

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

Supplementary Material

7. 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.

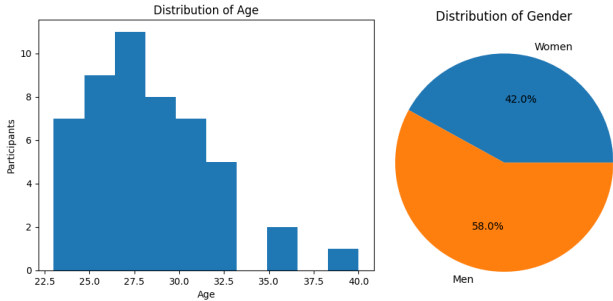


Figure 7. Statistics of the participants in our dataset.

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.

After filtering, we select 280 identities from RenderMe-360, each of which executes 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.

8. Depth Supervision with ProbeSDF

To supervise geometry, we leverage ground-truth depth maps I_{depth}^{gt} 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.

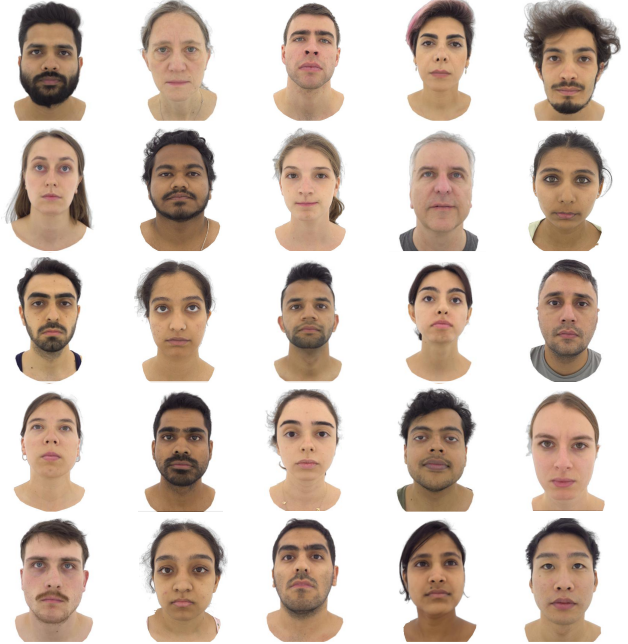


Figure 8. Examples of 25 distinct identities from the EXPRESS-50 dataset, each aligned to the same neutral expression. This snapshot represents a subset of the 50 identities available in the dataset.

9. Mesh and UV Enhancement

Without explicit mouth–interior geometry, the model exploits a shortcut: Mouth interior geometry and appearance are 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 modelling of the mouth interior, we extend the original FLAME [21] mesh topology by adding vertex sets for the upper and lower teeth, similar to VHAP [29], 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 colour, opacity, and feature values to mouth-interior Gaussians. Figure 12 illustrates the enhanced UV

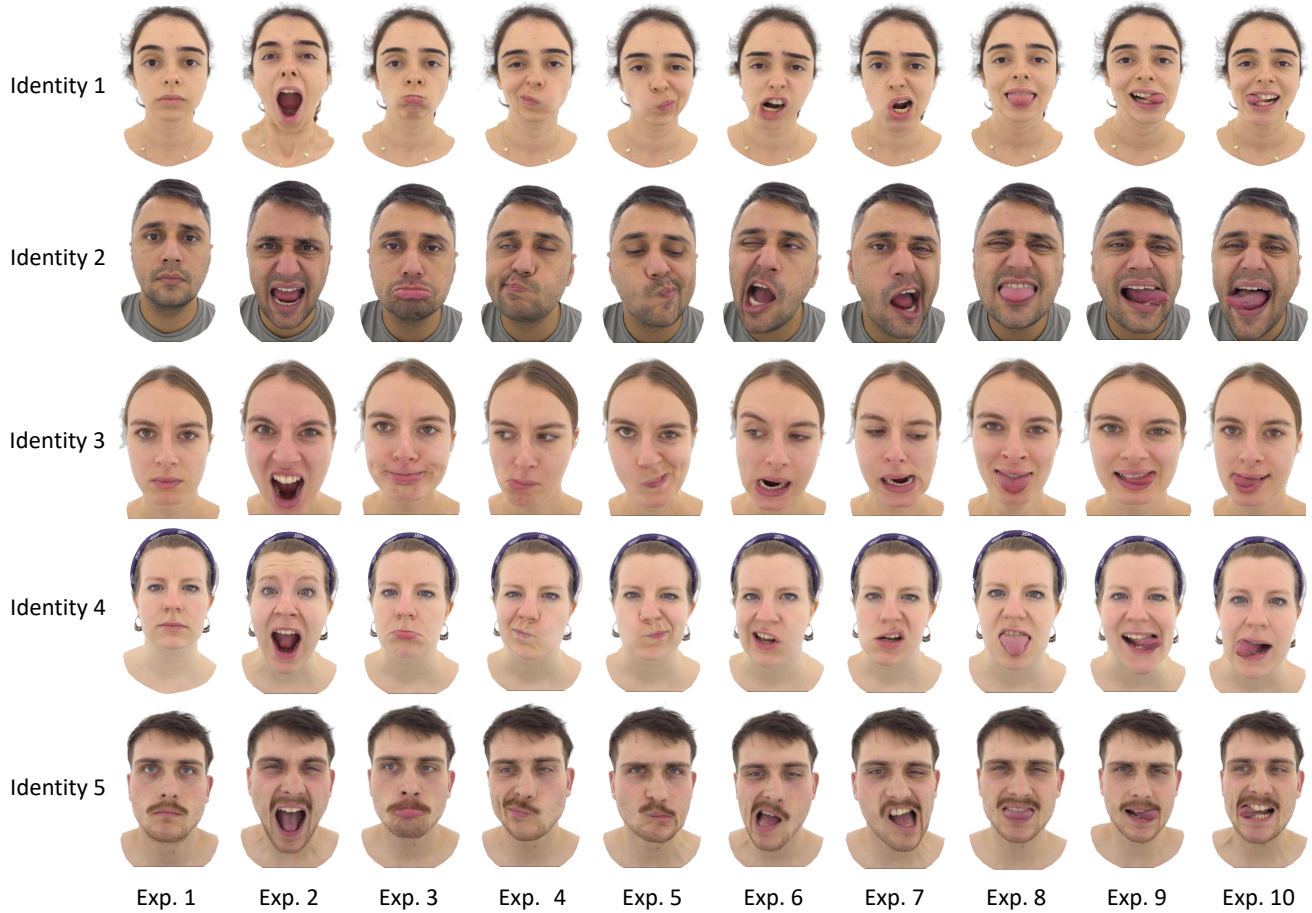


Figure 9. Diverse facial expressions from the EXPRESS-50 dataset are aligned consistently across all identities.



Figure 10. Examples of aligned expressions from the RenderMe360 dataset

layout.

10. User Study

We conducted a user study to qualitatively evaluate how well GRMM and Morphable Diffusion preserve (i) the ref-

erence person’s facial expression and (ii) identity. In each trial, participants saw three images: a frontal reference image (ground truth) on the left and two novel-view renderings from the two methods, labelled **A** and **B**. For each example, participants answered two questions: (1) which variant better preserves the expression, and (2) which variant better preserves the identity.

Responses were recorded using 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 fixed in advance and counterbalanced across examples to avoid label bias. We collected responses from **20** participants, each evaluating **15** examples, with two judgments per example (expression and identity), for a total of **600** judgments.

As summarised in Table 4, participants showed a clear overall preference for GRMM: **94.2%** of judgments favoured GRMM (combining strong and weak preferences), **4.5%** favoured Morphable Diffusion, and **1.3%** were ties.



Figure 11. Examples of reconstructed depth from ProbeSDF.

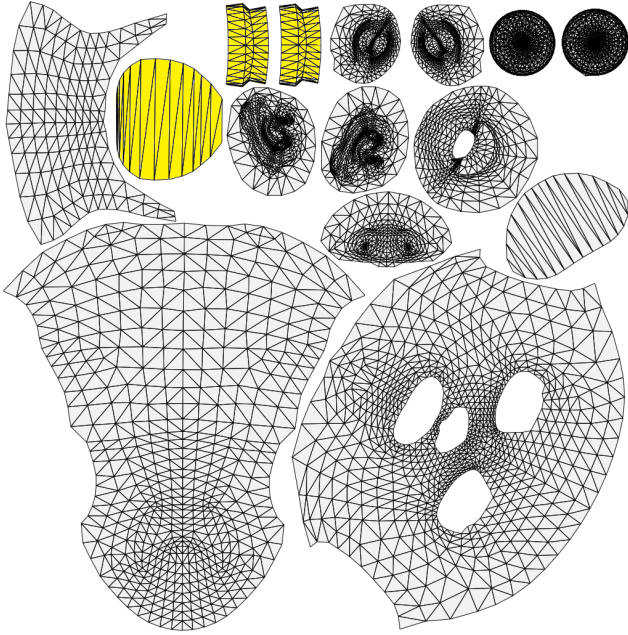


Figure 12. **Enhanced UV layout with mouth interior.** We extend the FLAME mesh with upper and lower teeth and an inner mouth surface. The UV map is expanded to cover these regions, allowing the network to assign colour, opacity, and features to mouth-interior Gaussians. The interior of the mouth is highlighted in *yellow*.

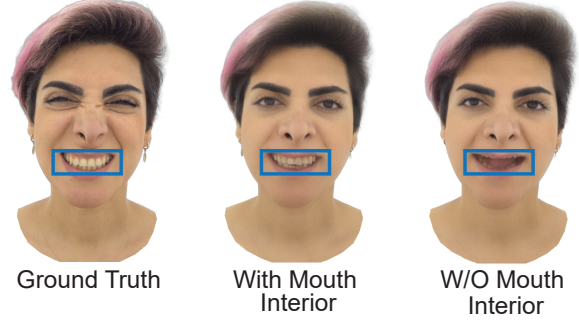


Figure 13. Zeroing the expression residual \mathbf{z}_{exp} exposes a shortcut: intra-oral appearance is entangled with expression, severely distorting the mouth.

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%.

11. Locality Regularisation for Inverse Rendering

Locality regularisation. During the second-stage refinement, we regularise the model to remain close to the pretrained solution for interpolations between the fine-tuned subject and dataset identities. Specifically, we sample a dataset identity r with known parameters $(\mathbf{z}_{id}^r, \mathbf{z}_{exp}^r, \alpha_{id}^r, \alpha_{exp}^r, \theta_{jaw}^r, \theta_{neck}^r)$ and consider the fine-tuned subject $(\mathbf{z}_{id}^*, \mathbf{z}_{exp}^*, \alpha_{id}^*, \alpha_{exp}^*, \theta_{jaw}^*, \theta_{neck}^*)$. In all experiments, we set $r = 5$ and construct interpolated parameters at a fixed $t = 0.3$:

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

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 regulariser 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 personalised subject, ensuring that refinements remain localised.

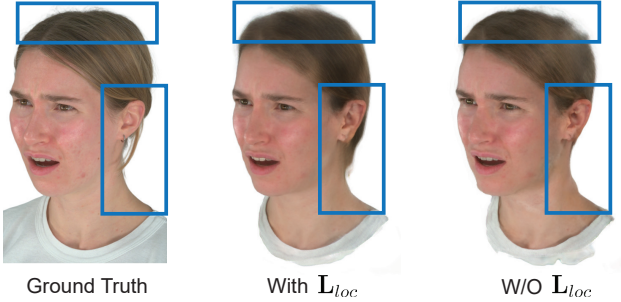


Figure 14. **Locality loss preserves the prior and improves novel view synthesis.** Qualitative ablation under *novel view synthesis*, comparing *With locality loss* vs. *W/O locality loss*. The locality loss preserves the prior and yields sharper details with fewer artefacts when rendering unseen viewpoints. PSNR (dB) on novel views: *With* 28.43 vs. *W/O* 27.75 (+0.68). *Please zoom in for details.*

12. 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.

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 rasterisation.

13. Implementation Details

Training. The model is trained for 250,000 iterations with a batch size of 1 on four NVIDIA A100 GPUs. We use the Adam [19] optimiser with a fixed learning rate of 1×10^{-4} for all learnable parameters. Gaussian primitives are initialised on the template mesh.

Inference and Inversion. At inference time, we render novel views using the learned GRMM representation. For inversion, we use a two-stage optimisation (Section 3.5). Stage 1 runs for 200 steps with a learning rate of 1×10^{-3} ,

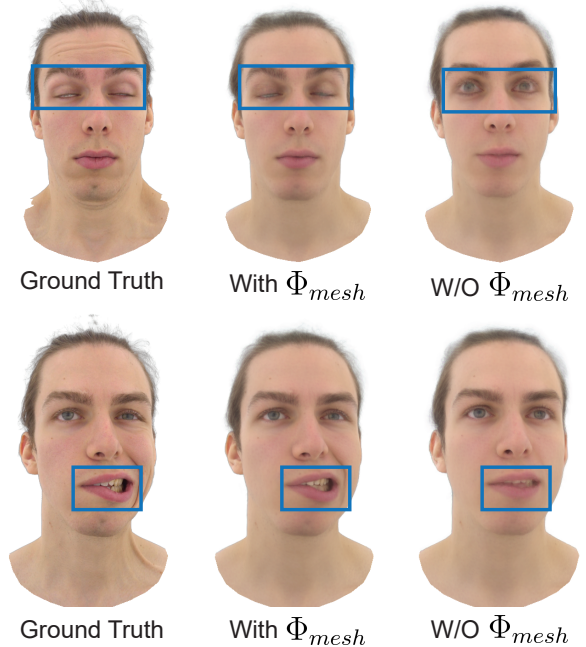


Figure 15. **No mesh decoder harms reconstruction expressivity and realism.** Qualitative ablation under the *reconstruction setting*, comparing *With mesh decoder* vs. *W/O mesh decoder*. Disabling Φ_{mesh} and learning only residual Gaussian offsets weakens mouth and cheek articulation and degrades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

followed by Stage 2 for 100 steps with a learning rate of 1×10^{-4} .

14. Additional Results

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

15. Limitations and Future Work

While significantly advancing the state of the art, our model is not without limitations. One notable limitation of the method is its difficulty 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 rendering inaccuracies. These issues underscore the need to improve the approach’s generalizability. A promising approach to address this limitation is to annotate a more diverse set of identities across a wide range of expressions and lighting conditions, ensuring the model can better accommodate subjects with varying appearances and environmental conditions, thereby enhancing

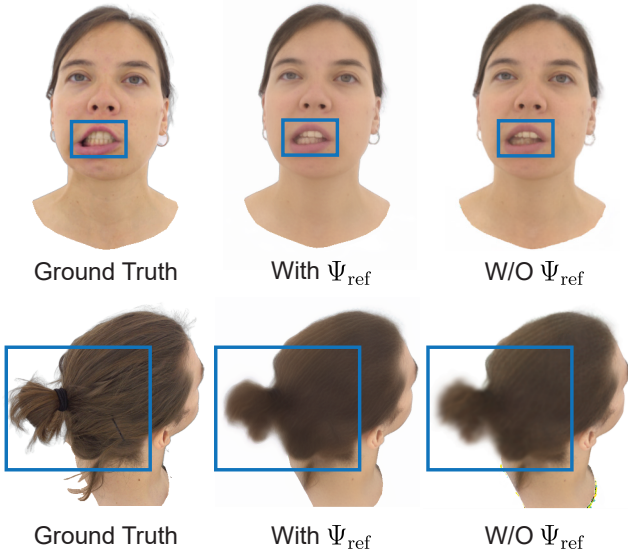


Figure 16. **Screen-space refinement improves reconstruction fidelity.** Qualitative ablation under the *reconstruction setting*, comparing *With* vs. *W/O* Ψ_{ref} . Removing Ψ_{ref} degrades hair texture, removes mouth details, and reduces overall sharpness and fidelity—even after 3D-aware rasterisation.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

the method’s ability to generalise to more challenging real-world scenarios.

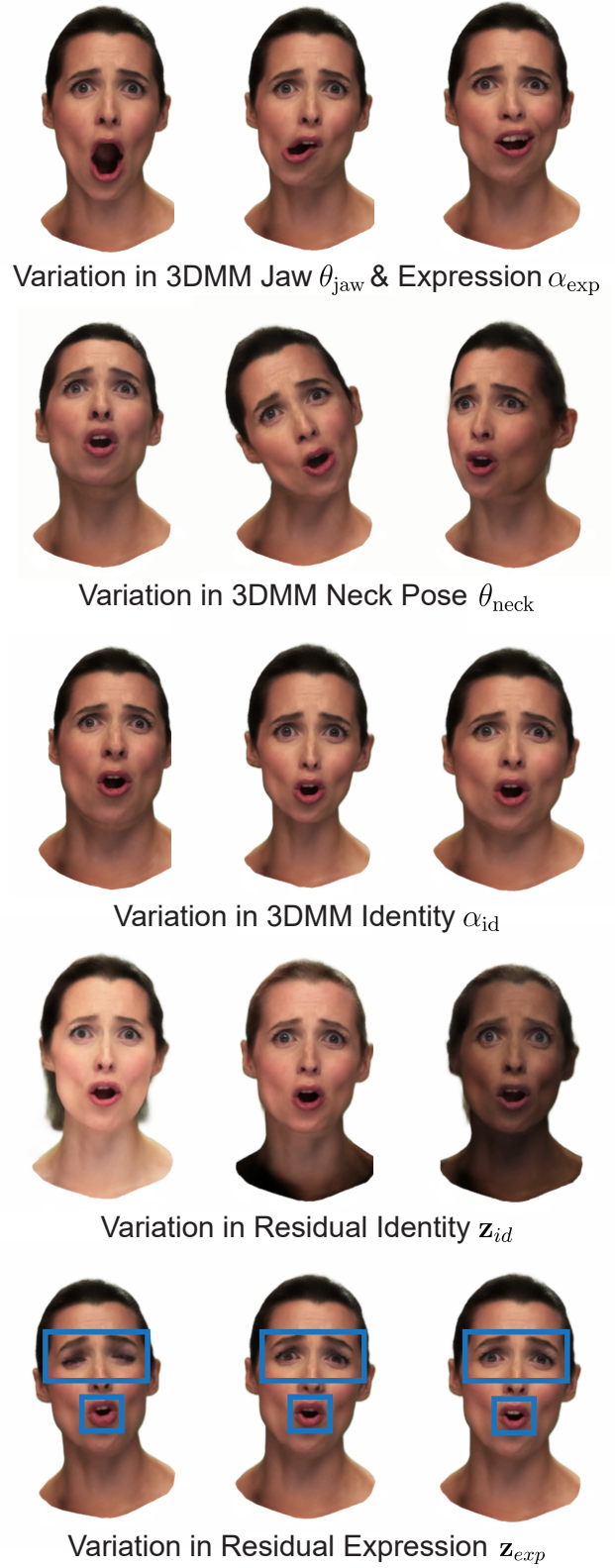


Figure 17. Distangled control of GRMM parameters.