3D Face Generation from Text Prompts via LoRA-Fine-Tuned Diffusion Models and Cross-Domain Diffusion

CSCI-677 Final Report

Minoo Ahmadi
minooahm@usc.edu

Armin Abdollahi
arminabd@usc.edu

Mehrshad Saadatinia
saadatin@usc.edu

Abstract

We present a two-stage framework for generating high-fidelity 3D facial models from natural language descriptions. In the first stage, we fine-tune a latent diffusion model, specifically Stable Diffusion, on the FFHQ dataset using Low-Rank Adaptation (LoRA). This fine-tuning enhances the model's ability to produce photorealistic facial images that accurately reflect the semantic content of diverse text prompts. In the second stage, we employ Wonder3D to convert the synthesized 2D facial images into detailed 3D representations through multi-view normal maps and color images. This approach combines the strengths of fine-tuned diffusion models with Wonder3D's efficient 3D reconstruction capabilities, bridging the gap between textual descriptions and 3D facial geometry. Our method facilitates intuitive and flexible 3D face generation, with potential applications in virtual reality, gaming, and digital avatar creation.

1. Introduction

3D face generation from natural language descriptions represents a fundamental challenge in computer vision and graphics with wide-ranging applications in virtual reality, gaming, digital avatar creation, and human-computer interaction [1]. While recent advancements in generative AI have revolutionized text-to-image synthesis, extending these capabilities to create high-fidelity 3D facial models remains complex, particularly when starting from text prompts alone. Traditional approaches to 3D face modeling typically rely on parametric models with limited expressivity or require multi-view inputs, making intuitive text-based generation difficult [1]. Recently, diffusion models have demonstrated remarkable success in generating high-quality 2D images from textual descriptions, presenting an opportunity to leverage these capabilities for 3D content creation [2]. However, significant challenges persist in translating the 2D generative power of these models into detailed and expressive 3D facial representations. In this work, we propose a two-stage framework that bridges this gap by combining the strengths of fine-tuned diffusion models and neural rendering techniques. Our approach first enhances a latent diffusion model, specifically Stable Diffusion, through Low-Rank Adaptation (LoRA) training on the FFHQ dataset [3]. This fine-tuning process significantly improves the model's ability to generate photorealistic facial images that accurately reflect diverse textual descriptions. The second stage leverages Wonder3D [1], a cross-domain diffusion framework that generates multi-view normal maps and corresponding color images, followed by geometry-aware normal fusion to reconstruct detailed 3D facial models from the generated 2D representations. Our contributions can be summarized as follows:

- We present an integrated pipeline combining LoRA-fine-tuned diffusion models with Wonder3D's cross-domain diffusion technique to enable high-fidelity 3D face generation from text prompts.
- We demonstrate how domain-specific fine-tuning enhances the semantic alignment and visual quality of facial images generated from diverse text descriptions.
- We show that our approach effectively captures the detailed geometric features of faces, producing expressive 3D models suitable for various applications.

This approach offers a flexible and intuitive method for text-driven 3D face generation, with significant advantages in both visual quality and semantic fidelity compared to existing techniques.

1.1. Change of Scope

Our project was initially titled *"3D Face Generation from Text Prompts via LoRA-Fine-Tuned Diffusion Models and Neural Rendering"*, with the goal of leveraging the EG3D framework [4] to generate 3D outputs. EG3D is a 3D-aware GAN architecture that combines StyleGAN2 with a tri-plane representation to produce high-resolution, multi-view-consistent images and corresponding 3D geometry.

However, integrating EG3D into our pipeline required substantial modifications, including training the model on our dataset and incorporating neural rendering (NeRF-based) mechanisms and GAN inversion techniques. Due to time constraints and limited computational resources, we were unable to implement these changes effectively.

Consequently, we shifted our approach to utilize a diffusion-based architecture for single-view 3D generation, specifically adopting the Wonder3D model. This model is better suited to our available resources and time frame, allowing us to achieve high-quality 3D reconstructions without the extensive training and architectural modifications that EG3D would have necessitated. Accordingly, we updated the project title to reflect this change in methodology.

2. Related Works

Several studies have explored the generation of 3D faces from textual descriptions, integrating natural language processing with computer vision techniques. Wu et al. [5] introduced a two-stage framework that first generates a 3D face matching concrete descriptions and then refines it using abstract descriptions. Zhang et al. [6] proposed DreamFace, a progressive scheme to generate personalized 3D faces guided by textual descriptions, employing a coarse-to-fine approach to create neutral facial geometry and utilizing latent diffusion models for detailed texture synthesis. Liu et al. [7] introduced Sherpa3D, a text-to-3D framework that achieves high-fidelity, generalizability, and geometric consistency simultaneously by employing guiding strategies derived from coarse 3D priors. Yu et al. [1] developed TG-3DFace, a framework that generates realistic 3D faces guided by textual descriptions, utilizing only text-2D face data for training and employing global contrastive learning and a fine-grained alignment module to ensure semantic consistency between the generated 3D faces and input texts. Kumar et al. [8] presented a plug-and-play, 3D-aware face editing framework that uses attribute-specific prompt learning, introducing a text-driven latent attribute editor (LAE) to enable the generation of facial images with controllable attributes across various poses. Another notable contribution to the field is the EG3D framework introduced by Chan et al. in 2022 [4]. A notable contribution relevant to our approach is the Wonder3D framework introduced by Long et al. [9] in 2023. Wonder3D introduces a cross-domain diffusion model that generates multi-view normal maps and corresponding color images from a single image. By ensuring consistency through a multi-view cross-domain attention mechanism, Wonder3D facilitates high-quality 3D reconstruction. The framework employs a geometry-aware normal fusion algorithm to extract detailed 3D geometry from the generated normal maps and color images, achieving high-fidelity reconstruction in just 2-3 minutes. This efficiency and quality make it particularly suitable for our text-to-3D facial modeling pipeline, where we leverage Wonder3D for the second stage of our framework. For the first stage of our pipeline, we draw inspiration from approaches that fine-tune diffusion models for specific domains. The Low-Rank Adaptation (LoRA) technique [10] offers an efficient way to specialize diffusion models for facial image generation while preserving their general capabilities. Our framework eliminates the need for complex latent space manipulations and GAN-inversion techniques commonly used in prior work. Unlike time-consuming neural rendering methods such as NeRF—which often require per-prompt optimization—we adopt a more efficient approach. Furthermore, we enhance the Stable Diffusion model, which by default performs poorly in detailed face generation. Collectively, these improvements address a significant gap in prior research.

3. Preliminaries

In this section, we discuss the preliminary concepts and background necessary to understand this work.

3.1. Diffusion Models

Diffusion models are a class of generative models that learn to generate data by reversing a gradual noising process. Starting from pure noise, these models iteratively denoise data to produce realistic samples. Notably, Stable Diffusion [2] has demonstrated impressive capabilities in generating high-quality images from textual prompts by operating in a latent space, which significantly reduces computational requirements while maintaining image fidelity.

3.2. Low-Rank Adaptation (LoRA)

Low-Rank Adaptation (LoRA) is a parameter-efficient fine-tuning technique designed for large-scale models [10]. Instead of updating all model parameters, LoRA introduces trainable rank decomposition matrices into each layer of the Transformer architecture, allowing for efficient adaptation to new tasks with reduced computational overhead. This approach is particularly beneficial when fine-tuning large models like Stable Diffusion, as it mitigates the risks of overfitting and catastrophic forgetting while requiring fewer resources.

3.3. Score Distillation Sampling (SDS)

Score Distillation Sampling (SDS) is the backbone of many text- or image-conditioned 3-D generation pipelines such as DreamFusion, Magic3D and Fantasia3D [11–13].Given a differentiable 3-D scene representation (e.g. a NeRF) with parameters θ, SDS renders an image $\mathbf{x}_\theta$ from a random camera, feeds it to a frozen 2-D diffusion model $\mathcal{D}$, and computes the *score* $\nabla_{\mathbf{x}} \log p_{\mathcal{D}}(\mathbf{x} \,|\, \text{condition})$ at a randomly sampled diffusion timestep. Minimizing the squared norm of that score w.r.t. θ,

$$\mathcal{L}_{\text{SDS}} = \Big\|\mathbf{s}_{\mathcal{D}}(\mathbf{x}_\theta, t, \text{cond}) - \mathbf{z}\Big\|^2,$$

(where $\mathbf{z} \sim \mathcal{N}(0, \sigma_t^2\mathbf{I})$ is the diffusion noise) aligns the rendered image distribution with the target distribution encoded by the diffusion prior. Because gradients are computed through the renderer, the 3-D scene gradually evolves so that its projections look realistic and condition-consistent across *all* sampled views. However, SDS requires thousands of gradient steps and often produces "Janus" two-face artefacts due to view-specific supervision; subsequent works propose coarse-to-fine schedules [12], re-consistency losses [14], or debiased scores [15] to mitigate these issues.

3.4. Cross-Domain Diffusion Single-Image 3D Reconstruction

Overview. Given a single RGB photograph, *Wonder3D* [9] produces a watertight, textured mesh in 2-3 minutes on one high-end GPU. The method (i) *jointly* synthesises multi-view normal maps and their aligned colour images with a *cross-domain diffusion* network, and (ii) fuses those normals into a signed-distance field (SDF) whose zero-level set is extracted as the final surface.

1) Cross-domain diffusion network. A single UNet—initialized from Stable Diffusion 1.5—operates on two domains: *normal* and *colour*. A learnable *domain-switcher* token, concatenated to the timestep embedding, tells the network which domain to predict at each denoising step, letting one backbone share priors across modalities without re-training separate models. During inference the model is called twice: once with the switcher = normal and once with switcher = colour, each time generating K=6–8 views under pre-sampled camera poses. Multi-view *cross-domain attention* layers let queries from one view (or domain) attend to keys/values from all other views and the opposite domain, enforcing geometric and appearance consistency and preventing the "Janus" multi-face artefacts that plague SDS pipelines.

2) Geometry-aware normal fusion. The predicted normals are back-projected into 3-D via known extrinsics, and a lightweight hash-grid SDF (two-layer instant-NGP) is optimised so that its gradient matches the fused normals. The optimisation employs (i) a *view-weighted normal loss* that assigns larger weights to normals forming a bigger angle with the viewing ray—these carry more reliable geometric information—and (ii) an *outlier-dropping* strategy that discards the top5% residuals at every iteration, shielding the SDF from hallucinated artefacts. After 60s the converged SDF is converted to a mesh with Marching Cubes, and the colour views are projected to a UV atlas to obtain detailed textures.

3) Advantages. Because geometry is supervised *explicitly* by normals rather than indirect 2-D photometric gradients, the reconstructions retain sharp edges and fine surface relief while running $>10\times$ faster than optimisation-heavy SDS baselines such as DreamFusion or Magic3D. Training on large-scale scanned objects further endows the model with strong out-of-distribution generalisation.

4. Methodology

We introduce a two-stage framework designed to generate high-fidelity 3D facial models from natural language descriptions. While diffusion models excel at generating high-fidelity 2D images from textual prompts, they often struggle to capture intricate facial details and perform suboptimally on close-up, photorealistic facial portraits. To address this, we first fine-tune a latent diffusion model—specifically, Stable Diffusion—on the high-quality FFHQ dataset using Low-Rank Adaptation (LoRA). This fine-tuning enables the model to specialize in generating high-quality facial images that accurately reflect the semantic content of diverse text prompts.

In the second stage, we apply a single-view 3D face reconstruction network—trained using multi-view diffusion priors—to convert each synthesized 2D image into a high-fidelity 3D mesh. By leveraging diffusion-based consistency across multiple virtual viewpoints during training, the model accurately recovers fine-scale facial geometry and dynamic expressions from just one input image

4.1. Stable Diffusion and Fine-Tuning

Stable Diffusion is a latent diffusion model (LDM) designed to generate high-quality images from textual descriptions. It operates by iteratively refining a noisy latent representation to produce coherent images. The model comprises three primary components: a variational autoencoder (VAE), a U-Net-based denoising network, and a text encoder. The VAE encodes images into a latent space, the U-Net performs denoising in this space, and the text encoder conditions the generation process on textual prompts [2].

Formally, the forward diffusion process adds Gaussian noise to a data sample $\mathbf{x}_0$ over T timesteps, producing noisy samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$. This process is defined as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}),$$

where β_t is a variance schedule controlling the noise level at each timestep. The reverse process aims to recover $\mathbf{x}_0$ by learning the conditional distribution $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, parameterized as:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)),$$

where μ_θ and Σ_θ are the mean and covariance predicted by a neural network.

To specialize Stable Diffusion for high-fidelity facial image generation, we employ Low-Rank Adaptation (LoRA), a parameter-efficient fine-tuning technique. LoRA freezes the pre-trained model weights and injects trainable low-rank matrices into the attention layers of the U-Net architecture, significantly reducing the number of trainable parameters [10]. Specifically, for a weight matrix $\mathbf{W} \in \mathbb{R}^{d\times k}$ in the original model, LoRA introduces two low-rank matrices, $\mathbf{A} \in \mathbb{R}^{d\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times k}$, such that the adapted weight is:

$$\mathbf{W}' = \mathbf{W} + \alpha \mathbf{A}\mathbf{B},$$

where α is a scaling factor and $r \ll \min(d, k)$, ensuring that the number of additional parameters is minimal. This adaptation allows the model to learn task-specific features without updating the entire weight matrix, making fine-tuning more efficient.

To fine-tune the diffusion model on the FFHQ dataset, we require aligned image-text pairs. However, FFHQ does not provide captions by default. Therefore, we generate captions for each image using an off-the-shelf image captioning model. Specifically, we use the BLIP (Bootstrapped Language-Image Pretraining) model by Salesforce [16], which combines vision-language pretraining with a Transformer-based encoder-decoder architecture. BLIP generates high-quality and semantically rich captions that are used as the textual supervision signal for conditioning during fine-tuning. This ensures that the model learns to associate facial features with accurate textual prompts, improving the semantic alignment and diversity of generated results.

4.2. Single-View 3D Reconstruction

For our second stage, we adopt the Wonder3D framework [9], which enables highly efficient high-quality 3D reconstruction from 2D inputs. Wonder3D first generates consistent multi-view normal maps and corresponding color images from a single image using a cross-domain diffusion model. This approach captures both the photorealism of diffusion models and the rich geometric details encoded in normal maps. The model facilitates information exchange across different views and modalities through a multi-view cross-domain attention mechanism [9]. This ensures that the generated normal maps and color images maintain geometric consistency across different viewpoints. Unlike methods that rely solely on color images, the explicit geometric information in normal maps significantly enhances reconstruction quality. To extract 3D geometry from the generated multi-view outputs, Wonder3D employs a geometry-aware normal fusion algorithm that robustly reconstructs surfaces from the multi-view 2D representations [9]. This algorithm is designed to handle inaccuracies in the generated views and produces clean, detailed geometries with high fidelity to the input image. The Wonder3D pipeline offers several advantages over previous approaches: it achieves high-quality reconstruction results while maintaining good efficiency (typically completing a reconstruction in 2-3 minutes), exhibits robust generalization to diverse input styles, and effectively addresses geometric inconsistencies that plague many single-view reconstruction methods [9].

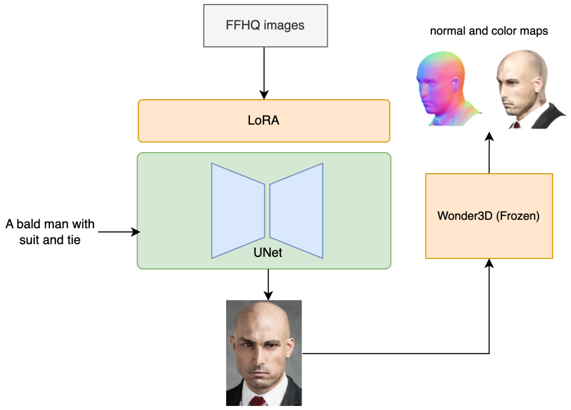


Figure 1. Overview of our face-to-3D pipeline. We fine-tune the Stable Diffusion model using images from the FFHQ dataset via a LoRA adapter. Given a text prompt (e.g., “A bald man with suit and tie”), the fine-tuned UNet generates a realistic face image. This image is passed to the pre-trained, frozen Wonder3D model, which predicts multi-view normal and color maps. These are used to reconstruct a high-fidelity, textured 3D mesh consistent with the input view.

4.3. Datasets

For training and fine-tuning our text-to-3D face generation pipeline, we utilize the Flickr-Faces-HQ (FFHQ) dataset. FFHQ is a high-quality dataset consisting of 70,000 images of human faces at a resolution of 1024×1024 pixels [3]. It includes a wide diversity in terms of age, ethnicity, facial expressions, accessories (e.g., glasses, hats), lighting conditions, and background variations, making it an ideal dataset for training generative models that require rich facial detail and diversity.

The dataset is curated to minimize artifacts and enhance realism, and it has become a standard benchmark for evaluating generative models such as StyleGAN and Stable Diffusion. In our project, we use FFHQ both to fine-tune the latent diffusion model for more semantically aligned facial image generation and to ensure high-fidelity outputs that are suitable for downstream 3D reconstruction.

5. Experiments and Results

We fine-tuned Stable Diffusion on a curated set of 5,000 images drawn from the 70,000-image Flickr-Faces-HQ (FFHQ) collection—just approximately 7% of the full dataset. Despite this limited sample, the resulting model already generates markedly sharper and more identity-consistent face renders than the original, unfine-tuned Stable Diffusion baseline.

5.1. Experiments

We conduct our experiments on a subset of the FFHQ dataset consisting of 5,000 high-resolution human face images. For computational efficiency, all images are downsampled to a resolution of 256×256. Each image is paired with an automatically generated caption using the BLIP image captioning model, which provides textual supervision during fine-tuning.

We fine-tune the `stable-diffusion-2-1` model using Low-Rank Adaptation (LoRA), which enables efficient adaptation of the large diffusion model with a minimal number of trainable parameters. Specifically, we apply LoRA to the cross-attention layers of the U-Net architecture, using a reduced rank of $r = 4$ and a scaling factor $\alpha = 8$.

To accelerate training, we reduce the number of denoising timesteps from the default 1,000 to 500, and fine-tune the model for 5 epochs with a batch size of 8 and a learning rate of 2e-5. Training is conducted on a single NVIDIA A100 GPU via Google Colab, using mixed-precision and memory-efficient techniques.

During inference, both the base and fine-tuned models are evaluated using identical prompts. The generated images are compared qualitatively to assess improvements in identity preservation, photorealism, and fidelity to facial attributes.

Additionally, we utilized the trained Wonder3D model to generate 3D faces from single-view inputs. This model is employed in the next phase to create 3D representations of the outputs generated by the fine-tuned diffusion model using textual prompts.

5.2. Final Results

In most typical synthetic-data generation studies, the Fréchet Inception Distance (FID) is used to evaluate the quality and diversity of generated samples. FID computes the Fréchet (Wasserstein-2) distance between two multivariate Gaussians fitted to the 2048-dimensional feature activations of a pre-trained Inception-v3 network—one Gaussian from real images and one from generated images; lower values indicate that the two feature distributions (and therefore the two image sets) are more similar. Some studies also employ downstream tasks as implicit metrics for generative-model performance. For the final evaluation of our project we adopt the FID metric, and supplement it with a qualitative visual inspection of the generated outputs.

As shown in Table 1, the FID between real FFHQ images (held-out images) and those generated by the original Stable Diffusion model is 150.7, indicating low visual fidelity and significant distribution mismatch. After fine-tuning, the FID improves substantially, reflecting a marked reduction in generative artifacts and closer alignment with real facial image statistics. Furthermore, the FID between the base and fine-tuned models is 83.6, suggesting that our model has undergone a meaningful distributional shift toward higher-quality generation. These results quantitatively confirm that even with limited data, fine-tuning significantly enhances facial realism in Stable Diffusion outputs and solidifies the need for such model.

Table 1. Fréchet Inception Distance (FID) between different model outputs

Comparison	**FID ↓**
Real vs. Base	150.7
Real vs. Finetuned	108.5
Base vs. Finetuned	83.6

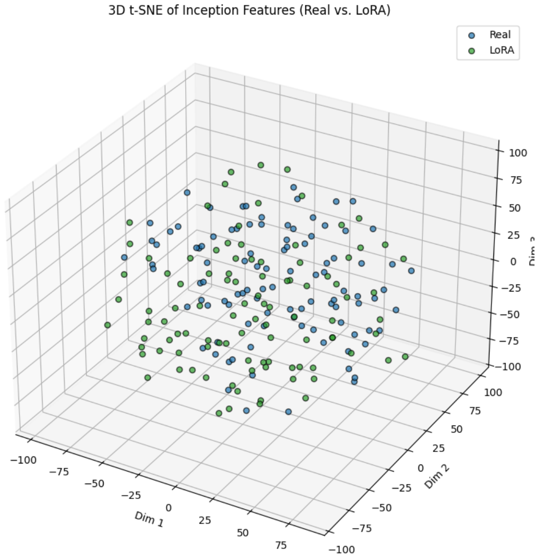
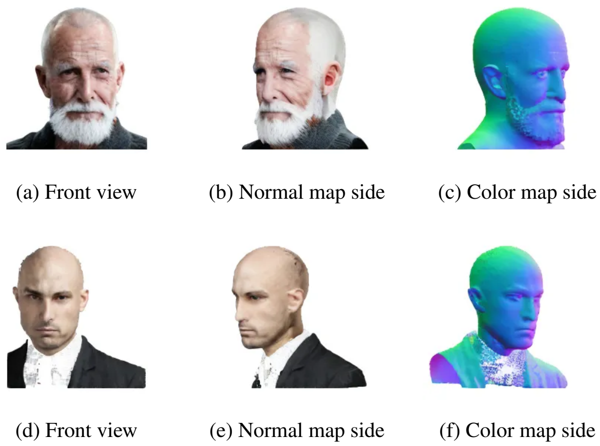


Figure 2. 3D t-SNE visualization of InceptionV3 features comparing real images to LoRA-generated images. The two distributions show the degree of similarity between synthetic and real samples in feature space on 300 images.

During the fine-tuning stage, we provided several prompts to both the base Stable Diffusion model and our fine-tuned version, then visually compared the outputs. Figures 4 and 5 clearly illustrate the superiority of the fine-tuned model in terms of identity preservation, photorealism, and alignment with the same 4 prompts. It is important to note that these results were obtained using only a small subset of the full dataset and a low LoRA rank of 4. We anticipate that minor artifacts observed in the current outputs will be significantly reduced with larger-scale training and increased LoRA rank. However, the limited time and scope of the course project, and the limitations of the compute did not allow us to explore the full potential of our method.

Figure 3 shows the 3D outputs generated from single-view inputs, rendered from three different angles. The last column displays the color maps produced by Wonder3D. These results clearly demonstrate that the reconstructed 3D representations are fully consistent with the original view and do not exhibit ”Janus” artifacts.


(a) Front view (b) Normal map side (c) Color map side

(d) Front view (e) Normal map side (f) Color map side

Figure 3. Multi-view consistent normals maps and color maps from the single-view (front-view) images fed to Wonder3D model [9]

6. Conclusion

In this work, we fine-tuned Stable Diffusion on a small subset—just 7%—of the FFHQ dataset and achieved a significant leap in face generation quality. While the original Stable Diffusion often produced low-fidelity and artifact-heavy results for faces, our fine-tuned model generates highly realistic facial images with sharp features and identity consistency.

To reconstruct 3D geometry from these improved 2D outputs, we employed the Wonder3D framework, which uses a cross-domain diffusion network with multi-view attention and domain-switching to jointly predict normal maps and color images from a single-view input. This architecture ensures geometric and visual consistency across views. By fusing the predicted normals using a geometry-aware normal fusion module, we obtain high-fidelity, textured 3D meshes. Compared to prior neural rendering methods, Wonder3D offers superior speed—typically completing in 2–3 minutes—while matching or exceeding their visual quality.

Due to limited computational resources and the time constraints of the academic semester, we were unable to train on the full FFHQ dataset or experiment with more resource-intensive neural rendering-based methods. Instead, we opted for a diffusion-based 3D generation pipeline that is both efficient and well-aligned with our hardware limitations.

For future work, we plan to expand our fine-tuning dataset to cover the full FFHQ corpus and explore subject-specific fine-tuning to better preserve identity. Additionally, integrating depth supervision and applying Wonder3D to in-the-wild portraits with challenging lighting and occlusions may further improve reconstruction robustness.

References

[1] C. Yu, G. Lu, Y. Zeng, J. Sun, X. Liang, H. Li, Z. Xu, S. Xu, W. Zhang, and H. Xu, “Towards high-fidelity text-guided 3d face generation and manipulation using only images,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1532–1542, 2023. 1, 2

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” *arXiv preprint arXiv:2112.10752*, 2022. 1, 2, 3

[3] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019. 1, 4

[4] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. D. Mello, O. Gallo, L. Guibas, J. Tremblay, S. Khamis, T. Karras, and G. Wetzstein, “Efficient geometry-aware 3d generative adversarial networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[5] M. Wu, H. Zhu, L. Huang, Y. Zhuang, Y. Lu, and X. Cao, “High-fidelity 3d face generation from natural language descriptions,” *arXiv preprint arXiv:2305.03302*, 2023. 2

[6] L. Zhang, Q. Qiu, H. Lin, Q. Zhang, C. Shi, W. Yang, Y. Shi, S. Yang, L. Xu, and J. Yu, “Dreamface: Progressive generation of animatable 3d faces under text guidance,” *arXiv preprint arXiv:2304.03117*, 2023. 2

[7] F. Liu, D. Wu, Y. Wei, Y. Rao, Y. Duan, and D. Lin, “Sherpa3d: Boosting high-fidelity text-to-3d generation via coarse 3d prior,” *arXiv preprint arXiv:2403.00825*, 2024. 2

[8] A. Kumar, M. Awais, S. Narayan, H. Cholakkal, S. Khan, and R. M. Anwer, “Efficient 3d-aware facial image editing via attribute-specific prompt learning,” in *European Conference on Computer Vision (ECCV)*, 2024. 2

[9] X. Long, Y.-C. Guo, C. Lin, Y. Liu, Z. Dou, L. Liu, Y. Ma, S.-H. Zhang, M. Habermann, C. Theobalt, and W. Wang, “Wonder3d: Single image to 3d using cross-domain diffusion,” *arXiv preprint arXiv:2310.15008*, 2023. 2, 3, 4, 6

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021. 2, 4

[11] B. Poole *et al.*, “Dreamfusion: Text-to-3d using 2d diffusion,” *arXiv:2209.14988*, 2022. 2

[12] C.-H. Lin *et al.*, “Magic3d: High-resolution text-to-3d content creation,” *arXiv:2211.10440*, 2023. 3

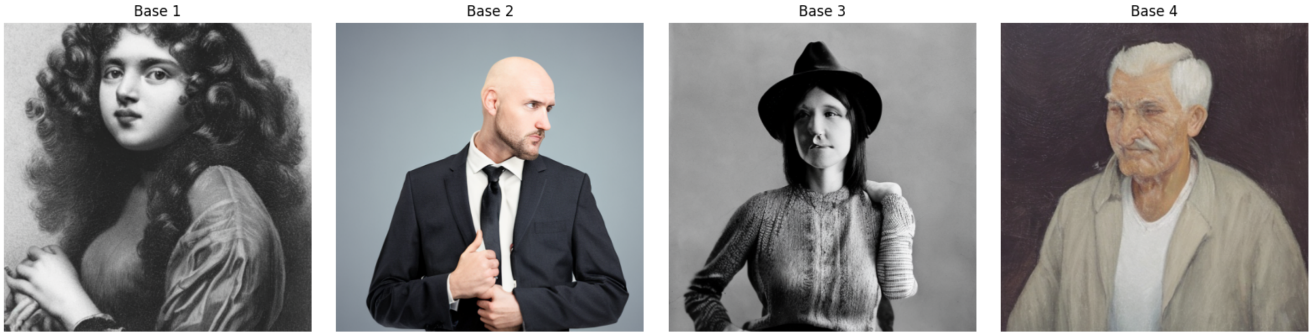


Figure 4. Outputs generated by the base Stable Diffusion model for the following prompts: (1) ”a young woman with curly hair”, (2) ”a bald man with a tie”, (3) ”a high fidelity photo of a woman with a hat”, (4) ”an old man with short hair”.

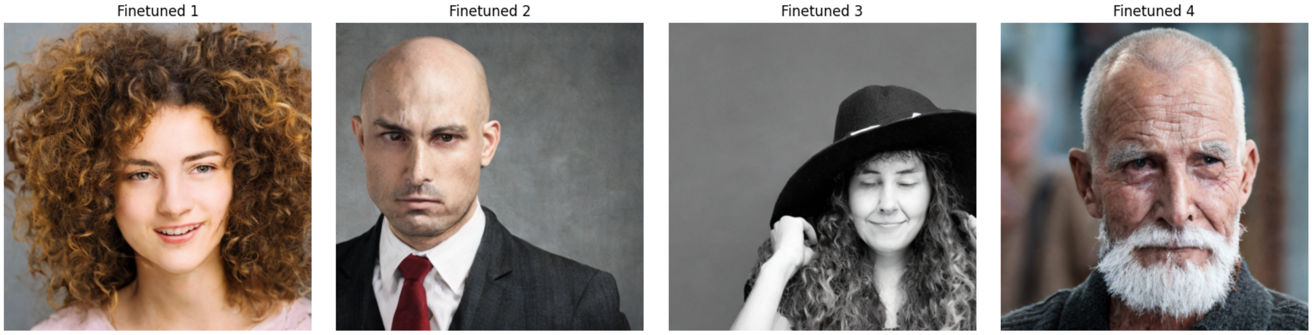


Figure 5. Outputs generated by the fine-tuned Stable Diffusion model for the same prompts as in Figure 4.

[13] R. Chen *et al.*, “Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d,” in *ICCV*, 2023. 2

[14] Z. Li *et al.*, “Debiasing scores and prompts of 2d diffusion for view-consistent text-to-3d,” *NeurIPS*, 2024. 3

[15] S. Hong *et al.*, “Debiased score distillation sampling,” *CVPR*, 2024. 3

[16] J. Li, D. Li, C. Xiong, and S. C. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 12888–12900, PMLR, 2022. 4