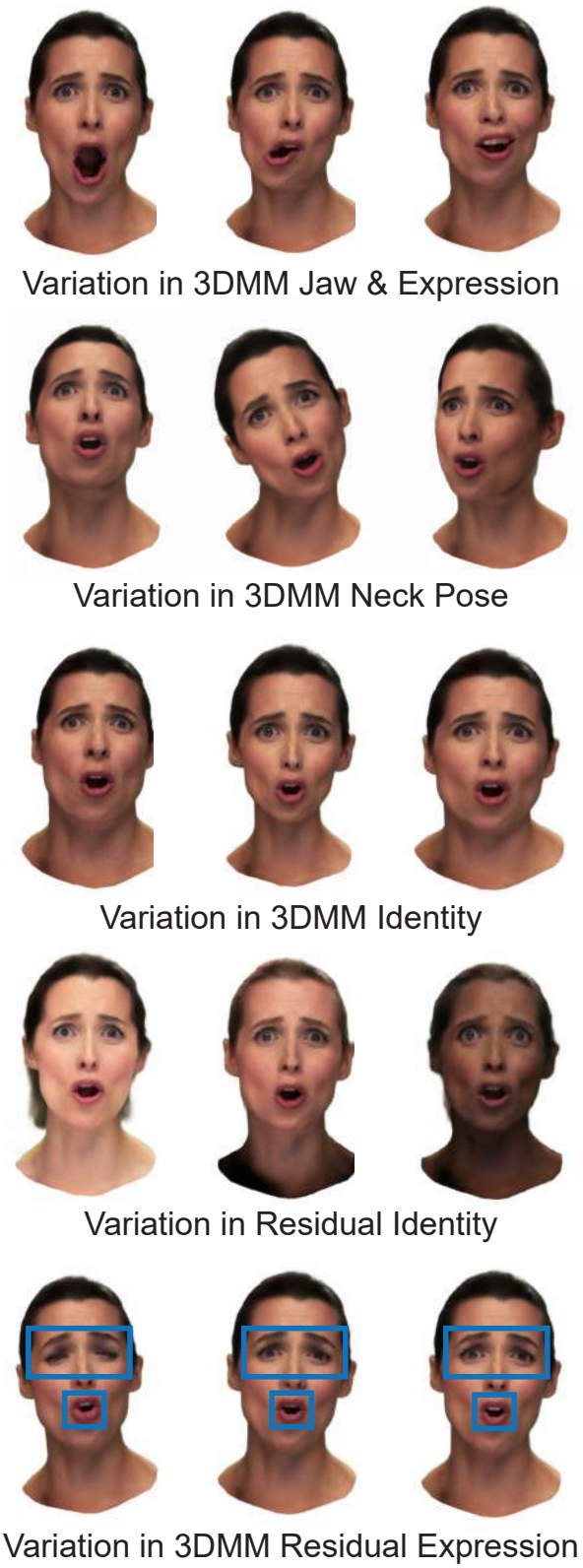

Figure 17. Distangled control of GRMM parameters.