LCM-Lookahead for Encoder-based Text-to-Image Personalization - Supplementary Materials

1 Additional comparisons

1.1 Celebrity Comparisons

In fig. 1 we show additional comparisons against the baseline methods, using celebrity inputs. Most baselines succeed in preserving the identities of all celebrities. The baseline IP-Adapter [17] variants still struggle with stylization prompts, and show background leak (though the latter can be fixed through appropriate background masking). Our results show good identity preservation while providing high editability.

1.2 Comparisons on InstantID paper results

In fig. 2 we expand Fig. 5 of the InstantID paper [16] with the results of PhotoMaker [5] and our own method. Here too, our method shows improvement over the prior art, and particularly the IP-Adapter variations on which it is based. While we were unable to verify this due to the LAION [11] dataset being withdrawn, a search through `https://haveibeentrained.com/` indicates that some of these individuals (e.g., Yann Lecun) may be included dozens of times in InstantID's reported training set. IP-Adapter did not report their training set for their FaceID versions. However, their diminished performance on these individuals hints that they did not observe them.

2 Maintaining model alignment

In the core paper, we note that applying the LCM-lookhead loss naively over extended training leads to a breaking of the alignment between the LCM [7] and non-LCM models. To avoid this, we investigated three options for preserving model alignment: (1) Applying a score distillation sampling (SDS) loss [9] to the LCM-model outputs, while using the non-LCM model to estimate the score function. (2) Adding the standard Consistency Model [13] loss to the objective of the LCM-path's outputs, and (3) scaling the LCM-LoRA weights during training. Finally, we also investigate the results when avoiding any alignment-preserving mechanism.

The results are provided in Tab. 1. Notably, without using any alignment-preserving objective, our method shows only mild improvement over not using an identity loss at all. Introducing the SDS objective improves matters, but

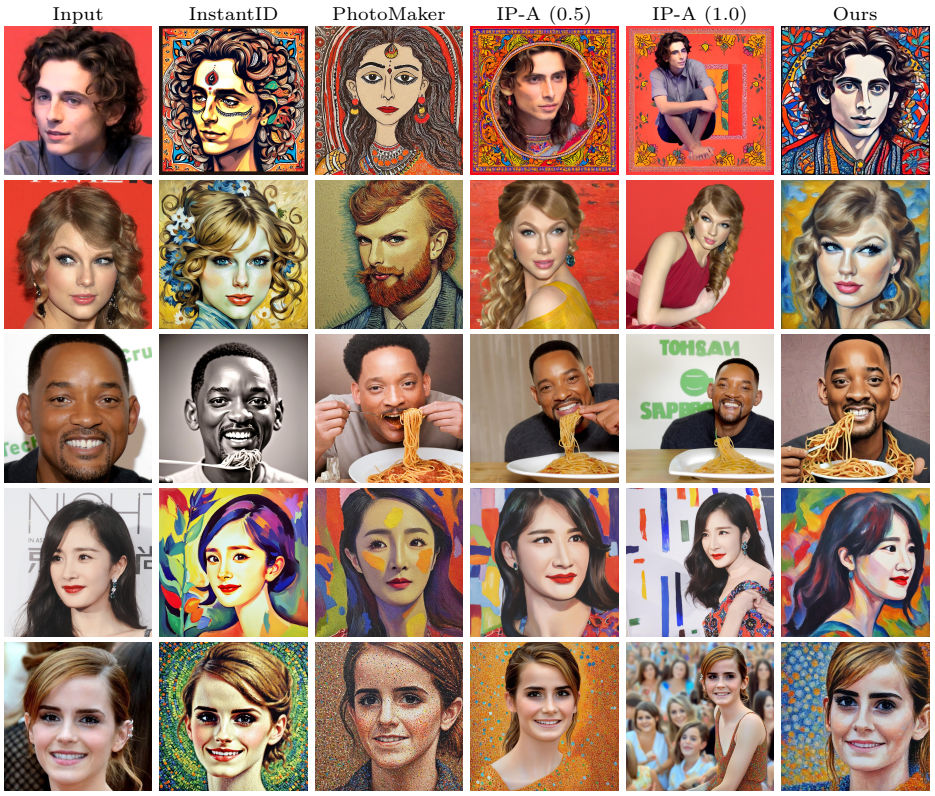

Fig. 1: Comparisons against prior and concurrent face-personalization encoders on celebrity data. IP-A (1.0) and (0.5) represent the IP-Adapter results with a scale of 1.0 and 0.5, respectively. IP-A (1.0) serves as the backbone which we fine-tune.

still under-performs the alternatives, including simply passing the identity loss through the single-step DDPM [2] approximation. Finally, using a consistency loss or the LoRA scaling approach leads to the best alignment preservation and hence the best downstream performance. Of the two, we settled with LoRA scaling because it provides improved results, is much simpler to implement in practice, and has negligible impact on compute requirements.

3 Choice of synthetic data

Here we report evaluation results for models trained on different datasets or synthetic data generation options. Specifically, we consider: (1) Synthetic identities created by leveraging SDXL Turbo's [10] mode collapse. (2) Celebrities created using SDXL's [8] prior knowledge. (3) ConsiStory [15], a training-free consistent subject generation method built on SDXL, and (4) the CelebA dataset [6]. The results on FFHQ [4] and Unsplash-50 are shown in Tab. 2.

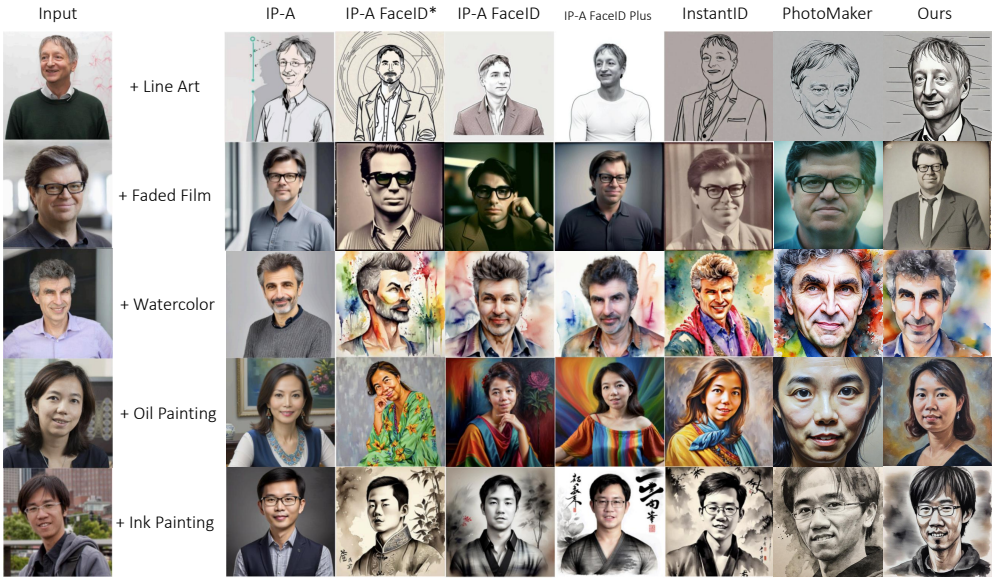

Fig. 2: We expand Fig. 5 of the InstantID [16] paper with PhotoMaker [5] and our own results. Here, we keep the column terminology employed by the original InstantID paper. Hence, IP-A refers to IP-Adapter-SDXL, IP-A FaceID* is the experimental version of IP-Adapter-SDXL-FaceID. IP-A FaceID is the IP-Adapter-SD1.5-FaceID model, and IP-A FaceID Plus is the IP-Adapter-SD1.5-FaceID-Plus model. Note that the last two models are based on SD1.5 and not on SDXL.

Notably, the CelebA dataset and the ConsiStory results contain only photo-realistic images (the latter because of its limitation in changing styles across the batch). This leads to diminished prompt-alignment, showing the importance of training on data for which the encoder is not pushed to adhere to a fixed output style.

4 Stylistic Diversity

While encoder-based personalization methods have achieved pleasing results, they still may under-perform optimization-based approaches on the diversity of styles that they can achieve. In fig. 3 we investigate how well our approach performs under more difficult style changes. As can be seen, our approach can successfully handle non photo-realistic renders and harder texture changes that are reflected in the structure of the image, such as mosaics or origami styles. However, our approach may struggle with styles that require large shape changes, such as the Minecraft style or PS1 style renders.

Table 1: Comparison of different alignment preservation approaches.

	FFHQ-5000		Unsplash-50	
	ID ↑	CLIP-T ↑	ID ↑	CLIP-T ↑
LoRA Scaling	0.345	26.33	0.308	26.79
LCM Loss	0.324	26.63	0.280	27.05
SDS	0.267	27.03	0.232	27.16
No Alignment	0.246	27.46	0.220	27.48

Table 2: Comparison of real and synthetic data generation options

	FFHQ-5000		Unsplash-50	
	ID ↑	CLIP-T ↑	ID ↑	CLIP-T ↑
SDXL Turbo	0.345	26.33	0.308	26.79
Synthetic Celebs	0.334	26.37	0.299	26.81
ConsiStory	0.358	24.87	0.303	25.44
CelebA	0.345	25.81	0.294	26.14

Fig. 3: Our method can generate more complex visual styles, but can struggle with significant shape changes (e.g., Minecraft's block-based style).

5 Limitations

As an encoder, our model must learn to generalize from its training data. This places limits on its ability to adapt to concepts which were sufficiently rare (or entirely unseen) during training, such as unusual makeup (fig. 4, right).

Additionally, our model posses unique limitations that may not be shared by prior art. Our training data contains a higher portion of non photo-realistic samples. Hence, the model may default to stylized results more often than the baseline (fig. 4, middle).

Finally, the model attempts to capture non-identity data from the image, such as accessories. Hence, providing it with a photo of a person wearing headphones will drive it to generate more photos with headphones. However, the model struggles to preserve the exact details of such accessories (fig. 4, left).

6 Additional implementation details

6.1 Classifier Free Guidance details

When using the full version of the model with attention-based key and value expansion, we find it useful to modify the classifier free guidance (CFG, [3]) formulation to:

$$\begin{aligned}\epsilon = \epsilon_{uncond} &+ s_{no-kv} \cdot (\epsilon_{no-kv} - \epsilon_{uncond}) \\ &+ s_{full} \cdot (\epsilon_{full} - \epsilon_{uncond}) \\ &+ s_{kv} \cdot (\epsilon_{kv} - \epsilon_{uncond}) ,\end{aligned} \tag{1}$$

Fig. 4: Limitations: **(left)** Our model captures accessories such as hats or headphones as part of the character. However, it does not accurately reproduce them in novel prompts. **(middle)** Our model may add stylization to output images even when not prompted for it. **(right)** The model fails to preserve tail concepts, such as excessive makeup.

where ϵ_{no-kv} is the denoiser's prediction when the keys and values are not extended (*i.e.* using $K^l_{z_{r,t}}$, $V^l_{z_{r,t}}$), and ϵ_{kv} is the denoiser's output when these keys and values are used, but all other conditioning codes (text, IP-Adapter) are set to their null value. ϵ_{full} includes all conditioning inputs, and ϵ_{uncond} includes null conditions for all. The scale parameters were set empirically to $s_{no-kv} = 3.0$, $s_{full} = 2.0$, $s_{kv} = 2.0$.

To facilitate CFG, we follow IP-Adapter and continue dropping the adapter and text conditions in 10% of iterations. For the expanded attention path, we draw inspiration from ConsiStory [14] and randomly drop 5% of keys and values at every iteration.

6.2 Additional parameters

When expanding the cross attention mechanism, we follow prior art [1, 15] and enable FreeU [12] at inference time. We use the following parameters: $b1 = 1.1$, $b2 = 1.1$, $s1 = 0.9$, $s2 = 0.2$. Without FreeU, blur or low resolution artifacts are more often inherited from the conditioning image.

7 Evaluation details

For all Unsplash-50 evaluations and figures in the core paper, we generated a single image of each combination of identity and prompt. In the quantitative evaluations, we compute the automatic metrics over the entire set. For the user study, we randomly sampled an identity and prompt combo for each question. For qualitative figures, here and in the core paper, we manually selected 14 of these identities.

For FFHQ evaluations, generating images with all prompts for every identity would take weeks. Hence, we randomly sampled one prompt from our set for each identity, and calculated the metrics over all $5,000$ image-prompt pairs. The same image-prompt pairs were used for all baselines and ablations.

We used the following prompts, to ensure coverage of both background changes and re-contextualization (which prior methods handle well), and style modifications (which they struggle with):

– “A photo of a face"
– “A pencil drawing of a face"
– “A face riding a bicycle"
– “A face as a Pixar character"
– “A face as a samurai in medieval japan"
– “An oil painting of a face"
– “A painting of a face in the style of Van Gogh"
– “A sculpture of a face"
– “A pop figure of a face"
– “A face on a billboard in times square"
– “A face in an astronaut suit"
– “A face piloting a fighter jet"
– “A face dressed as a superhero"
– “A face in papercraft style"
– “A digitial art painting of a face"
– “A cubism painting of a face"
– “A garden gnome of a face"
– “A grainy old time photo of a face"

For methods which require specific keywords in the prompt we replaced the subject word “face" with an appropriate reference to the keyword (e.g. “face img" for PhotoMaker).

References

1. Alaluf, Y., Garibi, D., Patashnik, O., Averbuch-Elor, H., Cohen-Or, D.: Cross-image attention for zero-shot appearance transfer (2023)
2. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
3. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021)
4. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
5. Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M.M., Shan, Y.: Photomaker: Customizing realistic human photos via stacked id embedding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
6. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
7. Luo, S., Tan, Y., Patil, S., Gu, D., von Platen, P., Passos, A., Huang, L., Li, J., Zhao, H.: Lcm-lora: A universal stable-diffusion acceleration module (2023)

8. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=di52zR8xgf
9. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=FjNys5c7VyY
10. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation (2023)
11. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
12. Si, C., Huang, Z., Jiang, Y., Liu, Z.: Freeu: Free lunch in diffusion u-net. arXiv preprint arXiv:2309.11497 (2023)
13. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023)
14. Tewel, Y., Gal, R., Chechik, G., Atzmon, Y.: Key-locked rank one editing for text-to-image personalization. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
15. Tewel, Y., Kaduri, O., Gal, R., Kasten, Y., Wolf, L., Chechik, G., Atzmon, Y.: Training-free consistent text-to-image generation (2024)
16. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A.: Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024)
17. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)