

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

Supplementary Material

1. Implementation Details

Camera settings. During the optimization, We employ a camera with fixed intrinsic parameters: near=0.1, far=10, fov=12.59, rendering image size=224. For the camera extrinsics, we defined a set of optional viewing angles and randomly selected one of these angles as the rendering viewpoint for optimization in each iteration. The elevation angle $x \in \{0, 10, 30\}$, the azimuth angle $y \in \{0, 30, 60, 300, 330\}$, and the camera distance $d \in \{1.5, 3\}$. We set these extrinsics to ensure that the rendering always includes the facial region.

Light settings. We utilize spherical harmonic (SH) to represent lighting. We pre-define 16 sets of spherical harmonic 3-band coefficients. In each iteration of rendering, we randomly select one set from these coefficients to represent the current lighting.

Prompt engineering. In the generation stage, for the face description prompt of a celebrity or a character, we add the prefix ‘a zoomed out DSLR photo of’. We also utilize the view-dependent prompt enhancement. For the azimuth in (0,45) and (315,360), we add a suffix ‘from the front view’, for the azimuth in (45,135) and (225,315), we add a suffix ‘from the side view’.

SDS Time schedule. Following the Dreamfusion [3], we set the range of t to be between 0.98 and 0.02 in the SDS computation process. Besides, we utilize the linearly decreasing schedule for t , which is crucial for the stability of synthesis. As the iteration progresses from 0 to the final (e.g. iteration 400), our t value linearly decreases from 0.98 to 0.02.

2. User survey as ablation

We conduct a user survey as ablation to further validate the effectiveness of our key design. A total of 100 volunteers participated in the experiment. We presented the results of our method and different degradation versions, alongside the text prompts. Then we invited the volunteers to rate the facial generation and editing. The ratings ranged from 1 to 5, with higher scores indicating higher satisfaction. The user rating results are shown in Tab. 1 and Tab. 2. The results indicate that removing any of our key designs during the face generation or face editing leads to a decrease in user ratings. This suggests that our key designs are necessary for synthesizing high-quality faces.

Generation			
ours	w/o L_{tex}^{yuv}	w/o L_{tex}^{pr}	w/o L_{tex}^{ga} & L_{tex}^{pr}
3.82	3.77	2.59	1.78

Table 1. Ablation study of face generation based on user ratings.

Editing		
ours	w/o SC-weight	w/o Reg
3.95	2.55	2.28

Table 2. Ablation study of face editing based on user ratings.



Figure 1. Relighting of our synthesized 3D faces.

3. More Relighting Results

We present some more relighting results in Fig 1. We recommend referring to the supplementary material video or project page, where the video results can better demonstrate our animation and relighting effects.

4. Generation with composed prompt

Our sequential editing can synthesize complex 3D faces, an alternative approach is to combine all editing prompts into a composed prompt and generate the face in one step.

In Fig.2, we showcase the results generated from a composed prompt with our generation stage. It can be observed that directly generating with the composed prompt leads to the loss of certain concepts and details present in the prompts (e.g., the cropped-made effect in row 1, or the black lips in row 2). This underscores the necessity of the editing technique we propose for synthesizing customized faces.

5. Computational cost.

We record and report the computational cost for our method and baseline methods. To generate a single face, Describe3D takes about 1 minute and Dreamface requires 5

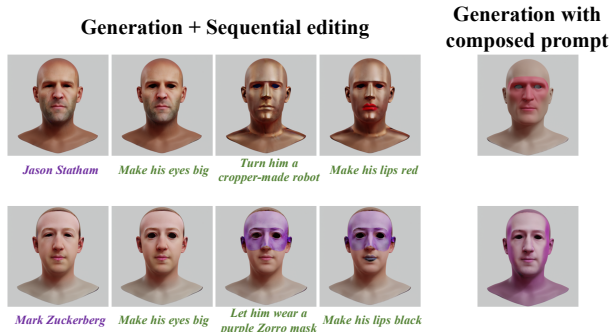


Figure 2. Generation with composed prompt leads to the loss of concepts in prompts.



Figure 3. Using grey/skin-color diffusion in the geometry phase.



Figure 4. Sequential editing on the same region.

minutes (1 minute with its web demo). TADA spends around 4 hours to synthesize an avatar. Our method costs 4 minutes in generation or single-round editing on an NVIDIA A30 GPU.

6. Default skin-color texture for the geometry phase.

In the geometry generation phase, one can employ a default skin-color texture instead of the gray texture, as seen in Fig. 3. The generated geometries are quite similar. This experiment indicates the geometry phase is not sensitive to the employed texture.

7. Sequential editing on the same region.

Our sequential editing can manipulate the same region multiple times, without causing extra artifacts. As shown in Fig 4, we first edit ‘Scarlett Johansson’ with the instruction ‘let her wear the Batman eyemask’, and then edit the face with the instruction ‘let her wear the Deadpool mask’.

8. More Comparison Results

We conduct more comparisons with more baseline methods. We add two baselines: a public implementation [1]

for the Dreamfusion, and AvatarCraft [2], a SOTA text-to-3D avatar method that utilizes the implicit neutral field representation. We compare text-guided 3D face generation, single-round 3D face editing, and sequential 3D face editing. Note that baseline methods are not capable of directly editing 3D faces with text instruction (e.g., ‘make her old’), so we let them perform the editing by generating a face with the composed prompt. For example, ‘an old Emma Watson’ is the composed prompt of ‘Emma Watson’ and ‘Make her old’.

We present the 3D face generation results in Fig 5 and Fig 6. The 3D face editing results are contained in Fig 7 and Fig 8. The comparisons on sequential editing are presented in Fig 9 and Fig 10. It should be noted that Dreamfusion [1] and Avatarcraft [2] occasionally fail to produce meaningful 3D shapes and instead output a white background for some prompts. This issue could potentially be addressed by resetting the random seed, however, due to time constraints, we did not attempt repeated trials. We have labeled these examples as ‘Blank Result’ in the figures.

References

- [1] Stable-dreamfusion. <https://github.com/ashawkey/stable-dreamfusion>, 2022. 2
- [2] 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
- [3] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1

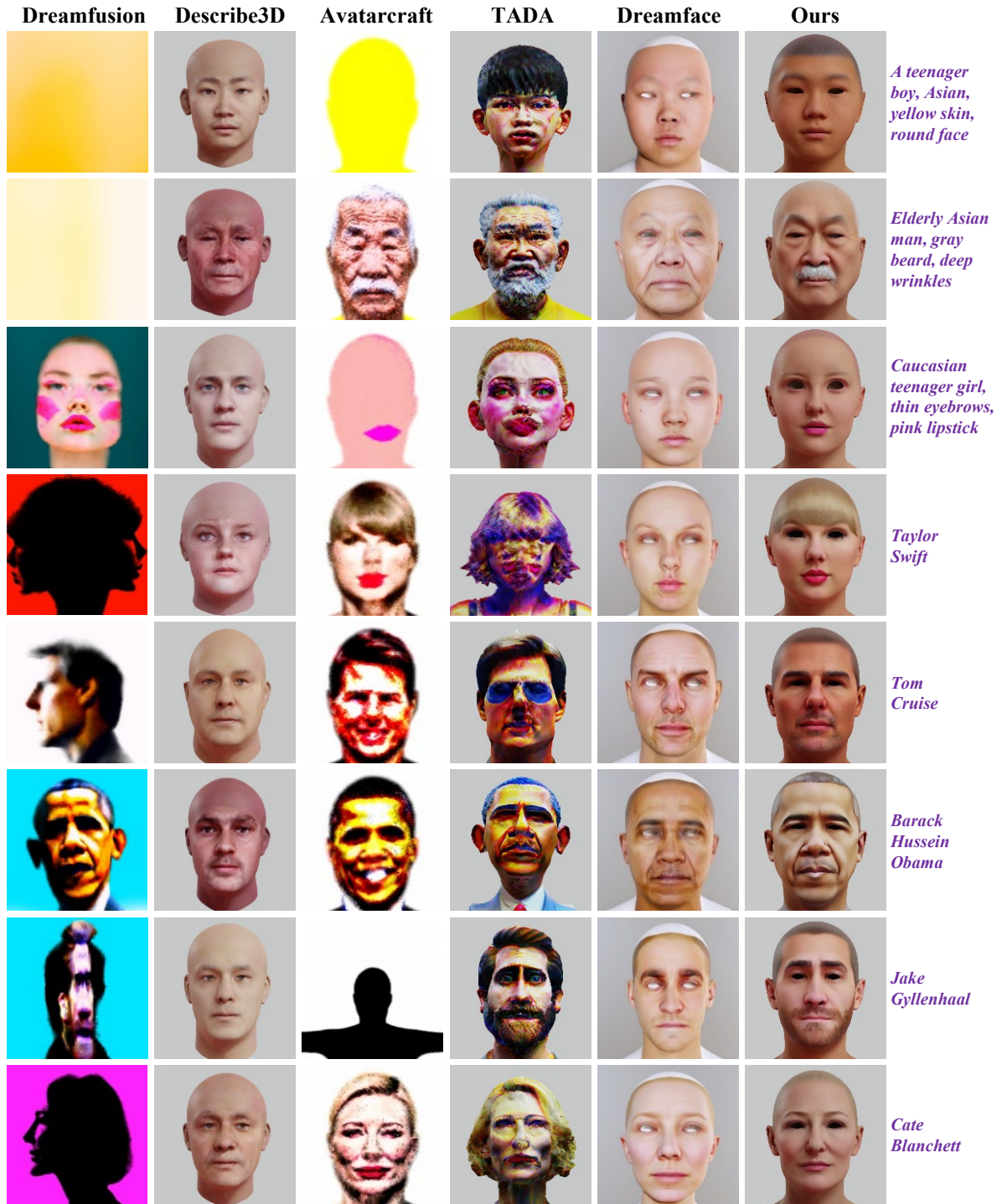


Figure 5. Comparison on text-guided 3D face generation.

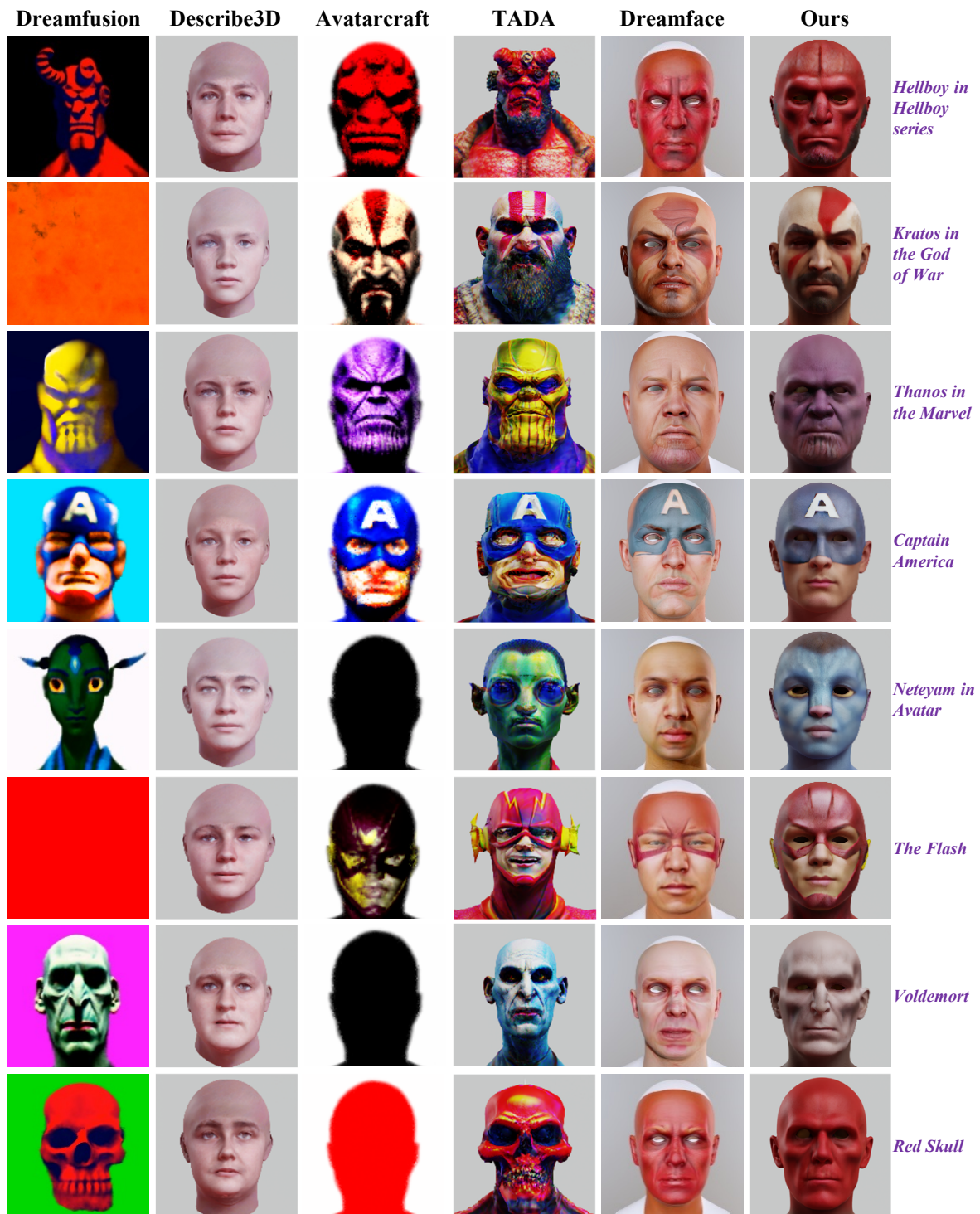


Figure 6. Comparison on text-guided 3D face generation.





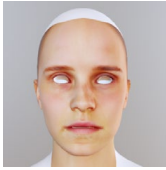
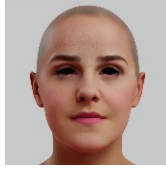



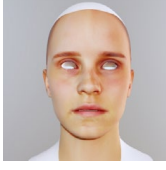

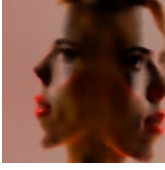


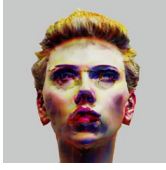
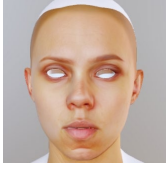
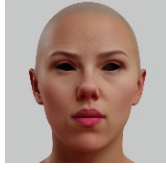
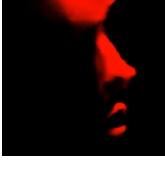


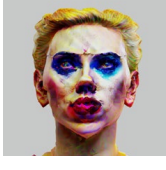
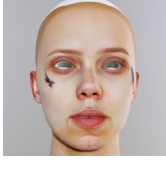
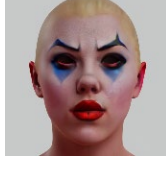
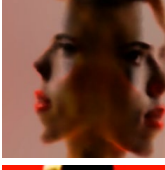


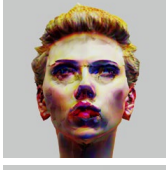
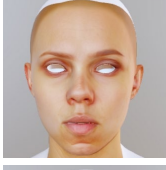
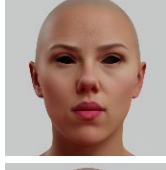

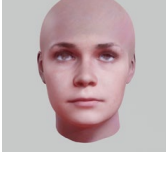
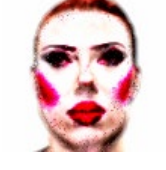
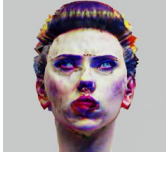
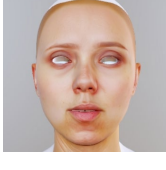
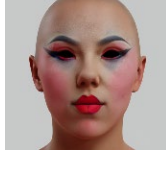


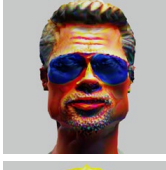
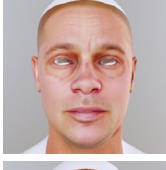
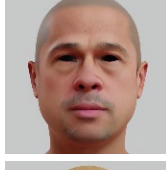





Dreamfusion	Describe3D	AvatarCraft	TADA	Dreamface	Ours	
						<i>Original: Emma Watson</i>
Blank Result						<i>Editing: Make her old</i>
						<i>Original: Scarlett Johansson</i>
						<i>Editing: Turn her into Harley Quinn</i>
						<i>Original: Scarlett Johansson</i>
						<i>Editing: Let her wear a Geisha makeup</i>
		Blank Result				<i>Original: Brad Pitt</i>
		Blank Result				<i>Editing: Make him made of wood</i>

Figure 7. Comparison on text-guided single-round 3D face editing.

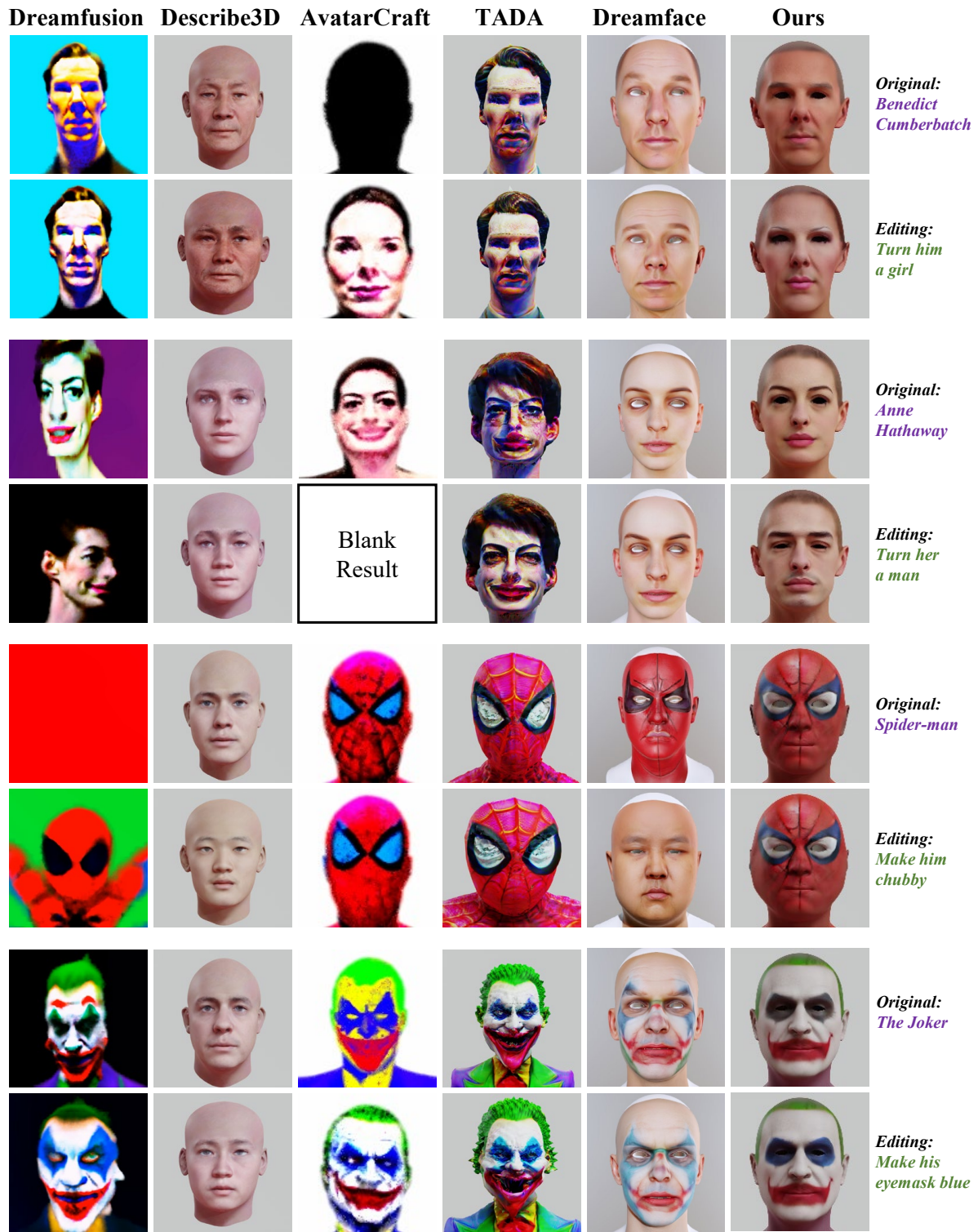


Figure 8. Comparison on text-guided single-round 3D face editing.

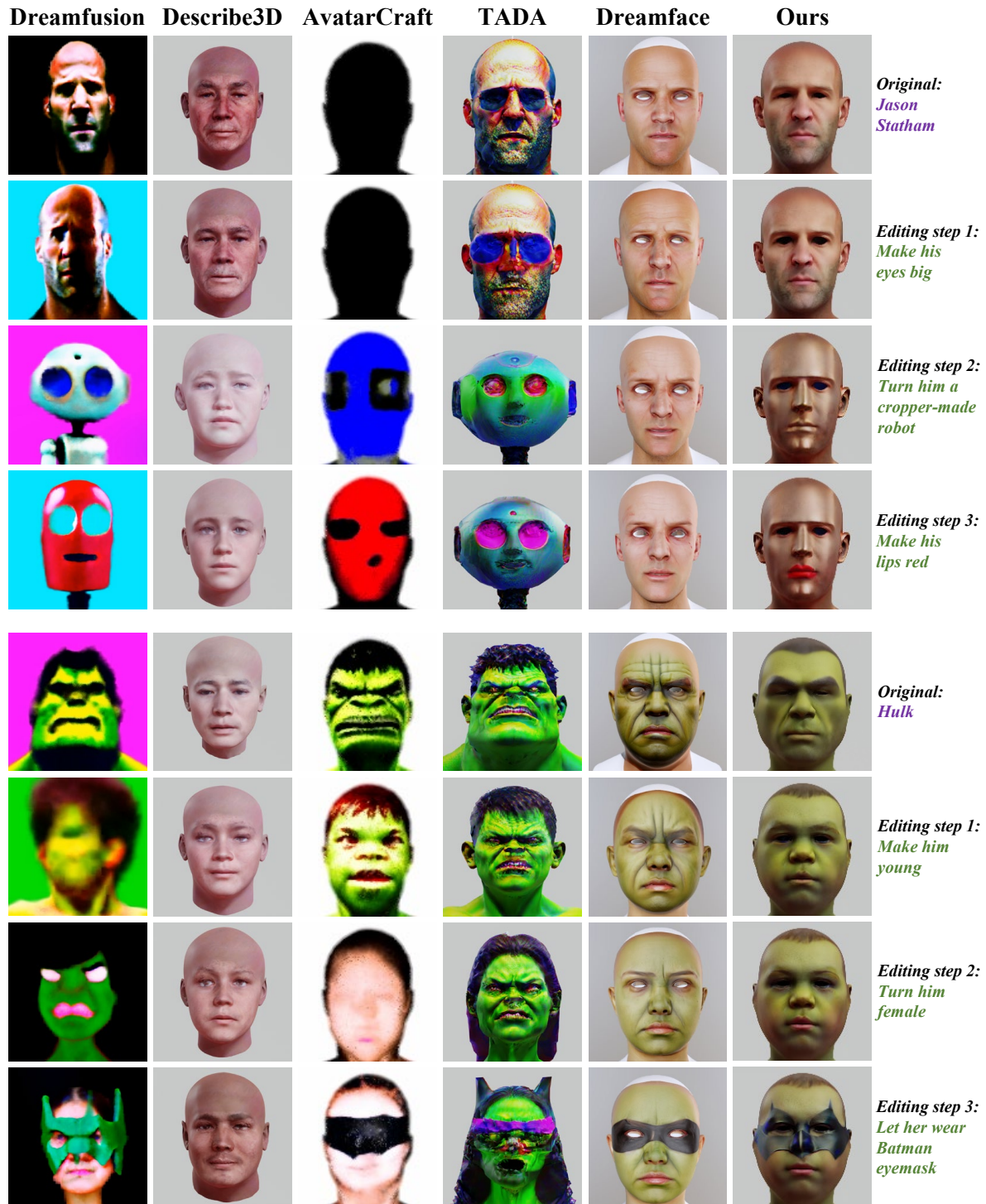


Figure 9. Comparison on text-guided sequential 3D face editing.

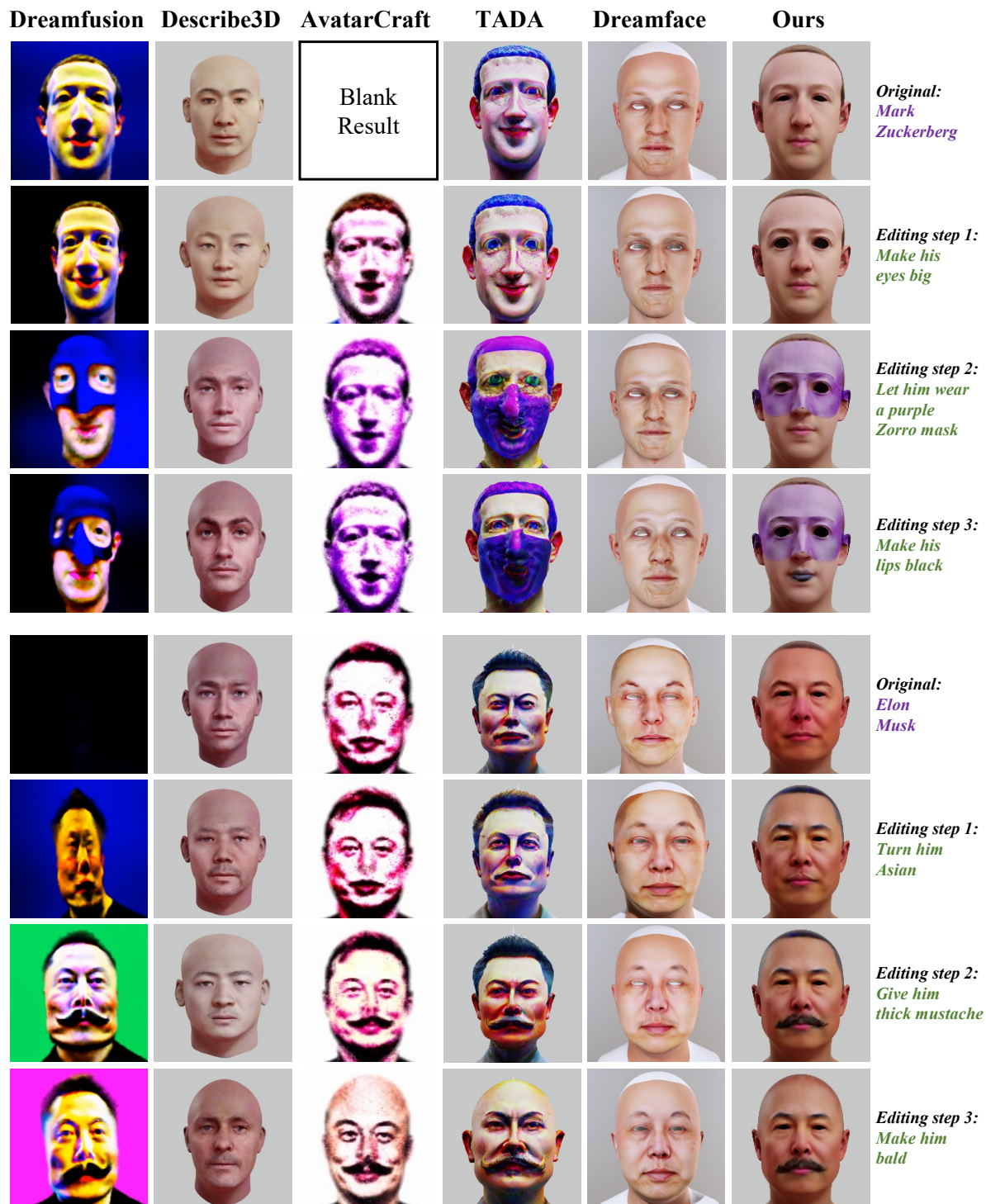


Figure 10. Comparison on text-guided sequential 3D face editing.