HyperLoRA: Parameter-Efficient Adaptive Generation for Portrait Synthesis

Supplementary Material

A. Appendix

A.1. Inference Speed

We evaluate the inference speed of IP-Adapter [6], InstantID [5], PuLID [3] and our HyperLoRA on a single NVIDIA V100 GPU. For each method, the same 512×512 image is adopted as the input of ID plug-in network, and we generate a 1024×1024 personalized image using the DPM-2M Karras sampler with 16 denosing steps. We repeat the same inference process four times and record the average inference time in Table. 1. The complete inference process consists of two parts: one is the preprocess and another is the real inference of base model. In the preprocess stage, we crop the input image, obtaining ID embedding and predicting the ID tokens (for HyperLoRA, generating the LoRA weights and merging into base model). The sequence length of LoRA coefficients is significantly larger than that of ID tokens. Thus, HyperLoRA is usually more time-consuming at this stage. However, HyperLoRA shows remarkable superiority in the second stage, since it doesn't introduce extra attention modules compared with adapter-based methods. Once merged into the base model, our HyperLoRA maintains the same inference performance as the original model.

Table 1. Inference speed of different methods (milliseconds).

Method	IP-Adapter	InstantID	PuLID	HyperLoRA
Preprocess	2996	758	236	1143
Inference	6148	8037	6616	4327

A.2. Interpolation

Benefit from the natural interpolability of LoRA, our HyperLoRA not only supports multiple input images easily but also enables smooth ID interpolation. Fig. 2 illustrates that the personalized images generated by the mixture of two Hyper ID-LoRAs can interpolate naturally from one ID to another. Furthermore, we find that the space of our generated LoRAs has similar properties as the $\mathcal{W}+$ space of StyleGAN [4]. Specifically, the slider LoRA [2] can be generated by HyperLoRA given only a pair of images, as shown in Fig. 1. The image pair consists of a normal face image and an edited one (modifying the attributes of face, e.g., age or eye size). By feeding this pair into HyperLoRA, we obtain two LoRAs, and then subtract the LoRA weights of the original image from the LoRA weights of the edited image. Fig. 1 demonstrates the difference weights between two LoRAs perform a similar behavior as a slider LoRA, possessing the ability to edit local attributes.

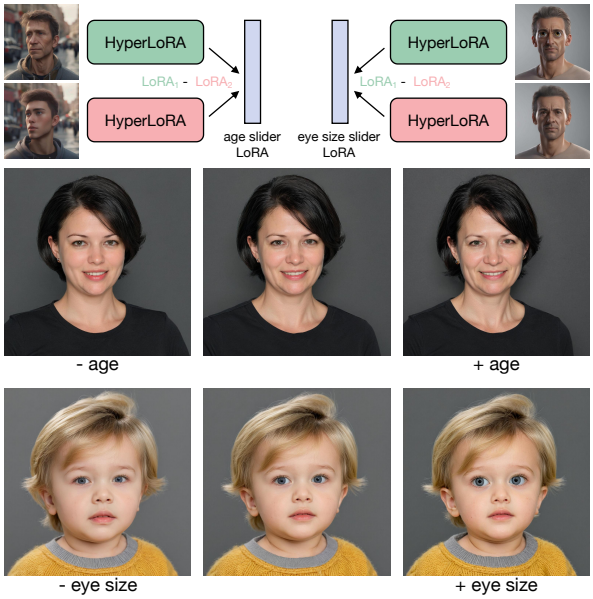

Figure 1. Using the slider LoRA generated by HyperLoRA from paired images, we are allowed to edit the local attrubutes (e.g. age or eye size) of synthesized portraits.

A.3. More Visual Results

Inference with ControlNet. ControlNet [7] is a practical network to control the diffusion models with additional condition images (edge, pose, depth and more). It helps us to generate satisfied images more easily, getting rid of spending effort trying different seeds, prompts and other parameters. Thus, it is crucial for an ID plug-in to be compatible with ControlNet. As shown in Fig. 3, our HyperLoRA can also generate portrait images with high photorealism and fidelity, when equipped with different ControlNets.

More Text-to-image Results. Fig. 4 and Fig. 5 present more portrait images synthesized by our HyperLoRA across various IDs and scenes, demonstrating its generalization.

A.4. Discussion on Fidelity and Editability

HyperLoRA relies on a large dataset with various facial features for training the LoRA basis, constructing a LoRA space covering sufficient face IDs, so that an unseen face can be reconstructed by the trained LoRA basis. Currently, our dataset comprises only 4.4 million images (in contrast, InstantID is trained with 60 million data). It is crucial to expand the dataset, aiming to further improve the fidelity. In the aspect of editability, it is a feasible solution to introduce the alignment loss of PuLID into HyperLoRA training, weakening the intrusion into the base model.

Figure 2. ID interpolation with HyperLoRA.

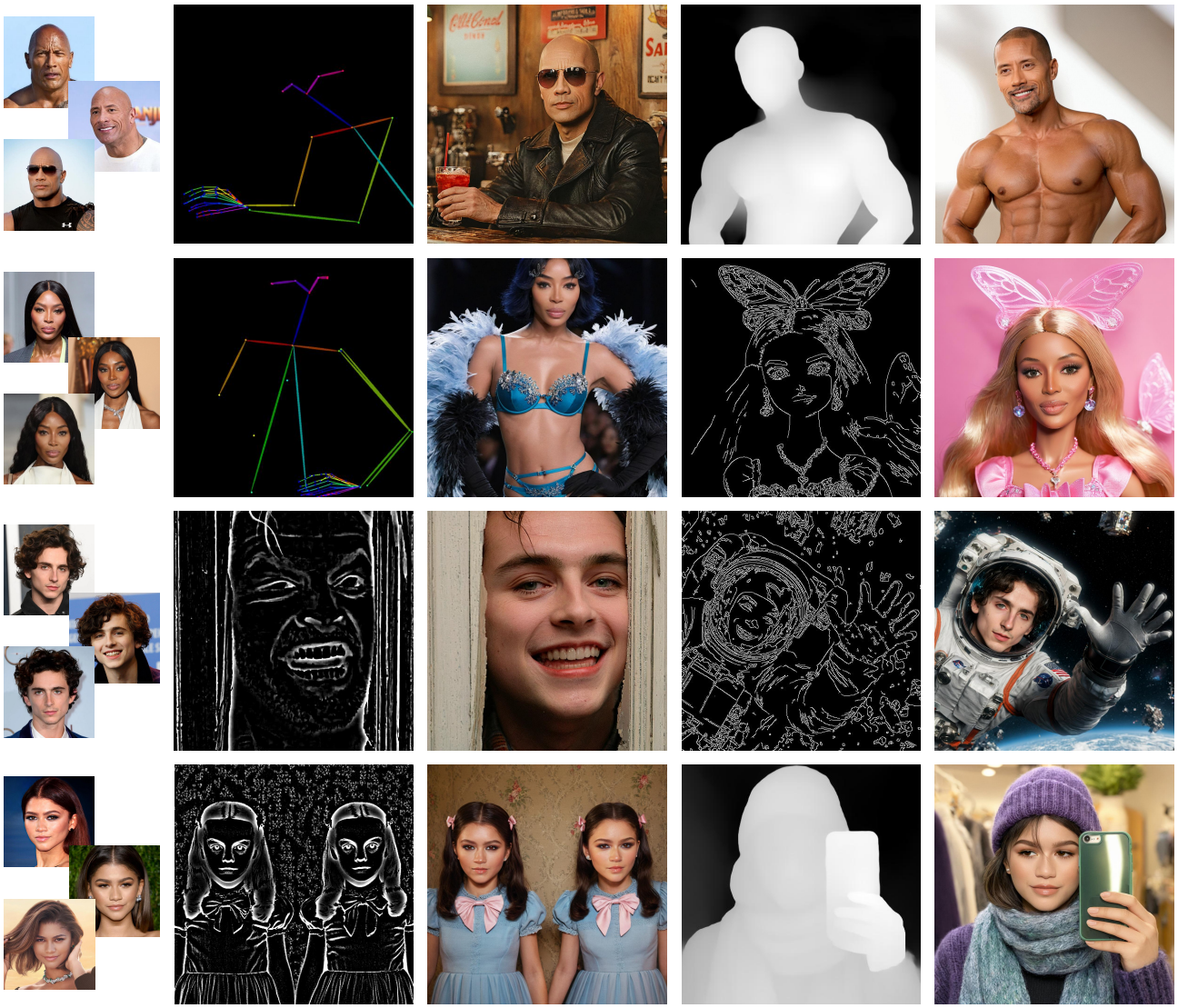

Figure 3. Portrait images generated by HyperLoRA and ControlNet.

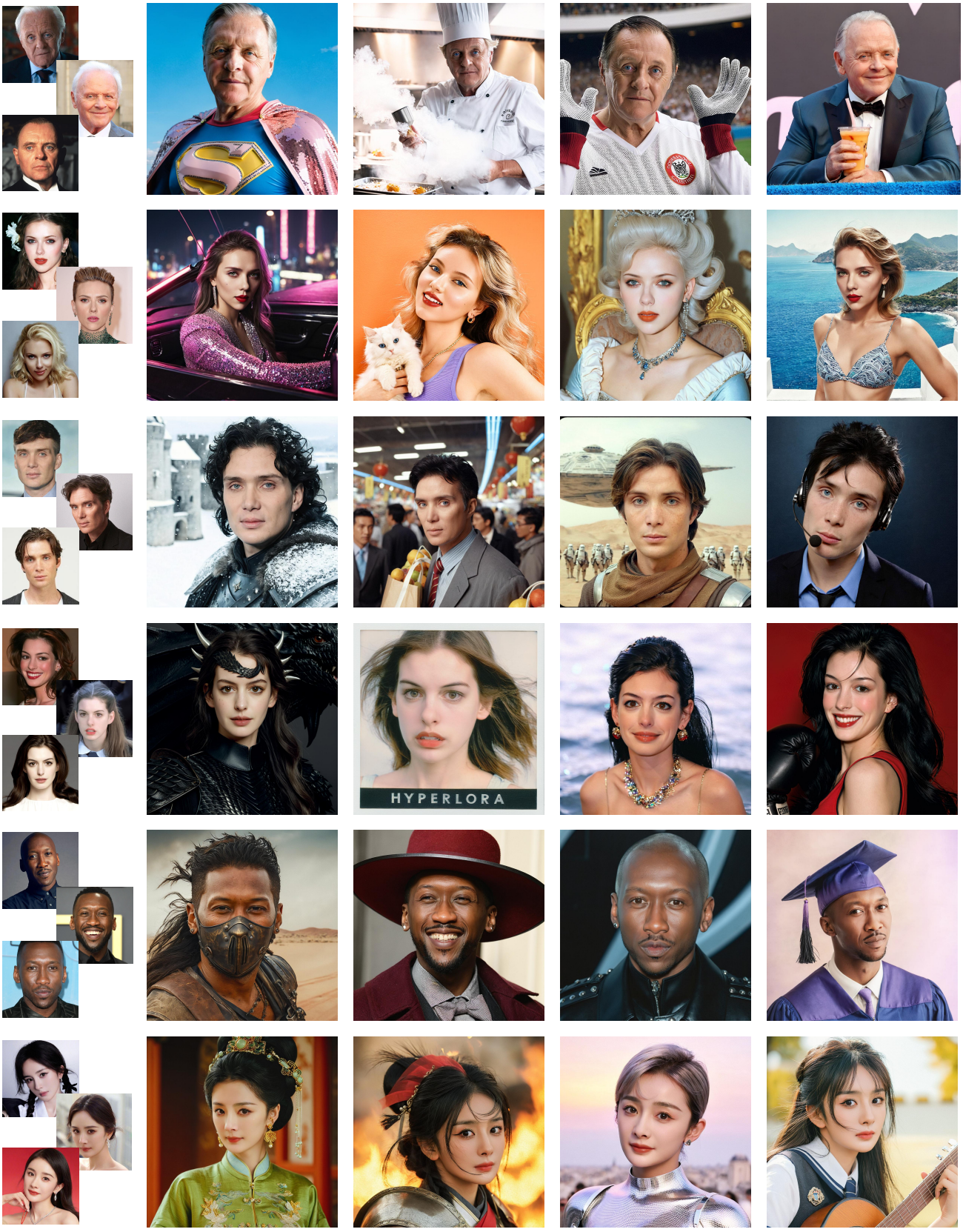

Figure 4. More Text-to-image results with HyperLoRA.

Figure 5. More Text-to-image results with HyperLoRA.

A.5. Details of Data Filtering

Our data filtering pipeline consists of four steps: **1)** We eliminate images with a resolution lower than 768×768; **2)** LAION Aesthetics Predictor [1] is utilized to assess the aesthetics score of images, and only those with a score higher than 5.5 are retained; **3)** We detect faces from images, and only keep those that contain a single face and the ratio of face region to the entire image is greater than 0.03; **4)** We further filter out those images where plural personal pronouns appear in the corresponding captions.

References

[1] Romain Beaumont and Christoph Schuhmann. Laion-aesthetics predictor. 4

[2] Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pages 172–188. Springer, 2024. 1

[3] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment. *Advances in neural information processing systems*, 2024. 1

[4] 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

[5] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 1

[6] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023. 1

[7] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1