Computers & Graphics 132 (2025) 104325

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on SMI 2025

Multimodal Conditional 3D Face Geometry Generation

Christopher Otto [a,b,*], Prashanth Chandran [b], Sebastian Weiss [b], Markus Gross [a,b], Gaspard Zoss [b], Derek Bradley [b]

[a] ETH Zürich, Switzerland
[b] DisneyResearch|Studios, Switzerland

ARTICLE INFO

Keywords:
Multimodal generation
3D face geometry
Deep learning

ABSTRACT

We present a new method for multimodal conditional 3D face geometry generation that allows user-friendly control over the output identity and expression via a number of different conditioning signals. Within a single model, we demonstrate 3D faces generated from artistic sketches, portrait photos, Canny edges, FLAME face model parameters, 2D face landmarks, or text prompts. Our approach is based on a diffusion process that generates 3D geometry in a 2D parameterized UV domain. Geometry generation passes each conditioning signal through a set of cross-attention layers (IP-Adapter), one set for each user-defined conditioning signal. The result is an easy-to-use 3D face generation tool that produces topology-consistent, high-quality geometry with fine-grain user control.

1. Introduction

The creation of 3D facial geometry for digital human characters is a modeling task that usually requires tremendous artistic skill. Digital sculpting with 3D modeling tools is a time-consuming and demanding process, especially when the target is as recognizable as a human face. This complexity has prompted research into data-driven sculpting methods [1] and other, more user-friendly, interactive interfaces [2].

Several common morphable 3D face models (e.g. FLAME [3]) simplify the facial modeling task by providing a shape subspace to operate in, as well as simple parameters to control the identity and expression geometry without the need for 3D modeling skills, but they are limited in expressiveness and offer only basic control knobs.

Recent methods can create high quality 3D geometry and textures from text prompts [4–8] via optimization, leveraging large pre-trained text-to-image diffusion models [9]. These methods allow layman users to create 3D faces through natural text descriptions. While this is a powerful approach, it can still be difficult to achieve a particular output through text description [10]. Some concepts like the specific curvature of a face or a unique facial expression are much easier to convey via sketches, edge contours or portrait photos than through text.

In the image domain, approaches like ControlNet [11] or T2I-Adapter [12] have demonstrated controllable image generation beyond text using sketches, images, or edge maps as conditioning signals. These methods provide users with much more fine-grained control over the generation process than text-based methods alone. Ye et al. [13] propose IP-Adapters to control Stable Diffusion [9] with image prompts by learning new cross-attention layers. However, image-based methods are not easy to extend to 3D facial geometry generation.

We present a flexible new method for 3D facial geometry generation that creates high quality faces from any one of various inputs, including sketches, 2D landmarks, Canny edges, FLAME face model parameters, portrait photos and text. Our approach is to train a conditional diffusion model on a high quality 3D facial dataset constructed from high resolution scans [14] represented in the 2D UV domain. Our model is trained from scratch, without the need for a pre-trained foundation model like Stable Diffusion. To condition our model we train one set of cross-attention layers for each type of conditioning input, following IP-Adapter [13]. First, the diffusion model learns to inject FLAME parameters via the original UNet cross-attention layers. We then freeze the diffusion model while training additional sets of cross-attention layers (e.g. one for artist sketches, 2D landmarks, portrait photos, etc.). Our FLAME-conditioned model allows us to re-interpret FLAME-parameterized faces in a generative sense - providing a space of high resolution stochastic variations on top of the traditional low resolution FLAME model.

Our method allows for fast and user-friendly creation of 3D digital character faces with expressions, generated with a consistent mesh topology, and controlled by the input mode preferred by the user, all within a single model. We demonstrate several applications of our method including sketch-based 3D face modeling, geometry from 2D facial landmarks, Canny edges, or portrait photos, text-to-3D facial geometry, and finally, extending the FLAME model space by allowing

* Corresponding author at: ETH Zürich, Switzerland.
E-mail address: christopher.otto@inf.ethz.ch (C. Otto).

https://doi.org/10.1016/j.cag.2025.104325
Received 15 July 2025; Accepted 22 July 2025
Available online 13 August 2025

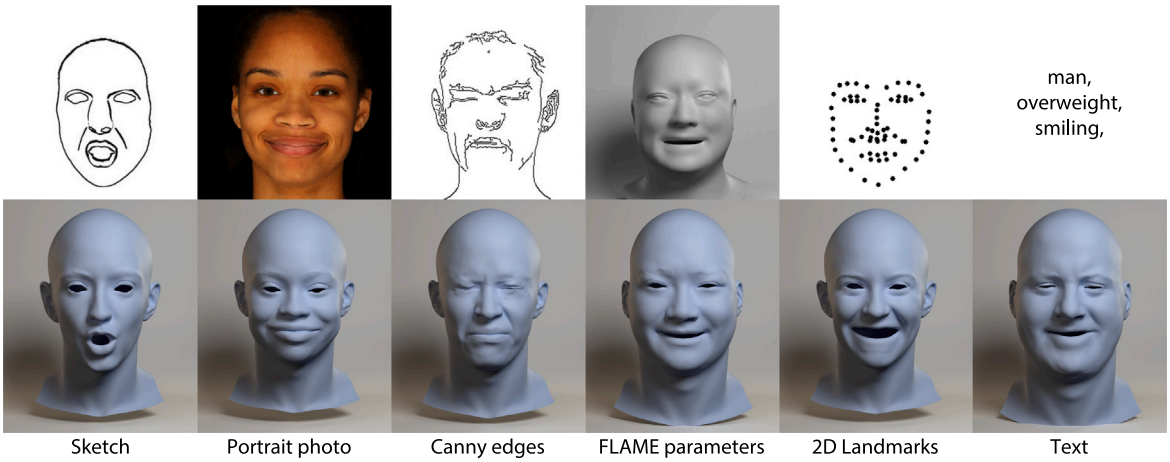

Fig. 1. We propose a novel method for diffusion-based controllable 3D face geometry generation that allows for controlling the results via several conditioning modes: artistic sketches, portrait photos, Canny edges, FLAME parameters, 2D facial landmarks, and text.

stochastic diffusion sampling conditioned on the same semantic FLAME parameters (Fig. 1). In summary, we make the following contributions:

- We present a new method for 3D face geometry generation from 6 different types of conditionings (prompts) within a single model.
- We propose a comprehensive solution for training such a method from scratch, with 3D geometry data augmentations and by representing 3D geometry as position maps to better fit existing diffusion pipelines.
- We show that our method supports face generation with expressions, sketch-based editing for 3D face design, stochastic variations of details conditioned on low resolution FLAME faces, generalization to in-the-wild data and dynamic face generation from videos.

2. Related work

In the following, we present relevant related work on 3D face geometry generation with diffusion models, as well as on injecting additional control modes into diffusion models.

2.1. 3D face geometry generation

Recent work uses diffusion models to control the generation of novel 3D face geometry. ShapeFusion [15] generates face geometry by running the diffusion process directly on the mesh input vertices. It allows unconditional and conditional face geometry generation and supports various editing operations on a given mesh, based on selected vertices (anchor points). However, it does not support conditioning signals beyond vertices. 4DFM [16] trains an unconditional diffusion model on a set of sparse 3D landmarks for facial expression generation. It can generate dynamic facial expression sequences based on 3D landmarks by retargeting the landmarks to a mesh after the diffusion process. While they support conditioning with different signals such as expression labels and text, they achieve control via classifier-guidance [17] which requires training additional classifiers on noisy data.

Other methods focus on 3D face or head avatars, which can generate 3D geometry and texture. Rodin [18] can generate triplane-based head geometry with text or image conditioning, but the resulting geometry is extracted with Marching Cubes [19] and thus not in the same topology across generations. HeadArtist [6], HeadSculpt [7] and HumanNorm [20] use Deep Marching Tetrahedra [21], text-prompts and a Score Distillation Loss (SDS) [22] to generate high-quality human heads. However, the topology of the extracted geometry differs across samples and generating a single geometry sample takes almost one hour on a single 3090 GPU. DreamFace [4], FaceG2E [5] and Bergman et al. [23] propose 3D Morphable Model (3DMM)-based [24] pipelines that generate topology-consistent 3D face geometry and textures from text. The geometry is created by optimizing 3DMM parameters using a SDS loss. During the optimization, the SDS loss uses the feedback from a pre-trained text-to-image latent diffusion model to update the 3DMM parameters given a geometry render. DreamFace [4] and FaceG2E [5] focus on optimizing 3DMM identity parameters and have therefore difficulties in directly generating facial expressions from text prompts. However, DreamFace does support generating faces with expressions from image prompts and FaceG2E results can be imported into the CG pipeline where facial expressions can be added in a separate step after the text-based generation. In general, SDS-based methods rely on the Stable Diffusion (SD) [9] prior which was pre-trained on billions of images [25] to guide their generations. SDS optimization is usually much slower in runtime compared to standard diffusion sampling, as it must backpropagate gradients from a diffusion model to a 3D model via a differentiable renderer for many optimization steps. Describe3D [26] can generate 3D face geometry from text-prompts without diffusion, by mapping CLIP text embeddings to 3DMM shape parameters. However, it does not support different facial expressions.

In our work, we generate controllable 3D face geometry in a single common topology from several different conditioning modes. Our method natively supports facial expressions and does not rely on SDS optimization, classifier-guidance or the SD prior.

2.2. Multimodal conditional image generation

Image generation with diffusion models can be controlled with conditioning modes (prompt types) that are different from text such as sketches [27], Canny edges [28], RGB images [13], expression parameters [29] or face shape [30,31]. To control pre-trained diffusion models with new modes, ControlNet [11] introduces a trainable copy of the diffusion model's UNet encoding blocks, which take the new conditioning as input. The output of the copied model is added to the skip-connections of the frozen pre-trained diffusion model. A separate trainable copy of the UNet encoding blocks (361M parameters) is created per conditioning mode. T2I-Adapter [12] aligns the internal knowledge of a pre-trained diffusion model with new control modes by proposing a small adapter network that achieves control similar to ControlNet, while requiring less parameters (77M). IP-Adapter [13] injects each conditioning via separate cross-attention layers [32] while requiring even less parameters (22M). It introduces new cross-attention layers whose outputs are added to those of the original UNet. Ye et al. [13] show that the diffusion model follows the added conditioning signal closely, when it comes through the newly trained cross-attention

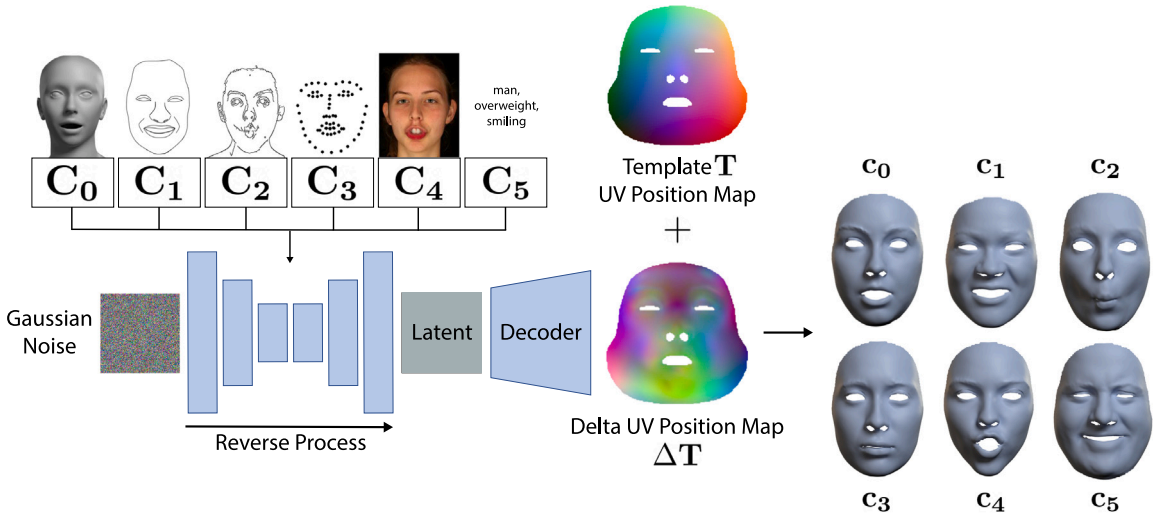

Fig. 2. Our pipeline for diffusion-based controllable 3D face geometry generation which uses a delta UV position map representation ΔT to generate results. We can control the results with several conditioning modes (i.e. FLAME parameters c_0, sketches c_1, Canny edges c_2, 2D landmarks c_3, portrait photos c_4 and text c_5).

layers. In general, adapters can add control with new conditioning modes even long after the training of the underlying base diffusion model is concluded. They can also add new conditioning modes for which only limited paired training data is available, because the underlying diffusion model is frozen, avoiding issues such as catastrophic forgetting [11,33]. While pre-trained text-to-image diffusion models understand the RGB image domain and can generate images conditioned on a large variety of input modalities, the domain gap to 3D face geometry is large. Therefore, many related works use the rather slow SDS optimization procedure to lift faces into 3D. To allow for faster inference-time sampling, we train our diffusion model from scratch using ground truth 3D face data and geometry data augmentations. We represent 3D face geometry in the 2D UV domain, which enables us to train a 2D diffusion model that can incorporate new conditioning modes using IP-Adapters.

3. Multimodal 3D face geometry generation

We propose a novel method for diffusion-based controllable 3D face geometry generation that allows for controlling the results with several conditioning modes (i.e. sketches, 2D landmarks, Canny edges, FLAME parameters, portrait photos and text).

Our method consists of four components: First, we create a dataset of 3D faces, where each face is represented as a UV position map describing the vertex positions (Section 3.1). This data representation can be easily processed with 2D convolutional neural networks. Second, a variational autoencoder (VAE), which compresses our UV position map face data into a latent space representation (Section 3.2). Third, a latent diffusion model (LDM), which learns a non-linear, deep controllable face model in latent space (Section 3.3). Fourth, we learn mode-specific cross-attention layers (IP-Adapters) with the ability to transform and inject conditioning modes into the LDM for controllable 3D face geometry generation (Section 3.4). Each of the components is explained in the following and visualized in Fig. 2.

3.1. Dataset and geometry representation

To represent our face geometries, we generate a novel dataset based on the 3D scan data acquired by Chandran et al. [14], where all faces are stabilized and in topological correspondence. In total, we use 323 identities in our training dataset, where each identity shows 24 different facial expressions (7752 examples). The full dataset contains 341 identities and we randomly choose around 5% (all expressions of 18 identities) as a validation set. We subtract a template face shape **T** from all faces in the dataset and thus represent each individual face as a delta from the template face. The computed delta representation Δ**T** reduces artifacts in the generated 3D face geometry, when compared to the full face representation. We transform each delta face into a vertex delta position map in UV space, which is suitable for being processed by neural networks [34–36]. This representation records the x, y, z coordinates of the face geometry within a 3-channel image, similar to traditional color texture maps, but instead of an RGB value at each pixel we store the x, y, z delta values. To improve generalization, we augment our existing geometry training data by synthetic identities which we generate via identity interpolation (50k examples) and by mixing face parts of different identities together [37] (150k examples). Adding the augmentations during LDM training improves the generalization to novel identities when conditioning on the FLAME parameters. We kindly refer to our supplementary material for ablation studies on the geometry representation and the use of geometry data augmentations. Combining original and augmented data leads to a total training dataset size of 207 752 examples. Thus our training dataset is slightly larger in size compared to related 2D image diffusion models that specialize on human faces (e.g. 30k images for the CelebA-HQ [38] face LDM [9], or the 70k images for the Diffusion Autoencoder in Preechakul et al. [39]), but still much smaller than the datasets required to train general foundation diffusion models that can represent various objects beyond faces. According to Kadkhodaie et al. [40] diffusion models trained on around 100k samples provide evidence of strong generalization in the face domain. We use the parameter space of a common 3D morphable face model (FLAME [3]) as a base conditioning because we can fit FLAME to the scan data and to the augmented data and thus generate a large dataset of paired geometry-FLAME parameter data. Additionally, we create paired training data for several conditioning modes that only have limited paired data available. For example, portrait photos are only available for the scans from the dataset of Chandran et al. [14], but not for the augmentation data. However, it is possible to inject new modalities with limited paired data by training new cross attention layers while keeping the LDM frozen as shown by Ye et al. [13] (and described in Section 3.4).

3.2. Variational autoencoder

To reduce the computational requirements for the diffusion model, we downsample our 256^2 UV position map data by a factor of four into a 64^2 latent space using a variational autoencoder (VAE) [41] consisting of an encoder $\mathcal{E}$ and a decoder D. We train the VAE from scratch following the autoencoder loss function and architecture as it is

presented in related work [9,42]. Specifically, we use the VQ-GAN [42] autoencoder loss:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{GAN}} + \mathcal{L}_{\text{reg}}. \tag{1}$$

$\mathcal{L}_{rec}$ consists of a pixel-wise L1 loss and a LPIPS [43] perceptual loss. It compares the input UV position maps to the reconstructions through the VAE. $\mathcal{L}_{GAN}$ evaluates inputs x and reconstructions $D(\mathcal{E}(x))$ with a patch-based discriminator [44] and $\mathcal{L}_{reg}$ employs a codebook loss which serves as a latent space regularizer. Please refer to Esser et al. [42] for more details.

3.3. Latent diffusion model

Next, we train a latent diffusion model (LDM) [9], that learns to generate latent UV position maps $\mathbf{z} = \mathcal{E}(\mathbf{x})$. To train the LDM, a forward diffusion process is defined as a Markov chain, which noises the latents $\mathbf{z}$ following a fixed noise schedule of T uniformly sampled timesteps. At the last time step T, the distribution is Gaussian. We can directly sample $\mathbf{z}_t$ at an arbitrary timestep t by:

$$\mathbf{z}_t(\mathbf{z}_0, \boldsymbol{\epsilon}) = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

where $1 - \bar{\alpha}_t$ describes the variance of the noise and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$ according to a fixed noise schedule.

We learn to predict the noise $\boldsymbol{\epsilon}$ that was added to a noisy latent image $\mathbf{z}_t$ following Ho et al. [45]:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathbf{z}_0, \mathbf{c}_0, t, \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}_0, t)\|_2^2 \right], \tag{3}$$

where t is the timestep, $\mathbf{c}_0 = \rho_\phi(\mathbf{y})$ is a FLAME parameter conditioning and $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}_0, t)$ is the UNet [46] neural network with parameters θ. The FLAME conditioning $\mathbf{c}_0$ is obtained by fitting FLAME to the face geometry encoded in $\mathbf{z_t}$ and mapping it through a MLP $\rho_\phi(\mathbf{y})$.

During inference (reverse diffusion process) we generate latent 2D UV position maps from the model distribution. We start from $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and iteratively compute less noisy latents until we reach a clean latent sample $\mathbf{z}_0$. Sampling following DDPM [45] or DDIM [47] computes $\mathbf{z}_{t-1}$ from $\mathbf{z}_t$ based on the UNet output.

3.4. Multimodal conditional generation

To control the generations with additional conditioning modes (beyond the FLAME parameters $\mathbf{c}_0$), we train different sets of cross-attention layers, following IP-Adapter [13]. The LDM itself is kept frozen. In this way, we can integrate novel conditioning modes post-LDM training even with limited paired mode-geometry data. We train one set for each of the following conditioning modes: sketches $\mathbf{c}_1$, Canny edges $\mathbf{c}_2$, 2D landmarks $\mathbf{c}_3$, portrait photos $\mathbf{c}_4$ and text $\mathbf{c}_5$. The output of the new cross-attention layers is added to the outputs of the existing LDM cross-attention layers, thereby injecting the new conditioning signal into the generation process:

$$\mathbf{Z} = \mathit{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + \mathit{Attention}'(\mathbf{Q}, \mathbf{K}'_m, \mathbf{V}'_m), \tag{4}$$

where $\mathbf{Q}$ are the intermediate UNet query features, $\mathbf{K}$ and $\mathbf{V}$ are keys and values for our FLAME conditioning $\mathbf{c}_0$ and $\mathbf{K}'_m$, $\mathbf{V}'_m$ are keys and values for the newly injected modality $\mathbf{c}_m$.

$$\begin{aligned} &\mathbf{K} = \mathbf{c}_0 \cdot \mathbf{W}_k, \mathbf{V} = \mathbf{c}_0 \cdot \mathbf{W}_v, \\ &\mathbf{K}'_m = \mathbf{c}_m \cdot \mathbf{W}'_{k,m}, \mathbf{V}'_m = \mathbf{c}_m \cdot \mathbf{W}'_{v,m} \end{aligned} \tag{5}$$

Here, $\mathbf{W}'_{k,m}$ and $\mathbf{W}'_{v,m}$ represent the newly added weights that are updated during training.

Prior to passing each of the above-mentioned conditioning modes to the cross-attention layers, we pass each of them through CLIP [48] and extract a 768 dimensional global CLIP feature vector, which serves as our conditioning representation. Following IP-Adapter, we train a small projection network consisting of one linear layer and layer normalization, designed to project the CLIP feature vector into several extra context tokens before injecting it into the cross-attention layers. We use 16 tokens for each of our conditionings to allow for meaningful attention computation.

At inference time, we can control the 3D face geometry generation with any of the modes using the respective set of cross-attention layers and classifier-free guidance [49]. The strength of the conditioning signal can be increased by increasing the hyperparameter w:

$$\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t, \mathbf{c}_0, \mathbf{c}_m, t) = w\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}_0, \mathbf{c}_m, t) + (1-w)\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t) \tag{6}$$

To condition only on a newly added mode or to generate geometry unconditionally, the FLAME conditioning $\mathbf{c}_0$ is set to its null embedding. Additionally, for unconditional generation, the new cross-attention in Eq. (4), which feeds $\mathbf{c}_m$ to the diffusion model, is simply not added to the original attention. Our full method pipeline is visualized in Fig. 2. For more implementation details, please refer to our supplementary material.

4. Results

We now show several results and applications of our new multimodal conditional 3D face geometry generation method. We begin by demonstrating control over the generated facial geometry using FLAME's identity and expression parameters (Section 4.1). We then demonstrate multimodal conditioned geometry generation by guiding the denoising process using sketches, sparse 2D landmarks, Canny edges, portrait photos and text. We evaluate the effectiveness of these different modalities in guiding the generated geometry in Section 4.2. We also compare our method to the state-of-the-art related work both on text and image prompts in Section 4.2. In addition to using different conditioning modes, we show how one can also spatially restrict the guidance to a particular face region to perform precise geometry edits in Section 4.3. We can also generate dynamic facial performances by guiding our model from video inputs and demonstrate that our method can produce facial animations that are stable across time (Section 4.4). Finally, we discuss the limitations of our method in Section 4.5.

4.1. Identity and expression conditioning

The base conditioning used to train our geometry generator are the identity and expression parameters from the FLAME model [3]. We use 300 identity parameters ($\boldsymbol{\beta}$), 100 expression parameters ($\boldsymbol{\psi}$) and 3 jaw pose ($\boldsymbol{\theta}$) parameters. We combine these FLAME parameters into a 403-dimensional conditioning vector which we pass through a 3-layer MLP with Leaky ReLU activation functions prior to injecting it into the diffusion model via cross-attention.

Recollect that we do not use the geometry from the FLAME model itself to train our diffusion model. Instead we fit the FLAME model to the high quality facial geometry from Chandran et al. [14] only to obtain identity and expression parameters, and train the diffusion model directly on the geometry captured by Chandran et al.

We visualize geometries generated by our model for unseen FLAME parameters in Fig. 3. As our underlying mesh topology is different from FLAME and represents ~10-times more vertices, it can express a greater level of detail that is not present in the lower resolution FLAME mesh. This high resolution detail is captured and reproduced by the denoising process. Therefore, by simply varying the noise seed, one can obtain variations of the FLAME-conditioned geometry, each of which contain different mid/high frequency details.

Disentangling Identity and Expression. Our diffusion model also preserves the disentanglement between facial identity and expression that is present in FLAME. In Fig. 4, we show how a smooth interpolation of FLAME's identity and expression coefficients results in a smooth, yet nonlinear, interpolation of our generated geometry. We can observe that the facial expression remains fixed when interpolating between identities, and vice versa (please also refer to the supplemental video). To eliminate the randomness in the generation, we used the same initial noise to generate the interpolated geometries along with DDIM sampling [47].

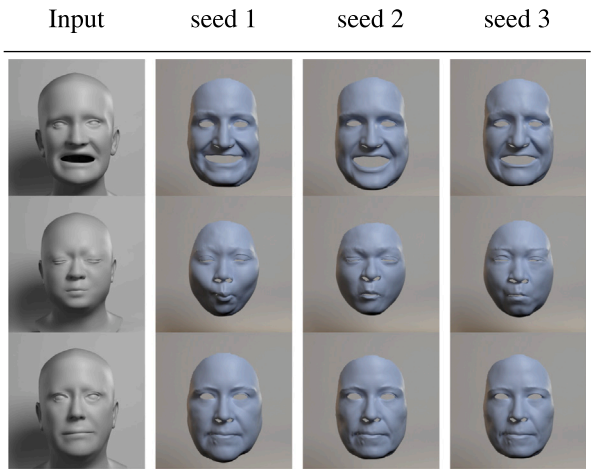

Fig. 3. Changing the noise seed while conditioning on the FLAME parameters does not affect the identity and expression of the generated geometry, but only the stochastic details that are added on top. Our model can capture richer geometric detail that is not present in the original FLAME mesh, while still respecting FLAME's identity and expression parameters. Each row shows a different set of FLAME parameters, with the corresponding FLAME mesh visualized in the first column.

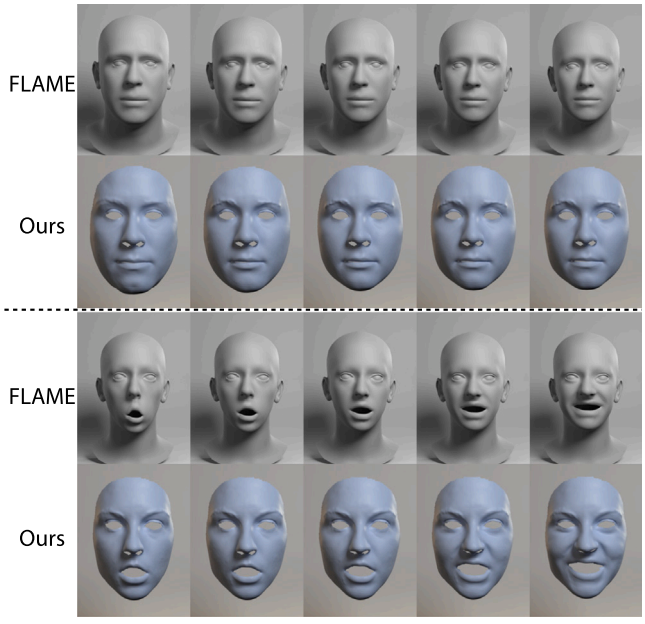

Fig. 4. Identity and expression disentanglement. Traversing the first dimension in FLAME's identity space leads to smooth changes in our generated face geometry (rows 1 and 2). Similarly a linear interpolation of FLAME's expression parameters results in a smooth, yet nonlinear, interpolation of facial expression as seen in rows 3 and 4. Please refer to our supplementary video to see the complete interpolation.

4.2. Multimodal conditioning

Beyond the underlying FLAME-based control, we introduce additional conditioning modes to control our diffusion model following Section 3.4. We now discuss the results of facial geometry generation by conditioning our diffusion model on sketches, sparse 2D landmarks, Canny edges, portrait RGB photos and text. Geometries generated for different conditionings from each of these additional modes are shown in Fig. 5.

Quantitative Comparison. While our generated geometry follows the identity and expression seen in the input conditioning signal, the degree to which the conditioning signal constrains the generated geometry varies from mode to mode. We evaluate the effectiveness of each of our conditioning modes in guiding the generated geometry towards ground truth scans, by computing the Euclidean error between the generated geometry and the ground truth geometry for each conditioning mode.

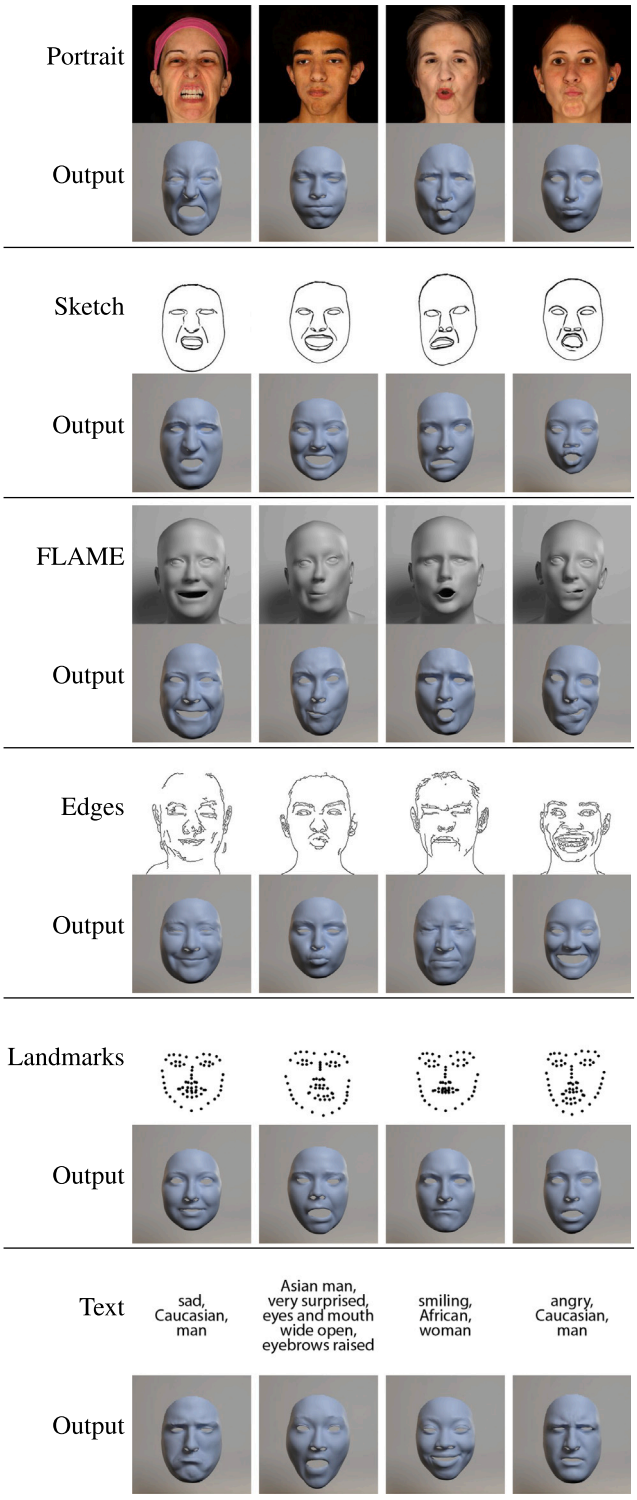

Fig. 5. Multimodal conditional generation results on our validation dataset, including conditioning on portrait images, sketches, FLAME parameters, Canny edges, 2D landmarks and text. The FLAME parameter inputs are visualized as meshes. Our model captures the facial identity and expression across all conditioning modes. We use classifier-free guidance of $w = 1$, except for some text-prompts where we use $w = 3$ for even stronger facial expressions.

Table 1
We report the vertex-to-vertex error (V2V) to the original scanned geometry in mm on our validation set of 432 shapes for generations from different types of conditioning signals. Conditions that are more descriptive of the end facial geometry like FLAME parameters, and portrait images achieve a lower error than others. Results are averaged over three different seeds.

V2V error	Mean ↓	Median ↓	Std ↓
2D landmarks	6.390	5.950	1.945
Canny edges	6.007	5.701	1.672
Sketch	5.521	5.106	1.792
Portrait photo	5.207	4.942	**1.471**
FLAME parameters	**5.008**	**4.737**	1.498

Table 2
We report the CLIP score (ViT-B/32) for text-to-geometry generation to related work. We prompt each method with 10 text prompts specifying a neutral expression. Additionally, we compare the average inference-time per sample between different methods on a single 3090 GPU. *Describe3D inference-time is measured on a single 1080 GPU and DreamFace-V2 results are exported from their web interface (N/A). We run *FaceG2E fast* for only 15/200 optimization steps to match the inference-time of our method for comparison. The best score per column is marked in bold and the second-best score is underlined.

Method	CLIP (ViT-B/32) ↑	Time ↓
Describe3D [26]	32.69	80.62*
DreamFace-V2 [4]	33.85	N/A
FaceG2E fast [5]	32.41	<u>6.07</u>
FaceG2E [5]	**35.03**	42.87
Ours	<u>34.15</u>	**5.48**

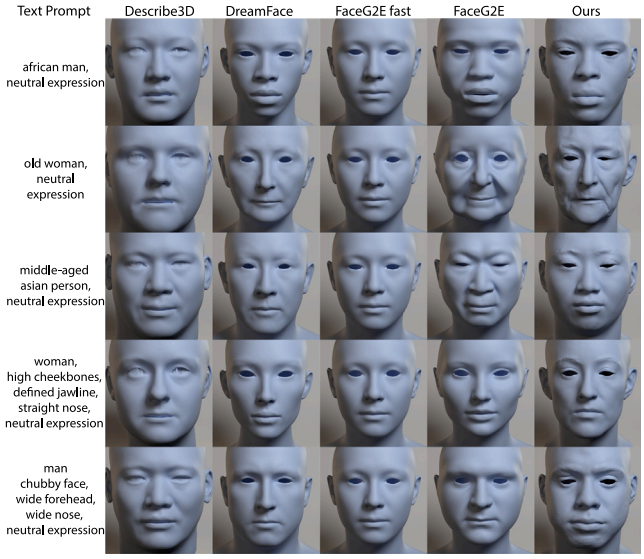

Fig. 6. Comparison to related work on text-to-geometry generation with neutral expression prompts. We sample from our model using classifier-free guidance $w = 3$.

In Table 1, we report the vertex-to-vertex (V2V) error of 432 shapes from our validation set for each type of conditioning. We observe that conditioning signals that are more descriptive, such as FLAME parameters or portrait photos, obtain a lower error when compared to signals that are less descriptive of the final geometry (2D landmarks, Canny edges and sketches). Please refer to our supplementary material for error maps on our validation set.

Next, we compare our method to the state-of-the-art related work methods. First, we look at the text conditioning mode, because prompts from this mode are most commonly supported by related work. We compare our method to Describe3D [26], DreamFace [4] and two variants of FaceG2E [5], over 10 different text prompts. We report the average CLIP score (ViT-B/32) between the CLIP embeddings from the text prompt and the CLIP embeddings extracted from the rendered generated face geometry. Specifically, we evaluate text prompts that

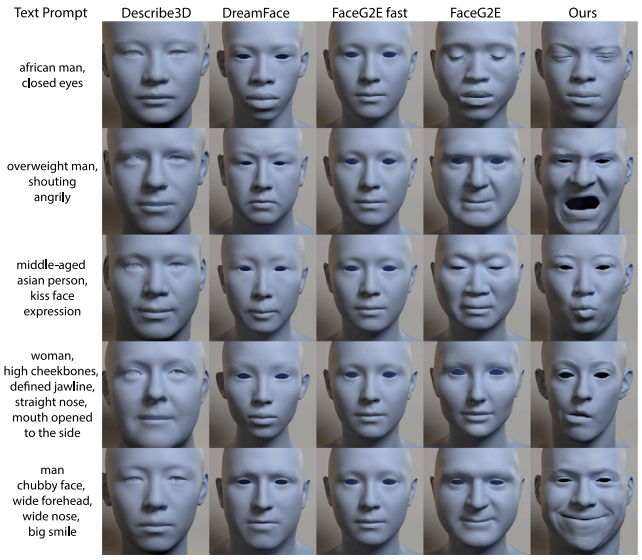

Fig. 7. Comparison to related work on text-to-geometry generation with expression prompts.

specify varied identities all with a neutral expression, because neutrals are supported by all methods. Each prompt is prepended with "*A shaded, textureless 3D face model of*", although for legibility, we shorten the text prompts when we add them to figures (e.g. Figs. 6 and 7). Please refer to our supplementary material for the exact text prompts. Among the compared methods, our method has the second highest CLIP score on text prompts that specify a neutral expression (Table 2).

Qualitative Comparison. We first show qualitative comparisons to state-of-the-art text-to-geometry methods on *neutral* text prompts in Fig. 6. Our method produces realistic faces that are subjectively on par with other techniques. Furthermore, it is important note that our method also natively supports generating faces with *expressions* by providing an expression description in the text prompt. This is a situation that other methods struggle with, as shown in Fig. 7.

Second, besides text prompts, DreamFace and our method also support RGB image prompts (Fig. 8). Our method achieves results comparable to DreamFace without relying on SDS optimization as part of the geometry generation. Note, that to aid the visual comparisons with previous methods, we complete the head in our results by deforming a template head to match our generated face. Despite training on purely studio data, we show how our model responds to conditionings derived from in-the-wild data in Fig. 9. For direct visual comparison with the identity and facial expression in the conditioning signal, we overlay the generated meshes onto validation images from whom the respective conditionings were extracted in Fig. 10.

Combining Two Conditionings. As our method involves learning additional modes of conditioning on top of an underlying FLAME conditioned diffusion model, we can also use more than one conditioning signal at inference time to guide the generation. In Fig. 11, we show how combining both FLAME parameter and portrait image conditioning lowers the vertex error on a validation sample, as the denoising UNet now has access to more information about the desired identity and expression. Note that only our base conditioning (FLAME parameters) is present when training any one of the other modalities cross-attention layers. Therefore, we can expect complementary results only when combining the FLAME parameter conditioning with another modality.

Inference time. We show that our method has significant inference-time benefits over its competitors that are mainly based on SDS optimization. It has the fastest average inference speed per sample on text-to-geometry generation, because once it is trained, it can directly

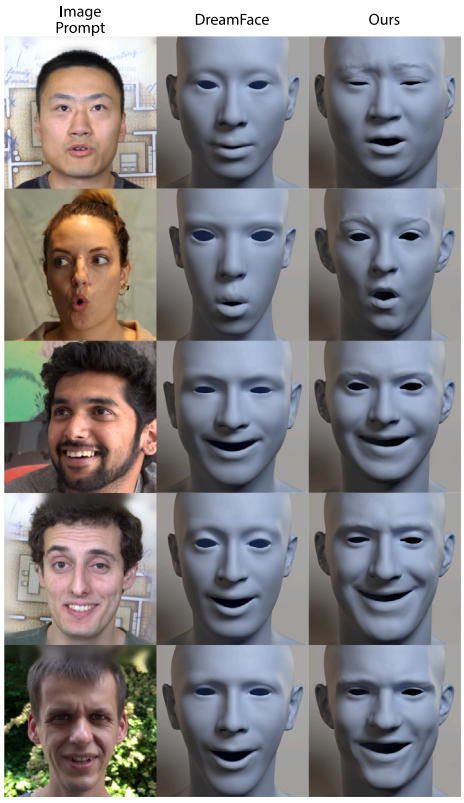

Fig. 8. Image prompt to face geometry generation. Comparison between DreamFace [4] and our method on in-the-wild test data.

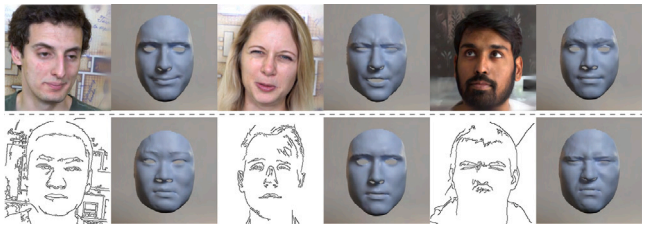

Fig. 9. Generation using conditioning signals obtained from in-the-wild test data (Portrait images top row, Canny edge maps bottom row). Our model produces reasonable facial geometry from in-the-wild conditions despite being trained only on studio data.

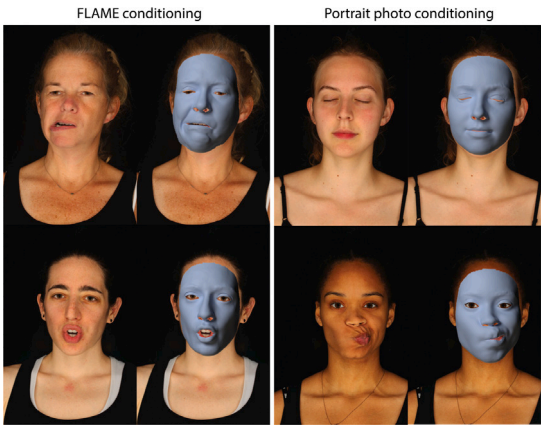

Fig. 10. We overlay our generated meshes on top of the images that display the identities (and expressions) from whom the respective conditioning signals were extracted. We show results for FLAME and portrait photo conditioning.

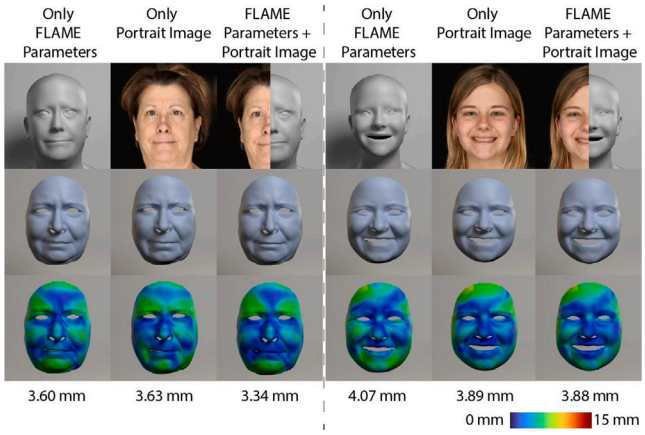

Fig. 11. Multimodal conditioning using portrait photo and FLAME parameter conditioning separately and simultaneously. The first row shows the conditioning inputs. The second row shows the generated face geometry. The third row shows the error map when compared with the original geometry from our validation set.

sample 3D faces from the diffusion model. This setup alleviates the need for compute intensive iterative SDS optimization of a 3D face representation. Specifically, our method is more than seven times faster than its closest competitor FaceG2E (5.48 s vs. 42.87 s). To compare results with similar inference speeds, we run *FaceG2E fast* for approximately the same time as our own method and evaluate the CLIP score results (see Table 2).

Controlling the Conditioning Strength. To control the strength w of the guiding condition, we make use of classifier-free guidance [49] following Eq. (6). Increasing the guidance strength increases the effect that the input conditioning (prompt) has on the resulting geometry. Stronger guidance can lead to increased level of detail in the generated face geometry and greater resemblance with the input prompt. For example, in Fig. 12, the expression of the generated geometry of the subject in the first row displays stronger wrinkles, and a closer match to the portrait image when setting $w = 3$ compared to setting it to $w = 1$. Unless specified differently, we use $w = 1$ for all our conditional generation results.

4.3. Geometry editing

The latent space of our autoencoder preserves the spatial layout of the original UV position map, much like how the latent space of the image autoencoder in text-to-image models [9] preserves the spatial layout of the encoded image. As a consequence, by masking regions in the latent UV position map corresponding to regions we wish to modify, and by denoising the masked regions, one can apply intuitive edits to particular regions of the facial geometry. Please refer to RePaint [50] for more details on the masking process. Even when using masks with sharp boundaries, the denoising process can take care of smoothly interpolating at the mask boundaries. We show results of guiding the editing of facial geometry with user conditions in Fig. 13. Specifically, we show an interactive sketching workflow (~6 s/sample), where an artist can progressively edit a generated geometry by modifying one region at a time.

4.4. Dynamic generation

Although our model is only trained with static face shapes, we find that it can generate temporally stable 3D facial geometries when conditioned on per-frame FLAME parameters derived from animation sequences or on CLIP embeddings obtained from individual frames from in-the-wild videos. In Fig. 14, we show the generated 3D face geometry

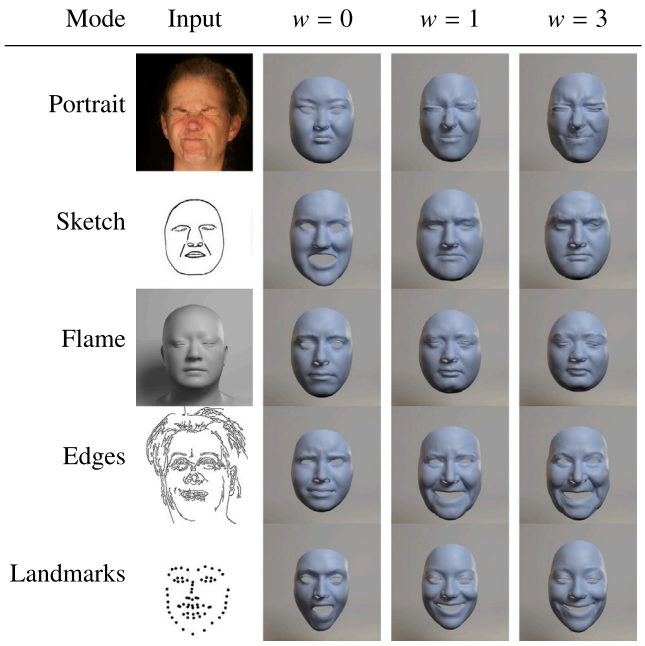

Fig. 12. By varying the guidance strength w, we can control the extent to which our conditioning signals affect the generated geometry. Setting $w = 0$ results in unconditional generation, while $w \geq 1$ results in conditional generation.

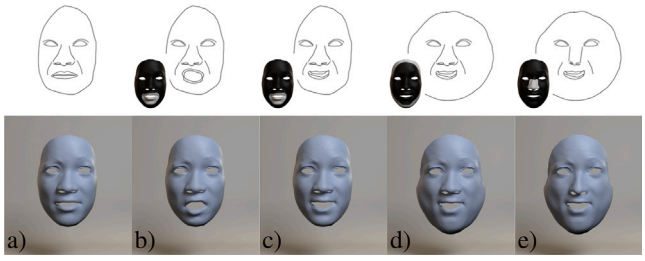

Fig. 13. Conditional generation from an input sketch (a), followed by local edits of the mouth (b, c), the face shape (d) and the nose (e). The masks used to constrain the region of modification are shown in the insets.

produced by our method when conditioned on various signals derived from videos. To demonstrate the use of sketches as dynamic conditioning, we use a recent face reconstruction technique [51] to track the facial geometry in 3D from an in-the-wild video and then render out 2D sketches using a hand-painted texture map. We identify that the only pre-processing required to obtain dynamically stable generations from CLIP embeddings is to temporally smooth them before using them as the conditioning signal. To further ensure stable generations across time, we use the same noise seed and DDIM sampling.

4.5. Limitations and future work

As limitations, we identify that our model can produce geometric artifacts for extreme expressions, especially when controlled using FLAME's jaw pose parameters. This problem is mainly a data limitation and could be resolved by sourcing a larger dataset of extreme expressions, by oversampling expressions during training or by weighting the loss towards focusing more on extreme expressions. Additionally, we identify a limitation of extremely similar CLIP-based conditionings for left/right mirrors of asymmetrical face expressions, leading to a direction ambiguity in the geometry output. We refer to our supplementary material for further discussion and visualization of these failure cases. Beyond addressing those limitations, future work could incorporate

Fig. 14. Dynamic geometry generation results given sketch, landmark, FLAME parameters or portrait photos from 4 different input videos as conditionings. Our results change smoothly across time while maintaining a consistent identity rather well.

facial appearance information into our method, enabling multimodal control over 3D faces with corresponding texture.

5. Conclusion

We propose a new framework for 3D facial geometry generation based on a latent diffusion model that can be guided using multiple types of conditionings (prompts). Our conditional geometry generator operates in a latent geometry space. It can produce high quality geometry at comparably fast inference speeds using a UV position map representation. It can be seamlessly conditioned on hand-drawn sketches, 2D landmarks, Canny edges, FLAME-parameters, RGB portrait photos and text; resulting in a comprehensive facial geometry generator that supports many applications. For example, stochastic detail variation in the generated geometry or local geometry edits. We train our model from scratch on only static face shapes captured in a studio setting and yet demonstrate that our model can generalize reasonably to in-the-wild conditioning signals, and can also generate facial performances when conditioned on frames from video data.

CRediT authorship contribution statement

Christopher Otto: Conceptualization, Data curation, Formal analysis, Investigation, Software, Validation, Visualization, Writing – original draft, Writing – review & editing, Project administration. **Prashanth Chandran:** Conceptualization, Data curation, Formal analysis, Investigation, Software, Validation, Visualization, Writing – original draft, Writing – review & editing, Project administration. **Sebastian Weiss:** Conceptualization, Data curation, Formal analysis, Investigation, Software, Validation, Visualization, Writing – original draft, Writing – review & editing, Project administration. **Markus Gross:** Conceptualization, Project administration, Supervision. **Gaspard Zoss:** Conceptualization, Data curation, Formal analysis, Investigation, Software, Validation, Visualization, Writing – original draft, Writing – review & editing, Project administration. **Derek Bradley:** Conceptualization, Data curation, Formal analysis, Investigation, Software, Validation, Visualization, Writing – original draft, Writing – review & editing, Project administration, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cag.2025.104325.

Data availability

The authors do not have permission to share data.

References

[1] Gruber A, Fratarcangeli M, Zoss G, Cattaneo R, Beeler T, Gross M, Bradley D. Interactive sculpting of digital faces using an anatomical modeling paradigm. Comput Graph Forum 2020;93–102. http://dx.doi.org/10.1111/cgf.14071.

[2] Kim H-J, Öztireli AC, Shin I-K, Gross M, Choi S-M. Interactive generation of realistic facial wrinkles from sketchy drawings. Comput Graph Forum 2015;34(2):179–91. http://dx.doi.org/10.1111/cgf.12551.

[3] Li T, Bolkart T, Black MJ, Li H, Romero J. Learning a model of facial shape and expression from 4D scans. ACM Trans Graph (ToG) (Proc SIGGRAPH Asia) 2017;36(6):194:1–194:17, URL https://doi.org/10.1145/3130800.3130813.

[4] Zhang L, Qiu Q, Lin H, Zhang Q, Shi C, Yang W, Shi Y, Yang S, Xu L, Yu J. DreamFace: Progressive generation of animatable 3D faces under text guidance. ACM Trans Graph (ToG) 2023;42(4). http://dx.doi.org/10.1145/3592094.

[5] Wu Y, Meng Y, Hu Z, Li L, Wu H, Zhou K, Xu W, Yu X. Text-guided 3D face synthesis – from generation to editing. 2023, arXiv, arXiv:2312.00375.

[6] Liu H, Wang X, Wan Z, Shen Y, Song Y, Liao J, Chen Q. 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; 2024, http://dx.doi.org/10.1145/3641519.3657512.

[7] Han X, Cao Y, Han K, Zhu X, Deng J, Song Y-Z, Xiang T, Wong K-YK. HeadSculpt: Crafting 3D head avatars with text. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, editors. Advances in neural information processing systems. Vol. 36, Curran Associates, Inc.; 2023, p. 4915–36, URL https://proceedings.neurips.cc/paper_files/paper/2023/file/0fb98d483fa580e0354bcdd3a003a3f3-Paper-Conference.pdf.

[8] Wang D, Meng H, Cai Z, Shao Z, Liu Q, Wang L, Fan M, Shan Y, Zhan X, Wang Z. HeadEvolver: Text to head avatars via locally learnable mesh deformation. 2024, arXiv preprint arXiv:2403.09326.

[9] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. CVPR, 2022, p. 10684–95.

[10] Koley S, Bhunia AK, Sekhri D, Sain A, Chowdhury PN, Xiang T, Song Y-Z. It's all about your sketch: Democratising sketch control in diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. CVPR, 2024, p. 7204–14.

[11] Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. ICCV, 2023, p. 3836–47.

[12] Mou C, Wang X, Xie L, Wu Y, Zhang J, Qi Z, Shan Y, Qie X. T2I-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. 2023, arXiv, arXiv:2302.08453.

[13] Ye H, Zhang J, Liu S, Han X, Yang W. IP-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023, arXiv, arXiv:2308.06721.

[14] Chandran P, Bradley D, Gross M, Beeler T. Semantic deep face models. In: 2020 international conference on 3D vision. 3DV, Los Alamitos, CA, USA: IEEE Computer Society; 2020, p. 345–54. http://dx.doi.org/10.1109/3DV50981.2020.00044, URL https://doi.ieeecomputersociety.org/10.1109/3DV50981.2020.00044.

[15] Potamias RA, Ploumpis MTS, Zafeiriou S. ShapeFusion: A 3D diffusion model for localized shape editing. 2024, arXiv, arXiv:2403.19773.

[16] Zou K, Faisan S, Yu B, Valette S, Seo H. 4D facial expression diffusion model. ACM Trans Multimed Comput Commun Appl 2024. http://dx.doi.org/10.1145/3653455.

[17] Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang P, Vaughan JW, editors. Advances in neural information processing systems (neurIPS). Vol. 34, Curran Associates, Inc.; 2021, p. 8780–94.

[18] Wang T, Zhang B, Zhang T, Gu S, Bao J, Baltrusaitis T, Shen J, Chen D, Wen F, Chen Q, Guo B. RODIN: A generative model for sculpting 3D digital avatars using diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. CVPR, 2023, p. 4563–73.

[19] Lorensen WE, Cline HE. Marching cubes: A high resolution 3D surface construction algorithm.. In: SIGGRAPH. ACM; 1987, p. 163–9.

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

[21] Shen T, Gao J, Yin K, Liu M-Y, Fidler S. Deep marching tetrahedra: a hybrid representation for high-resolution 3D shape synthesis. In: Advances in neural information processing systems. NeurIPS, 2021.

[22] 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.

[23] Bergman AW, Yifan W, Wetzstein G. Articulated 3D head avatar generation using text-to-image diffusion models. 2023, arXiv, arXiv:2307.04859.

[24] Blanz V, Vetter T. A morphable model for the synthesis of 3D faces. In: Proceedings of the 26th annual conference on computer graphics and interactive techniques. SIGGRAPH '99, USA: ACM Press/Addison-Wesley Publishing Co.; 1999, p. 187–94. http://dx.doi.org/10.1145/311535.311556.

[25] Schuhmann C, Beaumont R, Vencu R, Gordon CW, Wightman R, Cherti M, Coombes T, Katta A, Mullis C, Wortsman M, Schramowski P, Kundurthy SR, Crowson K, Schmidt L, Kaczmarczyk R, Jitsev J. LAION-5B: An open large-scale dataset for training next generation image-text models. In: Thirty-sixth conference on neural information processing systems datasets and benchmarks track. 2022.

[26] Wu M, Zhu H, Huang L, Zhuang Y, Lu Y, Cao X. High-fidelity 3D face generation from natural language descriptions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. CVPR, 2023, p. 4521–30.

[27] Voynov A, Aberman K, Cohen-Or D. Sketch-guided text-to-image diffusion models. In: ACM SIGGRAPH 2023 conference proceedings. SIGGRAPH '23, New York, NY, USA: Association for Computing Machinery; 2023, http://dx.doi.org/10.1145/3588432.3591560.

[28] Canny J. A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell 1986;PAMI-8(6):679–98.

[29] Kirschstein T, Giebenhain S, Nießner M. DiffusionAvatars: Deferred diffusion for high-fidelity 3D head avatars. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. CVPR, 2024, p. 5481–92.

[30] Ding Z, Zhang X, Xia Z, Jebe L, Tu Z, Zhang X. DiffusionRig: Learning personalized priors for facial appearance editing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. CVPR, 2023, p. 12736–46.

[31] Gu J, Gao Q, Zhai S, Chen B, Liu L, Susskind J. Control3Diff: Learning controllable 3D diffusion models from single-view images. Int Conf 3D Vis (3DV) 2024.

[32] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Proceedings of the 31st international conference on neural information processing systems. NIPS '17, Red Hook, NY, USA: Curran Associates Inc.; 2017, p. 6000–10.

[33] Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, Milan K, Quan J, Ramalho T, Grabska-Barwinska A, Hassabis D, Clopath C, Kumaran D, Hadsell R. Overcoming catastrophic forgetting in neural networks. Proc Natl Acad Sci 2017;114(13):3521–6. http://dx.doi.org/10.1073/pnas.1611835114.

[34] Feng Y, Wu F, Shao X, Wang Y, Zhou X. Joint 3D face reconstruction and dense alignment with position map regression network. In: Proceedings of the European conference on computer vision. ECCV, 2018.

[35] Otto C, Naruniec J, Helminger L, Etterlin T, Mignone G, Chandran P, Zoss G, Schroers C, Gross M, Gotardo P, Bradley D, Weber R. Learning dynamic 3D geometry and texture for video face swapping. Comput Graph Forum 2022;41(7):611–22. http://dx.doi.org/10.1111/cgf.14705.

[36] Gu X, Gortler SJ, Hoppe H. Geometry images. ACM Trans Graph 2002;21(3):355–61.

[37] Funkhouser T, Kazhdan M, Shilane P, Min P, Kiefer W, Tal A, Rusinkiewicz S, Dobkin D. Modeling by example. In: ACM SIGGRAPH 2004 papers. SIGGRAPH '04, New York, NY, USA: Association for Computing Machinery; 2004, p. 652–63. http://dx.doi.org/10.1145/1186562.1015775.

[38] Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANs for improved quality, stability, and variation. In: International conference on learning representations. ICLR, 2018.

[39] Preechakul K, Chatthee N, Wizadwongsa S, Suwajanakorn S. Diffusion autoencoders: Toward a meaningful and decodable representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. CVPR, 2022, p. 10619–29.

[40] Kadkhodaie Z, Guth F, Simoncelli EP, Mallat S. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In: The twelfth international conference on learning representations. ICLR, 2024.

[41] Kingma DP, Welling M. Auto-Encoding Variational Bayes. In: 2nd international conference on learning representations. ICLR, 2014.

[42] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. CVPR, 2021, p. 12873–83.

[43] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. CVPR, 2018.

[44] Isola P, Zhu J-Y, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: 2017 IEEE conference on computer vision and pattern recognition. CVPR, 2017, p. 5967–76. http://dx.doi.org/10.1109/CVPR.2017.632.

[45] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H, editors. Advances in neural information processing systems (neurIPS). Vol. 33, Curran Associates, Inc.; 2020, p. 6840–51.
[46] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical image computing and computer-assisted intervention – MICCAI 2015. Cham: Springer International Publishing; 2015, p. 234–41.
[47] Song J, Meng C, Ermon S. Denoising diffusion implicit models. In: International conference on learning representations. ICLR, 2021.
[48] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In: Meila M, Zhang T, editors. Proceedings of the 38th international conference on machine learning. PMLR, Proceedings of machine learning research, vol. 139, PMLR; 2021, p. 8748–63.
[49] Ho J, Salimans T. Classifier-free diffusion guidance. In: NeurIPS 2021 workshop on deep generative models and downstream applications. 2021.
[50] Lugmayr A, Danelljan M, Romero A, Yu F, Timofte R, Van Gool L. RePaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. CVPR, 2022, p. 11461–71.
[51] Chandran P, Zoss G, Gotardo P, Bradley D. Continuous landmark detection with 3D queries. In: 2023 IEEE/CVF conference on computer vision and pattern recognition. CVPR, Los Alamitos, CA, USA: IEEE Computer Society; 2023, p. 16858–67.