FaceChain: A Playground for Identity-Preserving Portrait Generation

Yang Liu [1*], Cheng Yu [1*], Lei Shang [1], Ziheng Wu [1], Xingjun Wang [1], Yuze Zhao [1], Lin Zhu [1],
Chen Cheng [1], Weitao Chen [1] Chao Xu [1], Haoyu Xie [1], Yuan Yao [1], Wenmeng Zhou [1],
Yingda Chen [1] ✉ , Xuansong Xie [1] ✉, Baigui Sun [1] ✉
[1]Alibaba Group
{ly261666, yucheng.yu, sl172005, zhoulou.wzh, xingjun.wxj, yuze.zyz, lin.zhu, chengchen.cc,
weitao.cwt, xc264362, xiehaoyu.xhy, ryan.yy, wenmeng.zwm,
yingda.chen, xingtong.xxs, baigui.sbg}@alibaba-inc.com

Abstract: Recent advancement in personalized image generation have unveiled the intriguing capability of pre-trained text-to-image models on learning identity information from a collection of portrait images. However, existing solutions can be vulnerable in producing truthful details, and usually suffer from several defects such as (i) The generated face exhibit its own unique characteristics, *i.e*. facial shape and facial feature positioning may not resemble key characteristics of the input, and (ii) The synthesized face may contain warped, blurred or corrupted regions. In this paper, we present FaceChain, a personalized portrait generation framework that combines a series of customized image-generation model and a rich set of face-related perceptual understanding models (*e.g*., face detection, deep face embedding extraction, and facial attribute recognition), to tackle aforementioned challenges and to generate truthful personalized portraits, with only a handful of portrait images as input. Concretely, we inject several SOTA face models into the generation procedure, achieving a more efficient label-tagging, data-processing, and model post-processing compared to previous solutions, such as DreamBooth [9] , InstantBooth [11] , or other LoRA-only approaches [3] . Through the development of FaceChain, we have identified several potential directions to accelerate development of Face/Human-Centric AIGC research and application. We have designed FaceChain as a framework comprised of pluggable components that can be easily adjusted to accommodate different styles and personalized needs. We hope it can grow to serve the burgeoning needs from the communities. Note that this is an ongoing work that will be consistently refined and improved upon. FaceChain is open-sourced under Apache-2.0 license at `https://github.com/modelscope/facechain`.

1. Introduction

Recent years have witnessed a remarkable progress in the field of text-to-image generation, with large models [5,8] emerging as the powerful foundation for creating high-fidelity and diverse images. Given a text prompt, these models have demonstrated the impressive ability to create realistic and detailed image, showcasing their potential for a wide range of applications, *e.g*. content generation, virtual reality and augmented reality. However, for human-centric content generation, pre-trained text-to-image models often struggle to produce satisfactory portrait images that retain identities of individuals. This can be frustrating for individuals who wish to generate self-portraits. The imperfection arises due to the inherent limitations of these models, which are not designed to accurately preserve identity information. To this end, several recent efforts have started with the emphasis on tackling faithful personalized text-to-image generation. These efforts aim to learn the identity information from a collection of portrait images, then generate new scenes or styles corresponding to the target human beings under the guidance of text prompt.

The existing human-centric personalized text-to-image generation methods can be categorized into (i) LoRA-based Framework, which leverages the LoRA (Low-Rank Adaptation) fine-tune [3] technology on text-to-image model (*e.g*. Stable Diffusion [8]) to generate identity-preserved images. (ii) Identifier-based methods [9, 11], which aim to learn a unique identifier relevant to identity information. Although these methods can synthesize identity-similar images, they still suffer from several defects, *e.g*., facial shape and facial feature positioning may differ significantly from the input face; the synthesized face may contain several warped and corrupted regions. In this paper we present FaceChain, a identifying preserving framework that not only preserves the distinguishing features of faces but also allows for versatile control over stylistic elements. Specifically, we inject

[*] Equal Contribution, ✉ Corresponding Author

two LoRA models into Stable Diffusion model, imparting it with the ability to integrate personalized style and identity information simultaneously. In particular, FaceChain is rooted in the ModelScope (https://modelscope.cn), an open-source community that seeks to bring together models from different areas and offer them via a unified interface. This allows FaceChain to integrate a comprehensive suite of face-related models, in addition the foundation model, to build the framework that generate identity-preserving portraits, the process of which we detail later in Sec. 2.

The rest of the paper is organized as follows. In Sec. 2, we describe how FaceChain is built around a pluggable framework that offers the versatility needed to generate identity-preserving personal portraits. Sec. 3 present a thorough discussion on how future research and applications can stem from and flourish on FaceChain. As a relatively new open-source project, we believe FaceChain has shown its potentials. Other than the practical functionalities, we also aspire to expanding it into the benchmark and playground, that inspires innovations in personalized text-to-image generation.

2. Architecture

FaceChain encapsulates the process of personalized portrait generation within an atomic pipeline, and is built upon Stable Diffusion [8] model. To improve the style stability and ID consistency of text-to-image generation, we adopt LoRA [3], a parameter-efficient strategy to fine-tune Stable Diffusion model. With the *composability* of multiple LoRA models, we learn the information for portrait style and human identities with different LoRA models, namely the style-LoRA model and face-LoRA model respectively. These two models are trained separately via text-to-image training on images of given style and human identities. Specifically, we choose to train the style-LoRA model offline, which we describe in Section 2.2, while the face-LoRA model is trained online using the images uploaded by users – which are of the same human identity. Since the quality of the user-uploaded images may vary, FaceChain incorporates a rich set of face-related perceptual understanding models to ensure face images feeding into the training process are normalized to meet certain quality standards, such as appropriate size, good skin quality, correct orientation, as well as having accurate tags. The weights of multiple LoRA models are then merged into the Stable Diffusion model during inference to generate personalized portraits. Finally, the details of the generated portraits are further enhanced by a series of post-processing steps. The overall processing pipeline is illustrated in Fig. 1.

2.1. Data Processing

2.1.1 Face Extraction

To improve the training-stability for face-LoRA models, FaceChain chains a series of face-related data processing modules to extract faces with appropriate size, good skin quality, and correct orientation from the images uploaded by users. These modules leverage extensively, the various models available on ModelScope, which we list below.

Image Rotation. The orientation of human in images uploaded by users may not be suitable for training. To rectify this, orientation of the uploaded image is first determined and image rotation is perform if necessary. First, a rotation angle determination model is adopted to predict the probabilities of the image rotation with angle 0°, 90°, 180°, and 270°. Then, the image is rotated by the angle with the largest probability. The rotation angle judging model is available at https://modelscope.cn/models/Cherrytest/rot_bgr.

Face Rotation. After the initial image rotation, face orientation within the image may still fall short of the training requirement. Therefore, we tail it with a more accurate face rotation module, which perform the rotation according to the location of facial landmarks. In particular, DamoFD [7], a face detector using Network Architecture Search [12] (NAS) is used to obtain the detection result of five face landmarks. Rotation matrix is then computed for the coordinates of detected landmarks and standard face template, using the least square method. During this process, the two sets of coordinates are firstly normalized, by subtracting their mean and divided by their standard deviation, to eliminate the effect of translation and scaling. If we denote the normalized coordinates of the detected landmarks and the face template as $\mathbf{P}_1, \mathbf{P}_2 \in \mathbb{R}^{5\times2}$, respectively. the aim here is to minimize $\|\mathbf{R}\mathbf{P}_1^\top - \mathbf{P}_2^\top\|^2$, where $\mathbf{R}$ is the rotation matrix formulated as Eq. 1.

$$\mathbf{R} = \begin{pmatrix} \cos\theta & \text{-}\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{1}$$

As such, $\mathbf{R} = ((\mathbf{P}_1^\top\mathbf{P}_1)^{-1}\mathbf{P}_1^\top\mathbf{P}_2)^\top$, and we get the rotation angle $\theta = \arctan(\mathbf{R}_{21}/\mathbf{R}_{22})$. The image is then rotated accordingly, as illustrated in Fig. 1. The DamoFD model is available at https://modelscope.cn/models/damo/cv_ddsar_face-detection_iclr23-damofd.

Face Region Crop and Segmentation. After image and face rotation, the face regions are then cropped out and masked from the input images. Using the DamoFD model, the bounding box of the face is determined, we then crop the image and adjust the size and position of the face. As such, we keep the face centered horizontally, with its size between 0.35 and 0.45 times the whole image size. Then we use Masked-attention Mask Transformer [1] model for human

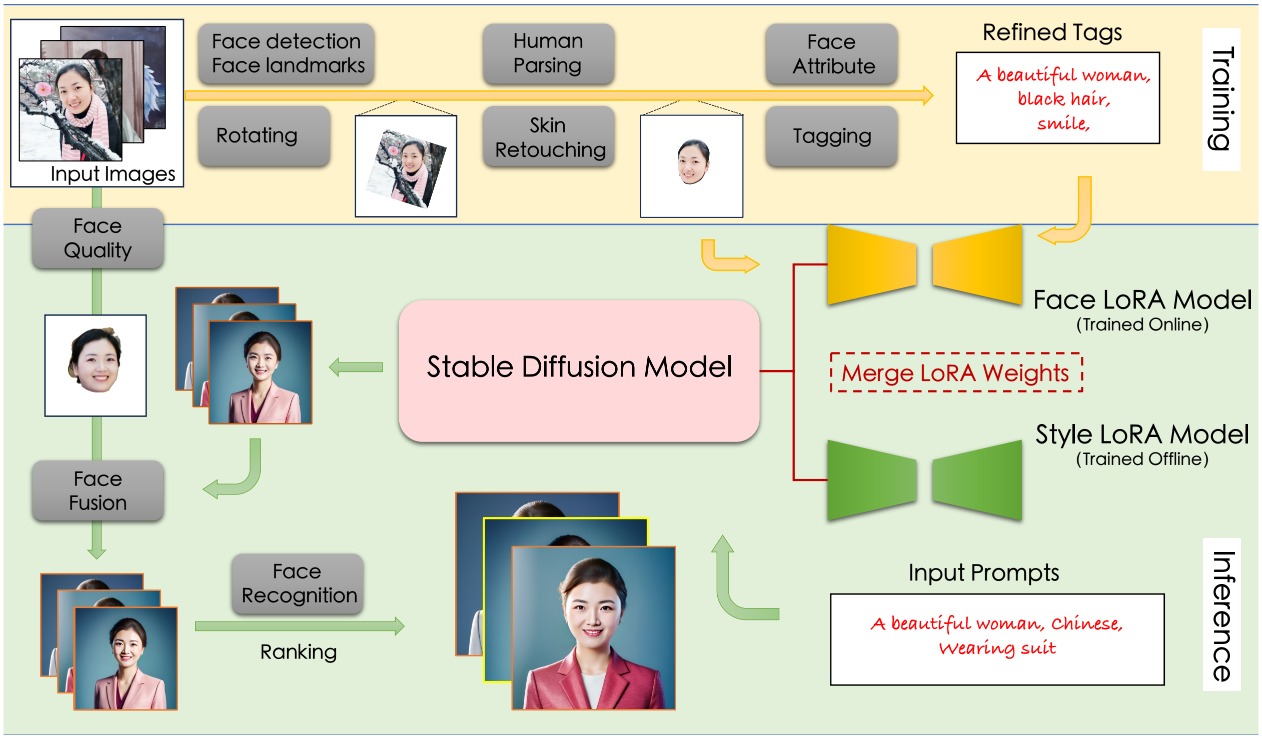

Figure 1: **Architectural overview of FaceChain personalized portrait generation.** During training, multiple data processing approaches are adopted to generate tagged face images to train face-LoRA model online. The weights of face-LoRA and style-LoRA models are then merged into foundation Stable-Diffusion model for text-to-image generation. The generated portraits go through post-face-fusion and ranking before returning to users.

parsing (M2FP) to generate the mask of the head region, and perform segmentation accordingly. The M2FP model is available at https://modelscope.cn/models/damo/cv_resnet101_image-multiple-human-parsing.

Skin Retouching. Since the skin quality of the images uploaded by users may be unpromising, we leverage the skin retouching module to improve the skin quality of the face images. In particular, the Adaptive Blend Pyramid Network [6] (ABPN) is used here for skin retouching, which is available at https://modelscope.cn/models/damo/cv_unet_skin-retouching.

2.1.2 Label Tagging

High-quality tagging is critical to facilitate text-to-image training. As such, we train face-LoRA model to learn the relationship between the face images and the generated tags. This allows the Stable Diffusion model to generate images with corresponding features during inference, when prompting with suitable tags. Therefore, the tags must be labeled appropriately to ensure that face-LoRA model can be triggered to produce stable output. We identified three major requirements for label tagging:

- Tags specific to a given image, such as facial expressions, jewelries and accessories, should be labeled accurately to retain the relationship between salient image features and the corresponding tags.

- Tags bound to human identities, such as eyes, lips and ears, can be removed. Instead, the LoRA model can be relied on to generate such features without any prompt words.

- In general, using one tag to describe the overall characteristics of the human identity works quite well in practice. For example, we may use a man/woman/boy/girl, as the trigger word for all images. As such, when adding the trigger word into input prompts, the features of the human identity can be generated more easily.

FaceChain combine different approaches for label tagging to satisfy the above requirements. First, the face images are fed into DeepDanbooru [1], a text annotation model, to get the preliminary tags. Then, we perform tag post-processing to choose tags corresponding to human identi-

[1] available at https://github.com/KichangKim/DeepDanbooru

Table 1: Trigger words describing different gender and age.

Gender Age	Male	Female
$0 \sim 20$	a boy, children	a girl, children
$20 \sim 40$	a handsome man	a beautiful woman
> 40	a mature man	a mature woman

ties, and remove them to meet the first two requirements. Finally, we use FairFace [4], a face attribute model, to predict the probabilities of the gender and age attributes for each image, and use the overall result as the final prediction for the human identity. Then we choose the trigger word according to the prediction of gender and age, as is shown in Table 1. The FairFace model is available at https://modelscope.cn/models/damo/cv_resnet34_face-attribute-recognition_fairface.

2.2. Model Training

Style-LoRA model acts as the anchor-stone for producing stable styles of portraits. It is important for personal portrait generation since it set the guardrails for image generation model at large. The main style-LoRA model used with FaceChain targets personal portraits, and is trained with a large number of portrait-like images of the same style, such as ID photos. We share here the hyper-parameters used for training the LoRA model here for full disclosure. The rank of LoRA model is set to 32. Learning rate is set to 1e-4, and cosine with restarts schedule is deployed. The LoRA model is trained for 20 epochs to produce the final model. We deploy 8bit AdamW optimizer [2] to save on training hardware. As to the training of face-LoRA model, the user-uploaded images are firstly rotated based on the angle predicted by the image rotation model. It is then followd by the face alignment method based on face detection and keypoint output, which obtains images containing forward-looking faces. Next, we use the human body parsing model and the human portrait beautification model to obtain high-quality face training images. Afterwards, we use a face attribute model and a text annotation model, combined with tag post-processing methods, to generate fine-grained labels for training images. Finally, we use the above images and label data to fine-tune the Stable Diffusion model to obtain the face-LoRA model.

2.3. Model Inference

During inference phase, we fuse the weights of the face-LoRA model and style-LoRA model into the Stable Diffusion model. The fusing weights are chosen to be 0.25 and 1.0, respectively. Next, we use Stable Diffusion‘s text-to-image generation pipeline to generate the preliminary personal portraits with preset input prompt words. Then we further improve the face details of the above generated portrait image and facial similarities with the input faces using a face fusion model. The template face used for fusion is selected from the training images via the face quality evaluation model. Finally, we use the face recognition model to calculate the similarity between the generated portrait image and the template face, the resulting portrait images are sorted and ranked accordingly before final output.

2.4. Model Post Processing

After generating preliminary portraits by the Stable Diffusion model, FaceChain integrates several post process modules listed below to improve the face details and facial similarities of the portraits. The overall pipeline for post process is illustrated in Fig. 1.

Template Face Selection. We adopt the Face Quality Assessment (FQA) model to evaluate the quality score for all faces from the user-uploaded images. The face with the highest quality score is then chosen as the template-face for face fusion. The FQA model is available at https://www.modelscope.cn/models/damo/cv_manual_face-quality-assessment_fqa.

Face Fusion. We perform face fusion for the generated portraits using the selected template-face to improve facial details. This allows the output portrait to retain major appearance features, while displaying more refined facial details. The face fusion model is available at https://www.modelscope.cn/models/damo/cv_unet-image-face-fusion_damo.

Similarity Ranking. The final output portraits are selected by comparing their facial similarities to the template face. Given the inherent statistical difference between generated portraits and the input images, we adopt Random Temperature Scaling [10] (RTS), a robust face recognition model for both in-distribution and out-of-distribution samples, to calculate the facial similarities . Finally, the portraits with high facial similarity are selected as the output. The RTS model is available at https://www.modelscope.cn/models/damo/cv_ir_face-recognition-ood_rts.

3. Future Work

Given multiple images depicting the same individual, FaceChain can generate a diverse collection of high-fidelity, identity-preserving portraits with distinct stylistic variations. These variations encompass a wide spectrum, ranging from classy identification photos and human portraits, to photos of futuristic aesthetics of the cyberpunk genre. Still, we acknowledge current work on FaceChain is merely scratching the tip of the iceberg, and an immerse universe of applications along the line is waiting to be explored. In this Section, we put forward a few directions we consider worthwhile to explore.

- Personalized generation framework capable of handling multiple subjects of different ages and genders.
- Improved data processing mechanism to retain stature impeccably, which will require more diverse training data.
- Support adaptive weight-selection for style and face LoRA models during model fusion process.
- Encode diverse styles information into a unified model that can be activated with specific triggering prompts.
- Develop tailored similarity ranking and face Fusion models for FaceChain.
- Explore train-free framework for customized portrait generation. Current approach used in FaceChain requires a new model to be trained for each human id, which can be computational expensive for wide adoption.

Acknowledgments. We wish to thank Dr. Peiran He and Miaomiao Cui for many helpful discussions.

References

[1] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2

[2] Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. In *International Conference on Learning Representations*, 2022. 4

[3] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 1, 2

[4] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021. 4

[5] Sangyun Lee. Dalle-2. 1

[6] Biwen Lei, Xiefan Guo, Hongyu Yang, Miaomiao Cui, Xuansong Xie, and Di Huang. Abpn: adaptive blend pyramid network for real-time local retouching of ultra high-resolution photo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2108–2117, 2022. 3

[7] Yang Liu, Jiankang Deng, Fei Wang, Lei Shang, Xuansong Xie, and Baigui Sun. Damofd: Digging into backbone design on face detection. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2

[9] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1

[10] Lei Shang, Mouxiao Huang, Wu Shi, Yuchen Liu, Yang Liu, Wang Steven, Baigui Sun, Xuansong Xie, and Yu Qiao. Improving training and inference of face recognition models via random temperature scaling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15082–15090, 2023. 4

[11] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 1

[12] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016. 2