Spherical Geometry Diffusion: Generating High-quality 3D Face Geometry via Sphere-anchored Representations

Junyi Zhang[1], Yiming Wang[1], Yunhong Lu[1], Qichao Wang[1], Wenzhe Qian[1], Xiaoyin Xu[1], David Gu[2], Min Zhang[1,3,4*]

[1]Zhejiang University
[2]Stony Brook University
[3]Shanghai Institute for Advanced Study-Zhejiang University
[4]Shanghai Institute for Mathematics and Interdisciplinary Sciences

Abstract

A fundamental challenge in text-to-3D face generation is achieving high-quality geometry. The core difficulty lies in the arbitrary and intricate distribution of vertices in 3D space, making it challenging for existing models to establish clean connectivity and resulting in suboptimal geometry. To address this, our core insight is to simplify the underlying geometric structure by constraining the distribution onto a simple and regular manifold, a topological sphere. Building on this, we first propose the Spherical Geometry Representation, a novel face representation that anchors geometric signals to uniform spherical coordinates. This guarantees a regular point distribution, from which the mesh connectivity can be robustly reconstructed. Critically, this canonical sphere can be seamlessly unwrapped into a 2D map, creating a perfect synergy with powerful 2D generative models. We then introduce Spherical Geometry Diffusion, a conditional diffusion framework built upon this 2D map. It enables diverse and controllable generation by jointly modeling geometry and texture, where the geometry explicitly conditions the texture synthesis process. Our method's effectiveness is demonstrated through its success in a wide range of tasks: text-to-3D generation, face reconstruction, and text-based 3D editing. Extensive experiments show that our approach substantially outperforms existing methods in geometric quality, textual fidelity, and inference efficiency.

Introduction

High-quality 3D face generation is essential for many applications in the computer graphics and movie industry, including virtual reality, computer games, and movie production. Although recent advances have enabled impressive text-driven control, a fundamental challenge persists: achieving high-quality and topologically sound mesh geometry. Many existing methods still produce meshes with significant artifacts and require excessive inference time.

The ultimate goal of 3D face synthesis is to generate a high-quality triangle mesh, the industry-standard representation renowned for capturing intricate detail. However, the explicit and unstructured nature of the mesh topology makes it exceptionally challenging for generative models to learn directly. This has led to two main approaches, each with significant limitations. The first approach involves simplifying the problem using the 3D Morphable Model (3DMM) (Wu et al. 2023; Aneja et al. 2023; Kirschstein et al. 2024; Zhang et al. 2023; Chen and Kim 2021). By operating in a low-dimensional parametric space, 3DMMs offer easy control but at a steep price: the generated geometry is fundamentally limited by the expressive capacity of the 3DMM prior, precluding high geometric accuracy. To break free from these constraints, a second and more recent approach uses flexible implicit representations to represent facial geometry (Liu et al. 2024; Zheng et al. 2022, 2024; Yenamandra et al. 2021; Wang et al. 2023). These methods offer a more accurate geometric approximation by learning the signed distance from dense points to the surface. However, to produce a usable mesh, these methods must employ a post-processing algorithm such as Marching Cubes (Lorensen and Cline 1998). This final step reintroduces the problem of defining *clean vertex connectivity*, often resulting in noisy artifact meshes.

The *connectivity* problem lies in the arbitrary and unstructured distribution of vertices in the 3D space, where unpredictable spatial relationships make clean topology reconstruction fundamentally challenging. **Our key insight**

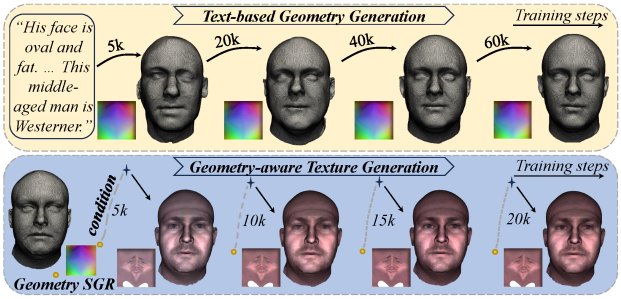

Figure 1: We present **Spherical Geometry Diffusion** as a novel framework for 3D faces generation using **Spherical Geometry Representation**. For *Text-based Geometry Generation*, we achieve flexible control through text conditions. For *Geometry-aware Texture Generation*, we create texture conditioning on the geometry through SGR's alignment.

*Corresponding author: Min Zhang (min_zhang@zju.edu.cn)

is to profoundly simplify this problem by constraining the geometry to a regular manifold. Instead of learning points in an arbitrary 3D space, we map the facial geometry to a canonical sphere. By anchoring geometric signals to uniformly distributed spherical coordinates, connectivity between points becomes trivial and robust to establish. This strategy has a two-fold advantage. First, it efficiently solves the connectivity problem. The spherical anchoring inherently defines the mesh topology, eliminating complex extraction and ensuring a high-quality mesh. Second, and just as importantly, this canonical sphere can be seamlessly unwrapped into a 2D map. This creates a structured grid-like domain for the geometry, unlocking the potential to leverage powerful 2D generative models.

Building on this insight, we propose **Spherical Geometry Representation (SGR)**. This representation parameterizes the 3D face onto a sphere and then unwraps it into a unified 2D map, where each pixel corresponds to a point signal on the sphere. Crucially, this unified map is versatile, capable of encoding not only the 3D vertex positions (geometry), but also the corresponding texture and fine-grained displacement in a unified format. The nature of this map provides an implicit definition of vertex connectivity, allowing a high-quality mesh to be reconstructed by efficient Delaunay spherical triangulation (Fortune 2017).

With geometry and texture elegantly aligned in corresponding 2D structures, we introduce **Spherical Geometry Diffusion**, a framework designed to efficiently generate high-quality facial geometry, as demonstrated in Fig. 1 and Fig. 2. Our core innovation is a latent diffusion model that synthesizes a complete geometry in a single forward pass. This native approach stands in stark contrast to conventional text-to-3D methods that require iterative optimization and grants our framework a significant advantage in efficiency. The unified nature of SGR enables seamless extension to diverse applications beyond pure geometry generation. The framework naturally supports geometry-aware texture synthesis by directly conditioning on the geometry through SGR's shared spherical coordinates, bypassing the complex depth rendering required by traditional pipelines. Furthermore, the architecture enables new face reconstruction and intuitive text-based editing, making it a comprehensive solution for 3D face manipulation.

Beyond the core framework, our work introduces two technical innovations that ensure geometric integrity and training stability. First, to preserve the topology, we develop a novel **Center-symmetric Padding** scheme. This technique respects the inherent spatial continuity of SGR, effectively eliminating the geometric cracks and boundary artifacts caused by standard 2D convolutions. Second, we propose **Geometric Regularization** to stabilize the training process. Our regularization leverages SGR's efficient reconstruction properties to provide direct geometric feedback during training, ensuring that the learned latent representations correspond to valid and high-quality facial geometries. This geometric grounding significantly improves both the training efficiency and the final output quality.

We conducted comprehensive experiments on large-scale datasets, evaluating our performance across text-to-geometry generation, face reconstruction, and geometry-aware texture synthesis. The results demonstrate that our method achieves superior performance across all evaluations, producing high-quality geometry that closely matches text prompts and delivering high visual quality.

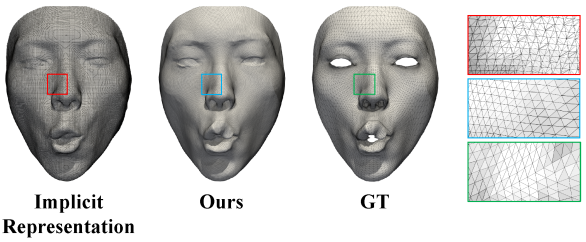

Figure 2: Qualitative comparison of reconstructed mesh quality, where the mesh is reconstructed from the implicit function obtained by ImFace++. A magnified view of the marked region is shown on the right.

In summary, our main contributions are as follows.

- We propose *Spherical Geometry Representation*, which enables faithful bidirectional mapping between 3D and 2D spaces, allowing us to leverage mature diffusion models for the 3D face while ensuring high-quality geometry.
- We introduce *Spherical Geometry Diffusion*, an efficient framework for controllable 3D face generation, taking advantage of SGR to simultaneously achieve high-quality geometry and texture alignment.
- We develop two technical innovations, *Center-symmetric Geometric Padding* and *Geometric Regularization*, which ensure geometric quality and stable training.

Related Work

3DMM-based Face Generation. 3DMM-based Face Generation focuses on learning the mapping between conditions and 3DMM parameters (Wu et al. 2025; Zhuang et al. 2025). The Describe3D method, proposed by Wu (Wu et al. 2023), learns a mapping function that translates text information in natural language into 3DMM parameters. Many approaches use trained 3DMM parameters as the initial state for a face model and refine geometric details through Score Distillation Sampling (SDS) (Poole et al. 2022). Methods such as FaceG2E (Wu et al. 2024), DreamFace (Zhang et al. 2023), HumanNorm (Huang et al. 2024), and AvatarCraft (Jiang et al. 2023) treat the parametric model as the starting geometric configuration and employ SDS techniques to improve the geometry. This improvement is achieved by incorporating 2D depth information and performing a coarse-to-fine optimization process to refine geometric details. **Spherical Geometry Diffusion**, on the other hand, learns directly from the original 3D data, allowing it to achieve high-quality geometry without the limitation imposed by 3DMM.

Implicit Function for Face Reconstruction. Implicit representations offer an alternative to traditional irregular mesh representations. These representations model 3D shapes by learning continuous deep implicit functions, which can capture shapes at any resolution by querying the occupancy

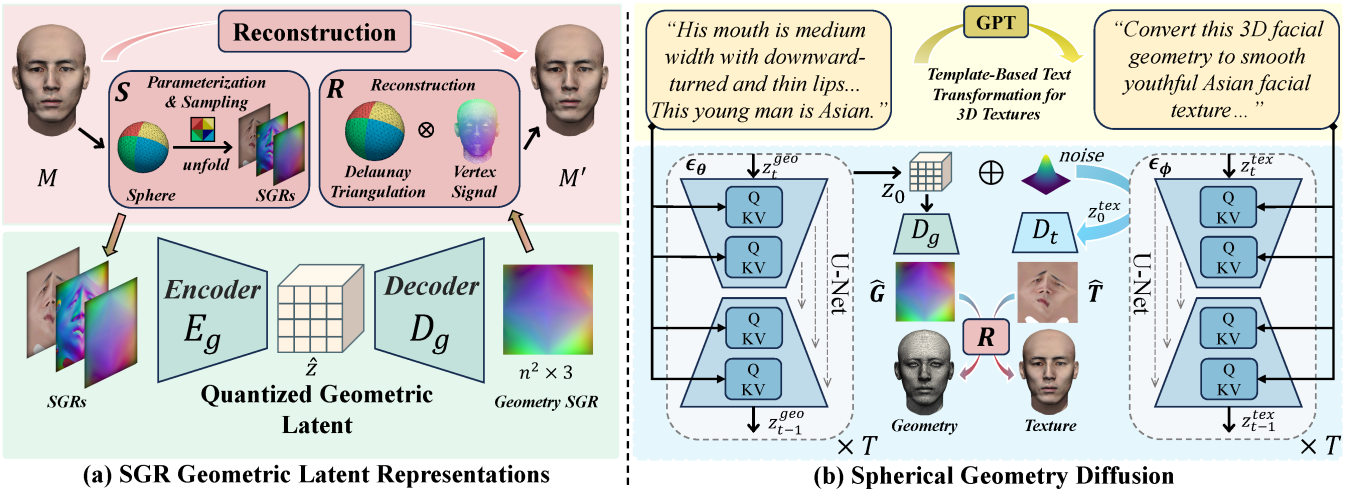

Figure 3: **Overview**. Our pipeline comprises two stages. **(a)** In the first stage, we construct and compress the ***Spherical Geometry Representation*** into ***Geometric Latent Representations***. *Geometric Regularization* and *Center-symmetric Padding* are introduced to enhance geometric quality and accelerate convergence. **(b)** Using these compact latent space, we train a conditional diffusion model through two sequential phases: (1) *Text-based Geometry Generation*, which generates 3D faces from text prompts, and (2) *Geometry-aware Texture Generation*, which creates textures conditioned on both text and the geometry SGR. The final latent codes are then decoded to reconstruct 3D meshes of high-quality.

value of points. This approach holds significant promise for high-precision modeling. In recent years, implicit neural representations have been incorporated into 3D face modeling. For example, i3DMM (Yenamandra et al. 2021) was the first implicit representation model specifically designed for human faces, although its reconstruction accuracy remains suboptimal. Subsequent work, including ImFace (Zheng et al. 2022), ImFace++ (Zheng et al. 2024), and other recent approaches (Giebenhain et al. 2023; Hong et al. 2022), has further explored implicit representations for 3D face modeling.Despite these advancements, implicit neural representations still depend on post-processing to generate explicit meshes, which results in suboptimal geometric quality. **Spherical Geometry Diffusion** addresses these challenges by producing high-quality explicit meshes through efficient transformations, ensuring superior geometry.

Method

Given a text description, our task is to generate high-fidelity 3D facial meshes. Our innovations are a novel representation for 3D face and modified Latent Diffusion Models (LDMs) tailored for efficient generation. The overview of the proposed method is illustrated in Fig. 3. The following sections detail our approach. We begin by introducing *SGR* for representing 3D facial geometry and texture. We then describe our Vector Quantized Variational Autoencoder (VQVAE) tailored for geometry SGR. Finally, we present a conditional diffusion strategy for generating high-quality 3D faces.

Preliminaries

Latent Diffusion Models. Diffusion models are powerful generative models that learn to synthesize data by reversing a fixed forward process that incrementally adds Gaussian noise ϵ to an input x_0. A noisy sample x_t at any time step t can be directly formulated as $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, where $\bar{\alpha}_t$ follows a predefined noise schedule. The reverse process is driven by a network $\epsilon_\theta(x_t, t)$, which is trained to predict the noise ϵ from x_t.

LDMs (Rombach et al. 2022) perform this entire diffusion process in a compressed latent space to mitigate the prohibitive computational cost. An autoencoder first maps the data x to a latent code $z_0 = \mathcal{E}(x)$. For conditional synthesis, text embedding C is injected into the denoising U-Net ϵ_θ through cross attention. This leads to the objective:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\mathcal{E}(x),C,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, C)\|_2^2\right]. \quad (1)$$

Spherical Geometry Representation

To constrain the complex distribution of sampling points in the 3D space, we leverage spherical parameterization and uniform sampling to obtain uniformly distributed sampling points in the sphere. The sampling point values are derived via barycentric interpolation of different signals.

Parameterization and Sampling. Given a 3D face mesh $\mathcal{M}$, we first map $\mathcal{M}$ to a spherical mesh $\mathcal{S}$ using spherical parameterization $\psi : \mathcal{M} \rightarrow \mathcal{S}$ (Praun and Hoppe 2003). Formally, let $\mathcal{T} = \{\Delta_k\}_{k=1}^{M}$ be the set of triangular faces in $\mathcal{M}$, where each Δ_k is defined by its vertices $\mathbf{v}_{k1}, \mathbf{v}_{k2}, \mathbf{v}_{k3}$. The spherical parameterization maps each triangle Δ_k to a spherical triangle Δ'_k on $\mathcal{S}$ with vertices $\mathbf{s}_{k1}, \mathbf{s}_{k2}, \mathbf{s}_{k3}$:

$$\Delta'_k = \psi(\Delta_k) = \{\psi(\mathbf{v}_{k1}), \psi(\mathbf{v}_{k2}), \psi(\mathbf{v}_{k3})\}, \quad (2)$$

$$\mathbf{s}_{ki} = \psi(\mathbf{v}_{ki}), \quad i = 1, 2, 3. \quad (3)$$

SGR is structured as a 2D image $\mathbf{G} = \{\mathbf{G}_{ij}\}_{i=1,j=1}^{H,W}$, where W, H is the resolution R of the SGR. We perform uniform

sampling in the rectangular domain $\mathcal{U}$, producing a discrete position $\mathbf{u}_{ij} = \left(\frac{i}{W}, \frac{j}{H}\right)$ for $i \in [1, W], j \in [1, H]$. Each sample point $\mathbf{u}_{ij}$ is then assigned to $\mathcal{S}$ via an area-preserving bijective mapping $\phi : \mathcal{U} \to \mathcal{S}$ (Clarberg 2008), resulting in a corresponding position $\mathbf{s}^{ij} \in \mathbb{R}^3$ on the sphere:

$$\mathbf{s}^{ij} = \phi(\mathbf{u}_{ij}), \quad \forall i \in [1, W],\ j \in [1, H]. \tag{4}$$

$\mathbf{G}_{ij}$ at each position $\mathbf{u}_{ij}$ represents a weighted surface geometry calculated by barycentric interpolation. Specifically, let $\mathbf{s}^{ij}$ lie within a spherical triangle Δ'_k on $\mathcal{S}$ with vertices $\mathbf{s}_{k1}, \mathbf{s}_{k2}, \mathbf{s}_{k3}$, where the signals are $\mathbf{p}_{k1}, \mathbf{p}_{k2}, \mathbf{p}_{k3}$, respectively. The barycentric coordinates $(\lambda_0, \lambda_1, \lambda_2)$ of $\mathbf{s}^{ij}$ with respect to these vertices are:

$$\lambda_c = \frac{\text{Area}(\mathbf{s}^{ij}, \mathbf{s}_{k(c+1)}, \mathbf{s}_{k(c+2)})}{\text{Area}(\mathbf{s}_{k1}, \mathbf{s}_{k2}, \mathbf{s}_{k3})}, \quad c = 0, 1, 2. \tag{5}$$

where the indices are taken modulo 3 and Area$(\cdot)$ denotes the spherical area formed by the points. With these barycentric coordinates, the interpolated value $\mathbf{G}_{ij}$ is given by:

$$\mathbf{G}_{ij} = \lambda_0 \mathbf{p}_{k1} + \lambda_1 \mathbf{p}_{k2} + \lambda_2 \mathbf{p}_{k3}. \tag{6}$$

It should be noted that signals $\mathbf{p}_{k1}, \mathbf{p}_{k2}, \mathbf{p}_{k3}$ are optional, allowing SGR to represent not only geometry but also diverse attributes such as texture, colors, and displacement maps through adaptable sampling. Fig. 6 demonstrates SGR's 2D representation of geometry, textures, and displacement map with the corresponding 3D meshes.

Furthermore, SGR supports arbitrary resolution. We can improve efficiency or preserve topology by reducing resolution or mapping only the original vertices. Reconstruction results and error at various resolutions are presented in Fig. 6. Limiting the mapping to the original vertices preserves the input topology (as shown in Fig. 6). In this work, we employ a consistent resolution to align meshes with different vertex counts and ensure a uniform point distribution, which is crucial for high-quality reconstruction. For details of spherical mapping, please refer to Appendix A.

Mesh Reconstruction. When reconstructing SGR into an explicit mesh, we first use $\phi : \mathcal{U} \to \mathcal{S}$ from Eq. 4 to map $\mathbf{G} = \{\mathbf{G}_{ij}\}_{i=1,j=1}^{H,W}$ to $\{\mathbf{S}^{ij}\}_{i=1,j=1}^{H,W}$, and then compute the triangulation on the sphere. The Delaunay triangulation of spherical points is equivalent to computing the convex hull function $f_{hull}(\cdot)$ (Fortune 2017). Eventually, the triangular face reconstruction $\mathcal{T}''$ of $\mathbf{G}$ is efficiently given by:

$$\mathcal{T}'' = f_{hull}(\{\mathbf{S}^{ij}\}_{i=1,j=1}^{H,W}). \tag{7}$$

Geometric Latent Representations

Being a 2D image structure, our SGR naturally aligns with conditional latent diffusion frameworks. To exploit this property, we first train an encoder E_g and a decoder D_g to compress the geometry SGRs into **Geometric Latent Representations** for effective generation. Although a standard VQVAE (Esser et al. 2021) architecture can be applied, we significantly enhance its performance by tailoring the model to the unique properties of our SGR. We introduce two targeted modifications: a novel padding scheme that preserves the spatial continuity of the sphere, and a geometric regularization term that leverages SGR's efficient reconstruction capabilities to accelerate convergence.

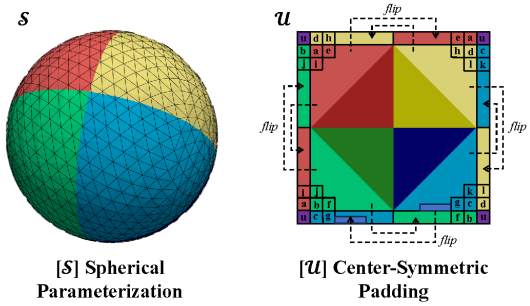

Figure 4: ($\mathcal{S}$) We obtain the spherical domain $\mathcal{S}$ via Spherical Parameterization $\mathcal{M} \to \mathcal{S}$, where ($\mathcal{U}$) denotes the unfold SGR grid. Identical colors indicate corresponding regions. ($\mathcal{U}$) The proposed *Center-symmetric Padding* mirrors the values symmetrically about the center of each edge. The purple color indicates averaging, with identical values marked by the same letters and colors.

Center-symmetric Padding. When the SGR is unwrapped in a 2D image, it inherently retains spherical continuity. We observed that directly applying traditional zero padding leads to cracks in 3D shapes, as illustrated in Fig. 5. To address this issue, we propose a new padding, *Center-symmetric Padding*, to better adapt to the spherical continuity of the SGR. Specifically, as illustrated in Fig. 4, each edge of $\mathcal{U}$ is centrally symmetric in $\mathcal{S}$, and the four corners of $\mathcal{U}$ are averaged in $\mathcal{S}$. The pseudocode for *Center-symmetric Padding* is shown in Appendix B. We use centrally-symmetric pixels as outer padding to convey 3D connectivity, thereby mitigating discontinuities.

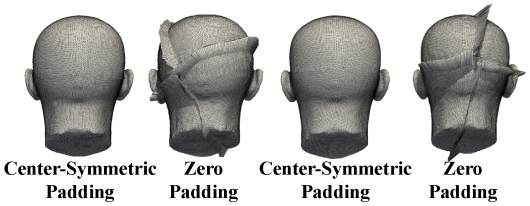

Figure 5: Impact of *Center-symmetric Padding*.

Geometric Regularization. Inspired by VQVAE (Esser et al. 2021), we incorporate pixel $\mathcal{L}_{\text{pix}}$, perceptual $\mathcal{L}_{\text{per}}$, and adversarial losses $\mathcal{L}_{\text{adv}}$ in our model. The loss function $\mathcal{L}_{rec}$ is a weighted sum of :

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{pix}} + \mathcal{L}_{\text{per}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}} \tag{8}$$

λ_{adv} controls the weight of $\mathcal{L}_{\text{adv}}$. For details of $\mathcal{L}_{\text{pix}}$, $\mathcal{L}_{\text{per}}$, and $\mathcal{L}_{\text{adv}}$, please refer to Appendix B.

However, traditional VQVAE suffers from high training costs and slow convergence. Leveraging the efficient conversion of SGR to 3D meshes, we introduce geometric regularization from a mesh perspective, which accelerates convergence and significantly boosts performance.

Given a reconstructed mesh $M = (V, F)$, where V denotes the vertices and F represents the faces, we define regularization as a weighted sum of the following losses.

Normals consistency loss, $\mathcal{L}_{nor}$, encourages the alignment of the normals of adjacent faces in the mesh. For two neighboring faces f_0 and f_1 with respective normals n_0 and n_1, the loss is calculated as:

$$\mathcal{L}_{nor} = \sum_{\langle f_0, f_1 \rangle} \left(1 - \frac{n_0 \cdot n_1}{\|n_0\| \|n_1\|}\right) \tag{9}$$

Laplacian smoothing loss, $\mathcal{L}_{lap}$, promotes smoothness across the mesh by penalizing large variations in vertex positions. It is defined as:

$$\mathcal{L}_{lap} = \sum_{i=1}^{N_v} \|Lv_i\|^2 \tag{10}$$

where N_v is the number of vertices in the mesh, and L is the Laplacian matrix applied to the vertex v_i.

Edge length regularization loss, $\mathcal{L}_{edge}$, penalizes the formation of distorted vertices. For each edge e in the mesh, the loss compares its actual edge length $\|e\|$ to a target length e_0:

$$\mathcal{L}_{edge} = \frac{1}{N_m} \sum_{m=1}^{N_m} \frac{1}{E_m} \sum_{e \in m} (\|e\| - e_0)^2 \tag{11}$$

where N_m is the number of meshes, E_m is the number of edges in each mesh, and $\|e\|$ represents the length of edge e.

The total loss of our geometric VQVAE is given by

$$\mathcal{L}_{total} = \alpha_{nor} \cdot \mathcal{L}_{nor} + \alpha_{lap} \cdot \mathcal{L}_{lap} + \alpha_{edg} \cdot \mathcal{L}_{edge} + \mathcal{L}_{rec} \tag{12}$$

where $\alpha_{nor}, \alpha_{lap}$, and α_{edg} are the respective weights for each loss term, allowing controlled emphasis on different aspects of *Geometric Regularization*.

Spherical Geometry Diffusion

By representing 3D assets as structured 2D maps, our **Spherical Geometry Diffusion** bypasses the complexities of 3D convolutions and irregular mesh processing. This reframes 3D generation as a 2D synthesis problem, allowing us to harness the pre-trained weights and optimization strategies of mature image diffusion models to boost performance.

We implement this through a decoupled two-stage process. First, *Text-based Geometry Generation* provides fine-grained control over facial structure from text prompts. Second, *Geometry-aware Texture Generation* leverages SGR's intrinsic alignment, using the geometry map as a perfect pixel-aligned condition for texture synthesis. This strategy guarantees geometric-textural consistency, a common failure in end-to-end models.

Text-based Geometry Generation. To efficiently generate the geometry SGR G from text embedding C, we model the distribution $p(G|C)$ within the latent space of our geometric VAE. The geometric UNet ϵ_θ is trained to denoise the latent z_t^{geo}, conditioned on the timesteps t and CLIP embeddings C of the text prompt. C is injected through cross-attention to minimize our objective in Eq. 1. At inference time, this trained model reverses the process. Starting from a random Gaussian latent noise, ϵ_θ iteratively denoises the latent vector, conditioned on the text embedding C. The resulting clean latents are then decoded by the geometric VAE decoder D_g to produce the final geometry SGR $\hat{G}$.

Geometry-aware Texture Generation. For the second stage, we learn to synthesize a corresponding texture SGR T. We frame this as a conditional image-to-image translation task, which is motivated by the intrinsic pixel-level correspondence of SGR. This key decision allows us to bypass depth rendering for alignment and instead leverage the pre-trained image diffusion checkpoint.

Specifically, we implement this by fine-tuning a texture generator ϵ_ϕ, to denoise a latent texture z_t^{tex}. Following InstructPix2Pix (Brooks et al. 2023), we first distill the comprehensive prompt y into a texture-specific description y_{tex}, which is then embedded to embedding c_{tex}. This is achieved by template prompts and GPT (Brown et al. 2020). Crucially, a strong geometric prior is supplied by concatenating the latent representation of the geometry, $E_g(G)$, with the noised texture latent z_t^{tex} along the channel dimension. We minimize the following latent diffusion objective:

$$\mathcal{L}_{\text{tex}} = \mathbb{E}_{\substack{E_t(T), E_g(G), C_{tex}, \\ \epsilon \sim \mathcal{N}(0,1), t}} \left[\|\epsilon - \epsilon_\phi(z_t^{\text{tex}}, t, E_g(G), C_{tex})\|_2^2 \right]. \tag{13}$$

Experimental Results

Implementation Details. We define the geometry SGR resolution (W, H) as 256x256, yielding reconstructed meshes with 65,536 vertices. For the corresponding texture SGRs, we employ a higher resolution of 512x512. Geometry SGRs are stored as 16-bit, three-channel PNGs, whereas all other SGRs utilize an 8-bit, three-channel format. For the text-based geometry generation stage, we train the Latent Diffusion Model for 500,000 steps on four NVIDIA A100 GPUs with a learning rate of 1e-4. The geometric regularization loss weights are set to $\alpha_{\text{nor}} = 0.1$, $\alpha_{\text{lap}} = 0.5$, and $\alpha_{\text{edg}} = 0.1$. To ensure training stability, we introduce an adversarial loss ($\lambda_{\text{adv}} = 0.1$) after the first 100 epochs. We employ the DPM-Solver++ scheduler (Lu et al. 2022) with Classifier-Free Guidance (Ho and Salimans 2022). For the geometry-aware texture generation stage, the model is fine-tuned from the official Stable-Diffusion-XL (SDXL) checkpoint (Podell et al. 2023). Further implementation details are available in Appendix C.

Datasets. We evaluate the performance of our method on various datasets, including Describe3D (Wu et al. 2023), FaceScape (Yang et al. 2020), and COMA (Ranjan et al. 2018). Describe3D contains 1,627 high-quality 3D face meshes paired with fine-grained textual descriptions, covering a wide age range and diverse ethnic backgrounds. FaceScape is a large-scale dataset that contains 16,572 3D scans of 847 individuals, each captured with 20 different expressions. The COMA dataset contains diverse sequential expression sequences from 12 individuals, totaling 17,794 meshes. Since FaceScape and COMA lack fine-grained text descriptions, we constructed structured text based on their attribute annotations, such as age, gender, and expression.

Method	DIV↑	DIV-ID↑	SP (mm)↓
Training data	1.00	1.00	-
3DMM	0.72	0.59	2.30
MAE	0.79	0.28	2.00
CoMA	0.69	0.52	2.47
FacialGAN	**0.96**	0.58	2.01
D3FSM	0.77	<u>0.81</u>	<u>0.84</u>
Ours	<u>0.93</u>	**0.88**	**0.82**

Table 1: Quantitative results of conditional geometry generation with respect to Diversity and Specificity.

Method	CD (mm) ↓	F-score@1mm ↑	AR ↓
i3DMM	0.759	83.87	3.18
FLAME	0.662	88.18	<u>1.84</u>
FaceScape	0.731	79.38	1.93
NPHM	0.637	91.67	3.90
ImFace	0.553	94.67	4.10
ImFace++	<u>0.511</u>	<u>96.11</u>	4.34
Ours	**0.466**	**98.77**	**1.35**

Table 2: Quantitative results of 3D face reconstruction.

As these attributes are not related to appearance, we primarily use these two datasets to evaluate the face geometry.

Quantitative Comparison

Conditional 3D Face Generation. To assess the geometric generation performance of our method, we conduct a quantitative evaluation on the FaceScape and COMA dataset. Following prior work (Taherkhani et al. 2023), we employ two metrics: Specificity (SP), which assesses how closely generated faces fit with training data distribution, and Diversity (DIV), which assesses the variety among generated samples. We compared our method with other mesh generation approaches, including 3DMM (Amberg et al. 2008), MAE (Abrevaya et al. 2018), CoMA (Ranjan et al. 2018), FacialGAN (Abrevaya et al. 2019a),and D3FSM (Abrevaya et al. 2019b). Our method demonstrates impressive performance across both metrics,effectively generating diverse and high-fidelity facial geometries, as shown in Table 1. Best performance are **bolded**, second-best are <u>underlined</u>.

Conditional 3D Face Reconstruction. We quantitatively evaluated the reconstruction of 3D face geometry on the FaceScape dataset. Following the official data split (Zheng et al. 2022), we train our model and then reconstruct the geometry from the test set using DDIM inversion (Dhariwal and Nichol 2021). We benchmark our approach against state-of-the-art (SOTA) methods employing alternate geometric representations, including FLAME (Li et al. 2017), FaceScape (Yang et al. 2020), i3DMM (Yenamandra et al. 2021), NPHM (Giebenhain et al. 2023), ImFace (Zheng et al. 2022), and ImFace++ (Zheng et al. 2024). We report Chamfer Distance (CD) and F-score as metrics for geometric accuracy, while Aspect Ratio (AR) (Parthasarathy et al. 1994) evaluates geometric quality, as shown in Table 2. The results demonstrate that our method achieves superior performance, outperforming all competing baselines in both geometric accuracy and mesh quality.

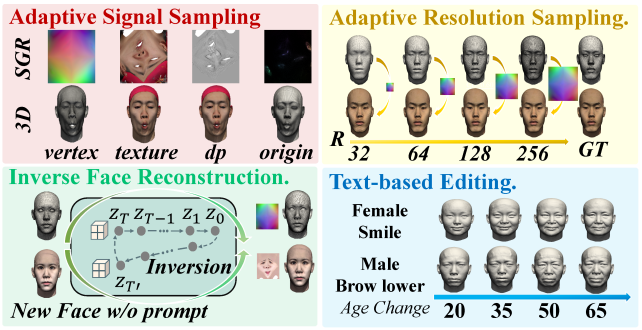

Figure 6: Our SGR enables adaptive signals and resolution. Leveraging the diffusion framework, we also enable inverse face reconstruction for new faces without prompts, along with text-based editing for gradual identity aging.

Method	Score ↑	Ranking-1 ↑	Time ↓
Describe3D	28.01	1	<u>0.7 m</u>
DreamFace	28.63	1	1.0 m
HumanNorm	29.42	3	85 m
FaceG2E	<u>29.74</u>	**8**	5.0 m
Ours	**29.80**	<u>7</u>	**0.2 m**

Table 3: CLIP evaluation results on the synthesized faces. Best performance are bolded, second-best are underlined.

Conditional Texture Generation. We quantitatively evaluate the visual performance on the Describe3D dataset using a CLIP-based assessment. This involves comparing our method against Describe3D (Wu et al. 2023), DreamFace (Zhang et al. 2023), HumanNorm (Huang et al. 2024) and FaceG2E (Wu et al. 2024), all tested on a shared set of 20 text prompts. In Table 3, we report the CLIP score (Sanghi et al. 2023) for text-mesh alignment, inference time (in minutes) for efficiency, and Ranking-1 for the times a method was ranked first according to CLIP. We use the official DreamFace website for its time calculations. The results demonstrate that our method achieves high fidelity and efficiency. As a native 3D generative model, it offers a significant advantage in computational speed over iterative SDS-based methods.

Qualitative Comparison

Conditional 3D Face Generation. In Fig. 7, we present qualitative results of our method compared to other methods. The results demonstrate that our method responds better to the fine-grained details of the text prompt.

Conditional 3D Face Reconstruction. We present qualitative results for Fig. 8. The results show that our method produces more accurate and detailed reconstructions. Moreover, Fig. 2 shows that our method generates uniform meshes, while the SDF-based methods exhibit noticeable artifacts.

Further Analysis. To demonstrate the versatility of our SGR representation, we conducted experiments on two ad-

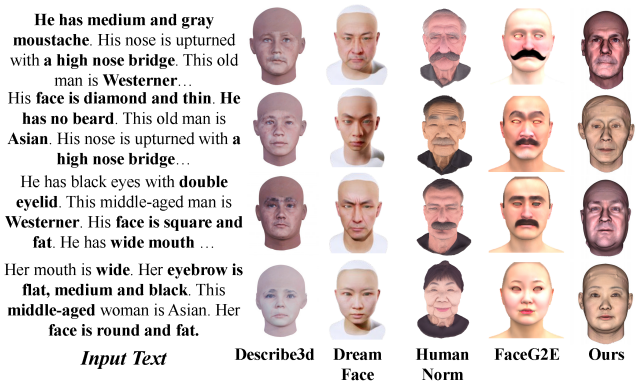

Figure 7: Qualitative results of Geometry-aware Texture Generation. Bold text indicates key aspects.

vanced tasks: direct text-based editing and out-of-domain synthesis using an SDS framework (Poole et al. 2022). For text-based editing, we show that modifying semantic attributes in the prompt, such as age or expression, enables precise corresponding edits to the facial geometry while preserving personal identity. As demonstrated in Fig. 6, increasing progressively the age in text prompt results in more pronounced facial wrinkles. This highlights our method's high degree of control over geometric features and its ability to capture fine details. More qualitative results, including those from our SDS experiments, can be found in Appendix D.

Ablation Study

Center-symmetric Padding. We conduct a qualitative ablation study on the effectiveness of *Center-symmetric Padding*. As shown in Fig. 5, *Center-symmetric Padding* helps to eliminate cracks and improve performance.

Geometric Regularization. We performed a quantitative experiment comparing model with and without *Geometric Regularization*, as shown in Table 4. *VQVAE-mesh* (with regularization) and *VQVAE-base* (without regularization) are equally trained for 30,000 steps. The results show that *Geometric Regularization* helps penalize deformed meshes, enabling faster convergence and improving efficiency.

Inversion steps. The quality of reconstruction during the inversion process is influenced by the number of steps taken. Generally, more steps lead to better quality, but this also increases computational time. To investigate this trade-off, we conduct ablation experiments. As illustrated in Table 4, the performance improves with more inversion steps. However, we observed diminishing returns, with additional steps beyond a certain point yielding only marginal gains. Consequently, we chose 600 inversion steps for our experiments to strike a balance between efficiency and quality.

Inference steps and guidance scale. In diffusion models, the inference step and the guidance scale are hyperparameters that regulate the generation performance. We conducted ablation studies to examine the impact of these variables. The results of this analysis, along with further ablation experiments, are detailed in Appendix D.

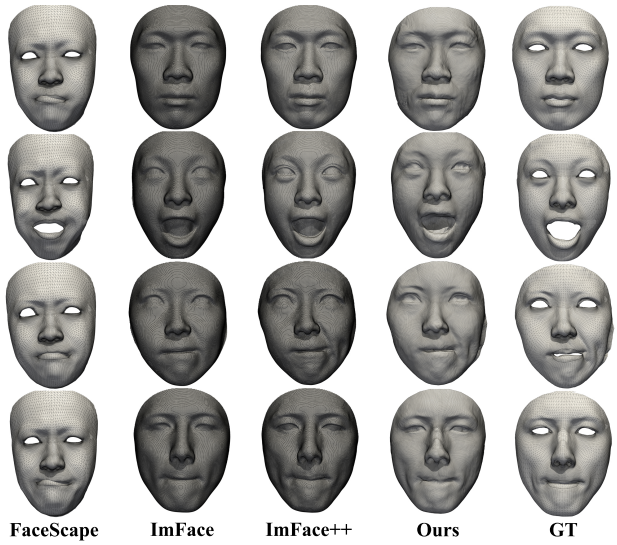

Figure 8: Qualitative results of 3D Face Reconstruction.

Setting	CD(mm) ↓	F-score@1mm ↑
Inversion Steps		
100 steps	1.235	66.70
200 steps	0.924	81.59
300 steps	0.693	93.00
400 steps	0.571	97.23
500 steps	0.571	97.24
600 steps	0.465	**98.77**
700 steps	0.465	98.76
800 steps	0.465	98.77
900 steps	0.465	98.77
1000 steps	**0.466**	98.76
Geometric Regularization		
VQVAE-base	0.810	74.83
VQVAE-mesh	**0.666**	**85.16**

Table 4: Ablation study on the number of inversion steps and effectiveness of geometric regularization.

Conclusion

In this paper, we introduced Spherical Geometry Diffusion, a framework that tackles the critical challenge of poor geometric fidelity in text-to-3D face generation. By representing facial structure on a canonical sphere, our method ensures clean mesh topology and seamlessly leverages powerful 2D diffusion models, establishing a new state of the art in geometric quality and textual alignment. We believe this principle of regularizing complex 3D structures onto simple manifolds offers a robust and scalable paradigm for future high-fidelity content creation.

Limitation. Although our method marks a significant advance, its performance is strained by data availability. Public text-3D face datasets, constrained by privacy concerns, are much smaller than 2D image datasets. We anticipate performance improvements as this data gap narrows.

Acknowledgments

This work was supported by the National Major Science and Technology Projects (2022ZD0117000 to M. Zhang), the National Natural Science Foundation of China (62202426 to M. Zhang), and the National Institutes of Health (NIH) (R21EB029733 to X. Gu).

References

Abrevaya; et al. 2018. Multilinear autoencoder for 3d face model learning. In *2018 IEEE winter conference on applications of computer vision (WACV)*, 1–9. IEEE.

Abrevaya, V. F.; Boukhayma, A.; Wuhrer, S.; and Boyer, E. 2019a. A decoupled 3d facial shape model by adversarial training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9419–9428.

Abrevaya, V. F.; Boukhayma, A.; Wuhrer, S.; and Boyer, E. 2019b. A generative 3D facial model by adversarial training. In *Proc. International Conference on Computer Vision (ICCV)*.

Amberg; et al. 2008. Expression invariant 3D face recognition with a morphable model. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, 1–6. IEEE.

Aneja, S.; Thies, J.; Dai, A.; and Nießner, M. 2023. Clipface: Text-guided editing of textured 3d morphable models. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–11.

Brooks; et al. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18392–18402.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, Z.; and Kim, T.-K. 2021. Learning feature aggregation for deep 3d morphable models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13164–13173.

Clarberg, P. 2008. Fast equal-area mapping of the (hemi) sphere using simd. *Journal of Graphics Tools*, 13(3): 53–68.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Esser; et al. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Fortune, S. 2017. Voronoi diagrams and Delaunay triangulations. In *Handbook of discrete and computational geometry*, 705–721. Chapman and Hall/CRC.

Giebenhain, S.; Kirschstein, T.; Georgopoulos, M.; Rünz, M.; Agapito, L.; and Nießner, M. 2023. Learning neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21003–21012.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hong, Y.; Peng, B.; Xiao, H.; Liu, L.; and Zhang, J. 2022. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20374–20384.

Huang, X.; Shao, R.; Zhang, Q.; Zhang, H.; Feng, Y.; Liu, Y.; and Wang, Q. 2024. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4568–4577.

Jiang, R.; Wang, C.; Zhang, J.; Chai, M.; He, M.; Chen, D.; and Liao, J. 2023. Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14371–14382.

Kirschstein; et al. 2024. Diffusionavatars: Deferred diffusion for high-fidelity 3d head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5481–5492.

Li, T.; Bolkart, T.; Black, M. J.; Li, H.; and Romero, J. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.*, 36(6): 194–1.

Liu, H.; Wang, X.; Wan, Z.; Shen, Y.; Song, Y.; Liao, J.; and Chen, Q. 2024. Headartist: Text-conditioned 3d head generation with self score distillation. In *ACM SIGGRAPH 2024 Conference Papers*, 1–12.

Lorensen, W. E.; and Cline, H. E. 1998. Marching cubes: A high resolution 3D surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, 347–353.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*.

Parthasarathy; et al. 1994. A comparison of tetrahedron quality measures. *Finite Elements in Analysis and Design*, 15(3): 255–261.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Praun, E.; and Hoppe, H. 2003. Spherical parametrization and remeshing. *ACM transactions on graphics (TOG)*, 22(3): 340–349.

Ranjan, A.; Bolkart, T.; Sanyal, S.; and Black, M. J. 2018. Generating 3D faces using convolutional mesh autoencoders. In *Proceedings of the European conference on computer vision (ECCV)*, 704–720.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent

diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Sanghi, A.; Fu, R.; Liu, V.; Willis, K. D.; Shayani, H.; Khasahmadi, A. H.; Sridhar, S.; and Ritchie, D. 2023. Clip-sculptor: Zero-shot generation of high-fidelity and diverse shapes from natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18339–18348.

Taherkhani, F.; Rai, A.; Gao, Q.; et al. 2023. Controllable 3d generative adversarial face model via disentangling shape and appearance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 826–836.

Wang, T.; Zhang, B.; Zhang, T.; Gu, S.; Bao, J.; Baltrusaitis, T.; Shen, J.; Chen, D.; Wen, F.; Chen, Q.; et al. 2023. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4563–4573.

Wu, H.; Zhao, M.; Hu, Z.; Fan, C.; Li, L.; Chen, W.; Zhao, R.; and Yu, X. 2025. ICE: Interactive 3D Game Character Facial Editing via Dialogue. *IEEE Transactions on Multimedia*.

Wu, M.; Zhu, H.; Huang, L.; Zhuang, Y.; Lu, Y.; and Cao, X. 2023. High-fidelity 3d face generation from natural language descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4521–4530.

Wu, Y.; Meng, Y.; Hu, Z.; Li, L.; Wu, H.; Zhou, K.; Xu, W.; and Yu, X. 2024. Text-Guided 3D Face Synthesis-From Generation to Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1260–1269.

Yang, H.; Zhu, H.; Wang, Y.; Huang, M.; Shen, Q.; Yang, R.; and Cao, X. 2020. 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*, 601–610.

Yenamandra, T.; Tewari, A.; Bernard, F.; Seidel, H.-P.; Elgharib, M.; Cremers, D.; and Theobalt, C. 2021. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12803–12813.

Zhang, L.; Qiu, Q.; Lin, H.; Zhang, Q.; Shi, C.; Yang, W.; Shi, Y.; Yang, S.; Xu, L.; and Yu, J. 2023. Dreamface: Progressive generation of animatable 3d faces under text guidance. *arXiv preprint arXiv:2304.03117*.

Zheng, M.; Yang, H.; Huang, D.; and Chen, L. 2022. Imface: A nonlinear 3d morphable face model with implicit neural representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20343–20352.

Zheng, M.; Zhang, H.; Yang, H.; Chen, L.; and Huang, D. 2024. ImFace++: A Sophisticated Nonlinear 3D Morphable Face Model with Implicit Neural Representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhuang, Y.; Lv, J.; Wen, H.; Shuai, Q.; Zeng, A.; Zhu, H.; Chen, S.; Yang, Y.; Cao, X.; and Liu, W. 2025. Idol: Instant photorealistic 3d human creation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 26308–26319.