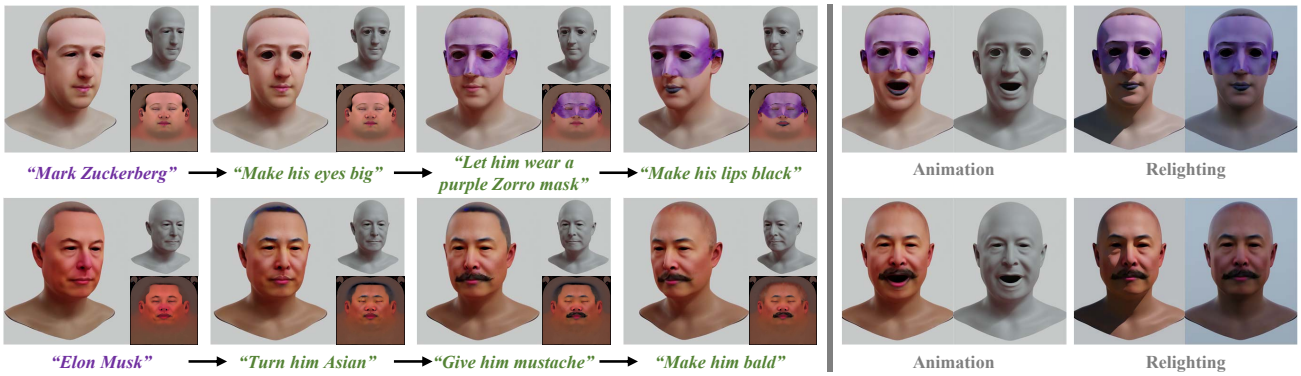
Text-Guided 3D Face Synthesis - From Generation to Editing

Yunjie Wu[†1], Yapeng Meng[†1,2], Zhipeng Hu[†1], Lincheng Li[*1]
Haoqian Wu[1], Kun Zhou[3], Weiwei Xu[3], Xin Yu[4]
[1]Netease Fuxi AI Lab [2]Tsinghua University [3]Zhejiang University [4]University of Queensland



(a) High-fidelity generation and flexible editing (b) Semlessly integrated to existing CG pipeline

Figure 1. (a) Our approach enables the high-fidelity generation and flexible editing of 3D faces from textual input. It facilitates sequential editing for creating customized details in 3D faces. (b) The produced 3D faces can be seamlessly integrated into existing CG pipelines.

Abstract

Text-guided 3D face synthesis has achieved remarkable results by leveraging text-to-image (T2I) diffusion models. However, most existing works focus solely on the direct generation, ignoring the editing, restricting them from synthesizing customized 3D faces through iterative adjustments. In this paper, we propose a unified text-guided framework from face generation to editing. In the generation stage, we propose a geometry-texture decoupled generation to mitigate the loss of geometric details caused by coupling. Besides, decoupling enables us to utilize the generated geometry as a condition for texture generation, yielding highly geometry-texture aligned results. We further employ a fine-tuned texture diffusion model to enhance texture quality in both RGB and YUV space. In the editing stage, we first employ a pre-trained diffusion model to update facial geometry or texture based on the texts. To enable sequential editing, we introduce a UV domain consistency preservation regularization, preventing unintentional changes to irrelevant facial attributes. Besides, we propose a self-guided consistency weight strategy to improve editing efficacy while preserving consistency. Through comprehensive experiments, we showcase our method's superiority in face synthesis. Project page: https://faceg2e.github.io/.

[†]Equal contribution
[*]Corresponding author

1. Introduction

Modeling 3D faces serves as a fundamental pillar for various emerging applications such as film making, video games, and AR/VR. Traditionally, the creation of detailed and intricate 3D human faces requires extensive time from highly skilled artists. With the development of deep learning, existing works [7, 9, 46, 55] attempted to produce 3D faces from photos or videos with generative models. However, the diversity of the generation remains constrained primarily due to the limited scale of training data. Fortunately, recent large-scale vision-language models (e.g., CLIP [32], Stable Diffusion [34]) pave the way for generating diverse 3D content. Through the integration of these models, numerous text-to-3D works [22, 27, 28, 49, 51] can create 3D content in a zero-shot manner.

Many studies have been conducted on text-to-3D face synthesis. They either utilize CLIP or employ score distillation sampling (SDS) on text-to-image (T2I) models to guide the 3D face synthesis. Some methods [45, 52] employ neural fields to generate visually appealing but low-quality geometric 3D faces. Recently, Dreamface [53] has demonstrated the potential for generating high-quality 3D face textures by leveraging SDS on facial textures, but their geometry is not fidelitous enough and they overlooked the subsequent face editing. A few works [1, 11, 26] enable text-guided face editing, allowing coarse-grained editing (e.g. overall style), but not fine-grained adjustments (e.g., lips

color). Besides, the lack of design in precise editing control leads to unintended changes in their editing, preventing the synthesis of customized faces through sequential editing.

To address the aforementioned challenges, we present text-guided 3D face synthesis - from generation to editing, dubbed **FaceG2E**. We propose a progressive framework to generate the facial geometry and textures, and then perform accurate face editing sequentially controlled by text. To the best of our knowledge, this is the first attempt to edit a 3D face in a sequential manner. We propose two core components: (1) Geometry-texture decoupled generation and (2) Self-guided consistency preserved editing.

To be specific, our proposed ***Geometry-texture decoupled generation*** generates the facial geometry and texture in two separate phases. By incorporating texture-less rendering in conjunction with SDS, we induce the T2I model to provide geometric-related priors, inciting details (e.g., wrinkles, lip shape) in the generated geometry. Building upon the generated geometry, we leverage ControlNet to force the SDS to be aware of the geometry, ensuring precise geometry-texture alignment. Additionally, we fine-tune a texture diffusion model that incorporates both RGB and YUV color spaces to compute SDS in the texture domain, enhancing the quality of the generated textures.

The newly developed ***Self-guided consistency preserved editing*** enables one to follow the texts, performing efficient editing in specific facial attributes without causing other unintended changes. Here, we first employ a pre-trained image-edit diffusion model to update the facial geometry or texture. Then we introduce a UV domain consistency preservation regularization to prevent unexpected changes in faces, enabling sequential editing. To avoid the degradation of editing effects caused by the regularization, we further propose a self-guided consistency weighting strategy. It adaptively determines the regularization weight for each facial region by projecting the cross-attention scores of the T2I model to the UV domain. As shown in Fig. 1, our method can generate high-fidelity 3D facial geometry and textures while allowing fine-grained face editing. With the proposed components, we achieve better visual and quantitative results compared to other SOTA methods, as demonstrated in Sec. 4. In summary, our contributions are:

- We propose FaceG2E, facilitating a full pipeline of text-guided 3D face synthesis, from generation to editing. User surveys confirm that our synthesized 3D faces are significantly preferable than other SOTA methods.
- We propose the geometry-texture decoupled generation, producing faces with high-fidelity geometry and texture.
- We design the self-guided consistency preservation, enabling the accurate editing of 3D faces. Leveraging precise editing control, our method showcases some novel editing applications, such as sequential and geometry-texture separate editing.

2. Related Work

Text-to-Image generation. Recent advancements in visual-language models [32] and diffusion models [8, 13, 42] have greatly improved text-to-image generation [3, 33, 34, 37]. These methods, trained on large-scale image-text datasets [40, 41], can synthesize realistic and complex images from text descriptions. Subsequent studies have made further efforts to introduce additional generation process controls [16, 48, 54], fine-tuning the pre-trained models for specific scenarios [10, 15, 35], and enabling image editing capabilities [5, 12, 23]. However, generating high-quality and faithful 3D assets, such as 3D human faces, from textual input still poses an open and challenging problem.

Text-to-3D generation. With the success of text-to-image generation in recent years, text-to-3D generation has attracted significant attention from the community. Early approaches [14, 20, 30, 38, 50] utilize mesh or implicit neural fields to represent 3D content, and optimized the CLIP metrics between the 2D rendering and text prompts. However, the quality of generated 3D contents is relatively low.

Recently, DreamFusion [31] has achieved impressive results by using a score distillation sampling (SDS) within the powerful text-to-image diffusion model [37]. Subsequent works further enhance DreamFusion by reducing generation time [27], improving surface material representation [6], and introducing refined sampling strategies [18]. However, the text-guided generation of high-fidelity and intricate 3D faces remains challenging. Building upon DreamFusion, we carefully design the form of score distillation by exploiting various diffusion models at each stage, resulting in high-fidelity and editable 3D faces.

Text-to-3D face synthesis. Recently, there have been attempts to generate 3D faces from text. Describe3D [47] and Rodin [45] propose to learn the mapping from text to 3D faces on pairs of text-face data. They solely employ the mapping network trained on appearance descriptions to generate faces, and thus fail to generalize to out-of-domain texts (e.g., celebrities or characters). On the contrary, our method can generalize well to these texts and synthesize various 3D faces.

Other works [11, 17, 21, 26, 53] employ SDS on the pre-trained T2I models. Dreamface [53] utilizes CLIP to select facial geometry from candidates. Then they perform the SDS with a texture diffusion network to generate facial textures. Headsculpt [11] employs Stable Diffusion [34] and InstructPix2Pix [5] for computing the SDS, and relies on the mixture of SDS gradients for constraining the editing process. These approaches can perform not only generation but also simple editing. However, they still lack the design in precise editing control, and unintended changes in the editing results often occur. This prevents them from synthesizing highly customized 3D faces via sequential editing. On the contrary, our approach facilitates accurate editing of

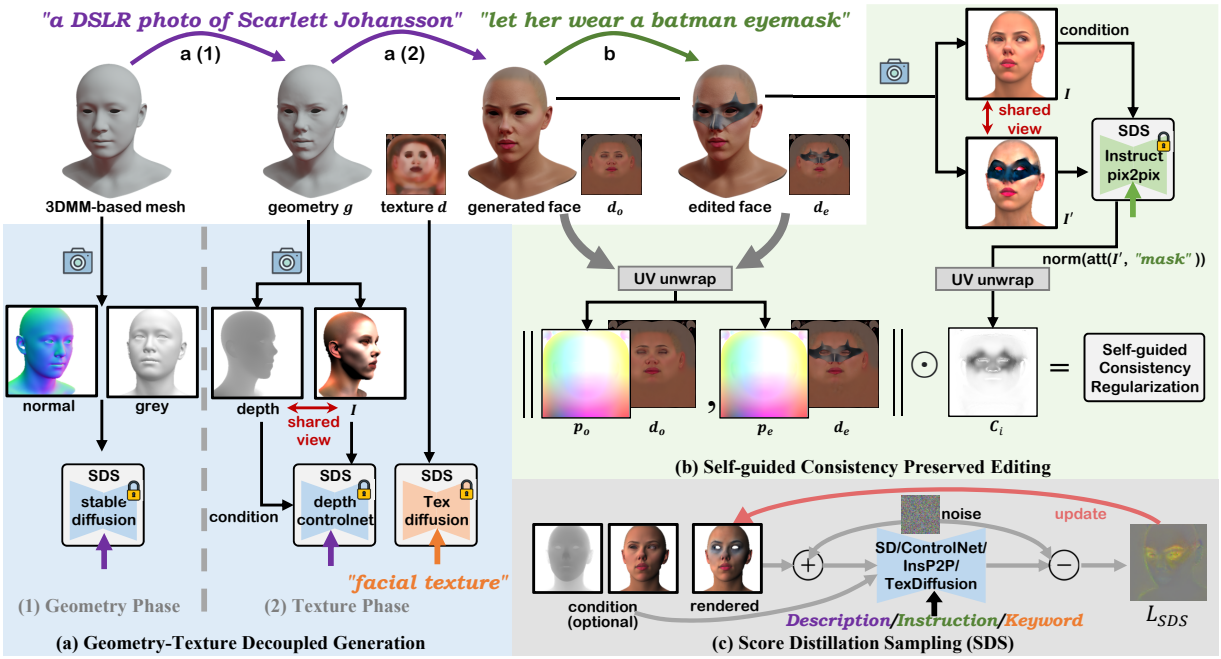


Figure 2. Overview of FaceG2E. (a) Geometry-texture decoupled generation, including a geometry phase and a texture phase. (b) Self-guided consistency preserved editing, in which we utilize the built-in cross-attention to obtain the editing-relevant regions and unwrap them to UV space. Then we penalize inconsistencies in the irrelevant regions. (c) Our method exploits multiple score distillation sampling.

3D faces, supporting sequential editing.

3. Methodology

FaceG2E is a progressive text-to-3D approach that first generates a high-fidelity 3D human face and then performs fine-grained face editing. As illustrated in Fig. 2, our method has two main stages: (a) Geometry-texture decoupled generation, and (b) Self-guided consistency preserved editing. In Sec. 3.1, we introduce some preliminaries that form the fundamental basis of our approach. In Sec. 3.2 and Sec. 3.3, we present the generation and editing stages.

3.1. Preliminaries

Score distillation sampling has been proposed in DreamFusion [31] for text-to-3D generation. It utilizes a pre-trained 2D diffusion model ϕ with a denoising function $\epsilon_\phi(z_t; y, t)$ to optimize 3D parameters θ. SDS renders an image $I = R(\theta)$ and embeds I with an encoder $\mathcal{E}(\cdot)$, achieving image latent z. Then it injects a noise ϵ into z, resulting in a noisy latent code z_t. It takes the difference between the predicted and added noise as the gradient:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(I) = \mathbb{E}_{t,\epsilon}\left[w(t)\left(\epsilon_\phi(z_t; y, t) - \epsilon\right)\frac{\partial z}{\partial I}\frac{\partial I}{\partial \theta}\right], \quad (1)$$

where $w(t)$ is a time-dependent weight function and y is the embedding of input text.

Facial Geometry and Texture is represented with parameters $\theta = (\beta, u)$ in FaceG2E. β denotes the identity coefficient from the parametric 3D face model HIFI3D [4], and u denotes a image latent code for facial texture. The geometry g can be achieved by the blendshape function $\mathbf{M}(\cdot)$:

$$g = \mathbf{M}(\beta) = T + \sum_i \beta_i \mathrm{S}_i, \quad (2)$$

where T is the mean face and S is the vertices offset basis. As to the texture, the facial texture map d is synthesized with a decoder: $d = \mathcal{D}(u)$. We take the decoder from VAE of Stable Diffusion [34] as $\mathcal{D}(\cdot)$.

3.2. Geometry-Texture Decoupled Generation

The first stage of FaceG2E is the geometry-texture decoupled generation, which generates facial geometry and texture from the textual input. Many existing works have attempted to generate geometry and texture simultaneously in a single optimization process, while we instead decouple the generation into two distinct phases: the geometry phase and the texture phase. The decoupling provides two advantages: 1) It helps enhance geometric details in the generated faces. 2) It improves geometry-texture alignment by exploiting the generated geometry to guide the texture generation.

Geometry Phase. An ideal generated geometry should possess both high quality (e.g., no surface distortions) and a

good alignment with the input text. The employed facial 3D morphable model provides strong priors to ensure the quality of generated geometry. As to the alignment with the input text, we utilize SDS on the network ϕ_{sd} of Stable Diffusion [34] to guide the geometry generation.

Previous works [21, 26, 52] optimize geometry and texture simultaneously. We observe this could lead to the loss of geometric details, as certain geometric information may be encapsulated within the texture representation. Therefore, we aim to enhance the SDS to provide more geometry-centric information in the geometry phase. To this end, we render the geometry g with texture-less rendering $\tilde{I} = \tilde{R}(g)$, e.g., surface normal shading or diffuse shading with constant grey color. The texture-less shading attributes all image details solely to geometry, thereby allowing the SDS to focus on geometry-centric information. The ***geometry-centric SDS*** loss is defined as:

$$\nabla_{\beta}\mathcal{L}_{\text{geo}} = \mathbb{E}_{t,\epsilon}\left[w(t)\left(\epsilon_{\phi_{sd}}\left(z_t; y, t\right) - \epsilon\right)\frac{\partial z_t}{\partial \tilde{I}}\frac{\partial \tilde{I}}{\partial g}\frac{\partial g}{\partial \beta}\right]. \quad (3)$$

Texture Phase. Many works [26, 53] demonstrate that texture can be generated by minimizing the SDS loss. However, directly optimizing the standard SDS loss could lead to geometry-texture misalignment issues, as shown in Fig .9. To address this problem, we propose the ***geometry-aware texture content SDS*** (GaSDS). We resort to the ControlNet [54] to endow the SDS with awareness of generated geometry, thereby inducing it to uphold geometry-texture alignment. Specifically, we render g into a depth map e. Then we equip the SDS with the depth-ControlNet ϕ_{dc}, and take e as a condition, formulating the GaSDS:

$$\nabla_{u}\mathcal{L}^{ga}_{\text{tex}} = \mathbb{E}_{t,\epsilon}\left[w(t)\left(\epsilon_{\phi_{dc}}\left(z_t; e, y, t\right) - \epsilon\right)\frac{\partial z_t}{\partial I}\frac{\partial I}{\partial d}\frac{\partial d}{\partial u}\right]. \quad (4)$$

With the proposed GaSDS, the issue of geometric misalignment is addressed. However, artifacts such as local color distortion or uneven brightness persist in the textures. This is because the T2I model lacks priors of textures, which hinders the synthesis of high-quality texture details.

Hence we propose ***texture prior SDS*** to introduce such priors of textures. Inspired by DreamFace [53], we train a diffusion model ϕ_{td1} on texture data to estimate the texture distribution for providing the prior. Our training dataset contains 500 textures, including processed scanning data and selected synthesized data [2]. Different from DreamFace, which uses labeled text in training, we employ a fixed text keyword (e.g., '*facial texture*') for all textures. Because the objective of ϕ_{td1} is to model the distribution of textures as a prior, the texture-text alignment is not necessary. We additionally train another ϕ_{td2} on the YUV color spaces to promote uniform brightness, as shown in Fig 3. We fine-tune both ϕ_{td1} and ϕ_{td2} on Stable Diffusion. The texture

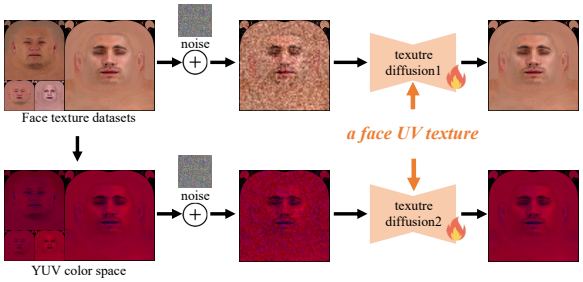


Figure 3. Training the texture diffusion model is performed on the collected facial textures in both RGB and YUV color space.

prior SDS is formulated with the trained ϕ_{td1} and ϕ_{td2} as:

$$\begin{aligned}
\nabla_{u}\mathcal{L}^{pr}_{\text{tex}} &= \mathcal{L}^{rgb}_{\text{tex}} + \lambda_{yuv}\mathcal{L}^{yuv}_{\text{tex}}, \\
\mathrm{L}^{rgb}_{\text{tex}} &= \mathbb{E}_{t,\epsilon}\left[w(t)\left(\epsilon_{\phi_{td1}}\left(z^{d}_t; y^*, t\right) - \epsilon\right)\frac{\partial z^{d}_t}{\partial d}\frac{\partial d}{\partial u}\right], \\
\mathcal{L}^{yuv}_{\text{tex}} &= \mathbb{E}_{t,\epsilon}\left[w(t)\left(\epsilon_{\phi_{td2}}\left(z^{d'}_t; y^*, t\right) - \epsilon\right)\frac{\partial z^{d'}_t}{\partial d}\frac{\partial d}{\partial u}\right],
\end{aligned} \quad (5)$$

where z^{d}_t and $z^{d'}_t$ denote the noisy latent codes of the texture d and the converted YUV texture d'. The y^* is the text embedding of the fixed text keyword. We combine the $\mathcal{L}^{ga}_{\text{tex}}$ and $\mathcal{L}^{pr}_{\text{tex}}$ as our final texture generation loss:

$$\mathcal{L}_{\text{tex}} = \mathcal{L}^{ga}_{\text{tex}} + \lambda_{pr}\mathcal{L}^{pr}_{\text{tex}}, \quad (6)$$

where λ_{pr} is a weight to balance the gradient from $\mathcal{L}^{pr}_{\text{tex}}$.

3.3. Self-guided Consistency Preserved Editing

To attain the capability of following editing instructions instead of generation prompts, a simple idea is to take the text-guided image editing model InstructPix2Pix [5] ϕ_{ip2p} as a substitute for Stable Diffusion to form the SDS:

$$\nabla_{\beta,u}\mathcal{L}_{\text{edit}} = \mathbb{E}_{t,\epsilon}\left[w(t)\left(\epsilon_{\phi_{ip2p}}\left(z'_t; z_t, y^*, t\right) - \epsilon\right)\frac{\partial z'_t}{\partial \beta, \partial u}\right], \quad (7)$$

where z'_t denotes the latent for the rendering of the edited face, and the original face is embedded to z_t as an extra conditional input, following the setting of InstructPix2Pix.

Note that our geometry and texture are represented by separate parameters β and u, so it is possible to independently optimize one of them, enabling separate editing of geometry and texture. Besides, when editing the texture, we integrate the $\mathcal{L}^{pr}_{\text{tex}}$ to maintain the structural rationality of textures.

Self-guided Consistency Weight. The editing SDS in Eq. 7 enables effective facial editing, while fine-grained editing control still remains challenging, e.g., unpredictable and undesired variations may occur in the results, shown as Fig.

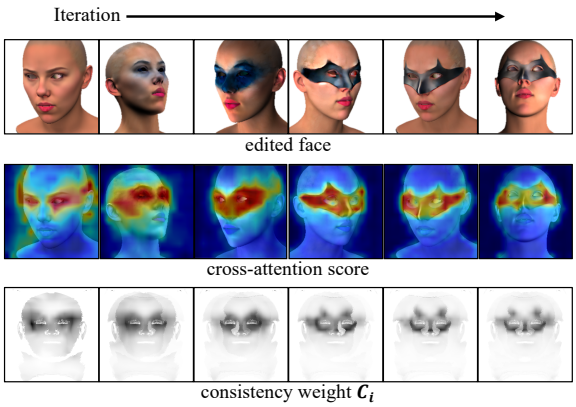


Figure 4. Visualization of the edited face, the cross-attention score for token "mask" and the consistency weight C_i during iterations in editing. Note the viewpoints vary due to random sampling in iterations.

10. This hinders sequential editing, as earlier edits can be unintentionally disrupted by subsequent ones. Therefore, consistency between the faces before and after the editing should be encouraged.

However, the consistency between faces during editing and the noticeability of editing effects, are somewhat contradictory. Imagine a specific pixel in texture, encouraging consistency inclines the pixel towards being the same as the original pixel, while the editing may require it to take on a completely different value to achieve the desired effect.

A key observation in addressing this issue is that the weight of consistency should vary in different regions: For regions associated with editing instructions, a lower level of consistency should be maintained as we prioritize the editing effects. Conversely, for irrelevant regions, a higher level of consistency should be ensured. For instance, given the instruction "*let her wear a Batman eyemask*", we desire the eyemask effect near the eyes region while keeping the rest of the face unchanged.

To locate the relevant region for editing instructions, we propose a self-guided consistency weight strategy in the UV domain. We utilize the built-in cross-attention of the InstructPix2Pix model itself. The attention scores introduce the association between different image regions and specific textual tokens. An example of the consistency weight is shown in Fig 4. We first select a region-indicating token T^* in the instruction, such as "*mask*". At each iteration i, we extract the attention scores between the rendered image I of the editing and the token T^*. The scores are normalized and unwrapped to the UV domain based on the current viewpoint, and then we compute temporal consistency weight $\tilde{C}_i$ from the unwrapped scores:

$$\tilde{C}_i = 1 - \left(\text{proj}\left(\text{norm}\left(\text{att}\left(I', T^*\right)\right)\right)\right)^2, \tag{8}$$

where $\text{att}(\cdot,\cdot)$ denotes the cross-attention operation to predict the attention scores, the $\text{norm}(\cdot)$ denotes the normalization operation, and the proj denotes the unwrapping projection from image to UV domain. As $\tilde{C}_i$ is related to the viewpoint, we establish a unified consistency weight C_i to fuse $\tilde{C}_i$ from different viewpoints. The initial state of C_i is a matrix of all 'one', indicating the highest level of consistency applied to all regions. The updating of C_i at each step is informed by the $\tilde{C}_i$. Specifically, we select the regions where the values in $\tilde{C}_i$ are lower than C_i to be updated. Then we employ a moving average strategy to get the C_i:

$$C_i = C_{i-1} * w + \tilde{C}_i * (1 - w), \tag{9}$$

where w is a fixed moving average factor. We take the C_i as a weight to perform region-specific consistency.

Consistency Preservation Regularization. With the consistency weight C_i in hand, we propose a region-specific consistency preservation regularization in the UV domain to encourage consistency between faces before and after editing in both texture and geometry:

$$\begin{aligned} \mathcal{L}_{\text{reg}}^{tex} &= \|(d_o - d_e) \odot C_i\|_2^2, \\ \mathcal{L}_{\text{reg}}^{geo} &= \|(p_o - p_e) \odot C_i\|_2^2, \end{aligned} \tag{10}$$

where d_o, d_e denote the texture before and after the editing, p_o, p_e denote the vertices position map unwrapped from the facial geometry before and after the editing, and $\odot$ denotes the Hadamard product.

With the consistency preservation regularization, we propose the final loss for our self-guided consistency preserved editing as:

$$L_{finalEdit} = L_{\text{edit}} + \lambda_{reg} L_{\text{reg}}, \tag{11}$$

where λ_{reg} is the balance weight.

4. Experiments

4.1. Implementation Details

Our implementation is built upon Huggingface Diffusers [44]. We use *stable-diffusion* [36] checkpoint for geometry generation, and *sd-controlnet-depth* [29] for texture generation. We utilize the official *instruct-pix2pix* [43] in face editing. The RGB and YUV texture diffusion models are both fine-tuned on the *stable-diffusion* checkpoint. We utilize NVdiffrast [25] for differentiable rendering. Adam [24] optimizer with a fixed learning rate of 0.05 is employed. The generation and editing for geometry/texture require 200/400 iterations, respectively. It takes about 4 minutes to generate or edit a face on a single NVIDIA A30 GPU. We refer readers to the supplementary material for more implementation details.

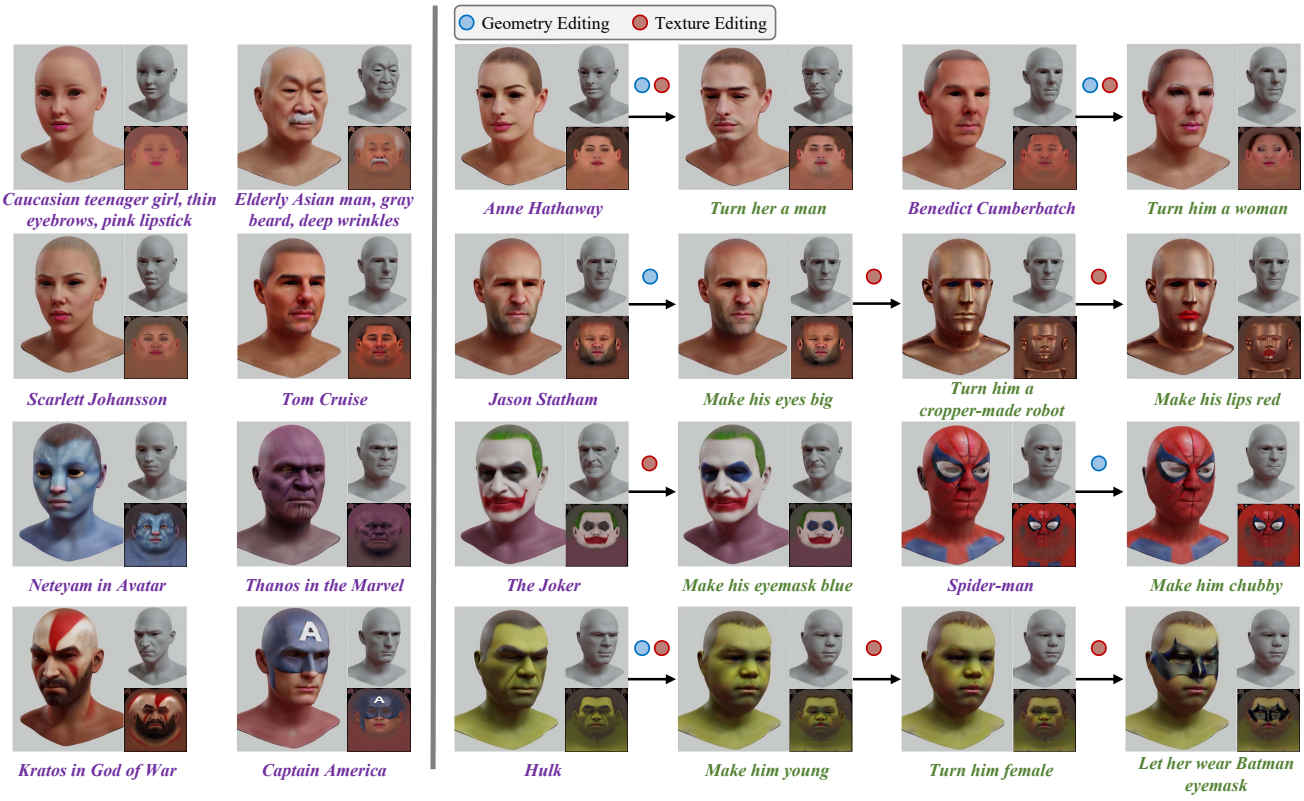


Figure 5. FaceG2E enables the generation of highly realistic and diverse 3D faces (on the left), as well as provides flexible editing capabilities for these faces (on the right). Through sequential editing, FaceG2E achieves the synthesis of highly customized 3D faces, such as 'A female child Hulk wearing a Batman mask'. Additionally, independent editing is available for geometry and texture modification.

4.2. Synthesis Results

We showcase some synthesized 3D faces in Fig. 1 and Fig. 5. As depicted in the figures, FaceG2E demonstrates exceptional capabilities in generating a wide range of visually diverse and remarkably lifelike faces, including notable celebrities and iconic film characters. Furthermore, it enables flexible editing operations, such as independent manipulation of geometry and texture, as well as sequential editing. Notably, our synthesized faces can be integrated into existing CG pipelines, enabling animation and relighting applications, as exemplified in Fig. 1. More animation and relighting results are in the supplementary material.

4.3. Comparison with the state-of-the-art

We compare some state-of-the-art methods for text-guided 3D face generation and editing, including Describe3D [47], DreamFace [53] and TADA [26]. Comparisons with some other methods are contained in the supplementary material.

4.3.1 Qualitative Comparison

The qualitative results are presented in Fig. 6. We can observe that: (1) Describe3D struggles to generate 3D faces following provided texts due to its limited training data and inability to generalize beyond the training set. (2) TADA produces visually acceptable results but exhibits shortcomings in (i) generating high-quality geometry (e.g., evident

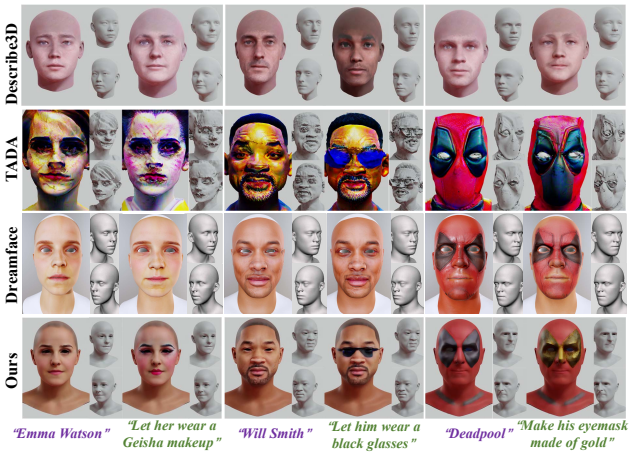


Figure 6. The comparison on text-guided 3D face synthesis. We present both the generation and editing results of each method.

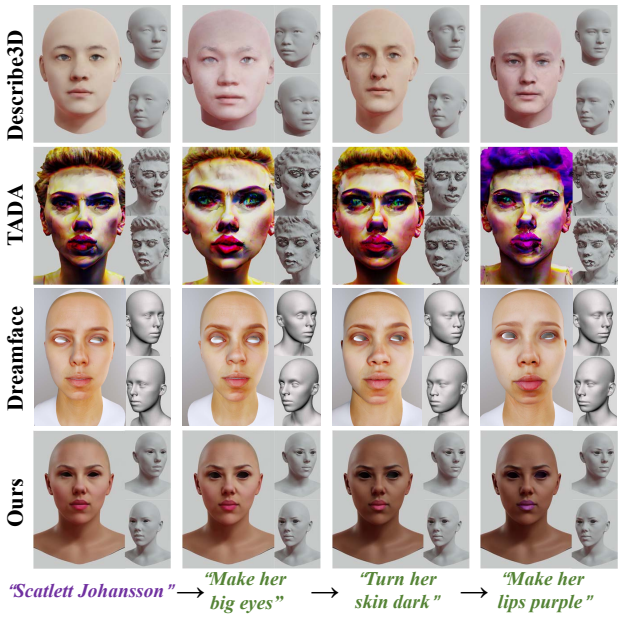


Figure 7. The comparison on sequential face editing.

Method	Generation		Editing	
	Score ↑	Ranking-1 ↑	Score ↑	Ranking-1 ↑
Describe3D [47]	29.81	0%	28.83	0%
Dreamface [53]	33.22	10%	33.14	10%
TADA [26]	34.85	10%	33.73	20%
Ours	**36.95**	**80%**	**35.50**	**70%**

Table 1. The CLIP evaluation results on the synthesized 3D faces.

geometric distortion in its outputs), and (ii) accurately following editing instructions (e.g., erroneously changing black glasses to blue in case 2). (3) Dreamface can generate realistic faces but lacks editing capabilities. Moreover, its geometry fidelity is insufficient, hindering the correlation between the text and texture-less geometry. In comparison, our method is superior in both generated geometry and texture and allows for accurate and flexible face editing.

We further provide a comparison of sequential editing in Fig. 7. Clearly, the editing outcomes of Describe3D and Dreamface in each round lack prominence. Although TADA performs well with single-round editing instructions, it struggles in sequence editing due to unintended changes that impact the preceding editing effects influenced by subsequent edits. For instance, in the last round, TADA mistakenly turns the skin purple. In contrast, our FaceG2E benefits from the proposed self-guided consistency preservation, allowing for precise sequence editing.

4.3.2 Quantitative Comparison

We quantitatively compare the fidelity of synthesized faces to text descriptions using the CLIP evaluation. We provide a total of 20 prompts, evenly split between generation and editing tasks, to all methods for face synthesis. All results are rendered with the same pipeline, except DreamFace, which takes its own rendering in the web demo [19]. A fixed prefix 'a realistic 3D face model of ' is employed for all methods when calculating the CLIP score. We report the CLIP Score [39] and Ranking-1 in Tab. 1. CLIP Ranking-1 calculates the ratio of a method's created faces ranked as top-1 among all methods. The results validate the superior performance of our method over other SOTA methods.

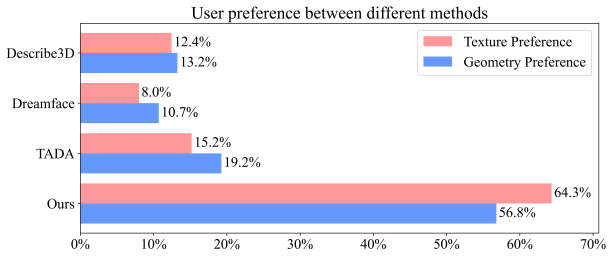


Figure 8. Quantitative results of user study. Our results are more favored by the participants compared to the other methods.

4.3.3 User Study

We perform a comparative user study involving 100 participants with Fuxi Youling Crowdsourcing[1] to evaluate our method against state-of-the-art (SOTA) approaches. Participants are presented with 10 face generation examples and 10 face editing examples, and are asked to select the best method for each example based on specific criteria. The results, depicted in Fig. 8, unequivocally show that our method surpasses all others in terms of user preference.

4.4. Ablation Study

Here we present some ablation studies. Extra studies based on user surveys are provided in the supplementary material.

4.4.1 Effectiveness of GDG

To evaluate the effectiveness of geometry-texture decoupled generation (GDG), we conduct the following studies.

Geometry-centric SDS (GcSDS). In Fig. 9(a), we conduct an ablation study to assess the impact of the proposed GcSDS. We propose a variation that takes standard textured rendering as input for SDS and simultaneously optimizes both geometry and texture variables. The results reveal that without employing the GcSDS, there is a tendency to generate relatively planar meshes, which lack geometric details such as facial wrinkles. We attribute this deficiency to the misrepresentation of geometric details by textures.

Geometry-aligned texture content SDS (GaSDS). In Columns 3 and 4 of Fig. 9(b), we evaluate the effective-

[1]https://fuxi.163.com/solution/data

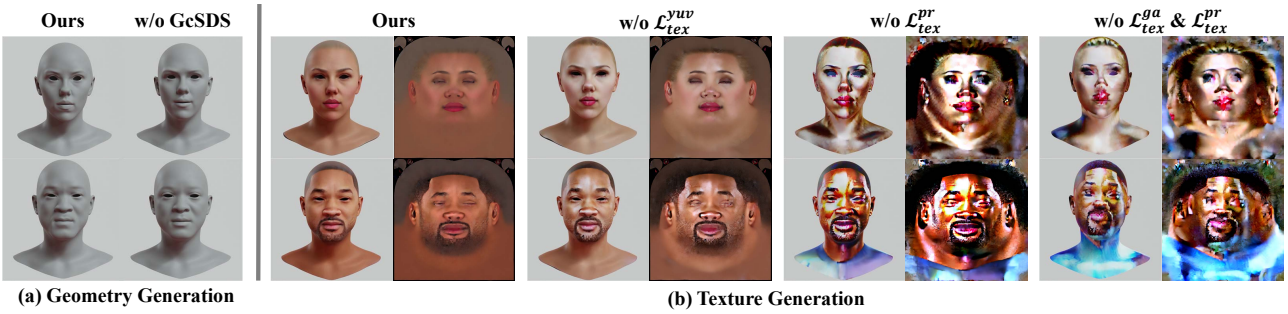


Figure 9. The ablation study of our geometry-texture decoupled generation. The input texts are 'Scarlett Johansson' and 'Will Smith'.

Figure 10. Analysis of the proposed self-guided consistency preservation (SCP) in 3D face editing.

ness of GaSDS. We replace the depth-ControlNet in GaSDS with the standard Stable-Diffusion model to compute L_{tex}^{ga}. The results demonstrate a significant problem of geometry-texture misalignment. This issue arises because the standard Stable Diffusion model only utilizes text as a conditional input and lacks perception of geometry.

Texture prior SDS. To assess the efficacy of our texture prior SDS, we compared it with two variants: one that solely relies on geometry-aware texture content SDS, denoted as ***w/o*** $\boldsymbol{L_{tex}^{pr}}$, and another that excludes the use of L_{tex}^{yuv}, denoted as ***w/o*** $\boldsymbol{L_{tex}^{yuv}}$. As shown in Columns 1,2 and 3 of Fig. 9(b), the results demonstrate that the ***w/o*** $\boldsymbol{L_{tex}^{pr}}$ pipeline generates textures with significant noise and artifacts. The ***w/o*** $\boldsymbol{L_{tex}^{yuv}}$ pipeline produces textures that generally adhere to the distribution of facial textures, but may exhibit brightness irregularities. The complete L_{tex}^{pr} yields the best results.

4.4.2 Effectiveness of SCP

To evaluate the effectiveness of the proposed self-guided consistency preservation (SCP) in editing, we conduct the following ablation study. We make two variants: One variant, denoted as ***w/o Reg***, solely relies on L_{edit} for editing without employing consistency regularization. The other variant, denoted as ***w/o SC-weight***, computes the consistency preservation regularization without using the self-guided consistency weight.

The results are shown in Fig. 10. While ***w/o Reg*** shows noticeable editings following the instructions, unexpected alterations occur, such as the skin and hair of Scarlett turning purple, and Hulk's skin turning yellow. This inadequacy can be attributed to the absence of consistency constraints. On the other hand, ***w/o SC-weight*** prevents undesirable changes in the results but hampers the effectiveness of editing, making it difficult to observe significant editing effects. In contrast, the full version of SCP achieves evident editing effects while preserving consistency in unaffected regions, thereby ensuring desirable editing outcomes.

5. Conclusion

We propose FaceG2E, a novel approach for generating diverse and high-quality 3D faces and performing facial editing using texts. With the proposed geometry-texture decoupled generation, high-fidelity facial geometry and texture can be produced. Despite achieving new state-of-the-art results, we notice some limitations in FaceG2E. (1) The geometric representation restricts us from generating shapes beyond the facial skin, such as hair and accessories. (2) Sequential editing enables the synthesis of customized faces, but it also leads to a significant increase in time consumption. Each round of editing requires additional time.

Acknowledgement

This work was supported in part by the National Key R&D Program of China (No. 2022YFF09022303).

References

[1] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. Clipface: Text-guided editing of textured 3d morphable models. *arXiv preprint arXiv:2212.01406*, 2022. 1

[2] Haoran Bai, Di Kang, Haoxian Zhang, Jinshan Pan, and Linchao Bao. Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction. *arXiv preprint arXiv:2211.13874*, 2022. 4

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2

[4] Linchao Bao, Xiangkai Lin, Yajing Chen, Haoxian Zhang, Sheng Wang, Xuefei Zhe, Di Kang, Haozhi Huang, Xinwei Jiang, Jue Wang, Dong Yu, and Zhengyou Zhang. High-fidelity 3d digital human head creation from rgb-d selfies. *ACM Transactions on Graphics*, 2021. 3

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023. 2, 4

[6] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 2

[7] Rahul Dey and Vishnu Naresh Boddeti. Generating diverse 3d reconstructions from a single occluded face image. In *CVPR*, pages 1547–1557, 2022. 1

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021. 2

[9] Abdallah Dib, Junghyun Ahn, Cedric Thebault, Philippe-Henri Gosselin, and Louis Chevallier. S2f2: Self-supervised high fidelity face reconstruction from monocular image. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2023. 1

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[11] Xiao Han, Yukang Cao, Kai Han, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, Tao Xiang, and Kwan-Yee K. Wong. Headsculpt: Crafting 3d head avatars with text. *arXiv preprint arXiv:2306.03038*, 2023. 1, 2

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2

[14] 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 TOG*, 41(4):1–19, 2022. 2

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[16] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023. 2

[17] Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. *arXiv preprint arXiv:2310.01406*, 2023. 2

[18] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv preprint arXiv:2306.12422*, 2023. 2

[19] Deemos. Inc. dreamface web demo. https://hyperhuman.deemos.com/, 2023. 7

[20] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *CVPR*, pages 867–876, 2022. 2

[21] Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control. *arXiv preprint arXiv:2303.17606*, 2023. 2, 4

[22] Zutao Jiang, Guansong Lu, Xiaodan Liang, Jihua Zhu, Wei Zhang, Xiaojun Chang, and Hang Xu. 3d-togo: Towards text-guided cross-category 3d object generation. In *AAAI*, pages 1051–1059, 2023. 1

[23] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, pages 6007–6017, 2023. 2

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[25] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. 5

[26] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J. Black. TADA! Text to Animatable Digital Avatars. In *International Conference on 3D Vision (3DV)*, 2024. 1, 2, 4, 6, 7

[27] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, pages 300–309, 2023. 1, 2

[28] Zhengzhe Liu, Yi Wang, Xiaojuan Qi, and Chi-Wing Fu. Towards implicit text-guided 3d shape generation. In *CVPR*, pages 17896–17906, 2022. 1

[29] lllyasviel. Controlnet. https://huggingface.co/runwayml/lllyasviel/sd-controlnet-depth, 2023. 5

[30] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *CVPR*, pages 13492–13502, 2022. 2

[31] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 3

[32] 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*, pages 8748–8763. PMLR, 2021. 1, 2

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2, 3, 4

[35] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 2

[36] RunwayML. Stable diffusion v1.5. https://huggingface.co/runwayml/stable-diffusion-v1-5, 2022. 5

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 35:36479–36494, 2022. 2

[38] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *CVPR*, pages 18603–18613, 2022. 2

[39] Aditya Sanghi, Rao Fu, Vivian Liu, Karl DD Willis, Hooman Shayani, Amir H Khasahmadi, Srinath Sridhar, and Daniel Ritchie. Clip-sculptor: Zero-shot generation of high-fidelity and diverse shapes from natural language. In *CVPR*, pages 18339–18348, 2023. 7

[40] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022. 2

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 2

[43] timbrooks. Instructpix2pix. https://huggingface.co/runwayml/timbrooks/instruct-pix2pix, 2023. 5

[44] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models, 2022. 5

[45] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4563–4573, 2023. 1, 2

[46] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 3d face reconstruction with dense landmarks. In *ECCV*, pages 160–177. Springer, 2022. 1

[47] 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*, pages 4521–4530, 2023. 2, 6, 7

[48] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. pages 7452–7461, 2023. 2

[49] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *CVPR*, pages 20908–20918, 2023. 1

[50] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *CVPR*, pages 20908–20918, 2023. 2

[51] Kim Youwang, Kim Ji-Yeon, and Tae-Hyun Oh. Clip-actor: Text-driven recommendation and stylization for animating human meshes. In *ECCV*, pages 173–191. Springer, 2022. 1

[52] Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Kang Du, and Min Zheng. Avatarverse: High-quality & stable 3d avatar creation from text and pose. *arXiv preprint arXiv:2308.03610*, 2023. 1, 4

[53] 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. 1, 2, 4, 6, 7

[54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2, 4

[55] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *ECCV*, pages 250–269. Springer, 2022. 1