InfiniteYou: Flexible Photo Recrafting While Preserving Your Identity

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

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

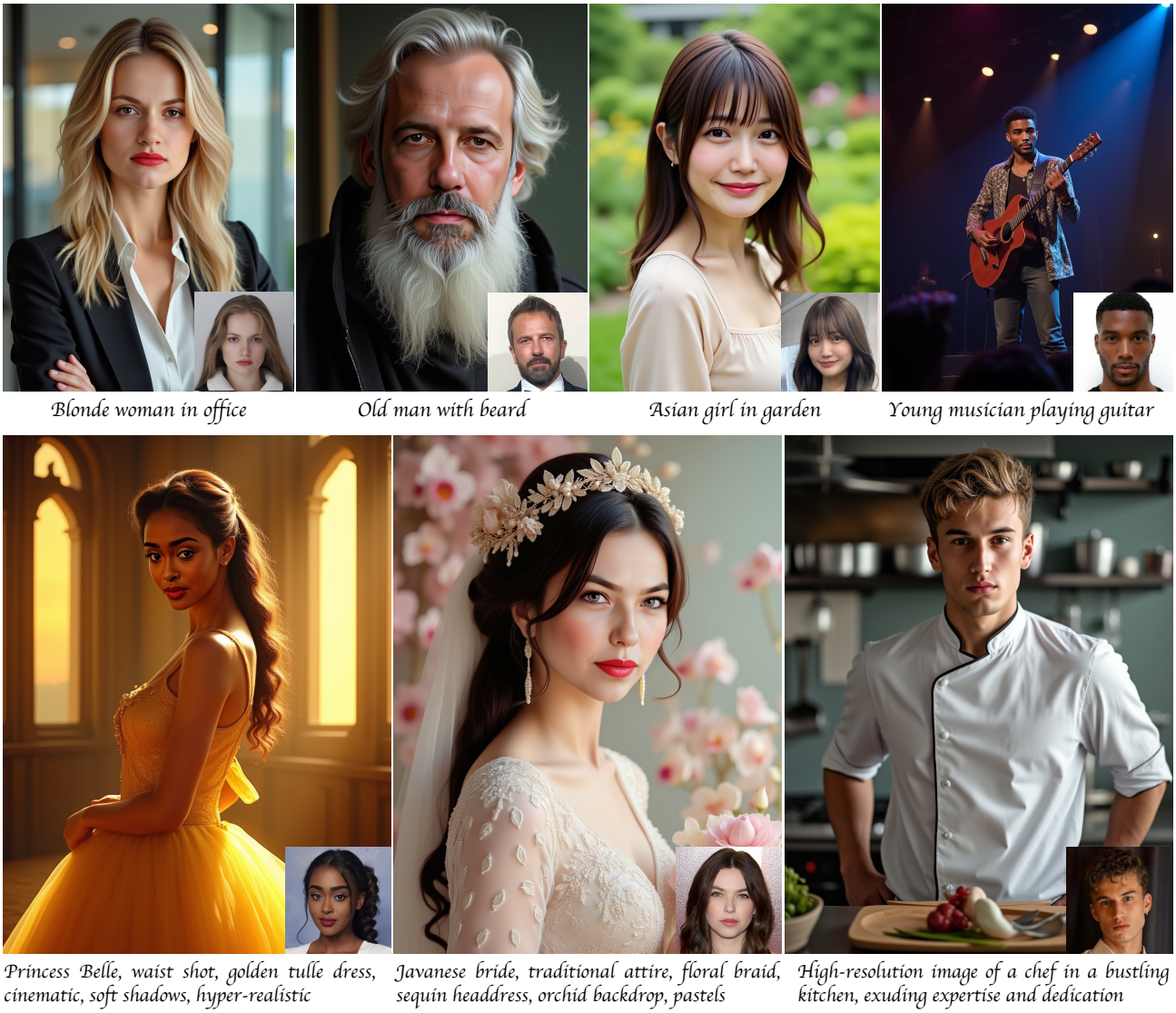

Figure 1. InfiniteYou generates identity-preserved images with exceptional identity similarity, text-image alignment, quality, and aesthetics.

Abstract

Achieving flexible and high-fidelity identity-preserved image generation remains formidable, particularly with advanced Diffusion Transformers (DiTs) like FLUX. We introduce ***InfiniteYou (InfU)****, one of the earliest robust frameworks leveraging DiTs for this task. InfU addresses significant issues of existing methods, such as insufficient identity similarity, poor text-image alignment, and low generation quality and aesthetics. Central to InfU is InfuseNet, a component that injects identity features into the DiT base model via residual connections, enhancing identity similarity while maintaining generation capabilities. A multi-stage training strategy, including pretraining and supervised fine-*

tuning (SFT) with synthetic single-person-multiple-sample (SPMS) data, further improves text-image alignment, ameliorates image quality, and alleviates face copy-pasting. Extensive experiments demonstrate that InfU achieves state-of-the-art performance, surpassing existing baselines. In addition, the plug-and-play design of InfU ensures compatibility with various existing methods, offering a valuable contribution to the broader community. Code and model: https://github.com/bytedance/InfiniteYou.

1. Introduction

Identity-preserved image generation aims to recraft a photograph of a specific person using free-form text descriptions while preserving facial identity. This task is challenging but highly beneficial. Previous methods [14, 51, 54] have been mainly developed on U-Net [42]-based text-to-image diffusion models [17, 41, 46], such as Stable Diffusion XL (SDXL) [38]. However, due to the limited generation capacity of the base model, the quality of generated images remains inadequate. Recent strides in Diffusion Transformers (DiTs) [12, 37] have made remarkable progress in content creation. In particular, the latest releases of state-of-the-art rectified flow DiTs, such as FLUX [26] and Stable Diffusion 3.5 [12], showcase incredible image generation quality. Consequently, it is crucial to explore solutions that can leverage the immense potential of DiTs for downstream applications like identity-preserved image generation.

DiT-based identity-preserved image generation remains formidable due to several factors: the absence of customized module designs, difficulties in model scaling, and a lack of high-quality data. Thus, effective solutions for this task using state-of-the-art rectified flow [4, 33] DiTs, such as FLUX [26], are currently scarce in both academia and industry. PuLID-FLUX, derived from PuLID [14], presented an initial attempt to develop an identity-preserved image generation model based on FLUX. Other open-source efforts, including FLUX.1-dev IP-Adapters from InstantX [20] and XLabs-AI [3], are not tailored for human facial identities. Nevertheless, existing methods fall short in three key aspects: **1)** The identity similarity is not sufficient; **2)** The text-image alignment and editability are poor, and the face copy-paste issue is evident; **3)** The generation capability of FLUX is largely compromised, resulting in lower image quality and aesthetic appeal.

To address the aforementioned challenges, we propose a simple yet effective framework for identity-preserved image generation, namely InfiniteYou (InfU). This framework is designed to be systematic and robust, enabling flexible text-based photo re-creation for diverse identities, races, and age groups across various scenarios. We introduce InfuseNet, a generalization of ControlNet [56], which ingests identity information along with the controlling conditions. The projected identity features are injected by InfuseNet into the DiT base model through residual connections, thereby disentangling text and identity injections. InfuseNet is both scalable and compatible, harnessing the powerful generation capabilities of DiTs. The scale-up injection network and the delicate architecture design, equipped with large-scale model training, effectively enhance identity similarity. To improve text-image alignment, image quality, and aesthetic appeal, we employ a multi-stage training strategy, including pretraining and supervised fine-tuning (SFT) [21, 49]. The SFT stage utilizes carefully designed synthetic single-person-multiple-sample (SPMS) data generation, leveraging our pretrained model itself and various off-the-shelf modules. This strategy enhances the quantity, quality, aesthetics, and text-image alignment of the training data, thus improving overall model performance and alleviating the face copy-paste issue. Thanks to the InfuseNet design, identity information is injected purely via residual connections between DiT blocks, unlike conventional practices [14, 51, 54] that directly modify attention [50] layers via IP-Adapter (IPA). Consequently, the generation capability of the base model remains largely intact, allowing for the generation of high-quality and aesthetically pleasing images. Moreover, InfU is plug-and-play and readily compatible with many other methods or plugins, thus offering significant value to the broader community.

Comprehensive experiments show that the proposed InfU framework achieves state-of-the-art performance (see Figure 1), significantly surpassing existing baselines in identity similarity, text-image alignment, and overall image quality. Our main contributions are summarized as follows:

- We propose InfiniteYou (InfU), a versatile and robust DiT-based framework for flexible identity-preserved image generation under various scenarios.
- We introduce InfuseNet, a generalization of ControlNet, which effectively injects identity features into the DiT base model via residual connections, enhancing identity similarity with minimal impact on generation capabilities.
- We employ a multi-stage training strategy, including pretraining and supervised fine-tuning (SFT), using synthetic single-person-multiple-sample (SPMS) data generation. This approach significantly improves text-image alignment, image quality, and aesthetic appeal.
- The InfU module features a desirable plug-and-play design, compatible with many existing methods, thus providing a valuable contribution to the broader community.

2. Related Work

Text-to-image Diffusion Transformers (DiTs). Diffusion models [17, 41, 46, 47] have become a standard paradigm given their incredible capability to produce high-fidelity images. Text-to-image generation aims to synthesize images through the denoising diffusion process [17, 41] from a