Supplementary Material - Multimodal Conditional 3D Face Geometry Generation

This supplementary document and our supplementary video provide additional insights into our method and results.

1. Ablation Studies

To validate the benefits of representing the face geometry as a delta from the template mesh, we train a latent diffusion model (LDM) on the full vertex map representation and a second LDM on the delta vertex map representation. We qualitatively compare both representations in Fig. 1 and can observe that the delta vertex map representation leads to less artifacts compared to the full vertex map representation (e.g.on the eyelids).

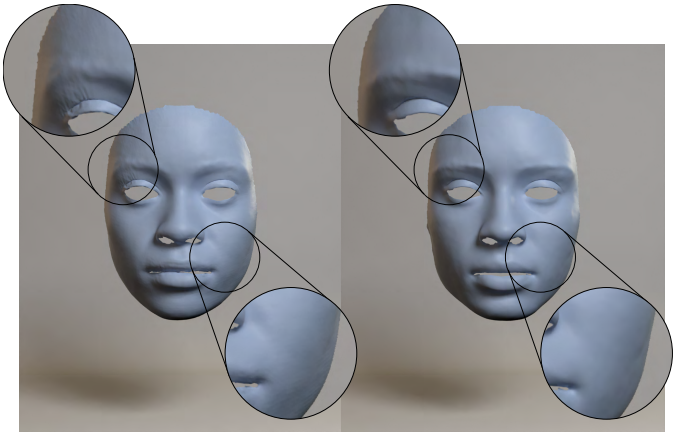

Fig. 1. Ablation study of the geometry representation. On the left is the generated geometry when training on the full vertex map representation, which shows visible artifacts (e.g., on the eyelid). Our proposed training with the delta vertex maps (right) removes such artifacts.

Next, we validate the benefits of adding geometry data augmentations to our training data. We compare FLAME parameter conditioned generations from two LDMs, where one was trained with and the other was trained without geometry data augmentations. As we illustrate in Table 1, our generations are closer to the ground truth geometry (lower vertex-to-vertex error), when using geometry data augmentations. This result indicates that using geometry data augmentations improves the models ability to capture unseen identities.

2. Implementation Details

For our dataset, we crop the full head face geometry to allocate more vertices to the face region, representing 50520 face vertices within each 256^2 UV position map. At 256^2 resolution we can represent reasonably high-resolution face geometry while being able to limit our VAE training time to 8 days using our training dataset of 7752 samples (~1.4 seconds/iteration; batch size 8). We use a learning rate of 4.5e-6 and a codebook size of 8192. Note that for training our VAE, we did not use the data augmentations described in Section 1 as the autoencoder was already able to reconstruct test geometries with high accuracy when trained only on the studio dataset. The vertex error between reconstructions and the original geometry is usually below 0.3 millimeters and only very high-frequency details are lost. We visualize the VAE reconstruction error in Fig. 2. Next, we train our LDM for 4 days (~1.6 seconds/iteration; batch size 12) with a learning rate of 1e-4 and diffusion timesteps $T = 1000$. We utilize geometry data augmentations with corresponding FLAME fits during training (+200k samples) to allow for better generalization across identities during generation. Afterwards, we train each set of cross-attention layers with a learning rate of 1e-4 for 6 days (~3.3 seconds/iteration; batch size 24) while keeping the LDM frozen. With a probability of 0.05, we randomly set either $\mathbf{c}_0$ or $\mathbf{c}_m$ or both to their null embeddings during training. This step enables classifier-free guidance at inference. Note that we do not add augmented geometry data to train the new cross-attention layers because we do not have access to paired mode-geometry data for modes such as portrait photos. We use the CLIP ViT-L/14 model [1] to extract 768 dimensional CLIP feature vectors as our conditioning representation for all modalities except the base FLAME parameter conditioning. We visualize the layers that pass our base FLAME parameter conditioning to the diffusion model in Fig. 3. These layers are trained jointly with the diffusion model parameters. Afterwards, the diffusion model and the base conditioning layers are frozen. For adding a new modality a newly added set of mode-specific layers are trained (linear layer, layer norm and a set of cross-attention layers). Sketches, portrait photos, Canny edges and landmarks are passed through a frozen CLIP image encoder before reaching their own mode-specific trainable layers. We train a different set of layers per mode (e.g. one for sketches, one for portrait photos etc.). Text is passed through a frozen CLIP text encoder before reaching text-specific trainable layers (Fig. 3). All training experiments were run on a single RTX A6000 GPU. Inference was run on single RTX A6000 GPUs, 3090 GPUs, and 1080 GPUs. Also note that the VAE and the LDM have to be trained only once and novel condition-

Table 1. We compare the 3D face geometry generated by diffusion models that were trained with and without 3D data augmentations. We measure the vertex-to-vertex error (V2V) in mm between FLAME parameter conditioned generations and ground truth geometry on neutral shapes from our validation set. The model trained using data augmentation is able to capture unseen identities better. Results are averaged over three different seeds.

V2V error	Mean ↓	Median ↓	Std ↓
No augmentations	4.093	3.692	2.170
With augmentations	**3.757**	**3.352**	**2.085**

ing modes can be added by training only the new cross-attention layers. Unless mentioned otherwise, we generate every result by running DDPM sampling steps $S = 50$ with conditioning strength $w = 1$. The average time to generate a geometry sample with our diffusion model is ~6 seconds on a single 3090 GPU. To aid the visual similarity for the comparison with the state-of-the-art methods, we complete the head by deforming a template head to match our generated face. We run the CLIP score evaluation for all methods on full head renders. We align all 3D faces to the same space, before rendering each with the same camera. We use the CLIP ViT-B/32 variant for the score calculation and report the average score for each method.

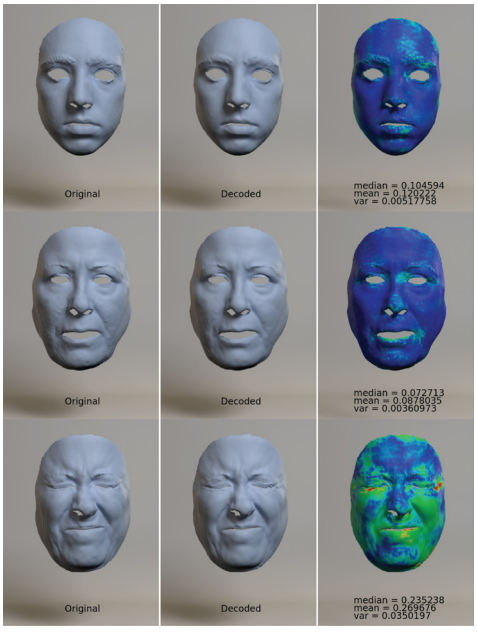

Fig. 2. The VAE reconstruction error is usually below 0.3 millimeters when compared to the ground truth geometry. Some high frequency details are lost after encoding and decoding the original geometry with the VAE due to VAE compression.

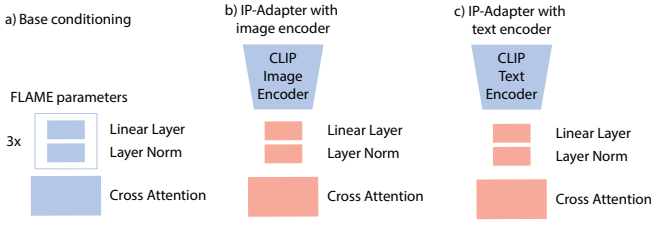

Fig. 3. Modality injection visualization. Our FLAME parameter base conditioning layers (a) are trained jointly with the diffusion model. Afterwards, both are frozen and only the new mode-specific layers are trained. Sketches, portrait photos, Canny edges and landmarks are passed through a frozen CLIP image encoder (b) before reaching their own set of mode-specific trainable layers (one set of layers per mode). Text is passed through a frozen CLIP text encoder (c) before reaching a set of text-specific trainable layers.

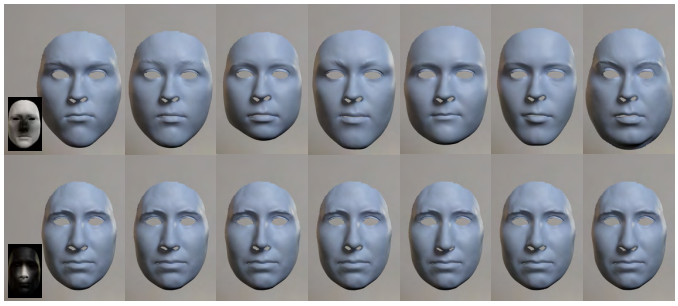

Fig. 4. Unconditional face shape editing (inpainting). In the top row, the nose is kept fixed, while we sample the remaining regions unconditionally. In the bottom row, we sample the nose region unconditionally while keeping the other regions fixed.

3. Exact Text Prompts

Table 2 lists the exact text prompts used for the CLIP score comparison and the figures in the main document.

4. Additional Geometry Editing Results

Further mask-based editing of facial geometries using our model is shown in Fig. 4. In the top row of Fig. 4 we mask the nose region of the latent position map, such that it remains fixed throughout the multiple steps of denoising. We then generate multiple geometry samples by varying the initial noise input to the diffusion model. The noise predicted at each denoising step is multiplied with the nose mask before being fed as input to the denoising UNet for the next time step. This denoising procedure leads to generations where the generated samples all share the same nose shape, but vastly differing facial identities. In the bottom row of Fig. 4, we show the result of inverse masking, where the face shape is held fixed while allowing the nose shape to change. Our model produces meaningful results in both cases.

5. Additional Quantitative and Qualitative Results

For the base FLAME parameter conditioning, we visualize the error maps to the ground truth scanned geometry in Fig. 5. Next, we compare the text-to-geometry generation results of HeadArtist [2] and HumanNorm [3] with our method. Both are based on Deep Marching Tetrahedra [4] and SDS optimization [5] and can represent face parts beyond the skin. The extracted face geometries differ in topology and optimizing for one sample takes around one hour on a 3090 GPU. In contrast, our method's inference speed is 1000-times faster on a 3090 GPU and produces results in a single common topology.

6. Failure Cases

We do observe geometric artifacts around the mouth region for extreme expressions (Figure 6, column 2), due to limited extreme expressions in our training data. Additionally, the generated face geometry can open the mouth to the wrong side when conditioning with CLIP embeddings (Figure 6, column

Table 2. The exact text prompts used in the comparison to the related work methods. Prompts 1 - 10 specify a neutral expression, while prompts 11 - 20 specify other facial expressions.

Nr.	Text prompt
1	A shaded, textureless 3D face model of an African woman with a neutral expression.
2	A shaded, textureless 3D face model of an overweight man with a neutral expression.
3	A shaded, textureless 3D face model of an old woman with a neutral expression.
4	A shaded, textureless 3D face model of a middle-aged Asian person with a neutral expression.
5	A shaded, textureless 3D face model of a middle-aged Caucasian woman with a neutral expression.
6	A shaded, textureless 3D face model of a woman with high cheekbones, a defined jawline, and a straight nose with a neutral expression.
7	A shaded, textureless 3D face model of a young woman with a round face, big eyes and small mouth with a neutral expression.
8	A shaded, textureless 3D face model of a man with a chubby face, a wide forehead and a wide nose with a neutral expression.
9	A shaded, textureless 3D face model of a young African man with a neutral expression.
10	A shaded, textureless 3D face model of a young Asian man with a neutral expression.
11	A shaded, textureless 3D face model of a smiling overweight man.
12	A shaded, textureless 3D face model of an overweight man shouting angrily.
13	A shaded, textureless 3D face model of a sad Caucasian man.
14	A shaded, textureless 3D face model of a middle-aged Asian person with a kiss face expression.
15	A shaded, textureless 3D face model of a smiling African woman.
16	A shaded, textureless 3D face model of a woman with high cheeckbones, a defined jawline, and a straight nose. Her mouth is opened to the side.
17	A shaded, textureless 3D face model of an angry Caucasian man.
18	A shaded, textureless 3D face model of a man with a chubby face, a wide forehead and a wide nose with a big smile on his face.
19	A shaded, textureless 3D face model of a young African man with a closed eyes facial expression.
20	A shaded, textureless 3D face model of a young Asian man with a very surprised facial expression. His eyes and mouth are wide open and the eyebrows raised.

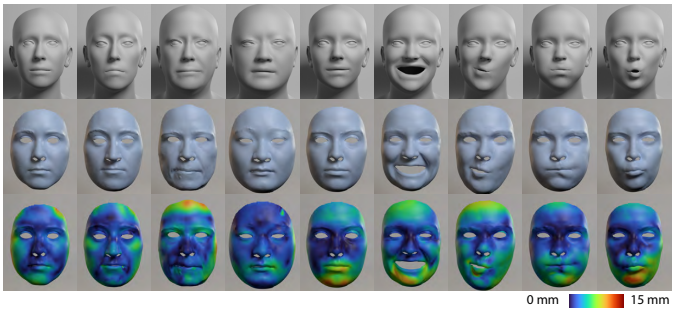

Fig. 5. Error maps on our validation set. The first row shows the FLAME mesh as generated by the FLAME face model from the input FLAME parameters. The second row shows the generated geometry from our model conditioned on the respective FLAME parameters. The third row visualizes the error from our conditional generations to the original scanned geometry in our validation set. The first four columns are various identities, while the last five columns are different expressions of the same subject.

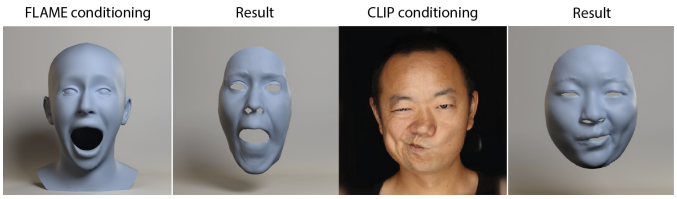

Fig. 6. Failure cases. Our results can display geometric artifacts around the mouth area for extreme expressions (column 2) and sometimes incorrect mouth opening sides when conditioned on CLIP embeddings (column 4).

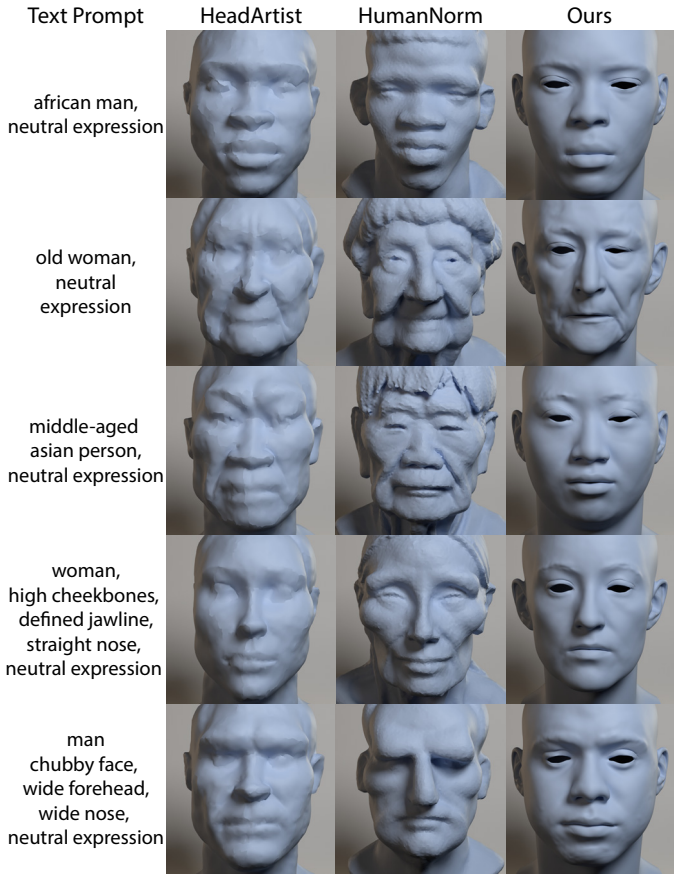

Fig. 7. Qualitative comparison with the text-to-geometry generation ability of HeadArtist [2] and HumanNorm [3]. For legibility, we shortened the text prompt. Please refer to Table 2 for the exact text prompts.

3 and 4). We identify that this behavior occurs, when the CLIP embeddings for both mouth opening directions (left/right) are extremely similar (cosine similarity close to 1). Thus, this behavior is caused by the similarity of specific conditionings and not by the diffusion model.

References

[1] Radford, A, Kim, JW, Hallacy, C, Ramesh, A, Goh, G, Agarwal, S, et al. Learning transferable visual models from natural language supervision. In: Meila, M, Zhang, T, editors. Proceedings of the 38th International Conference on Machine Learning (PMLR); vol. 139 of *Proceedings of Machine Learning Research*. PMLR; 2021, p. 8748–8763.

[2] Liu, H, Wang, X, Wan, Z, Shen, Y, Song, Y, Liao, J, et al. Headartist: Text-conditioned 3d head generation with self score distillation. In: ACM SIGGRAPH 2024 Conference Papers. SIGGRAPH '24; New York, NY, USA: Association for Computing Machinery. ISBN 9798400705250; 2024,URL: `https://doi.org/10.1145/3641519.3657512`. doi:`10.1145/3641519.3657512`.

[3] Huang, X, Shao, R, Zhang, Q, Zhang, H, Feng, Y, Liu, Y, et al. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. 2024.

[4] Shen, T, Gao, J, Yin, K, Liu, MY, Fidler, S. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In: Advances in Neural Information Processing Systems (NeurIPS). 2021,.

[5] Poole, B, Jain, A, Barron, JT, Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion. In: The Eleventh International Conference on Learning Representations (ICLR). 2023,.