Appendix

A. Implementation Details of Baselines

We compare our method with four baseline methods: Textual Inversion [17], DreamBooth [49], NeTI [2], and Celeb Basis [69]. For Textual Inversion, we use the diffusers implementation [63] with Stable Diffusion v2.1 as the base model. The textual embeddings are initialized with the embeddings of "human face". We perform 5,000 optimization steps using a learning rate of 5e-3 and a batch size of 8. For DreamBooth, we also use the diffusers implementation and tune the U-Net with prior preservation loss. We perform 800 fine-tuning steps using a learning rate of 2e-6 and a batch size of 1. For NeTI and Celeb Basis, we use their official implementations and follow the official hyperparameters described in their papers. Moreover, we apply the textual bypass and Nested Dropout [47] techniques for NeTI.

Table 3. The 20 prompts used in the quantitative evaluation.

a photo of a S^* person
a S^* person with a sad expression
a S^* person with a happy expression
a S^* person with a puzzled expression
a S^* person with an angry expression
a S^* person plays the LEGO toys
a S^* person on the beach
a S^* person piloting a fighter jet
a S^* person wearing the sweater, a backpack and camping stove, outdoors, RAW, ultra high res
a S^* person wearing a scifi spacesuit in space
a S^* person and Anne Hathaway are baking a birthday cake
a S^* person and Anne Hathaway taking a relaxing hike in the mountains
a S^* person and Anne Hathaway sit on a sofa
a S^* person and Anne Hathaway enjoying a day at an amusement park
a S^* person shakes hands with Anne Hathaway in news conference
cubism painting of a S^* person
fauvism painting of a S^* person
cave mural depicting a S^* person
pointillism painting of a S^* person
a S^* person latte art

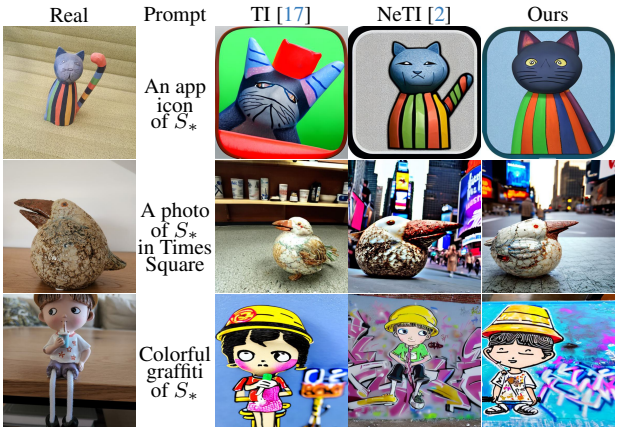

Figure 12. Results for general concepts.

B. Text Prompts

In Tab. 3, we list all 20 text prompts used in the quantitative evaluation. These prompts cover a range of modifications, including expression editing, background modification, individual interaction, and artistic style.

C. Results for General Concepts

While our primary focus is on the human face, we also present results for general concepts in Fig. 12. As illustrated, our method outperforms Textual Inversion [17] and shows promise for general concepts. We leave it as the future work.

D. Additional Qualitative Comparisons

In Fig. 13, we present further qualitative comparisons against baseline methods across a diverse set of prompts. Additionally, we compare our method with more baselines, including E4T [18], ProFusion [70], PhotoVerse [8], and FaceStudio [67], in Fig. 14.

E. Results of Our Fast Version Method

As illustrated in Sec. 5.2, we developed a fast version of our method with a learning rate of 0.08. This fast version enables learning of the new concept in 25 optimization steps, taking only 26 seconds. In Figs. 15 and 16, we provide qualitative results of applying this fast version to a variety of prompts. The results demonstrate that our fast version allows for high-quality personalized face generation within a remarkably short training time.

F. Additional Qualitative Results

In Fig. 17 and Fig. 18, we provide additional qualitative results obtained by our method on a diverse set of prompts.

G. Results on Synthetic Facial Images

Besides evaluating on real facial images, we also evaluate our method on synthetic facial images generated by StyleGAN. The results are shown in Fig. 19. As can be seen, our method achieves high-quality personalized face generation on synthetic facial images.

H. Additional Ablation Study Results

As illustrated in Sec. 5.3, our ablation study involves the individual removal of the following sub-modules: 1) Cross Initialization, 2) mean textual embedding, and 3) the regularization term. Additional ablation study results for each variant are presented in Fig. 20.

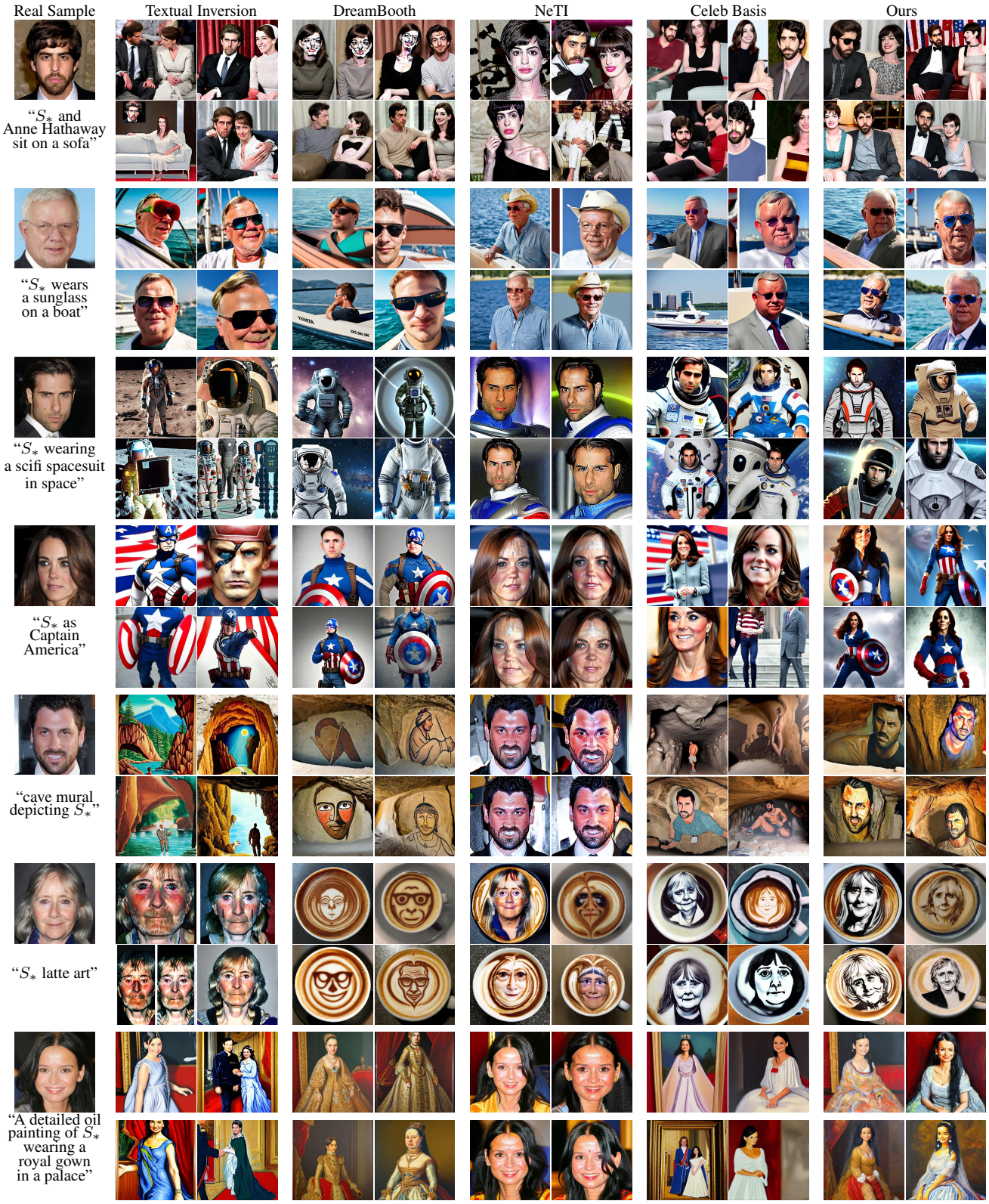

Figure 13. Additional qualitative comparisons. Given a single input image, we present four images generated by each method using identical random seeds. Our approach demonstrates superior performance in identity preservation and editability.

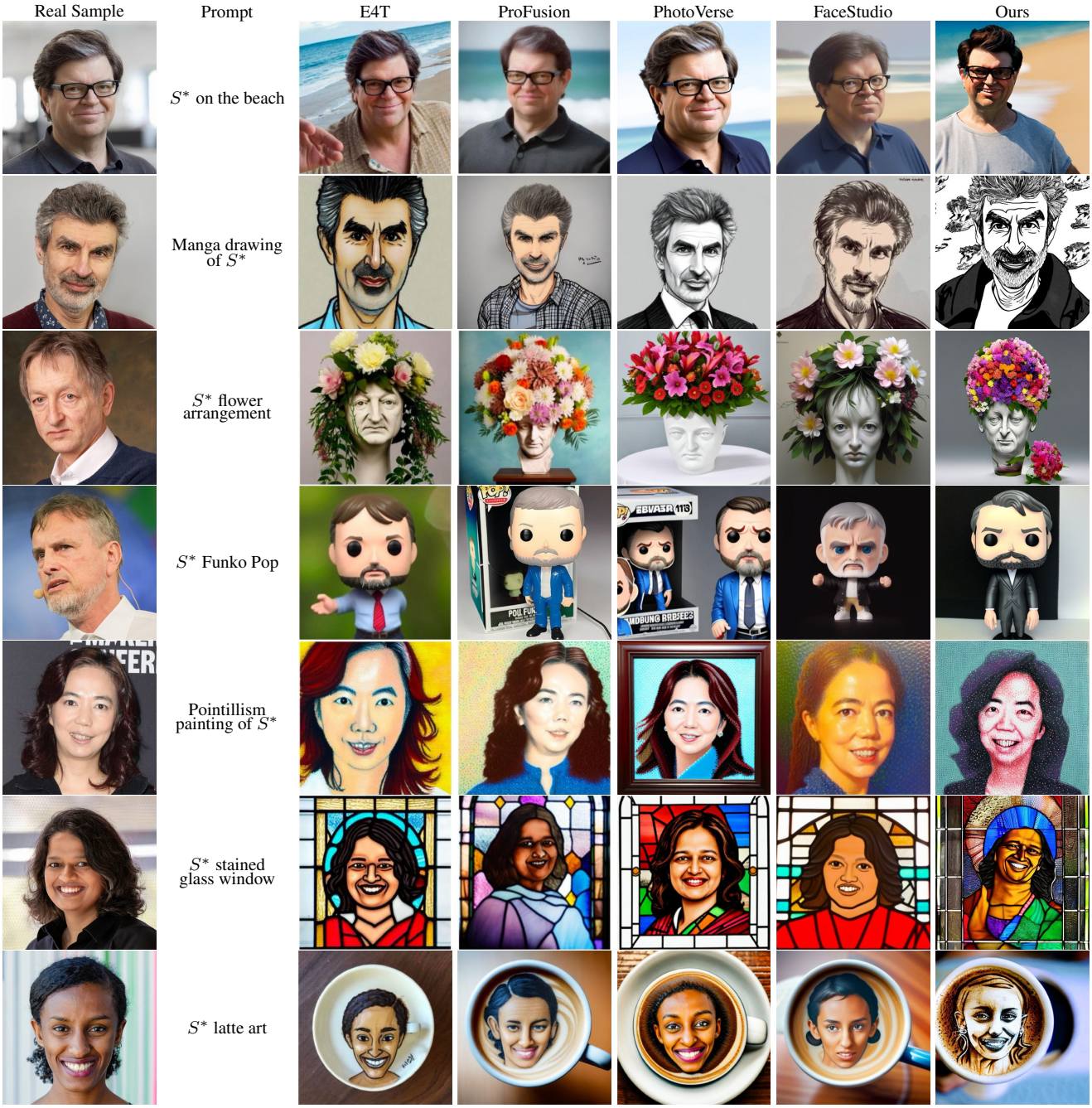

Figure 14. Qualitative comparisons to more baselines, including E4T [18], ProFusion [70], PhotoVerse [8], and FaceStudio [67].

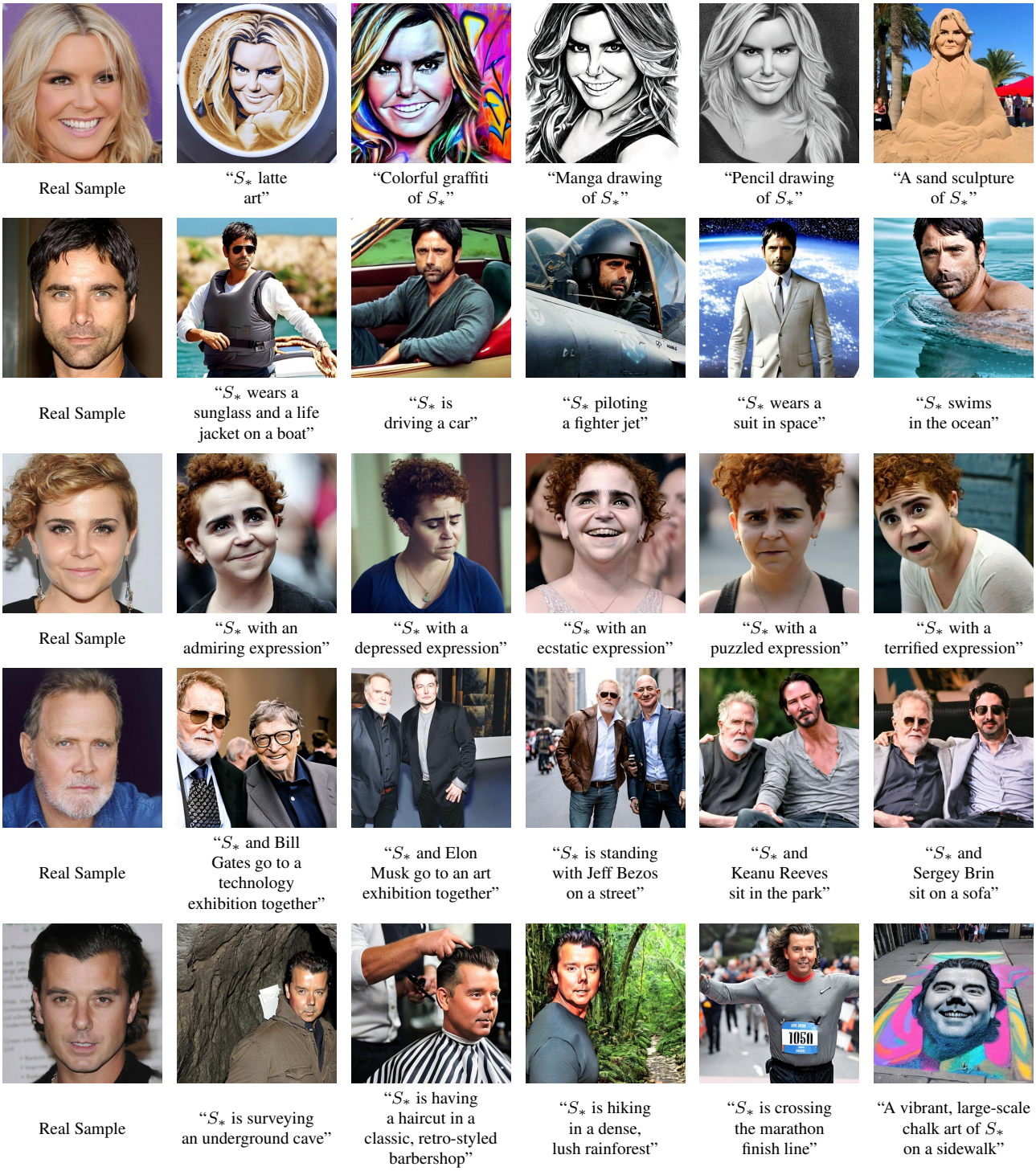

Figure 15. Images generated by our fast version method with a learning rate of 0.08. Results are obtained after 25 optimization steps, taking only 26 seconds.

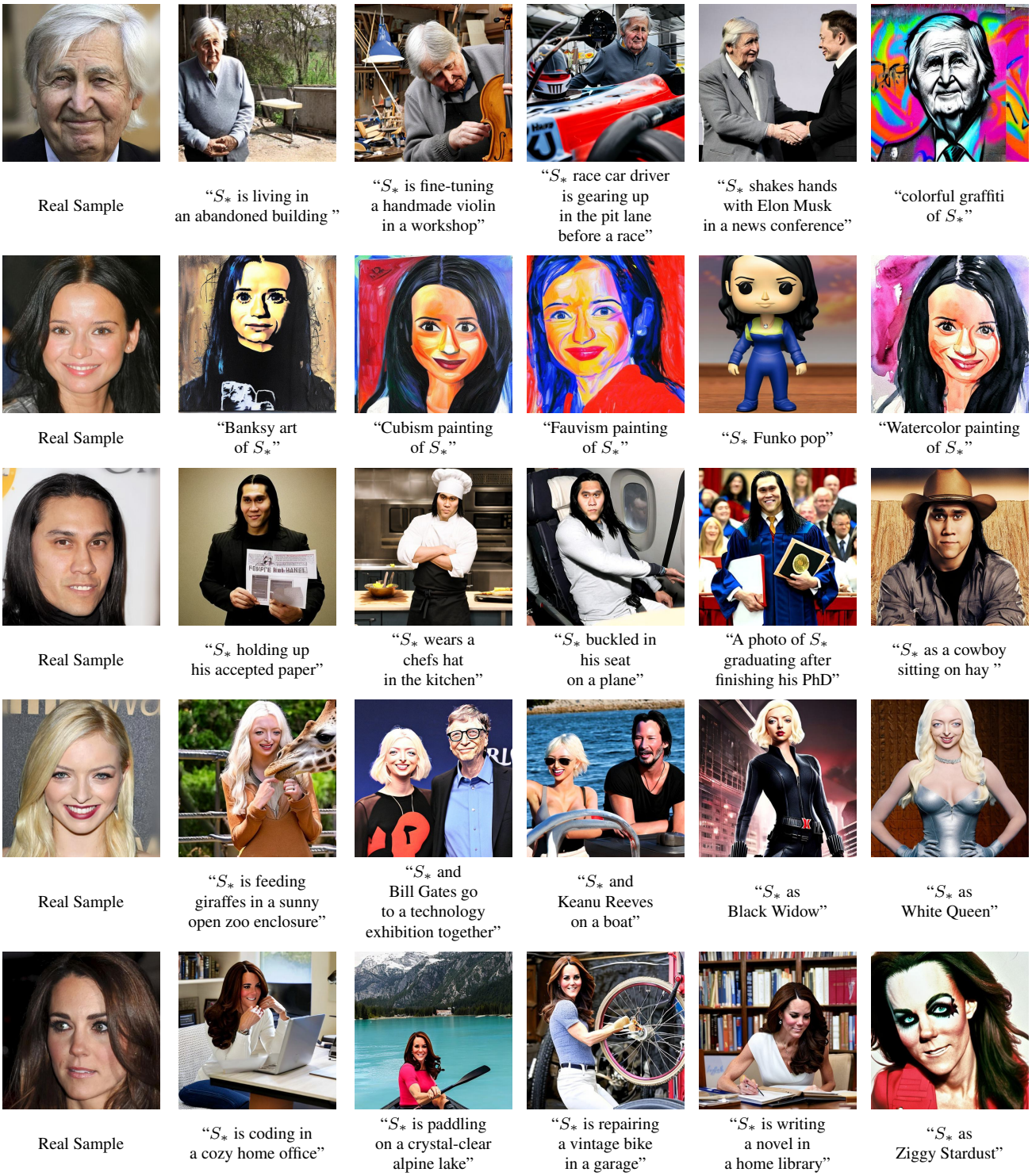

Figure 16. Images generated by our fast version method with a learning rate of 0.08. Results are obtained after 25 optimization steps, taking only 26 seconds.

Figure 17. Additional examples of personalized text-to-image generation obtained with Cross Initialization.

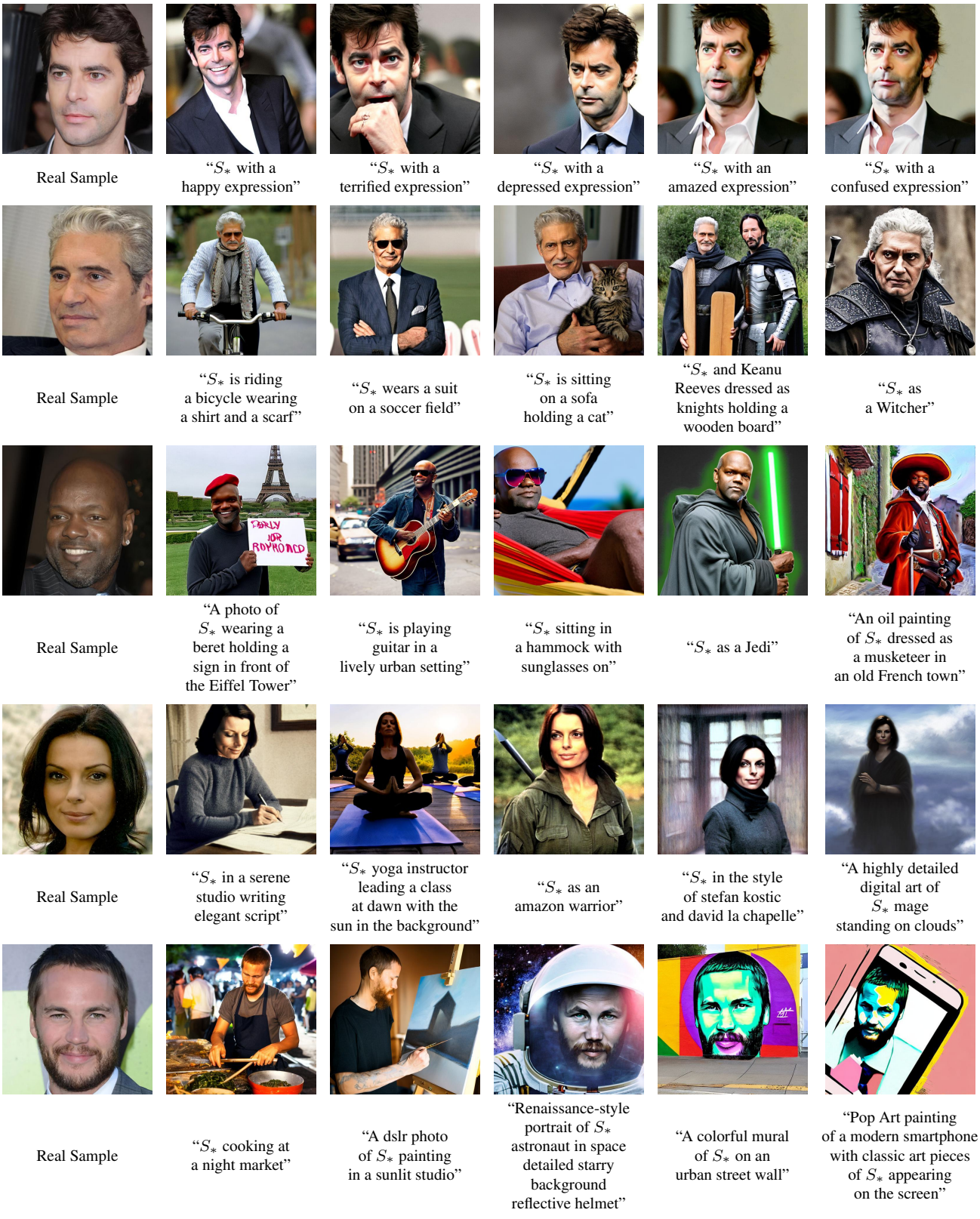

Figure 18. Additional examples of personalized text-to-image generation obtained with Cross Initialization.

Figure 19. Additional results on synthetic facial images generated by StyleGAN, where the input images are sourced from [69].

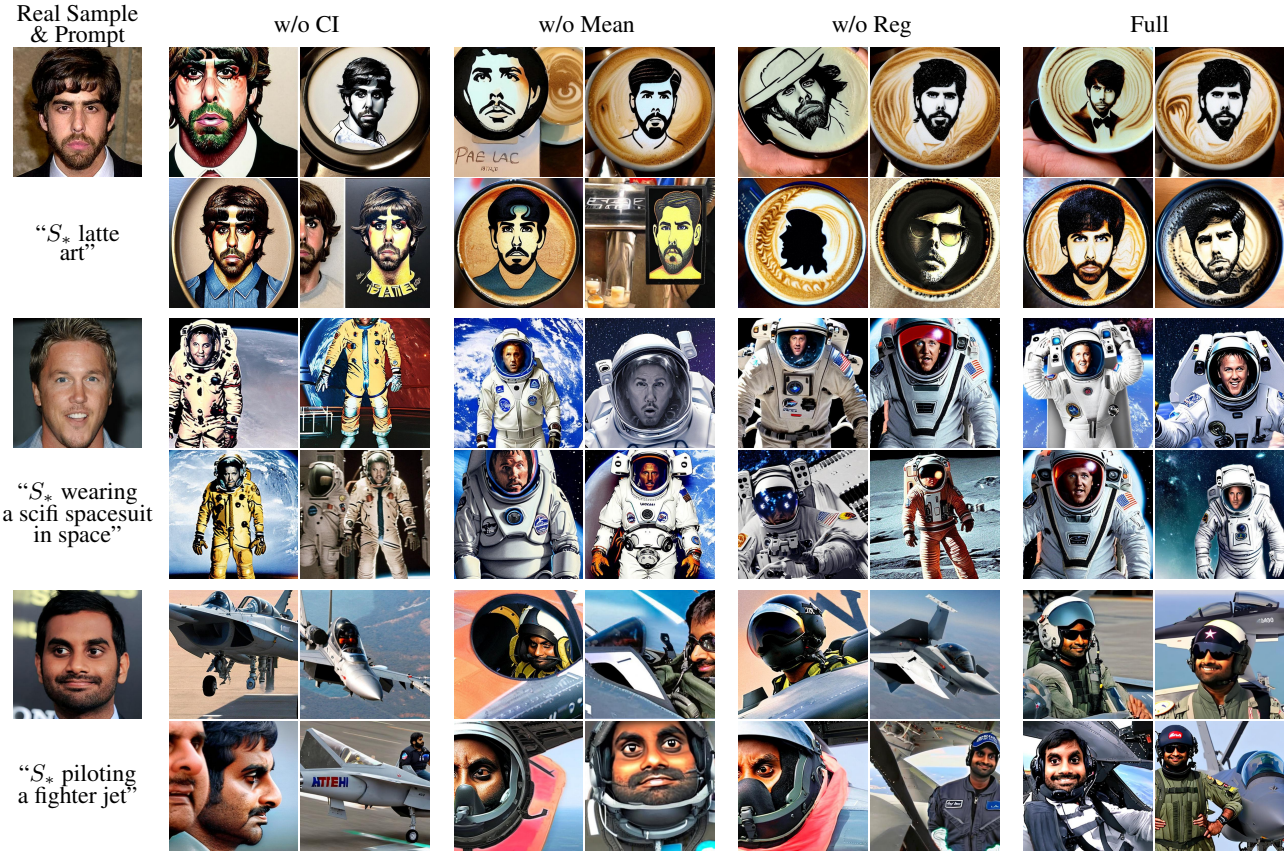

Figure 20. Additional ablation study. We compare the models trained without Cross Initialization (w/o CI), without mean textual embedding (w/o Mean), and without regularization (w/o Reg). As can be seen, all sub-modules are essential for achieving identity-preserved and prompt-aligned personalized face generation.