# Few-Shot 3D Face Generation via a Controllable Diffusion Model Guided by Text and Images

Jinfu Wei[1]*, Zheng Zhang[2]*, Qinchuan Zhang[2], Ran Liao[1]†, Duan Gao[2]†
[1]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2]Huawei Cloud Computing Technologies Co., Ltd., Shenzhen, China

***Abstract*—Recent advancements in text-to-3D generation have relied on large 3D datasets or expensive optimization processes during inference. In this paper, we introduce ControlFace, a novel framework designed for the creation of computer graphics-friendly 3D faces under the guidance of text and images. We utilize a controllable diffusion model to generate physically-based facial assets in texture space. The key to achieving few-shot generation lies in 3D-aware controls: a texture-space facial representation of geometry proxy. The main distinguishing feature of our framework is the effective integration of 3D facial priors with the diversity inherited from text-to-image diffusion models through few-shot learning, requiring only 36 3D faces for training. Once trained, ControlFace can generate diverse 3D faces in a feed-forward manner within 5 seconds and perform editing and stylization without 3D labeled data. We have demonstrated the effectiveness of our method in generating and editing various digital characters, guided by multi-modal controls.**

***Index Terms*—Generative Model, 3D Avatar, Physically-based Facial Assets**

## I. Introduction

Creating high-quality 3D digital humans is a challenging problem in computer graphics and computer vision. Traditional reconstruction-based methods either rely on expensive specialized hardware [1] or restricted parametric face models [2] to reconstruct the 3D face of real-world humans. Recent advancements in vision-language models [3], [4] have led to progress in text-to-3D generation [5], [6]. Current 3D face generation approaches can be classified into two main categories: classic inference-based methods and optimization-based methods. Classic inference-based methods utilize feed-forward generative models trained on 3D facial datasets to generate 3D faces with different representations, such as triangle mesh and Neural Radiance Fields (NeRF) [7]. Optimization-based methods incorporate 2D vision-language models to increase diversity and employ inference-time optimization techniques, typically based on CLIP [8], [9] or Score Distillation Sampling (SDS) [10], to generate 3D facial assets.

The main challenge in 3D face generation arises from the contradiction between the desired diversity and the limited amount of 3D facial data. A typical 3D facial dataset used in classic inference-based methods is not only costly to obtain, but also significantly smaller than the amount of training data for general vision-language models (VLMs). Therefore, the diversity of these methods is limited. Current optimization-based methods address this challenge through inference-time optimization, which is time-consuming and often suffers from artifacts such as oversaturation, over-smoothing, and diversity collapse. Consequently, the effective integration of 2D VLMs into 3D generation remains a challenge. To address these challenges, we introduce ControlFace, a novel controllable generative framework for high-quality and diverse 3D face generations guided by input prompts as shown in Fig. 1. Our key motivation is: there is a similarity between images and textures of faces, and we could restore this relationship thereby introducing geometry into diffusion models. To achieve this, we introduce a 3D-aware control module and use texture-space semantic maps as our 3D-aware control signals to ”control” the pre-trained diffusion model. This approach offers several advantages: 1) Few-shot learning: the number of trainable parameters in our control module is much smaller than that of the diffusion model, making it possible to train the control module with a compact dataset; 2) Feed-forward inference: the control module adapts the domain from natural images to UV texture space without compromising diversity, enabling 3D generation in a feed-forward manner within 5 seconds.

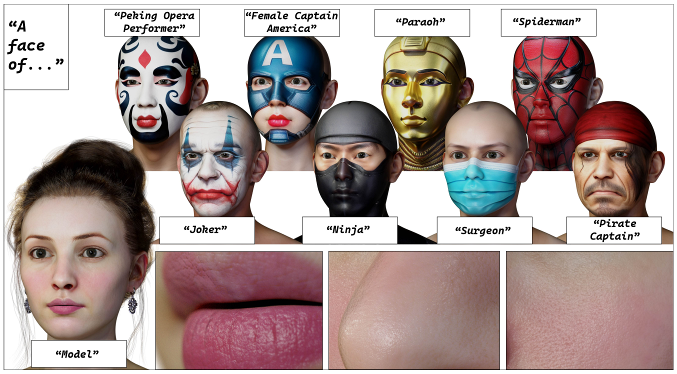


Fig. 1. ControlFace is capable of synthesizing various PBR textures with micro-structure-level skin details, which are compatible with existing CG pipelines. The hair, eyeballs, and jewelry are manually chosen.

Our generation pipeline consists of two main stages: Geometry Proxy Generation and Fine 3D Generation. In the first stage, we propose a geometry selection strategy, which combines the RGB image, rendered geometry image, and corresponding facial attribute, to select the optimal geometry. In the second stage, our albedo diffusion takes text descriptions as input and is controlled by the 3D-aware priors defined by the selected geometry proxy in texture space. This enables the generation of fine-grained geometry details (displacement

*Equal Contribution; †Corresponding Authors.

map) and physically-based rendering (PBR) appearance. ControlFace supports fine geometry control for synthetic textures including wrinkles, eyelids, and even topology. Furthermore, our model supports stylized 3D face generation and image guidance by seamlessly transferring 2D generation techniques into the 3D domain without 3D supervision.

In summary, our main contributions are: 1) A novel framework for generating high-quality digital faces, using text and image guidance. Our approach combines the diversity inherent in pretrained text-to-image diffusion models with the 3D facial priors; 2) A controllable diffusion generative model that can generate high-quality PBR facial maps, supporting few-shot training(requires only a minimal dataset consisting of 36 3D faces) and fast generation in a feed-forward manner within 5 seconds; 3) various applications of manipulations and controls, allowing for both localized and global editing (repainting and stylization) and precise adjustments to facial geometry at both detailed (e.g., wrinkles) and overall levels (e.g., topology).

## II. Related Work

### *A. General 3D Content Generation*

Image generation has made significant advancements in the recent decade, including GANs [11] and Diffusions [12], thanks to the development of generative models and the availability of large-scale datasets. These image generation models also contributed valuable insights for 3D generation. Inspired by the success of implicit representations such as NeRF [7], several works [13]–[15] effectively combine them with pre-trained diffusion models, enabling 3D generation by using SDS [6], although it suffers from over-smoothing, over-saturation, and long per-case optimization time. The following works leverage multi-view image generation in the way of feed-forward or improved optimization based on SDS [16], [17] to improve the quality and fidelity of 3D generation. The multi-view generation still faces the problem of inconsistent objects under different views. Overall, it's hard to find a balance between time-consuming and quality. Despite the rapid development of 3D datasets [18], the size of the 3D dataset is still smaller compared to the vast amount of data used in 2D large-scale model training.

### *B. 3D Face Generation*

As a sub-task of text-to-3d generation, text-guided 3D face generation has made significant progress. Current researches often use implicit representation to generate avatars, including feed-forward by GANs [19] or DMs [20], [21] and optimization-based methods [9], [22], [23]. Although They can create visually appealing avatars from text or images, they are not compatible with existing computer graphics pipelines and have limitations in animation, high-resolution rendering, and re-lighting than explicit 3D faces. For explicit representation, several 3D GANs work on a certain face dataset and suffer from limited diversity [24], [25]. Additionally, GAN models often face mode collapse issues, limiting their scalability for large-scale dataset training. Although CLIP-based optimization can be integrated to enhance generation diversity like [26], it is still restricted by dataset size. [10] achieve diverse generation through SDS-based optimization with fine-tuned vision-language models but rely on large, expensive datasets containing over 1,000 3D face assets and involve time-consuming iterations for each generation. [27] finetune the U-Net from Stable Diffusion on 188 super-high quality samples to generate UV maps. In contrast, we focus on generating computer graphics-friendly 3D digital humans via few-shot learning. We only rely on a compact dataset comprising 30+ 3D facial assets. Importantly, different from most, our method supports fast generation in a feed-forward manner, eliminating the need for time-consuming inference-time optimization, and supporting fine geometry control simultaneously.

## III. Method

ControlFace, as illustrated in Fig. 2, is a novel generative framework that integrates the Latent Diffusion Model (LDM) with 3D-aware controls to generate high-fidelity 3D faces. In our method, we utilize a combination of base shape $g_b$ and displacement map $v$ to represent the face geometry. The base shape is represented using a triangle mesh with unified topology, while the displacement map allows for adjustments to the position of each vertex on the facial surface. The appearance properties of faces are modeled by physically-based, spatially-varying bidirectional reflectance distribution functions (SVBRDFs) including diffuse albedo $k_d$, surface normal $n$, specular albedo $k_s$, and specular roughness $r$.

ControlFace consists of three main modules: geometry proxy generation, 3D-aware albedo diffusion, and fine-grained facial generation. In the geometry proxy generation module, we select the optimal base geometry $g_b$ from a set of pre-built face candidates, which are reconstructed from 2D facial images. To facilitate the controllable generation, we render the base geometry into UV texture space maps, serving as the 3D facial priors in the subsequent stage. In the albedo diffusion, our core generative network, conditioned by the texture-space 3D facial prior, generates a high-quality facial albedo map $k_d$ according to the input text. Finally, we generate the fine-grained geometry details, represented by displacement map $v$, and PBR material maps $\{k_s, n, r\}$ from the albedo map through image-to-image translation techniques.

### *A. Geometry Proxy Generation*

Recent approaches [10] rely on the similarity between text and rendered geometry images to guide the selection process. However, they face limitations in capturing detailed characteristics due to the lack of facial appearance details in the rendered images. To address this challenge, we incorporate additional information from portrait images and facial features to resolve ambiguities in matching text and mesh.

We construct the candidate pool with 8,425 samples using the CelebA-HQ. Specifically, for each image $I_i$ in CelebA-HQ, we utilize a single-image 3D face reconstruction method [28] to obtain the corresponding 3D face shape $g_i$. To select the optimal geometry proxy from candidate set $S = g_i, R_i, I_i, f_i$, where $R_i$ consists of 10 texture-less rendering images under

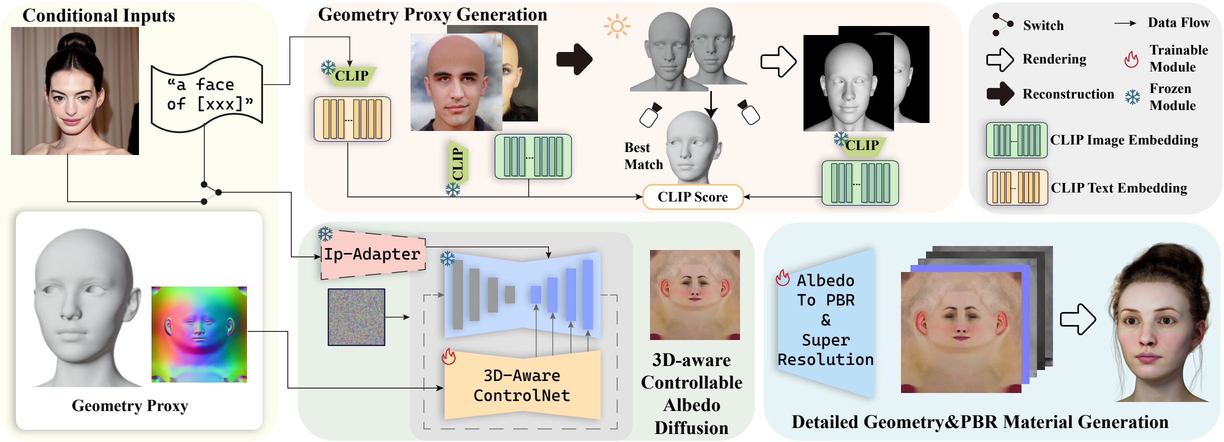


Fig. 2. Overview. Given a text description ("a face of [xxx]"), ControlFace can select the best matching geometry proxy by the Geometry Proxy Generation module. Then, the 3D-aware Controllable Albedo Diffusion generates correlated albedo with the text guidance or optional image guidance. Finally, high-quality PBR textures and detailed geometries are provided by the Albedo-To-PBR model together with the Super Resolution network.

different camera view, $f_i$ is the facial attributes provided by CelebA-HQ, our selection process is driven by a matching loss, including the text-to-render similarity $L_R$, the text-to-image similarity $L_I$, and the text-to-attribute similarity $L_f$:

$$\begin{aligned}
\mathcal{L} &= \mathcal{L}_R(t, R_i) + \mathcal{L}_I(t, R_i) + \mathcal{L}_f(t, f_i), \\
\mathcal{L}_R &= \lambda_d^R e_i(R_i) \cdot e_t(t) + \lambda_r \Delta e_i(R_i) \cdot \Delta e_t(t), \\
\mathcal{L}_I &= \lambda_d^I e_i(I_i) \cdot e_t(t), \\
\mathcal{L}_f &= \lambda_t e_t(f) \cdot e_t(t),
\end{aligned} \tag{1}$$

where $e_i(\cdot) = norm(\mathcal{E}_I(\cdot))$ is the normalized image features generated by CLIP's image encoder $\mathcal{E}_I$, $e_t(\cdot) = norm(\mathcal{E}_T(\cdot))$ is the normalized text features generated by CLIP's text encoder $\mathcal{E}_T$, $\Delta e_i = e_i - \bar{e}_i, \Delta e_t = e_t - \bar{e}_t$ are the relative similarity [9], $\bar{e}_t, \bar{e}_i$ represent the embedding of an anchor text (e.g. "a face of") and the renderings of the mean mesh $\bar{g}_i$. We select the highest matching score as our geometry proxy.

### B. 3D-aware Controllable Albedo Diffusion

A high degree of similarity was observed between facial texture and rendering images. This relationship can be effectively modeled through an underlying geometric control module. Inspired by ControlNet [29], we propose 3D-aware controllable albedo diffusion, which takes text descriptions $c_t$ as input and utilizes a texture space facial prior $c$ to condition our control module $\mathcal{C}$, enabling the generation of high-quality albedo maps $k_d$. Specifically, we render the geometry proxy $g_b$ to geometry normal map $c_n$ in UV texture space. Unlike general image generation, facial texture generation requires precise topological alignment to ensure that facial features are located in specific regions. The texture-space normal control offers dense, pixel-wise controls for geometrically related detail features such as wrinkles. Compared to generic diffusion models, our method excels in precise control over the generation process by the efficient incorporation of 3D prior.

The architecture of our control module follows the same design as described by [29]. It consists of two parts: a locked copy of a large pre-trained model and a trainable copy that is connected using zero-convolution layers. Specifically, our control module takes the 3D-aware control condition $c \in \mathbb{R}^{3\times512\times512}$ as input and uses a small network consisting of four convolution layers to encode the image space condition $c$ into a feature space $c_f \in \mathbb{R}^{4\times64\times64}$. This ensures that the resolution of the encoded condition features matches the resolution of the latent space in the stable diffusion model. To incorporate the encoded condition $c_f$ into the generative process, we clone the trainable parameters of encoder blocks and the middle block of the UNet architecture. These cloned blocks are used to extract features from the encoded condition $c_f$. These features are then added back to the decoding blocks and middle blocks of the UNet through skip connections.

The training objective $\mathcal{L}$ of our control module is:

$$\mathcal{L} = \mathbb{E}_{t, c_f, c_t, \epsilon \sim N(0,1), t}\left[ \left\| \epsilon - \epsilon_\theta\left(z_t, t, c_f, c_t\right) \right\|_2^2 \right], \tag{2}$$

where $t$ represents the time step, $c_t$ is the text prompt, $c_f$ is the latent feature of our 3D-aware condition $c$, $\epsilon_\theta$ denotes the denoising UNet, and $z_t$ is the noisy image.

*Non-uniform time step sampling:* We have observed that the impact of 3D-aware conditional control is more significant in the early phases of denoising sampling compared to the later stages. To enhance the control capabilities and accelerate convergence, we draw inspiration from [30] and utilize a cubic function to implement non-uniform time step sampling during training: $t = (1 - (\frac{t}{T})^3) \times T, t \in U(0, T)$. This approach allows us to allocate more time steps to the earlier stages of the denoising process and thus helps to maximize the impact of 3D-aware control.

### C. Detailed Geometry and PBR Material Generation

The detailed facial geometry, as captured by displacement maps, alongside the physically-based material maps, plays a crucial role in achieving realistic digital human renderings. Recognizing that both the displacement map and physically-based material maps exhibit strong pixel-level correlations with the albedo map, we introduce an image-to-image translation network to faithfully generate the detailed geometry $(v)$

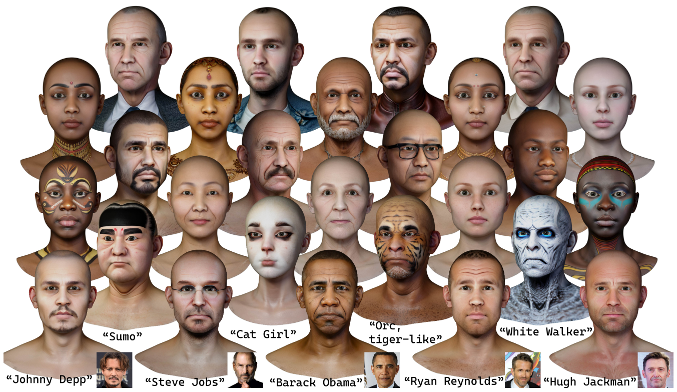
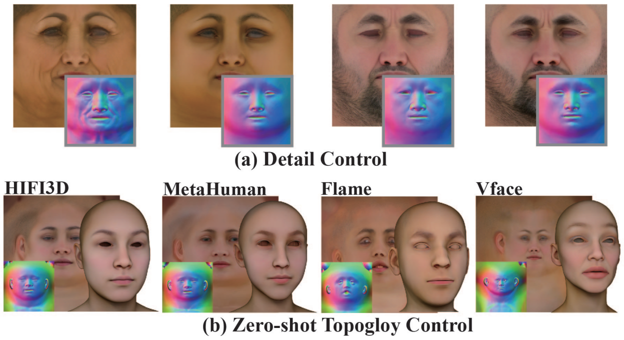


Fig. 3. Generation results. We present the generation results of ControlFace, showcasing a diverse range of avatars created using various prompts. The first 4 rows demonstrate individuals described by general prompts, while subsequent rows showcase fictional characters and celebrities.



(a) Detail Control

(b) Zero-shot Topogloy Control

Fig. 4. Examples generated under geometry proxy guidance. We fix the prompts and seeds in (a), and ControlFace can generate geometry-related textures, such as wrinkles(left) and eyelids(right). In (b), ControlFace can synthesize out-of-doimian topology texture by giving geometry normals.

and material maps ($\{k_s, n, r\}$) directly from the input diffuse albedo map ($k_d$). To keep diversity in PBR appearance, we utilized and finetuned the SVBRDF decoder in [31] on our facial assets. The training loss is the sum of three terms:

$$\mathcal{L} = \lambda_{map}\mathcal{L}_{map} + \lambda_p\mathcal{L}_p + \lambda_{gan}\mathcal{L}_{gan}, \tag{3}$$

where $\mathcal{L}_{map}$ is the $L_1$ loss on the reconstructed maps, $\mathcal{L}_p$ is a perceptual loss, and $\mathcal{L}_{gan}$ is the Patch GAN loss. In our setting, we set $\lambda_r = 2$, $\lambda_p = 0.5$, $\lambda_g = 0.5$. To further improve the quality, we introduce a texture alignment and super-resolution module. Texture alignment is used to fix pixel-level misalignments in texture space after albedo generation; the super-resolution module is used to produce pore-level details for PBR materials and further enhance the realism of rendering. For more details please refer to supplementary.

## IV. Results

### A. Implementation Details

*Dataset:* To generate high-quality 3D faces with PBR textures, a dataset of 36 text-asset pairs was constructed. This dataset comprised 32 randomly selected samples from the commercial 3DScanStore [32] and four manually created examples featuring accessories such as face masks and eye covers. Further data augmentation is detailed in the supplementary material.

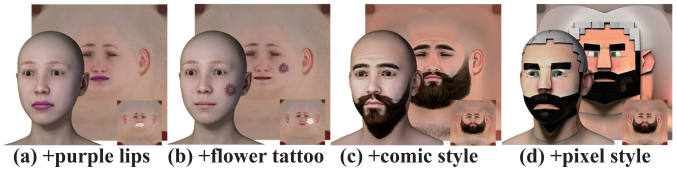


Fig. 5. Local manipulation and style transformation. Fig (a) and (b) showcase the local repainting through hand-made masks. In Fig (c), ControlFace converts the digital human into a comic style by adding the prompt "comic style". Fig (d) showcases personalized stylization using the pixel style LoRA.

*Setting:* Experiments were conducted on a single Nvidia RTX 3090 GPU, and the training employs the Adam optimizer. The 3D-aware generation model was trained for 20 epochs with a learning rate of 5e-6. For PBR material generation, individual UNet networks were trained for each specific output, using a learning rate of 1e-5 for 30 epochs.

### B. Generation Results

ControlFace exhibits remarkable diversity by generating avatars with different clothing, accessories, skin tones, or even non-human features such as orcs, which are out of our training distribution. Moreover, ControlFace is capable of generating avatars with a high degree of resemblance to the given name or detailed description, without additional optimization processes, as shown in Fig. 3. Furthermore, we provide extended results of diversity affected by initial random noise in supplementary.

*Geometry Proxy Control:* Geometric normal maps provide dense guidance enriched with 3D geometric priors, making accurate controlling possible as shown in Fig. 4a. Geometry topology defines the way to project 3D models' surface to textures. Leveraging geometry control, ControlFace can perform zero-shot control of topology with the guidance of the geometry normal map extracted from the target topology. As shown in Fig. 4b, we generate the texture on different out-of-domain topologies including Vface [32], Flame [33], HIFI3D [34], and MetaHuman [35].

*Local Manipulation and Style Transformation:* Apart from the generation process, ControlFace can further perform local manipulation based on image-to-image and inpainting. As shown in Fig. 5a&b, with only a rough mask, ControlFace can modify local appearance, such as recoloring lips or adding tattoos and decals with text prompts. Furthermore, we can transfer the texture style while maintaining the recognizable identities by image-to-image generation. As shown in Fig. 5c&d, we enable the personalized stylized generation or transformation, such as a comic style or pixel art, by simply replacing the text prompts or using Lora as a Style Adaptor.

*Image Guidance Generation:* With pre-trained IP-Adapter [37], ControlFace is capable of providing precise control over avatars' identity-specific styles with an image prompt as shown in Fig. 7. Moreover, to verify the effectiveness of image guidance compared with text guidance, we evaluate the semantic similarity of CLIP [3] and the facial similarity

TABLE I
QUANTITATIVE COMPARISON OF CLIP SCORE AND INFERNECE TIME.

| | CLIP Score ↑ | Inference Time ↓ |
|---|---|---|
| AvatarClip [9] | 0.284 | ∼ 5 hours |
| DreamMat [17] | 0.234 | ∼ 30 mins |
| ChatAvatar [36] | 0.282 | ∼ 30 sec |
| ClipFace [26] | 0.262 | ∼ 30 mins |
| **ControlFace(Ours)** | 0.312 | ∼ 5 sec |

TABLE II
QUANTITATIVE COMPARISON OF IMAGE GUIDANCE GENERATION.

| | 1-cosine↓ | l1 ↓ | l2↓ | CLIP Score↑ |
|---|---|---|---|---|
| w/o. image | 0.879 | 6.348 | 1.321 | 0.568 |
| w. image | 0.821 | 6.065 | 1.274 | 0.593 |

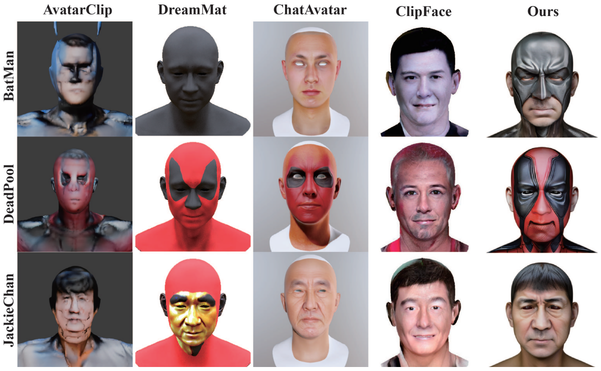


Fig. 6. Quanlitative comparison of generation quality with other methods.

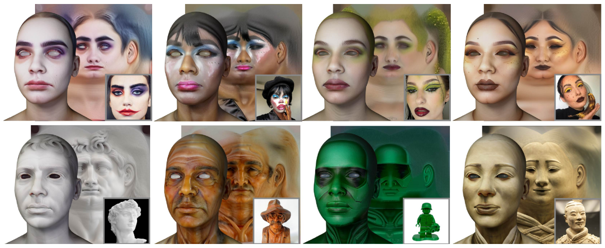


Fig. 7. Results of image guidance. ControlFace is capable of facilitating digital makeup imitation and identity injection with image (bottom-left).

of DeepFace [38] on FFHQ. Specifically, we sampled 1000 images from FFHQ and utilized CLIP Interrogator [1] to collect corresponding text prompts. Then, we generated 3D faces with and without image guidance separately and compared the similarity between rendering images and images in FFHQ. The comparison in Table II shows that the image guidance module enables more consistent features than text-only guidance.

### C. Comparison

We compare ControlFace with the other 3D digital human generation methods in Fig. 6, including AvatarClip [9], DreamMat [17], ClipFace [26] and ChatAvatar [36], in terms of Clip Score and inference time (see Table I). Following [10], we generated 10 different characters, including people generated by general descriptions, celebrities, and film characters. All prompts used for generation followed the same anchor: "a face of xxx". We use "ViT-L/14" as the pre-trained CLIP model. As a result, ControlFace produces results that exhibit a higher resemblance to text descriptions with less inference time.

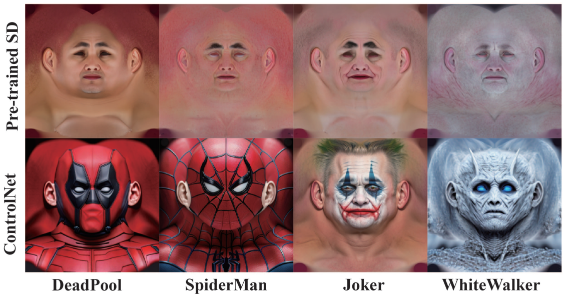


Fig. 8. The catastrophic forgetting of the finetuned SD model. The first row showcases the generation results of the fine-tuned SD model, while the second row presents the generation results of ControlFace using the same prompt.

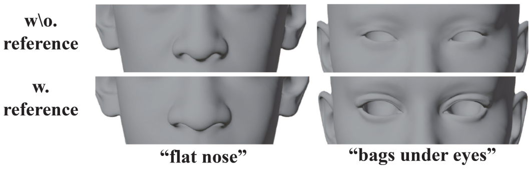


Fig. 9. Comparison of improved geometry selection.

### D. Ablation study

*Comparison with Finetuning:* Direct fine-tuning of the pre-trained Stable Diffusion model with limited training data leads to catastrophic forgetting caused by overfitting. As illustrated in Fig. 8, the Stable Diffusion model, fine-tuned with a dataset of merely over 30 samples, exhibits significant overfitting and fails to generate fictional characters. On the other hand, refraining from overriding the pre-trained parameters of Stable Diffusion, ControlFace maintains the model's generation and generalization abilities.

*Improved Geometry Selection:* To validate the effectiveness of the image-guided geometry proxy generation, we conduct the experiments with or without image guidance and select the one with the highest clip score for each method to ensure fairness. As shown in Fig. 9, when matching some description of facial details, the method without image guidance fails to select the best matching result. In contrast, our method obtains more accurate matching results. It demonstrates the effectiveness of our proposed method.

## V. CONCLUSION

In this paper, we propose ControlFace, a novel diffusion-based generative model that incorporates 3D-aware controls to enable the creation of customized, high-quality 3D facial assets. ControlFace can achieve few-shot learning, leveraging a compact 3D face dataset consisting of just 30+ samples. Our generative model is highly efficient, allowing for training within an hour and generating high-fidelity results in only seconds. We demonstrated the effectiveness of our ControlFace method in generating and editing a wide variety of digital characters, guided by multi-model controls including text

[1] huggingface.co/spaces/pharmapsychotic/CLIP-Interrogator

prompts, character portrait images, styled reference images, and 3D-aware controls.

## References

[1] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar, "Acquiring the reflectance field of a human face," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000, pp. 145–156.

[2] Volker Blanz and Thomas Vetter, "A morphable model for the synthesis of 3d faces," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 157–164. 2023.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695.

[5] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole, "Zero-shot text-guided object generation with dream fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 867–876.

[6] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv*, 2022.

[7] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[8] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Niessner, "Clipface: Text-guided editing of textured 3d morphable models," in *ACM SIGGRAPH 2023 Conference Proceedings*, New York, NY, USA, July 2023, SIGGRAPH '23, pp. 1–11, Association for Computing Machinery.

[9] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu, "Avatarclip: Zero-shot text-driven generation and animation of 3d avatars," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–19, 2022.

[10] Longwen Zhang, Qiwei Qiu, Hongyang Lin, Qixuan Zhang, Cheng Shi, Wei Yang, Ye Shi, Sibei Yang, Lan Xu, and Jingyi Yu, "Dreamface: Progressive generation of animatable 3d faces under text guidance," *arXiv preprint arXiv:2304.03117*, 2023.

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," 2014.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[13] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng, "Dreamgaussian: Generative gaussian splatting for efficient 3d content creation," *arXiv preprint arXiv:2309.16653*, 2023.

[14] Yiwen Chen, Chi Zhang, Xiaofeng Yang, Zhongang Cai, Gang Yu, Lei Yang, and Guosheng Lin, "It3d: Improved text-to-3d generation with explicit view synthesis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 1237–1244.

[15] Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang, "Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6796–6807.

[16] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu, "Clay: A controllable large-scale generative model for creating high-quality 3d assets," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 4, pp. 1–20, 2024.

[17] Yuqing Zhang, Yuan Liu, Zhiyu Xie, Lei Yang, Zhongyuan Liu, Mengzhou Yang, Runze Zhang, Qilong Kou, Cheng Lin, Wenping Wang, et al., "Dreammat: High-quality pbr material generation with geometry-and light-aware diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 4, pp. 1–18, 2024.

[18] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi, "Objaverse: A universe of annotated 3d objects," *arXiv preprint arXiv:2212.08051*, 2022.

[19] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al., "Efficient geometry-aware 3d generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16123–16133.

[20] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo, "Rodin: A generative model for sculpting 3d digital avatars using diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4563–4573.

[21] Xiaolong Shen, Jianxin Ma, Chang Zhou, and Zongxin Yang, "Controllable 3d face generation with conditional style code diffusion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 4811–4819.

[22] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K. Wong, "Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models," 2023.

[23] Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Daniel Du, and Min Zheng, "Avatarverse: High-quality & stable 3d avatar creation from text and pose," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 7124–7132.

[24] Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, et al., "Learning formation of physically-based face attributes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3410–3419.

[25] Menghua Wu, Hao Zhu, Linjia Huang, Yiyu Zhuang, Yuanxun Lu, and Xun Cao, "High-fidelity 3d face generation from natural language descriptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4521–4530.

[26] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner, "Clipface: Text-guided editing of textured 3d morphable models," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–11.

[27] Mingyuan Zhou, Rakib Hyder, Ziwei Xuan, and Guojun Qi, "Ultravatar: A realistic animatable 3d avatar diffusion model with authenticity guided textures," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1238–1248.

[28] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[29] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, "Adding conditional control to text-to-image diffusion models," 2023.

[30] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," *arXiv preprint arXiv:2302.08453*, 2023.

[31] Linxuan Xin, Zheng Zhang, Jinfu Wei, Wei Gao, and Duan Gao, "Dreampbr: Text-driven generation of high-resolution svbrdf with multi-modal guidance," *arXiv preprint arXiv:2404.14676*, 2024.

[32] 3DScanStore, "3d scan store," https://www.3dscanstore.com/, 2023.

[33] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero, "Learning a model of facial shape and expression from 4d scans.," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194–1, 2017.

[34] Linchao Bao, Xiangkai Lin, Yajing Chen, Haoxian Zhang, Sheng Wang, Xuefei Zhe, Di Kang, Haozhi Huang, Xinwei Jiang, Jue Wang, et al., "High-fidelity 3d digital human head creation from rgb-d selfies," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 1, pp. 1–21, 2021.

[35] Unreal Engine, "Metahuman," https://www.unrealengine.com/en-US/metahuman, 2023.

[36] Qixuan Zhang, Longwen Zhang, Lan Xu, Di Wu, and Jingyi Yu, "Chatavatar: Creating hyper-realistic physically-based 3d facial assets through ai-driven conversations," in *ACM SIGGRAPH 2023 Real-Time Live!*, New York, NY, USA, 2023, SIGGRAPH '23.

[37] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," 2023.

[38] Sefik Serengil and Alper Ozpinar, "A benchmark of facial recognition pipelines and co-usability performances of modules," *Journal of Information Technologies*, vol. 17, no. 2, pp. 95–107, 2024.