Appendix of FFCLIP

Yiming Zhu[1*] **Hongyu Liu**[2,4*] **Yibing Song**[3†] **Ziyang Yuan**[1] **Xintong Han**[4]
Chun Yuan[1†] **Qifeng Chen**[2] **Jue Wang**[3]

[1]Tsinghua Shenzhen International Graduate School
[2]Hong Kong University of Science and Technology
[3]Tencent AI Lab [4]Huya Inc
zym20@mails.tsinghua.edu.cn hliudq@cse.ust.hk
yibingsong.cv@gmail.com yuanc@sz.tsinghua.edu.cn

1 More results

In Fig. 1, we show more results for model in editing images with text prompts containing multiple semantics. Meanwhile, we display more results for the ability of free-form image manipulation. These visual results demonstrate the powerful performance of FFCLIP.

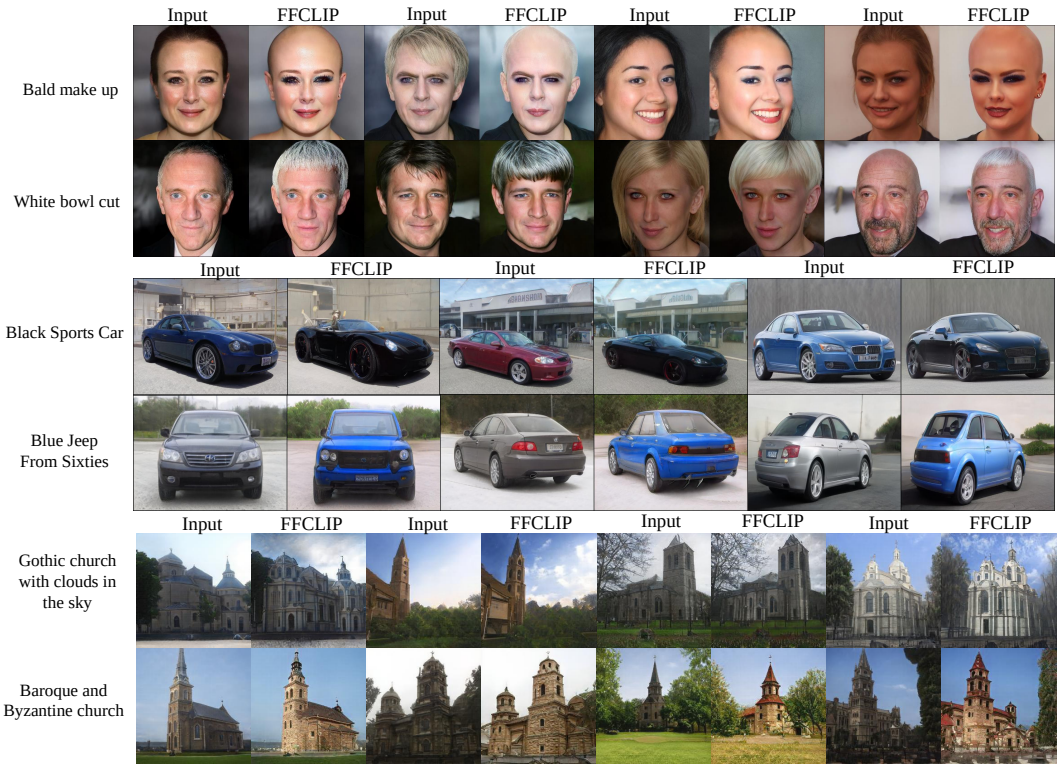

Figure 1: Our manipulation results with text prompts containing multiple semantic meanings.

*Y. Zhu and H. Liu contributes equally. †Y. Song and C. Yuan are corresponding authors. This work is done when Y. Zhu is an intern in Tencent AI Lab.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Figure 2: Our manipulation results of human portraits. All input images are inversions of the real images. The target manipulation semantic used in the text prompt is indicated above each column. FFCLIP has the capability of generating photo-realistic and text-relevant results.

2 Comparison with HairCLIP

As shown in Fig. 7 and Fig. 8, we compare with HairCLIP in hair manipulation. The HairCLIP cannot reflect the manipulation semantics with hair color and style well(e.g., see 'Red hair' and 'Bob cut hairstyle'). In contrast, our FFCLIP can modify the content more correctly with the capability of finding the latent subspace adaptively. The related numerical comparison as shown in Table 1, for the numerical text prompt 'Gray hair' and 'Blond hair,' we use the similar measurement method of 'Red hair.'

Table 1: Quantitative comparison between HairCLIP and FFCLIP.

Text Prompt	**Editing Performance**	
	HairCLIP	Ours
Bald	0.2481	**0.0279**
Red hair	1.0474	**0.7171**
Gray hair	0.6344	**0.4931**
Blond hair	1.5126	**1.350**

3 Human Subject Evaluation for More Text Prompts

We conduct another Human Subject Evaluation with four more text prompts in Table 2. Except for the text prompts, the other settings in this Human Subject Evaluation are the same as the main paper. We can find that our FFCLIP outperforms other baselines.

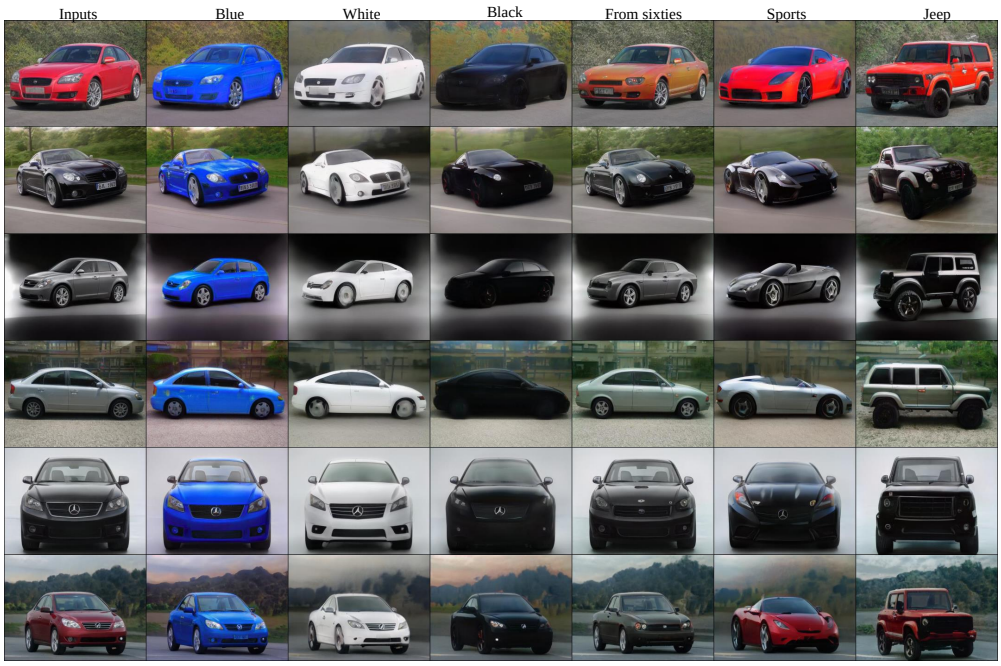

Figure 3: Our manipulation results of cars. All input images are inversions of the real images. The target manipulation semantic used in the text prompt is indicated above each column. FFCLIP has the capability of generating photo-realistic and text-relevant results.

Table 2: Quantitative evaluations on the CelebA-HQ dataset. FFCLIP is more effective to produce semantically relevant results for human subjects.

Text Prompt	**Human Subject Evaluation**		
	TediGAN	StyleCLIP	Ours
Blue eyes	14.3%	31.4%	**54.3** %
Disgust	4.2%	22.9%	**72.9** %
Dreadlocks hairstyle	1.4%	1.4%	**97.2**%
Jewfro hairstyle	1.4%	1.4%	**97.2**%

4 Discussion for StyleCLIP, HairCLIP, and TediGAN

As shown in Fig. 9, StyleCLIP and HairCLIP need experiential behaviors to match the text semantic and latent space. While the latent space W in StyleGAN is not disentangled completely, the experiential behaviors cannot find the correct latent subspace for the target text prompt, so the StyleCLIP needs to train a specific mapper to find the desired latent subspace for specific text. The HairCLIP can manipulate the image with different texts, but it just edits the hair regions. In contrast, FFCLIP proposes a semantic alignment module to align the text semantic and latent space, and we can find the target latent space with text adaptively after alignment. For TediGAN, it makes the text feature close to the latent code and uses the style-mixing operation in StyleGAN to manipulate the image content. Meanwhile, the style-mixing operation needs to replace the specific layer in latent code with text feature, the text semantic cannot align latent space well with this experiential behavior, so it cannot manipulate the image with some interesting text prompts (i.e., 'Taylor Swift' or 'Beard Blond.'). Moreover, TediGAN just manipulates the human portrait because the coarse correspondence between specific semantic and specific layers in latent code is hard to reproduce for other datasets.

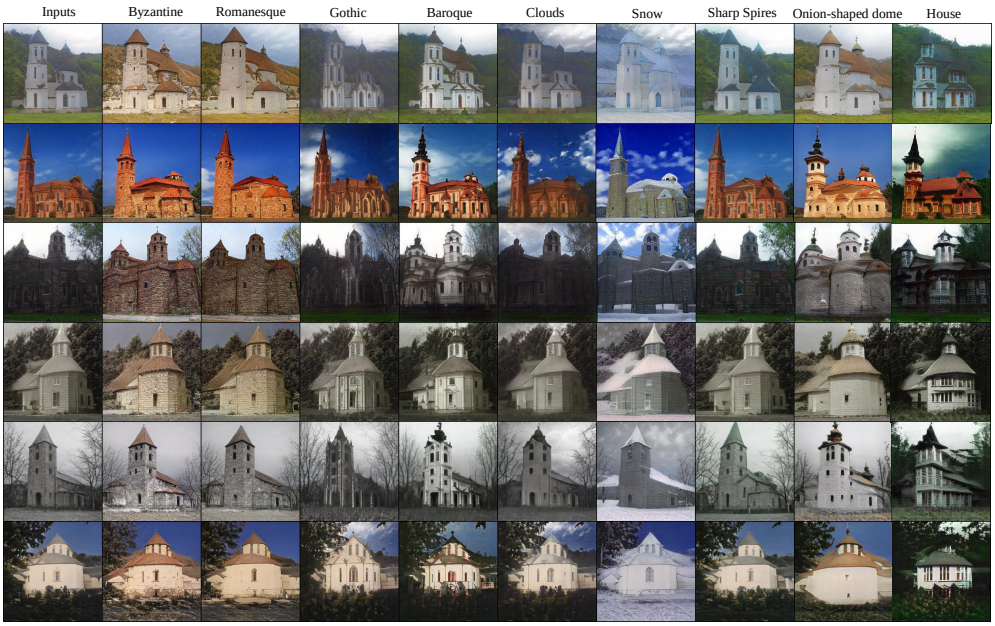

Figure 4: Our manipulation results of churches. All input images are inversions of the real images. The target manipulation semantic used in the text prompt is indicated above each column. FFCLIP has the capability of generating photo-realistic and text-relevant results.

5 Unseen text prompts and different inversion encoder

We show more visual results for unseen text prompts in Fig. 10 and Fig. 11. The FFCLIP performs well in these text prompts. Meanwhile, as shown in Fig. 12, although the FFCLIP is trained with e4e encoder, it can edit image correct with High-fidelity [2] and Restyle [1] encoders, respectively. These phenomenons prove the robustness of our model and the effectiveness of the semantic alignment module.

6 Large CLIP Model

We show the visual comparison results between different CLIP models in Fig. 13. We find that the large CLIP model ViT-L/14 has no noticeable impact on performance.

7 More visual comparison results

We show more visual comparison results between StyleCLIP, TediGAN and FFCLIP in Fig. 14 and Fig. 15. And we compare our model with StyleCLIP in hair color manipulation in Fig. 16. Our method can preserve the identity of images and ensure well manipulation performance at the same time.

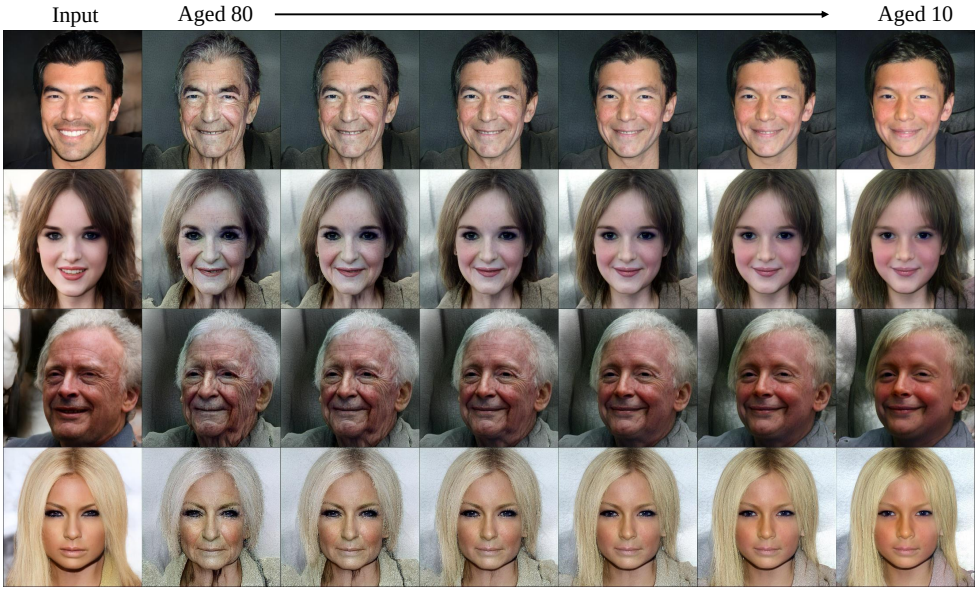

Figure 5: Expression interpolation results. We generate the intermediate results between ‘Aged 10’ and ‘Aged 80’.

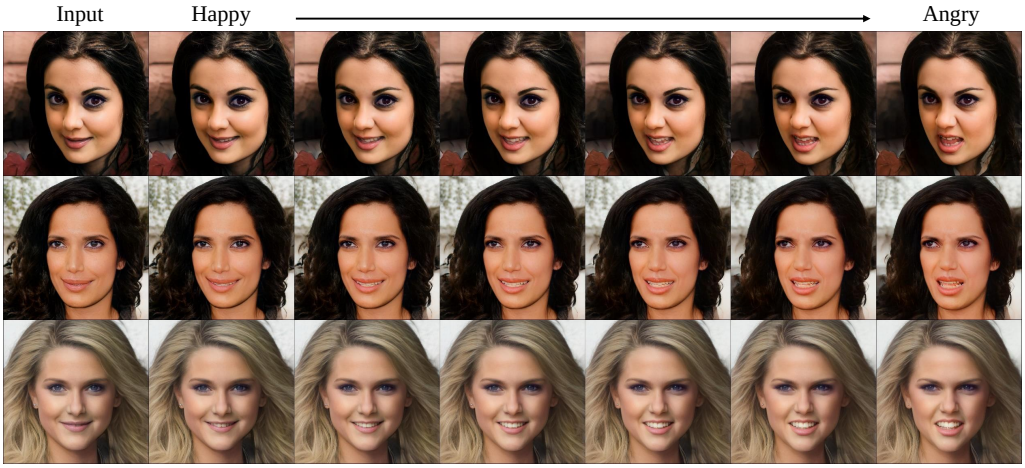

Figure 6: Expression interpolation results. We generate the intermediate results between ‘Happy’ and ‘Angry’.

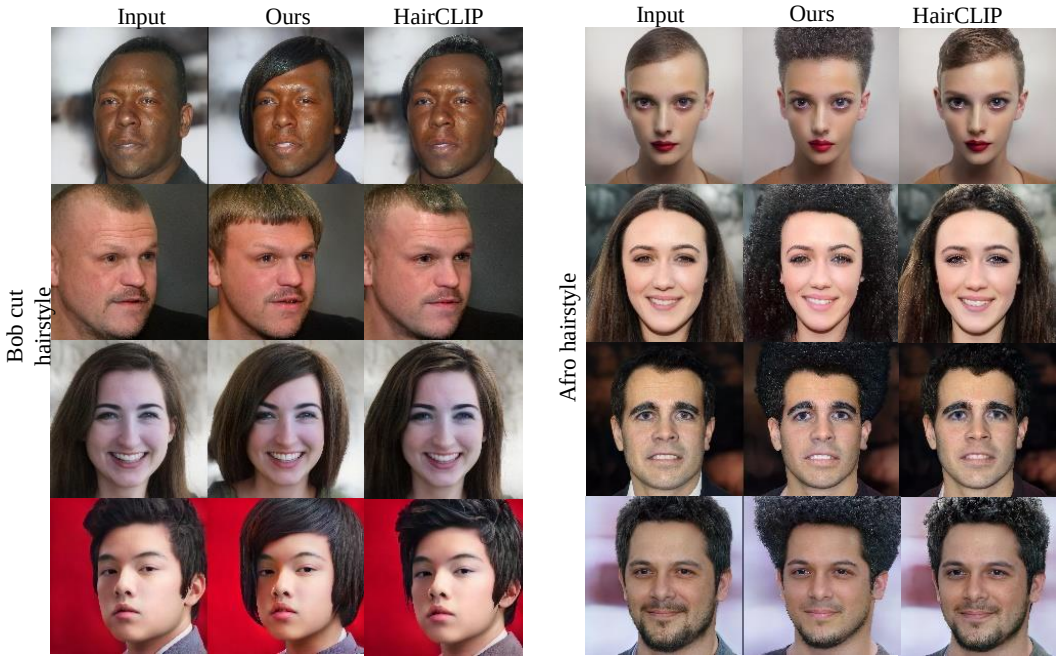

Figure 7: Visual comparison with HairCLIP [3] on the CelebA-HQ dataset.The text guidance is described on the left side. FFCLIP is more effective to produce semantic relevant and visually realistic results.

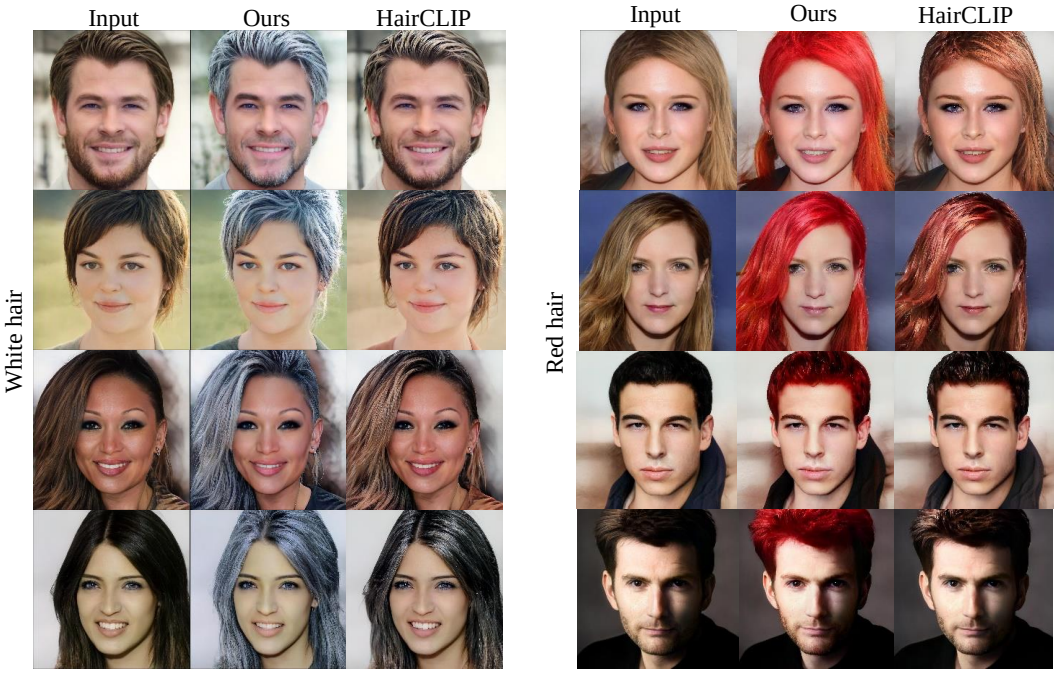

Figure 8: Visual comparison with HairCLIP [3] on the CelebA-HQ dataset. The text guidance is described on the left side. FFCLIP is more effective to produce semantic relevant and visually realistic results.

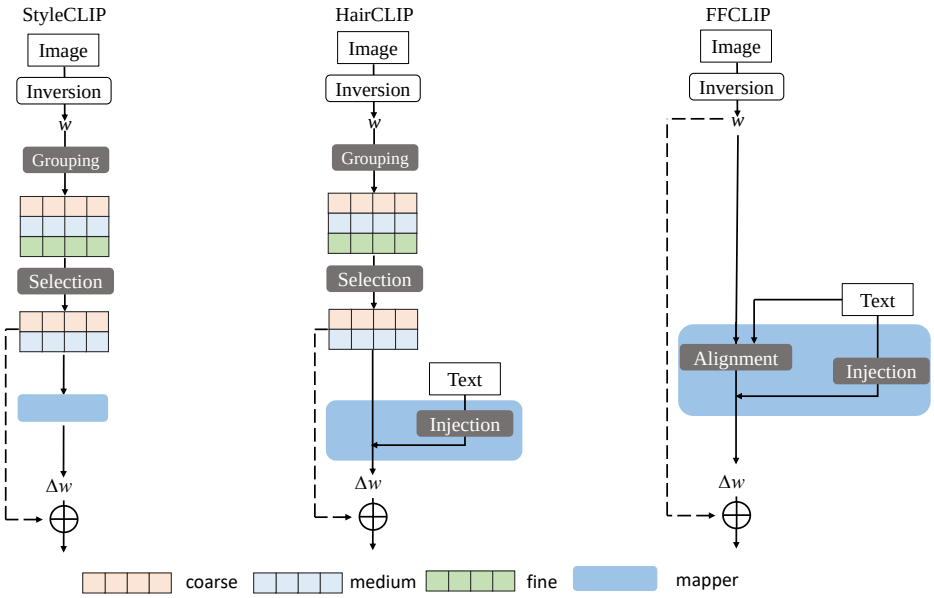

Figure 9: The grouping and selection processes are experiential behaviors. Our FFCLIP can adaptively align the semantic of text and latent space, so that we can edit images with different text prompts by a single model.

Figure 10: The visual results for unseen text prompts. The text prompts are on the left side of each group.

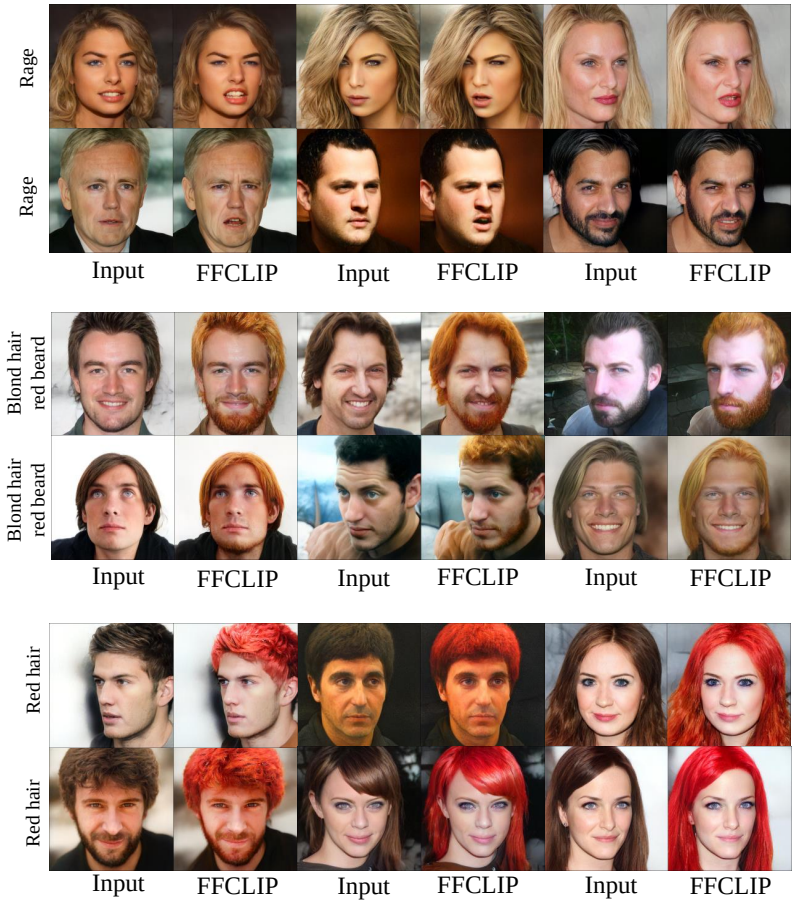

Figure 11: The visual results for unseen text prompts. The text prompts are on the left side of each group.

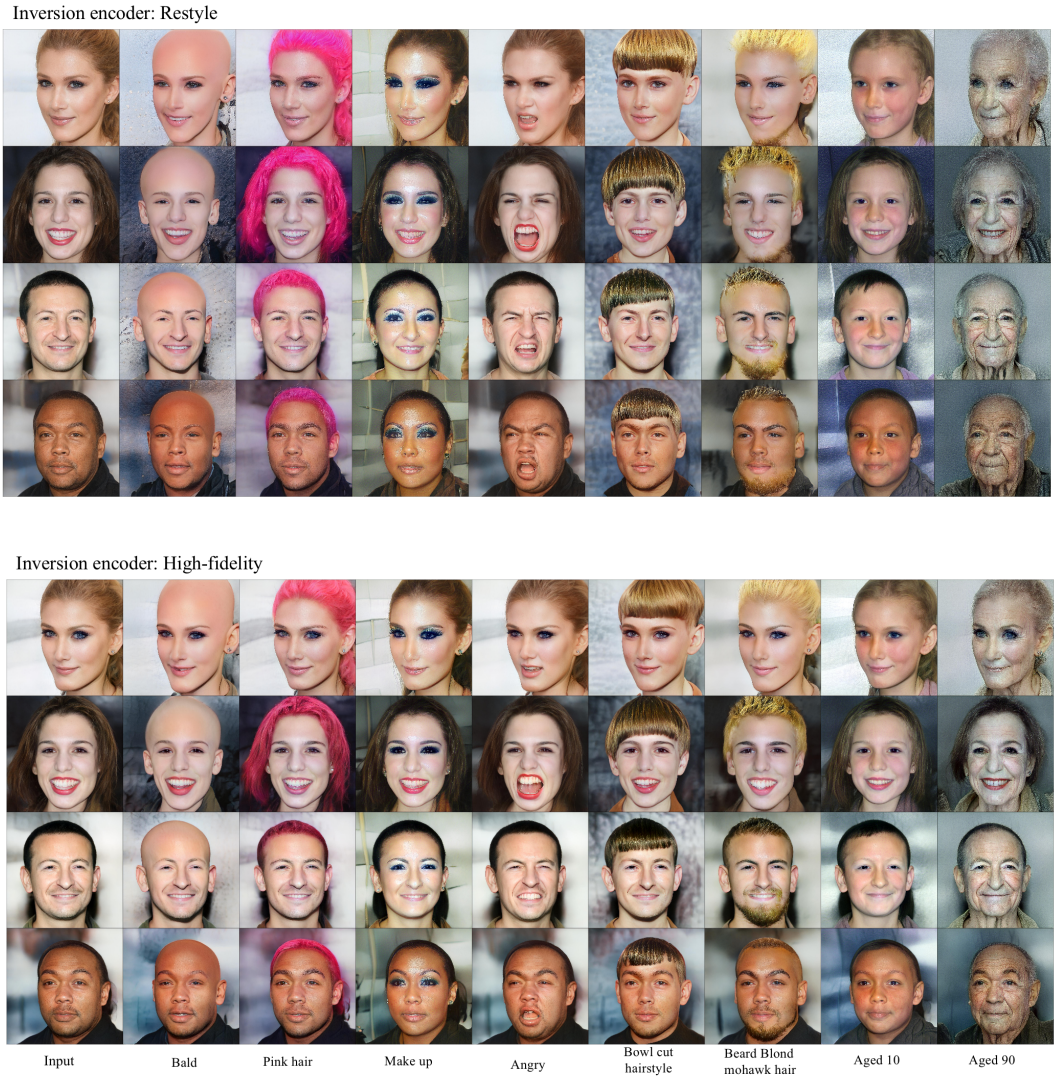

Figure 12: We train our model with e4e inversion encoder and test our model with Restyle [1] and High-fidelity [2] inversion encoders, respectively. The FFCLIP demonstrates good generalizability on these encoders.

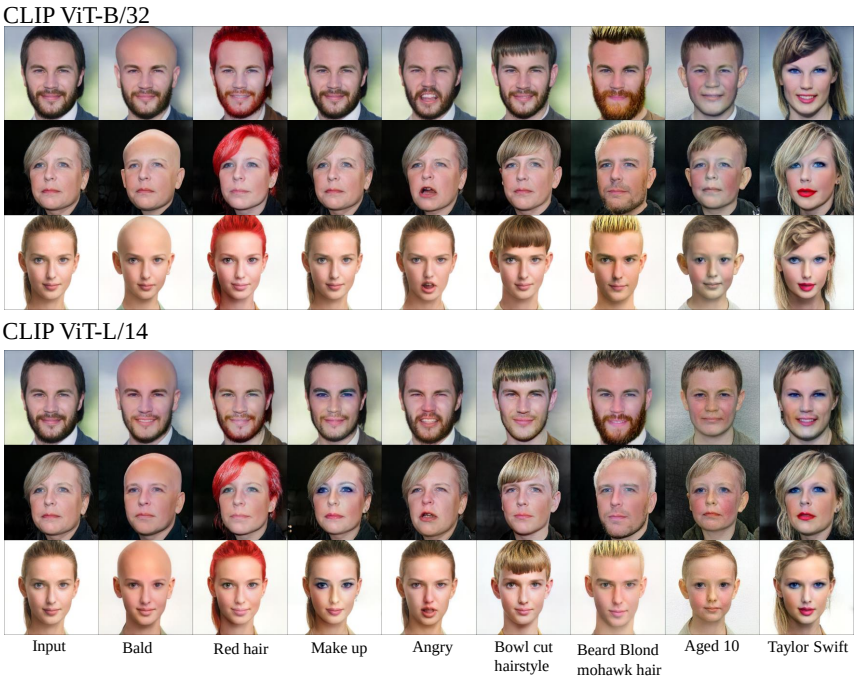

Figure 13: Visual comparison results between different CLIP models. We follow the StyleCLIP and use the ViT-B/32 in our method.

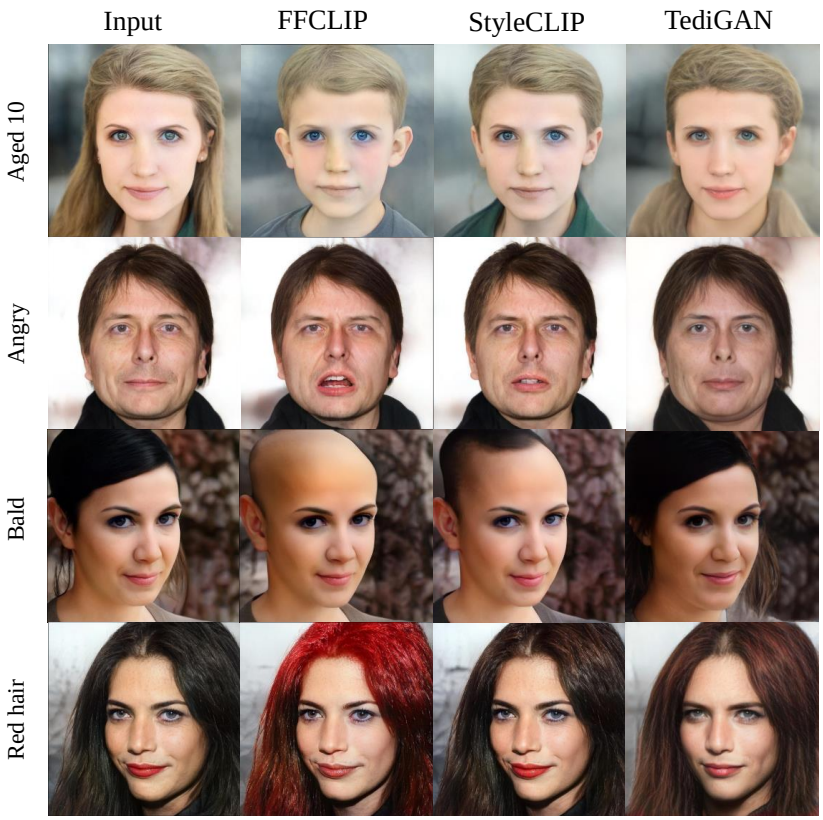

Figure 14: More visual comparison results.

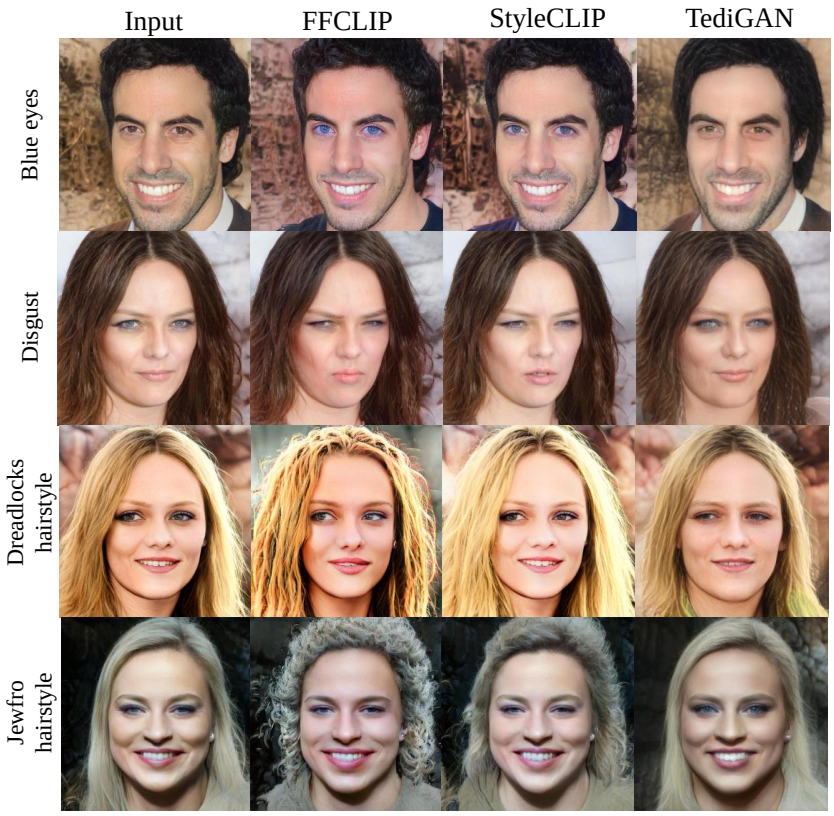

Figure 15: More visual comparison results.

Figure 16: Visual comparison results between StyleCLIP and our method in hair color manipulation.

References

[1] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[2] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[3] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. Hairclip: Design your hair by text and reference image. *arXiv preprint arXiv:2112.05142*, 2021.