InfiniteYou: Flexible Photo Recrafting While Preserving Your Identity

Supplementary Material

Liming Jiang Qing Yan Yumin Jia Zichuan Liu Hao Kang Xin Lu
ByteDance Intelligent Creation

Project Page: **https://bytedance.github.io/InfiniteYou**

A. Additional Implementation Details

We implement our InfiniteYou (InfU) framework using PyTorch and leverage the Hugging Face Diffusers library. The DiT base model is FLUX.1-dev [11]. We set the multiplication factor $i = 4$ for InfuseNet. The projection network is derived from [18], with the token number of the projected identity feature set to 8. All experiments are conducted using FSDP [21] on NVIDIA H100 GPUs, each with 80GB VRAM. We use the AdamW [16] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The weight decay is set to 0.01. We employ Conditional Flow Matching [5, 14] as the loss function with logit-normal sampling [5] of `rf/lognorm(0.00, 1.00)`. For stage-1 pretraining, the model is trained using an initial learning rate of 2×10^{-5} on 128 GPUs. The total batch size is set to 512, and stage-1 training spans 300k iterations. For stage-2 supervised fine-tuning, the model is trained with an initial learning rate of 1×10^{-5} on 64 GPUs, with a total batch size of 256. All other settings remain unchanged.

B. Dataset Details

For stage-1 pretraining, we use a total of nine open source datasets, including VGGFace2 [2], MillionCelebs [20], CelebA [15], CelebV-HQ [22], FFHQ [8], VFHQ [17], EasyPortrait [10], CelebV-Text [19], CosmicManHQ-1.0 [13], as well as several high-quality internal datasets. We perform careful data pre-processing and filtering, removing images with low-quality small faces, multiple faces, watermarks, or NSFW content. The data is pre-processed for training using aspect ratio bucketing [1]. The total amount of single-person single-sample (SPSS) real data for stage-1 pre-training reaches 43 million, which we consider sufficient for large-scale training of identity-preserved image generation models. For stage-2 supervised fine-tuning, the total quantity of single-person-multiple-sample (SPMS) synthetic data is 2 million. All data is generated by the stage-1 pretrained InfU model itself, equipped with useful off-the-shelf modules (see Section 3.3). High-quality synthetic data are also carefully processed and filtered to obtain image pairs with normal poses, high ID resemblance, and good aesthetics, ensuring their usefulness. In addition, we observe that training the model with a mixture of captions from multiple sources, *e.g.*, humans, small captioning models, and large vision-language models (VLMs), is beneficial. Besides the original captions in the datasets, we employ BLIP-2 [12] and InternVL2 [3] to obtain text captions from diverse sources for training.

C. Evaluation Details

We conduct evaluations on a portrait benchmark created by GPT-4o [7], comprising 200 prompts and corresponding gender information. This benchmark covers a variety of cases, including different prompt lengths, face sizes, views, scenes, ages, races, complexities, *etc*. We selected 15 representative identity samples and paired their gender with all appropriate prompts, resulting in $1,497$ testing outputs for systematic evaluations. We apply three representative and useful evaluation metrics, *i.e.*, ID Loss [4], CLIPScore [6], and PickScore [9]. ID Loss is defined as $1 - \mathrm{CosSim}\left(\mathrm{ID}_{\mathrm{gen}}, \mathrm{ID}_{\mathrm{ref}}\right)$, where CosSim is cosine similarity, and $\mathrm{ID}_{\mathrm{gen}}$ and $\mathrm{ID}_{\mathrm{ref}}$ are the generated and reference identity images, respectively. A lower ID Loss means higher similarity. We follow the original papers to use CLIPScore and PickScore. A higher CLIPScore indicates better text-image alignment, and a higher PickScore signifies better image quality and aesthetics.

D. Limitations and Societal Impacts

Despite promising results, the identity similarity and overall quality of InfU could be further improved. Potential solutions include additional model scaling and an enhanced InfuseNet design. On another note, InfU may raise concerns about its potential to facilitate high-quality fake media synthesis. However, we believe that developing robust media forensics approaches can serve as effective safeguards.

References

[1] NovelAI Aspect Ratio Bucketing. https://github.com/NovelAI/novelai-aspect-ratio-bucketing. Accessed: 2023-02-16. 1

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *FG*, 2018. 1

[3] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 1

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 1

[5] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 1

[6] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 1

[7] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint*, arXiv:2410.21276, 2024. 1

[8] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1

[9] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *NeurIPS*, 2023. 1

[10] Karina Kvanchiani, Elizaveta Petrova, Karen Efremyan, Alexander Sautin, and Alexander Kapitanov. EasyPortrait–face parsing and portrait segmentation dataset. *arXiv preprint*, arXiv:2304.13509, 2023. 1

[11] Black Forest Labs. FLUX.1 release: Announcing black forest labs. https://blackforestlabs.ai/announcing-black-forest-labs/, 2024. Accessed: 2024-08-01. 1

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1

[13] Shikai Li, Jianglin Fu, Kaiyuan Liu, Wentao Wang, Kwan-Yee Lin, and Wayne Wu. CosmicMan: A text-to-image foundation model for humans. *arXiv preprint*, arXiv:2404.01294, 2024. 1

[14] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023. 1

[15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 1

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1

[17] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. VFHQ: A high-quality dataset and benchmark for video face super-resolution. In *CVPRW*, 2022. 1

[18] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint*, arXiv:2308.06721, 2023. 1

[19] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. CelebV-Text: A large-scale facial text-video dataset. In *CVPR*, 2023. 1

[20] Yaobin Zhang, Weihong Deng, Mei Wang, Jiani Hu, Xian Li, Dongyue Zhao, and Dongchao Wen. Global-local GCN: Large-scale label noise cleansing for face recognition. In *CVPR*, 2020. 1

[21] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. PyTorch FSDP: experiences on scaling fully sharded data parallel. *arXiv preprint*, arXiv:2304.11277, 2023. 1

[22] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*, 2022. 1