When StyleGAN Meets Stable Diffusion: a $\mathcal{W}_+$ Adapter for Personalized Image Generation

Supplementary Material

This supplemental material mainly contains:

- Details of mapping network $\mathcal{F}_w$ in Section I
- Attribute editing in previous methods in Section II
- Analyses of two-stage training in Section III
- Visualization of residual cross-attention in Section IV
- Text prompt for training Stage I in Section V
- More generation and editing results in Figs. E and F
- Performance on other SD models in Figs. G and H
- More visual comparison with competing methods in Fig. I

I. Details of Mapping Network

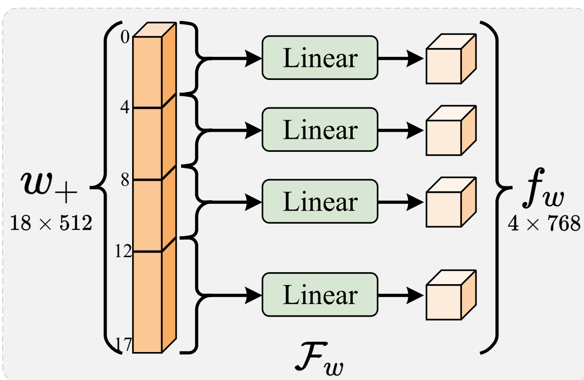

Figure A. Details of our mapping network $\mathcal{F}_w$.

The trainable modules of our $\mathcal{W}_+$ adapter consist of two parts, 1) the mapping network with 7.1 M parameters, and 2) residual cross-attention with 31.6 M parameters. Details of our mapping network $\mathcal{F}_w$ are shown in Fig. A. To align with the input dimension of Stable Diffusion, the w_+ vector is divided into four groups. Each group is projected to a token of dimension 768 for Stable Diffusion V1.*.

II. Attributes Editing in Previous Methods

Fig. B shows that CelebBasis encounters challenges in preserving the identity details. Both FastComposer and IP-Adapter-Face) tend to overlook the facial attribute descriptions provided in the text captions. In contrast, our method not only excels in preserving identity but also edits common attributes with a seamless outcome.

III. Analyses of Two-stage Training

The training of our $\mathcal{W}_+$ adapter contains two stages. In Stage I, the model learns a mapping from $\mathcal{W}+$ to the SD latent space. Subsequently, in Stage II, the mapping network is fixed, and only the residual cross-attention is fine-tuned to facilitate in-the-wild generation. It is noteworthy that adopting a one-stage training approach, which directly optimizes the mapping network and residual cross-attention for in-the-wild generation, results in challenges in preserving consistent layout when editing attributes (see the 1-*st* row in Fig. C). This observation suggests that directly embedding the w_+ vector into diverse and complex in-the-wild generation may struggle to align well with the $\mathcal{W}_+$ space, thereby showing the necessity of our two-stage training approach.

Figure B. Results of competing methods with text attributes.

IV. Visualization of Residual Cross-attention

To demonstrate the influence of our residual cross-attention on the hidden states $f'z$, we visualize the cross-attention score of the last layer by computing $\mathrm{Softmax}\left(\frac{\mathbf{Q}^{\dagger}\mathbf{K}^{\dagger\top}}{\sqrt{d}}\right)$ at time step $t = 1$. In Fig. D, we observe an obvious impact of our $w+$ vector on the facial region, with minimal impact on the background. This observation showcases the effectiveness of our $\mathcal{W}_+$ adapter for editable face generation.

V. Text Prompt for Training Stage I

The text prompt describing a human face in Stage I includes:

- a face
- a photo of a face
- a close-up of a face
- a depiction of a face
- a good photo of a face
- a photography of a face
- a cropped photo of a face
- a good photography of a face
- a close-up photography of a face

Figure C. Comparison between one- and two-stage training.

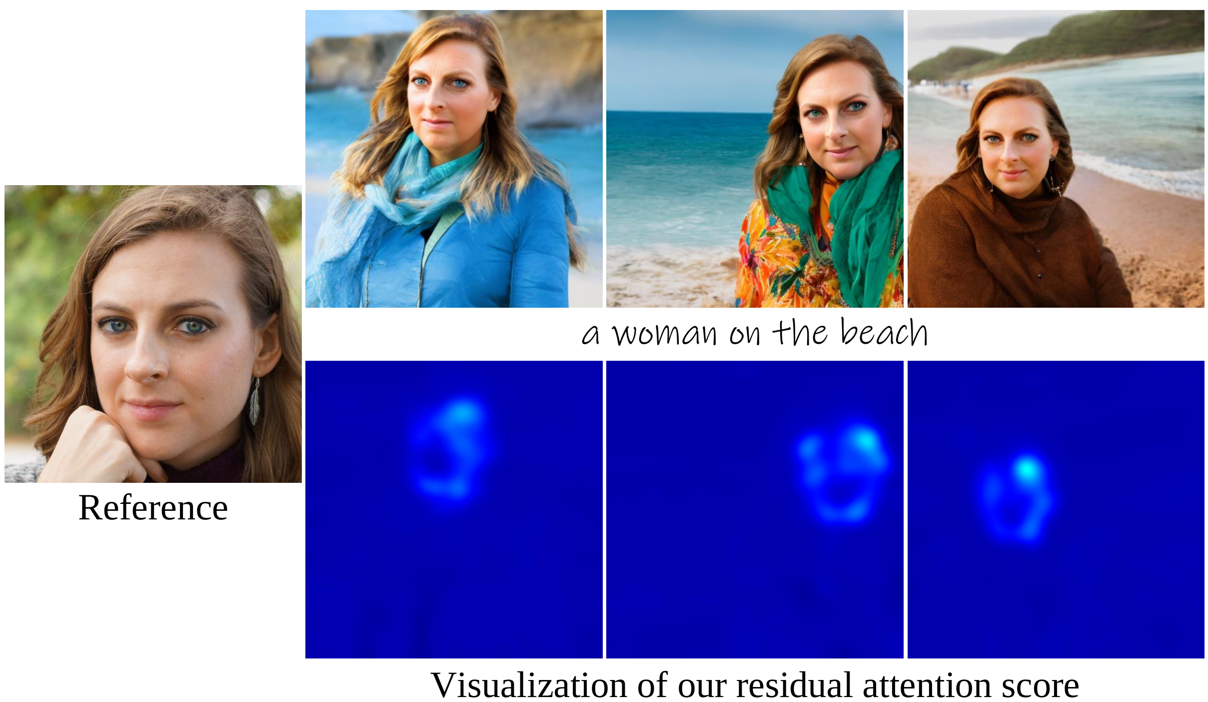

Figure D. Visualization of residual cross-attention score.

Figure E. More results of our $\mathcal{W}_+$ adapter for in-the-wild generation with different scenarios and attribute editing.

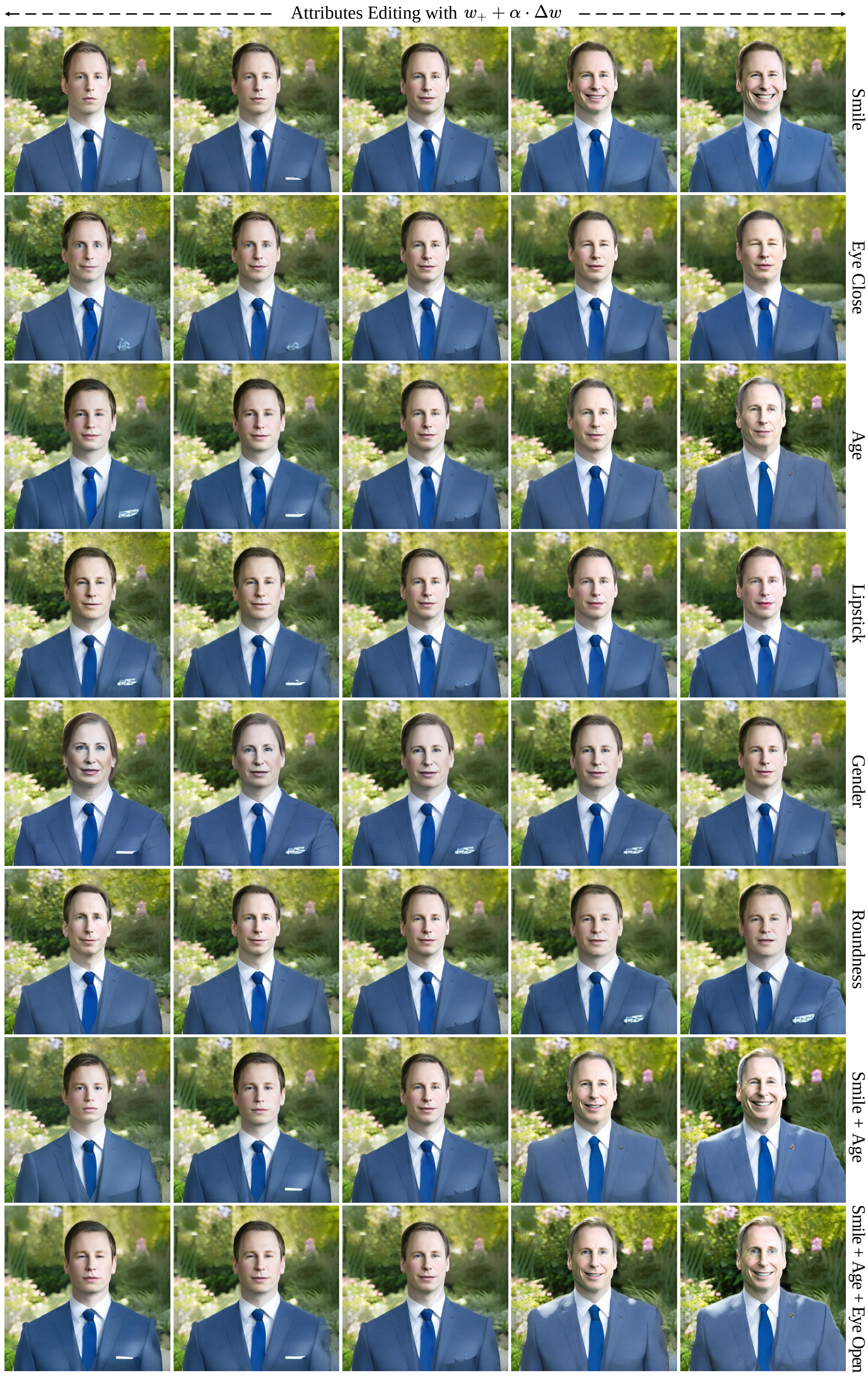

Figure F. More results of attribute editing for a single reference.

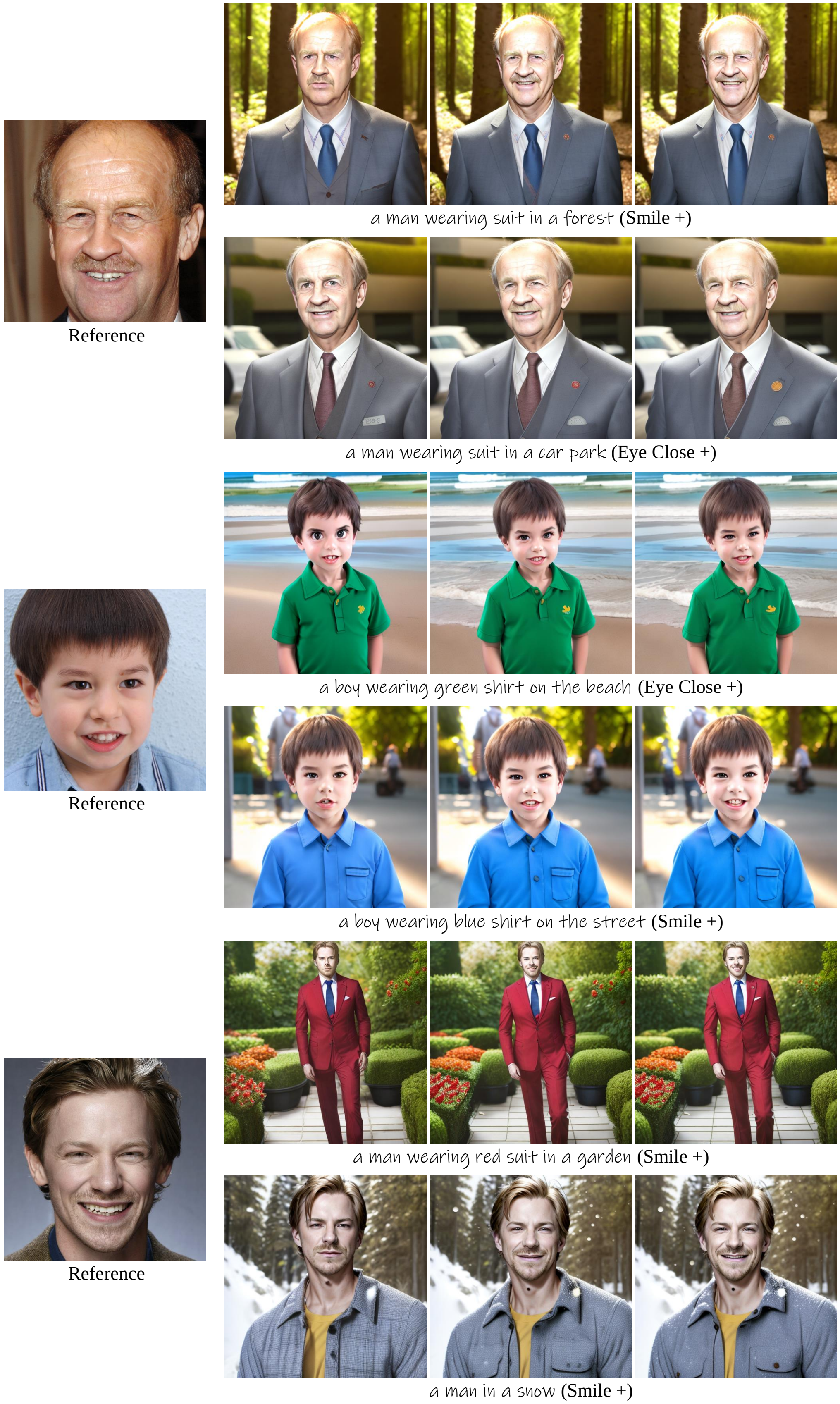

Figure G. Results of our $\mathcal{W}_+$ adapter using other SD model (*i.e.*, Protogen).

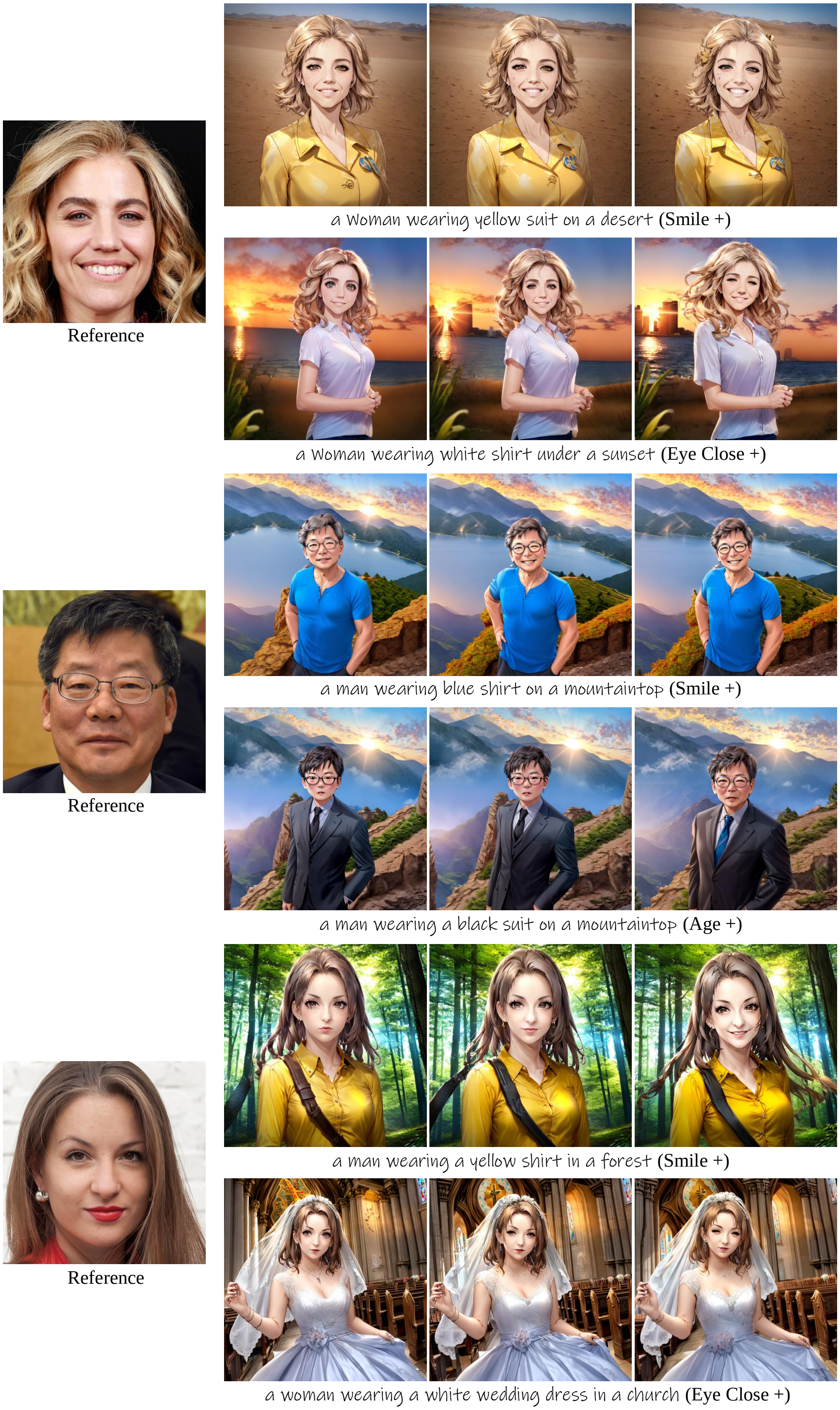

Figure H. Results of our $\mathcal{W}_+$ adapter using other SD model (*i.e.*, dreamlike-anime).

Figure I. More visual comparison with competing methods in different scenarios.