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

Anonymous submission

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 few-shot framework designed for the creation of computer graphics-friendly 3D faces under the guidance of input 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 a geometry proxy based on the input prompt. 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, our method can generate diverse 3D faces in a feed-forward manner within 5 seconds, without any optimization during inference. Moreover, our modular architecture enables 3D facial stylization without needing 3D labeled data. We have demonstrated and evaluated the effectiveness of our method in generating and editing a wide variety of digital characters, guided by multi-model controls such as text descriptions, a variable number of facial images, and style reference images.

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 (Debevec et al. 2000) or restricted parametric face models (Blanz and Vetter 1999; Paysan et al. 2009) to reconstruct the 3D face of real-world humans. Recent advancements in vision-language models (Radford et al. 2021; Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020) have led to progress in text-to-3D generation (Poole et al. 2022; Tang et al. 2023). 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 (Gecer et al. 2020; Li et al. 2020a) and Neural Radiance Fields (NeRF) (Wang et al. 2023a). Optimization-based methods incorporate 2D vision-language models to increase diversity and employ inference-time optimization techniques, typically based on CLIP (Aneja et al. 2023a; Wu et al. 2023; Hong et al. 2022) or Score Distillation Sampling (SDS) (Zhang et al. 2023a), to generate 3D facial assets.

Figure 1: ControlFace is capable of synthesizing various 4K PBR textures with micro-structure-level skin details, which are compatible with existing CG pipelines.

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 a general vision-language model. Therefore, the diversity of these methods is limited. It is natural to leverage the diversity of vision-language models to enhance the diversity of 3D generation, the real question is how to combine them effectively. 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 vision-language models 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 Figure 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 text-to-image diffusion models. To achieve this, we introduce a 3D-aware control module and use texture-space semantic maps (texture-space normal maps and landmarks) as our 3D-aware control signals to "control" the pre-trained, frozen text-to-image 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, without the need for any inference-time optimizations.

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 generative model takes text descriptions as input and is controlled by the 3D-aware priors in the UV texture space defined by the selected optimal geometry proxy. This enables the generation of fine-grained geometry details (displacement map) and physically-based rendering (PBR) appearance. ControlFace supports fine geometry control for synthetic textures including wrinkles, eyelids, and even topology. Additionally, our model supports stylized 3D face generation and image guidance by seamlessly transferring 2D generation techniques into the 3D domain without the need for 3D supervision and fine-tuning.

In summary, ControlFace makes significant advancements in the creation of 3D digital humans, enabling its practical use in a wide range of applications. Our main contributions can be summarized as follows:

- We introduce 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.
- We propose a controllable diffusion generative model that can generate $4K$ 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).
- ControlFace enables a wide range 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).

Related Work

General 3D Content Generation

Image generation has made significant advancements in the recent decade, including GANs (Goodfellow et al. 2014; Karras et al. 2018; Karras, Laine, and Aila 2019; Karras et al. 2020) and Diffusions (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Rombach et al. 2022), 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. Early works use CLIP to supervise the 3D generation (Mohammad Khalid et al. 2022; Michel et al. 2022). Inspired by the success of implicit representations such as NeRF (Mildenhall et al. 2021) and 3D Gaussian splatting (Kerbl et al. 2023), several works (Wang et al. 2023b; Tang et al. 2023; Chen et al. 2024; Yi et al. 2024) effectively combine them with pre-trained diffusion models, enabling 3D generation by using Score Distillation Sampling (SDS) (Poole et al. 2022), although it suffers from over-smoothing, oversaturation, 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 (Qiu et al. 2024; Shi et al. 2023b; Zeng et al. 2024; Voleti et al. 2024; Zhang et al. 2024b) 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 (Deitke et al. 2023; Liu et al. 2023; Shi et al. 2023a), the size of the 3D dataset is still smaller compared to the vast amount of data used in 2D large-scale model training. Therefore, creating CG-friendly assets remains a challenge.

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 avatar, including feed-forward by GANs (Chan et al. 2022; An et al. 2023) or DMs (Wang et al. 2023a; Shen et al. 2024) and optimization-based methods (Cao et al. 2023; Wang et al. 2023a; Hong et al. 2022; Zhang et al. 2024a; Han et al. 2024; Huang et al. 2024; Lei et al. 2024; Liu et al. 2024). 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 (Li et al. 2020b; Wu et al. 2023; Gruber et al. 2024). 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 Aneja et al. (2023b), it is still restricted by dataset size. Diffusion-based explicit avatar generation methods (Zhang et al. 2023a,c) 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. Liao et al. (2023) derive a high-resolution upsampled SMPL-X with displacement and texture. Based on the SMPL-X parameters model, they create 3D avatars from text through SDS optimization. Zhou et al. (2024) 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.

Method

ControlFace, as illustrated in Figure 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 selected base geometry into UV texture space maps, which serve 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 physically-based material maps $\{k_s, n, r\}$ from the albedo map through image-to-image translation techniques.

Geometry Proxy Generation

The goal of Geometry Proxy Generation is to produce a geometry mesh that satisfies the text description. Recent approaches (Zhang et al. 2023a) rely on the similarity between text and the rendered geometry image 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 corresponding facial features to resolve ambiguities in matching text and facial mesh.

Specifically, we propose to select the optimal geometry proxy g_b from a predefined candidates set $S = \{s_i\} = \{g_i, R_i, I_i, f_i\}$, consisting of the triangle mesh g_i, the rendering images $R_i(g_i, v_i, l_i)$ of g_i with camera v_i and lighting l_i, the reference portait image I_i, and facial attributes f_i. The selection process is driven by a matching loss, which is composed of three terms: 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 (Zhang et al. 2023a; Hong et al. 2022), $\bar{e_t}, \bar{e_i}$ represent the embedding of an anchor text (e.g. "a face of") and the renderings of the mean mesh $\bar{g_i}$. We compute the matching score between the input text and each candidate and then choose the candidate with the highest similarity as our geometry proxy g_i^*.

Candidate Pool We construct the candidate pool with 8,425 samples S using the CelebA-HQ dataset (Liu et al. 2015). Specifically, for each image I_i in CelebA-HQ, we utilize a single-image 3D face reconstruction method (Chai et al. 2022) to obtain the corresponding 3D face shape.

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 a generative network controlled by the underlying geometric topology. Inspired by ControlNet (Zhang, Rao, and Agrawala 2023), we propose 3D-aware controllable albedo diffusion, which takes textual descriptions c_t as input and utilizes a texture space facial prior $c = \{c_n, c_l\}$ 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 UV texture space maps, including a sparse representation (facial landmarks c_l) and a dense representation (geometry normal map c_n). Unlike general image generation, facial texture generation requires precise topological alignment to ensure that facial features are located in specific regions within the UV texture space. Our facial landmark control provides strong constraints on the facial semantic regions, while 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 Zhang, Rao, and Agrawala (2023). 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

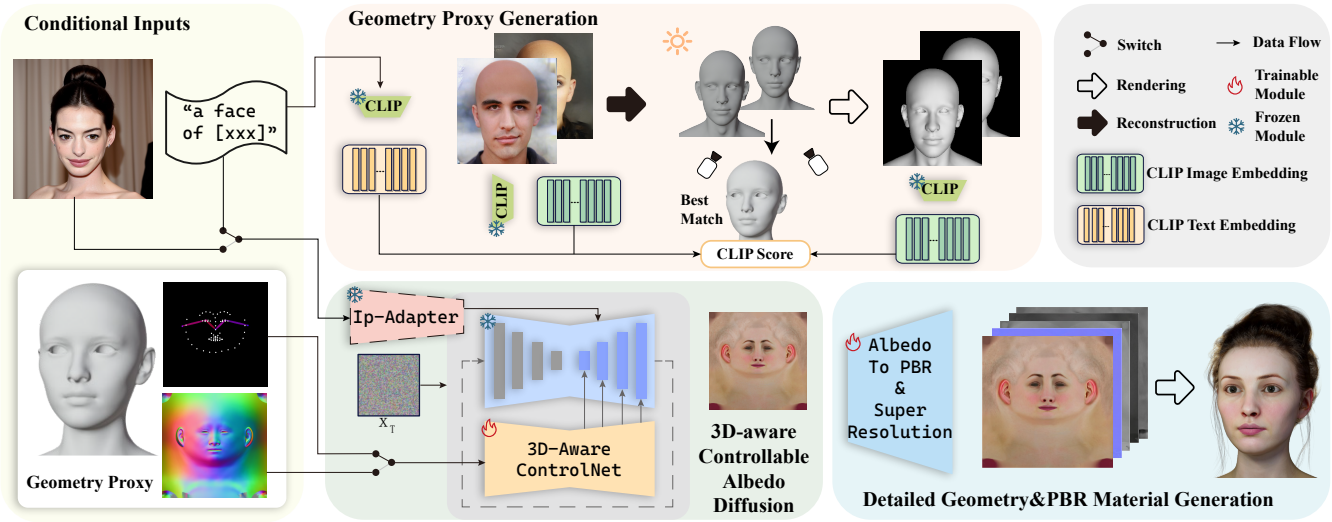

Figure 2: Overview of our proposed ControlFace. Given a text description ("a face of [xxx]"), ControlFace can select the best matching geometry proxy by the Geometry Proxy Generation module. Then, the 3D-aware Controllable Albedo Diffusion generates correlated albedo with the text guidance or optional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

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:

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

where t represents the time step, c_t is the text prompt, c_f is the latent feature of our 3D-aware condition c, ϵ_θ 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 T2I-adapter (Mou et al. 2023) and utilize a cubic function to implement non-uniform time step sampling during training:

$$t = (1 - (\frac{t}{T})^3) \times T \quad t \in U(0, T) \quad (3)$$

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.

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) and material maps ($\{k_s, n, r\}$) directly from the input diffuse albedo map (k_d).

Our network architecture is based on a UNet model, comprising an encoder and a decoder, together with an additional middle block. Within the encoder and decoder, there are four blocks. Each of these blocks consists of four convolution layers with residual blocks, dedicated to $2x$ upsampling or downsampling. The middle block includes four convolutional layers and two Vision Transformer blocks with self-attention mechanisms. The training loss is the sum of three terms:

$$\mathcal{L} = \lambda_{map}\mathcal{L}_{map} + \lambda_p\mathcal{L}_p + \lambda_{gan}\mathcal{L}_{gan}, \quad (4)$$

where $\mathcal{L}_{map}$ is the L_1 loss on the reconstructed maps, $\mathcal{L}_p$ is a perceptual loss based on LPIPS (Zhang et al. 2018), and $\mathcal{L}_{gan}$ is the Patch GAN loss (Li and Wand 2016). In our setting, we set $\lambda_r = 2$, $\lambda_p = 0.5$, $\lambda_g = 0.5$.

Texture Augmentation 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.

Results

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 database (3DScanStore 2023) and four manually created examples featuring accessories such as face masks and eye covers. Further data augmentation techniques are detailed in the supplementary material.

Figure 3: Text-guidance generation results of ControlFace. 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.

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.

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. The related results are shown in Figure 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 Figure 4a. 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 Figure 4b, we generate the texture on the other different topologies including Vface (3DScanStore 2023), Flame (Li et al. 2017), HIFI3D (Bao et al. 2021), and MetaHuman (Engine 2023), while ControlFace exclusively trained on our own facial topology(Master Model). We also verify the effectiveness of topology control by landmarks. Trained on the Master Model, ControlFace shows good fitness as shown in Figure 4c. However, as the traditional 68 landmarks can not provide full coverage of the face, this may result in misalignment, especially around the nostrils and eye contours.

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 Figure 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 Figure 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.

	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

Table 1: Quantitative comparison of image guidance generation. We compared the similarity of FFHQ with and without image guidance. 1-cosine, l1, and l2 are calculated based on DeepFace and CLIP Score measures the similarity in clip space of images.

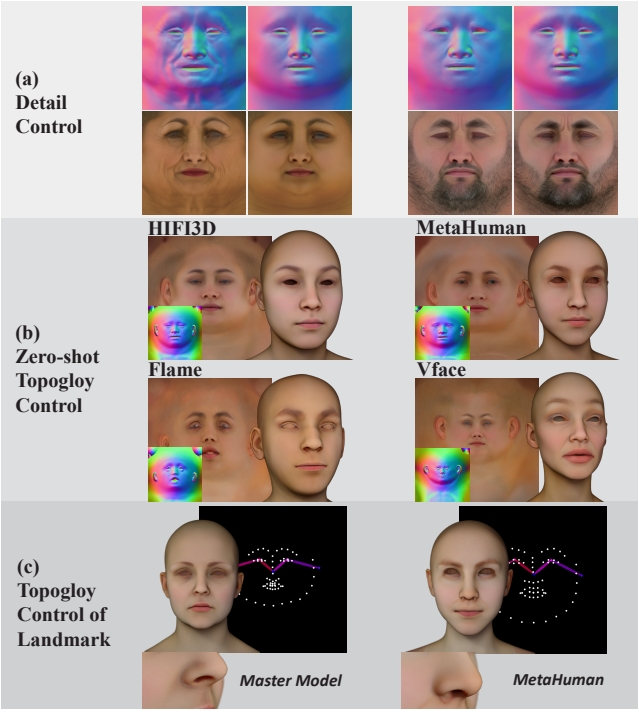

Figure 4: Examples generated under geometry proxy guidance. When we fix the input prompts and random seed, the 3D aware conditional control module can generate geometry-related textures, such as wrinkles(left) and eyelids(right) in Fig (a). ControlFace is able to control the topology by giving different geometry normals in Fig (b). While the topology control of landmarks degrades when tested on out-of-domain topology (MetaHuman) in Fig (c).

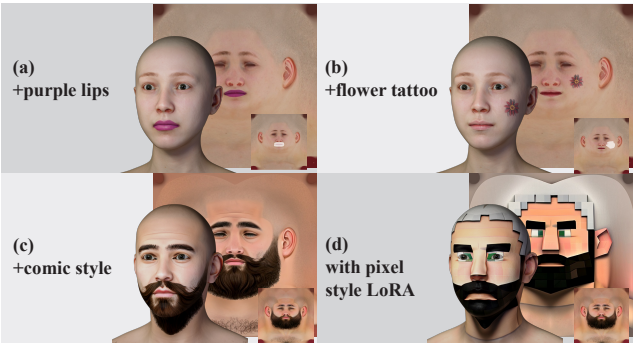

Figure 5: Local manipulation and style transformation. Fig (a) and (b) showcase the local manipulation of lip color and the addition of a flower tattoo with text guidance through hand-made masks. In Fig (c), ControlFace converts the digital human into a comic style by adding the prompt "comic style". Fig (d) showcases personalized stylized transformation in the style of "Minecraft" using the pixel style LoRA.

Image Guidance Generation With IP-Adapter(Ye et al. 2023), ControlFace is capable of image guidance to provide precise control over avatars' identity-specific styles with an image prompt. As shown in Figure 6a, we achieved makeup imitation on avatars with the reference images. Also, as shown in Figure 6b, ControlFace can reproduce avatars with similar characteristics to the input 2D image, such as "David made of plaster". Moreover, to verify the effectiveness of image guidance compared with text guidance, we evaluate the semantic similarity of CLIP (Radford et al. 2021) and the facial similarity of DeepFace (Serengil and Ozpinar 2024) on FFHQ (Karras, Laine, and Aila 2019). Specifically, we sampled 1000 images from FFHQ and utilized CLIP Interrogator [1] to collect their corresponding text prompts. With those prompts, 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 1 shows that the image guidance module enables more consistent features than text-only guidance.

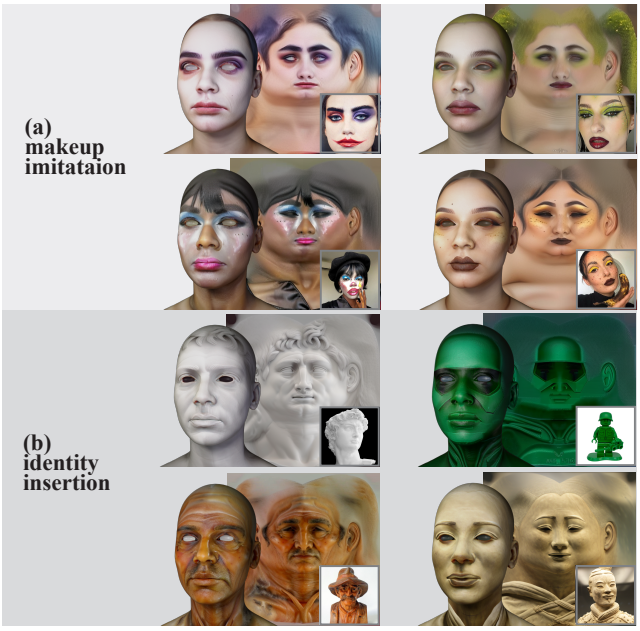

Figure 6: Results of Image Guidance. ControlFace is capable of facilitating digital makeup imitation and identity injection with image guidance (bottom-left). Fig (a) depicts "a face with makeup": Through image guidance, an avatar with a similar makeup style is generated. Fig (b) shows "a face": By inputting images of faces made of different materials, avatars resembling the features of the material are generated.

	CLIP Score ↑	Inference Time ↓
ClipFace	0.262	∼ 30 mins
AvatarClip	0.284	∼ 5 hours
ChatAvatar	0.282	∼ 30 sec
Ours	**0.312**	**∼ 5 sec**

Table 2: Comparison of ControlFace with other digital human generation methods in terms of CLIP Score and Inference Time. The best results are in **bold**.

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

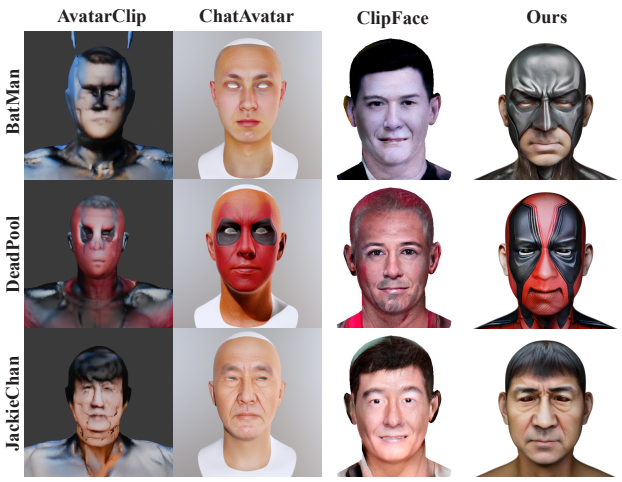

Figure 7: Comparison of generation quality. Our approach is capable of generating more diverse results which exhibit a higher resemblance to textual descriptions.

Comparison

As demonstrated in Figure 7, we compare ControlFace with the other 3D digital human generation methods, including AvatarClip (Hong et al. 2022), ClipFace (Aneja et al. 2023b) and CharAvatar (Zhang et al. 2023d) (an online, fast version of DreamFace (Zhang et al. 2023b)), in terms of Clip Score and inference time. Following DreamFace, we generated 10 different characters, including people generated by general descriptions, celebrities, and film characters. For ClipFace and AvatarClip, we use the official implementation. For CharAvatar, we generate the 3D asserts and obtain their rendering results from the official website. All prompts used for generation followed the same anchor: "a face of xxx". We use "ViT-L/14" as the pre-trained CLIP model.

The quantitative comparison results (see Table 2) indicate that our proposed ControlFace outperforms other methods in terms of clip score and inference time. As a result, ControlFace produces results that exhibit a higher resemblance to text descriptions with less inference time.

Figure 8: The catastrophic forgetting of the finetuned SD modelg. 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.

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 Figure 8, the Stable Diffusion model, fine-tuned with a dataset of merely over 30 samples, exhibits significant overfitting and fails to generate fictional characters such as Deadpool and Joker. On the other hand, refraining from overriding the pre-trained parameters of Stable Diffusion, ControlFace maintains the model's generation and generalization abilities.

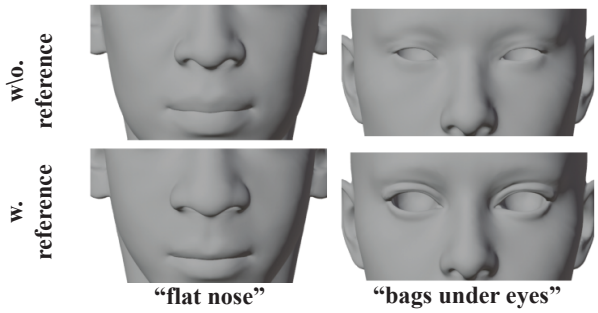

Figure 9: Comparation of improved geometry selection. When matching certain descriptions (e.g., "flat nose" and "bags under eyes") the method without reference of image guidance fails to select the best matching result.

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 Figure 9, when matching some description of facial details, the method without image guidance fails to select the best matching result. In contrast, the method with additional guidance obtains more accurate matching results. It demonstrates the effectiveness of our proposed method.

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

During the texture generation process, we did not separate accessories such as hair, clothing, glasses, etc., from the texture map, distorting the appearance of accessories affected by geometric curvature. In the future, we may construct a large-scale accessory library and utilize the clip search algorithm to generate accessories.

References

3DScanStore. 2023. 3D Scan Store. https://www.3dscanstore.com/.

An, S.; Xu, H.; Shi, Y.; Song, G.; Ogras, U. Y.; and Luo, L. 2023. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20950–20959.

Aneja, S.; Thies, J.; Dai, A.; and Niessner, M. 2023a. ClipFace: Text-Guided Editing of Textured 3D Morphable Models. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, 1–11. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701597.

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

Bao, L.; Lin, X.; Chen, Y.; Zhang, H.; Wang, S.; Zhe, X.; Kang, D.; Huang, H.; Jiang, X.; Wang, J.; et al. 2021. High-fidelity 3d digital human head creation from rgb-d selfies. *ACM Transactions on Graphics (TOG)*, 41(1): 1–21.

Blanz, V.; and Vetter, T. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, 187–194. USA: ACM Press/Addison-Wesley Publishing Co. ISBN 0201485605.

Cao, Y.; Cao, Y.-P.; Han, K.; Shan, Y.; and Wong, K.-Y. K. 2023. DreamAvatar: Text-and-Shape Guided 3D Human Avatar Generation via Diffusion Models. arXiv:2304.00916.

Chai, Z.; Zhang, H.; Ren, J.; Kang, D.; Xu, Z.; Zhe, X.; Yuan, C.; and Bao, L. 2022. REALY: Rethinking the Evaluation of 3D Face Reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; Karras, T.; and Wetzstein, G. 2022. Efficient Geometry-Aware 3D Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16123–16133.

Chen, Y.; Zhang, C.; Yang, X.; Cai, Z.; Yu, G.; Yang, L.; and Lin, G. 2024. It3d: Improved text-to-3d generation with explicit view synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1237–1244.

Debevec, P.; Hawkins, T.; Tchou, C.; Duiker, H.-P.; Sarokin, W.; and Sagar, M. 2000. Acquiring the Reflectance Field of a Human Face. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, 145–156. USA: ACM Press/Addison-Wesley Publishing Co. ISBN 978-1-58113-208-3.

Deitke, M.; Liu, R.; Wallingford, M.; Ngo, H.; Michel, O.; Kusupati, A.; Fan, A.; Laforte, C.; Voleti, V.; Gadre, S. Y.; VanderBilt, E.; Kembhavi, A.; Vondrick, C.; Gkioxari, G.; Ehsani, K.; Schmidt, L.; and Farhadi, A. 2023. Objaverse-XL: A Universe of 10M+ 3D Objects. arXiv:2307.05663.

Engine, U. 2023. MetaHuman. https://www.unrealengine.com/en-US/metahuman.

Gecer, B.; Lattas, A.; Ploumpis, S.; Deng, J.; Papaioannou, A.; Moschoglou, S.; and Zafeiriou, S. 2020. Synthesizing Coupled 3D Face Modalities by Trunk-Branch Generative Adversarial Networks. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, 415–433. Cham: Springer International Publishing. ISBN 978-3-030-58526-6.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Networks.

Gruber, A.; Collins, E.; Meka, A.; Mueller, F.; Sarkar, K.; Orts-Escolano, S.; Prasso, L.; Busch, J.; Gross, M.; and Beeler, T. 2024. GANtlitz: Ultra High Resolution Generative Model for Multi-Modal Face Textures. In *Computer Graphics Forum*, volume 43, e15039. Wiley Online Library.

Han, X.; Cao, Y.; Han, K.; Zhu, X.; Deng, J.; Song, Y.-Z.; Xiang, T.; and Wong, K.-Y. K. 2024. Headsculpt: Crafting 3d head avatars with text. *Advances in Neural Information Processing Systems*, 36.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239.

Hong, F.; Zhang, M.; Pan, L.; Cai, Z.; Yang, L.; and Liu, Z. 2022. AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars. *ACM Transactions on Graphics (TOG)*, 41(4): 1–19.

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.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. arXiv:1710.10196.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proc. CVPR*.

Kerbl, B.; Kopanas, G.; Leimkuehler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4): 139:1–139:14.

Lei, B.; Yu, K.; Feng, M.; Cui, M.; and Xie, X. 2024. DiffusionGAN3D: Boosting Text-guided 3D Generation and Domain Adaptation by Combining 3D GANs and Diffusion Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10487–10497.

Li, C.; and Wand, M. 2016. Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks. arXiv:1604.04382.

Li, R.; Bladin, K.; Zhao, Y.; Chinara, C.; Ingraham, O.; Xiang, P.; Ren, X.; Prasad, P.; Kishore, B.; Xing, J.; and Li, H. 2020a. Learning Formation of Physically-Based Face Attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3410–3419.

Li, R.; Bladin, K.; Zhao, Y.; Chinara, C.; Ingraham, O.; Xiang, P.; Ren, X.; Prasad, P.; Kishore, B.; Xing, J.; et al.

2020b. Learning formation of physically-based face attributes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3410–3419.

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 Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6): 194:1–194:17.

Liao, T.; Yi, H.; Xiu, Y.; Tang, J.; Huang, Y.; Thies, J.; and Black, M. J. 2023. Tada! text to animatable digital avatars. *arXiv preprint arXiv:2308.10899*.

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.

Liu, R.; Wu, R.; Hoorick, B. V.; Tokmakov, P.; Zakharov, S.; and Vondrick, C. 2023. Zero-1-to-3: Zero-shot One Image to 3D Object. arXiv:2303.11328.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Michel, O.; Bar-On, R.; Liu, R.; Benaim, S.; and Hanocka, R. 2022. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13492–13502.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Communications of the ACM*, 65(1): 99–106.

Mohammad Khalid, N.; Xie, T.; Belilovsky, E.; and Popa, T. 2022. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 conference papers*, 1–8.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; Shan, Y.; and Qie, X. 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*.

Paysan, P.; Knothe, R.; Amberg, B.; Romdhani, S.; and Vetter, T. 2009. A 3D Face Model for Pose and Illumination Invariant Face Recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, 296–301.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. DreamFusion: Text-to-3D Using 2D Diffusion. https://arxiv.org/abs/2209.14988v1.

Qiu, L.; Chen, G.; Gu, X.; Zuo, Q.; Xu, M.; Wu, Y.; Yuan, W.; Dong, Z.; Bo, L.; and Han, X. 2024. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9914–9925.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Serengil, S.; and Ozpinar, A. 2024. A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules. *Journal of Information Technologies*, 17(2): 95–107.

Shen, X.; Ma, J.; Zhou, C.; and Yang, Z. 2024. Controllable 3d face generation with conditional style code diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4811–4819.

Shi, R.; Chen, H.; Zhang, Z.; Liu, M.; Xu, C.; Wei, X.; Chen, L.; Zeng, C.; and Su, H. 2023a. Zero123++: a Single Image to Consistent Multi-view Diffusion Base Model. arXiv:2310.15110.

Shi, Y.; Wang, P.; Ye, J.; Long, M.; Li, K.; and Yang, X. 2023b. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Tang, J.; Ren, J.; Zhou, H.; Liu, Z.; and Zeng, G. 2023. DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation. *arXiv preprint arXiv:2309.16653*.

Voleti, V.; Yao, C.-H.; Boss, M.; Letts, A.; Pankratz, D.; Tochilkin, D.; Laforte, C.; Rombach, R.; and Jampani, V. 2024. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*.

Wang, T.; Zhang, B.; Zhang, T.; Gu, S.; Bao, J.; Baltrusaitis, T.; Shen, J.; Chen, D.; Wen, F.; Chen, Q.; and Guo, B. 2023a. 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.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2023b. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. *arXiv preprint arXiv:2305.16213*.

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.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models.

Yi, T.; Fang, J.; Wang, J.; Wu, G.; Xie, L.; Zhang, X.; Liu, W.; Tian, Q.; and Wang, X. 2024. 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*, 6796–6807.

Zeng, X.; Chen, X.; Qi, Z.; Liu, W.; Zhao, Z.; Wang, Z.; Fu, B.; Liu, Y.; and Yu, G. 2024. Paint3d: Paint anything

3d with lighting-less texture diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4252–4262.

Zhang, H.; Chen, B.; Yang, H.; Qu, L.; Wang, X.; Chen, L.; Long, C.; Zhu, F.; Du, D.; and Zheng, M. 2024a. Avatarverse: High-quality & stable 3d avatar creation from text and pose. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7124–7132.

Zhang, L.; Qiu, Q.; Lin, H.; Zhang, Q.; Shi, C.; Yang, W.; Shi, Y.; Yang, S.; Xu, L.; and Yu, J. 2023a. DreamFace: Progressive Generation of Animatable 3D Faces under Text Guidance. *ACM Transactions on Graphics*, 42(4): 138:1–138:16.

Zhang, L.; Qiu, Q.; Lin, H.; Zhang, Q.; Shi, C.; Yang, W.; Shi, Y.; Yang, S.; Xu, L.; and Yu, J. 2023b. DreamFace: Progressive Generation of Animatable 3D Faces under Text Guidance. *arXiv preprint arXiv:2304.03117*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models.

Zhang, L.; Wang, Z.; Zhang, Q.; Qiu, Q.; Pang, A.; Jiang, H.; Yang, W.; Xu, L.; and Yu, J. 2024b. CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets. *ACM Transactions on Graphics (TOG)*, 43(4): 1–20.

Zhang, Q.; Zhang, L.; Xu, L.; Wu, D.; and Yu, J. 2023c. ChatAvatar: Creating Hyper-realistic Physically-based 3D Facial Assets through AI-Driven Conversations. In *ACM SIGGRAPH 2023 Real-Time Live!*, SIGGRAPH '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701580.

Zhang, Q.; Zhang, L.; Xu, L.; Wu, D.; and Yu, J. 2023d. ChatAvatar: Creating Hyper-realistic Physically-based 3D Facial Assets through AI-Driven Conversations. In *ACM SIGGRAPH 2023 Real-Time Live!*, SIGGRAPH '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701580.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Zhou, M.; Hyder, R.; Xuan, Z.; and Qi, G. 2024. 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*, 1238–1248.

Reproducibility Checklist

This paper:

- Includes a conceptual outline and/or pseudocode description of AI methods introduced **yes**
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results **yes**
- Provides well marked pedagogical references for less-familiare readers to gain background necessary to replicate the paper **yes**

Does this paper make theoretical contributions? **no**

This paper mainly contributes to application in CG-friendly 3D face field.

Does this paper rely on one or more datasets? **yes**

If yes, please complete the list below.

- A motivation is given for why the experiments are conducted on the selected datasets **yes**
- All novel datasets introduced in this paper are included in a data appendix. **NA**

  No novel datasets
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. **NA**

  No novel datasets
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. **yes**
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available.**no**

  Because high-quality 3D faces are scarce, the 36 faces we used for few-shot training are from commercial "https://www.3dscanstore.com/".
- All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. **yes**

Does this paper include computational experiments? **no**

This paper mainly focuses on qualitative comparisons. The CLIP Score and DeepFace are used as quantitative metrics to evaluate the similarity, and they are publicly available.