Supplementary Material for HyperDreamBooth: HyperNetworks for Fast Personalization of Text-to-Image Models

Additional Experiments

Qualitative Comparisons We provide more extensive qualitative comparisons over 10 identities of SFHQ and different stylistic prompts for our method, E4T [1] and concurrent work Face0 [4]. We show results in Figure 1. Our method typically demonstrates high editability and maintains identity effectively in the single-reference scenario, outperforming other methods. While E4T and Face0 yield impressive results, they often suffer from overfitting to realistic faces, which restricts their editability. Additionally, these methods, and especially Face0, tend to produce frontal outputs that closely resemble the pose of the reference input image.

Quantitative Comparisons and Ablations In this section, we present a quantitative comparison of our method with the E4T method [1] on a subset of 10 identities from the SFHQ dataset, with E4T kindly provided by the authors. We evaluate performance in generating stylistic portraits for 20 style prompts. Our evaluation metrics include face identity preservation ("Face Rec." from a VGGFace2 Inception ResNet), subject fidelity (assessed using DINO and CLIP-I metrics), and prompt fidelity (evaluated using CLIP-T). The results of this comparison are detailed in Table 1. Note that our method is trained on 15k identities vs. 100k identities for E4T. WE note that these metrics can be relatively suboptimal in the stylized portrait scenario, with face recognition metrics overfitting to the realistic domain, and DINO and CLIP-I metrics strongly preferring very similar structure or high-level semantics instead of focusing on specific subject and style details of images. In the main paper we include what we would consider a stronger user study experiment, where users choose the method that they prefer in terms of subject and style fidelity.

Our analysis reveals that our method obtains higher face identity and prompt fidelity metrics compared to E4T. In terms of similarity to the reference subject image, measured by DINO and CLIP-I metrics, our method shows lower scores compared to E4T. This can be due to the fact that E4T outputs often have very strong structural resemblance to the reference face image (very similar pose, size of face in the image, realistic face). It is known that DINO and CLIP-I metrics favor structure and semantics to original images respectively, instead of nuanced subject and style features. This is not necessarily something we want in the generation of a new stylized portrait of a face, with variation in pose, expression and style of a face being preferred in many applications.

Limitations

Given the statistical nature of HyperNetwork prediction, some samples that are OOD for the HyperNetwork due to lighting, pose, or other reasons, can yield subotpimal results. Specifically, we identity three types of errors that can occur. There can be (1) a semantic directional error in the HyperNetwork's initial prediction which can yield erroneous semantic information of a subject (wrong eye color, wrong hair type, wrong gender, etc.) (2) incorrect subject detail capture during the fast finetuning phase, which yields samples that are close to the reference identity but not similar enough and (3) underfitting of both HyperNetwork and fast finetuning, which can yield low editability with respect to some styles.

Societal Impact

This work aims to empower users with a tool for augmenting their creativity and ability to express themselves through creations in an intuitive manner. However, advanced methods for image generation can affect society in complex ways [3]. Our proposed method inherits many possible concerns that affect this class of image generation, including altering sensitive personal characteristics such as skin color, age and gender, as well as reproducing unfair bias that can already be found in pre-trained model's training data. The underlying open source pre-trained model used in our work, Stable Diffusion, exhibits some of these concerns. All concerns related to our work have been present in the litany of recent personalization work, and the only augmented risk is that our method is more efficient and faster than previous work. In particular, we haven't found in our experiments any difference with respect to previous work on bias, or harmful content, and we have qualitatively found that our method works equally well across different ethnicities, ages, and other important personal characteristics. Nevertheless, future research in generative modeling and model personalization must continue investigating and revalidating these concerns.

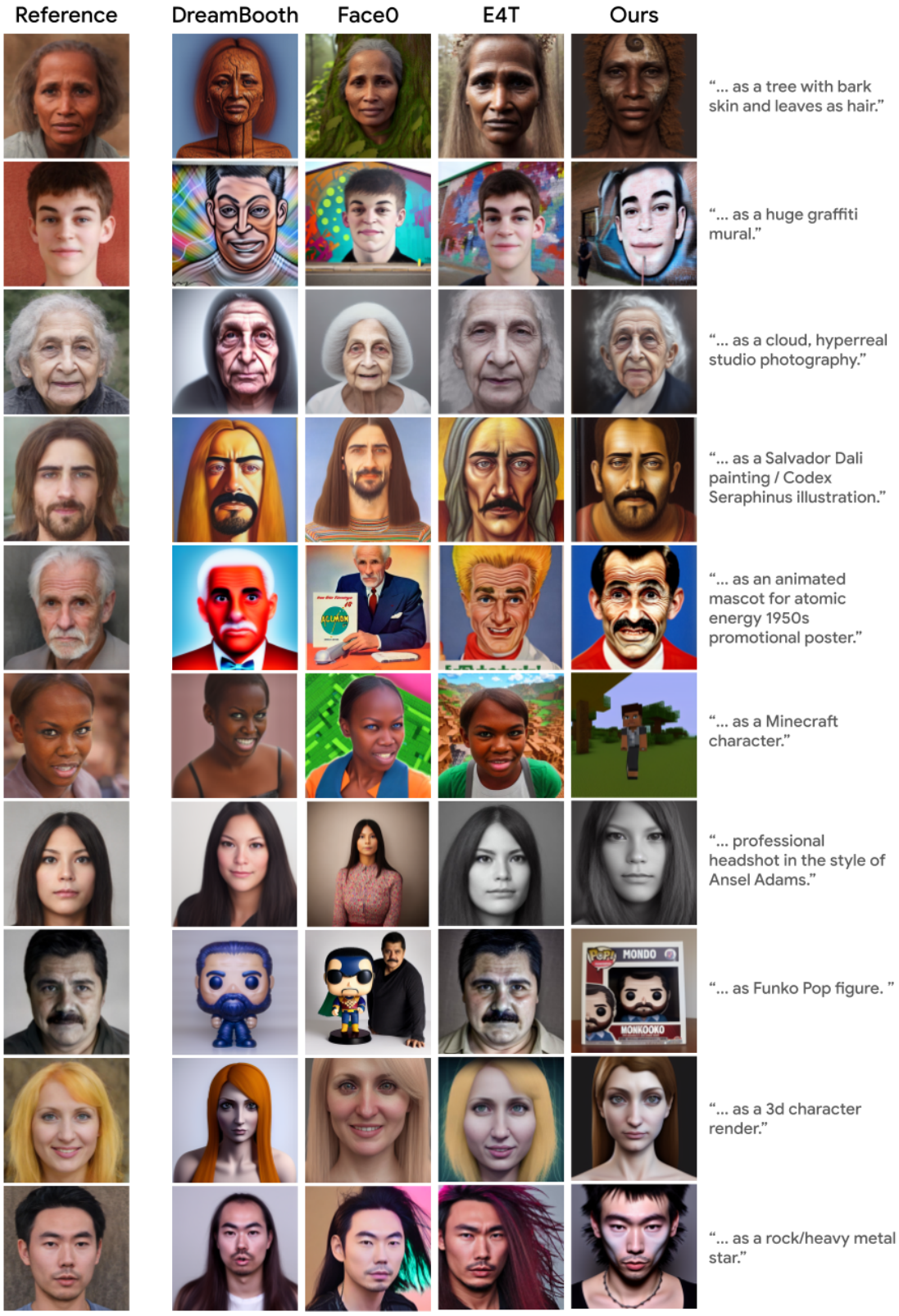

Figure 1. **Qualitative Comparison:** We compare random samples from DreamBooth [2], Face0 (concurrent work) [4], the state-of-the-art published work E4T [1] and our method (HyperDreamBooth).

Table 1. **Comparisons with E4T.** This table presents a comparative analysis of our method against E4T for face identity preservation (Face Rec.), subject fidelity (DINO, CLIP-I), and prompt fidelity (CLIP-T). Our method demonstrates higher performance in face identity preservation and prompt fidelity, indicating a closer adherence to the original identity and text prompts.

Method	Face Rec. ↑	DINO ↑	CLIP-I ↑	CLIP-T ↑
Ours	**0.7076**	0.528	0.628	**0.284**
E4T	0.693	**0.634**	**0.679**	0.280

References

[1] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023. 1, 2

[2] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022. 2

[3] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[4] Dani Valevski, Danny Wasserman, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. *arXiv preprint arXiv:2306.06638*, 2023. 1, 2