Appendix

Experiment Details

Data Preparation. We synthesized 500k images with StyleGAN2 [4] and scored 6 attributes(gender,smile,eyeglasses,age,lipstick,beard) with CelebA attribute classifiers [3]. In **Figure 2a**, for each attribute, we compute the average of 1000 images in which the corresponding classifier predicts the attribute class with the highest confidence. After applying our method to [1] and editing such sets of images to achieve the opposite attribute class, we invert them back to the $\mathbf{W}$ space and re-generate the self-corrected samples for **Figure 3b**. Similarly, for **Figure 3c**, we sample 10k latent codes corresponding to images with the highest classifier confidence for predicting eyeglasses and gender.

Latent Interpolation Methods. For **Figure 4**, we train both [5] and [1] on the same set of 1000 latent code samples with the highest classifier confidence for each attribute. For [2] and [6], we use the original directions as presented in the original paper,and we use the channel for "grey hair" as the Age+ channel for [6].

Attribute Dependency. First, we sample 3000 test images with all attributes of interest(gender, smile, eyeglasses, age, lipstick, beard) lying around the attribute classifiers' decision boundaries. We split the images into 5 test sets AD calculation. We present the full procedure to calculate AD on each attribute a as follows:

- For each set of images with target attribute $a \in A$, where A stands for all attributes, we interpolate the original latent codes following [5] and [1] for $d = 6$ in 9 steps.
- For each interpolation result at step s, we compute $x = \frac{\Delta l_s^a}{\sigma l^a}$, which stands for the absolute change in the target attribute logit, normalized by the population standard deviation and obtain the x-values for plotting AD.
- For each interpolation result at step s, we also compute $y = \frac{1}{|A|-1} \Sigma_{i \in A \setminus a} \frac{\Delta l_s^i}{\sigma l^i}$, which stands for the mean of the absolute change in the other attribute logits, normalized by each population standard deviation, and obtain the y-values for plotting AD.
- We then group the (x,y) pairs with their x values into buckets of $(0, 0.25], (0.25, 0.5], \ldots, (1.75, 2]$, and plot the midpoint for each bucket as the final x-value, mean of y values within each bucket as the final y-value.

References

[1] Zikun Chen, Ruowei Jiang, Brendan Duke, Han Zhao, and Parham Aarabi. Exploring gradient-based multi-directional controls in gans. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII*, pages 104–119. Springer, 2022. 1

[2] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020. 1

[3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[4] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1

[5] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1

[6] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021. 1