"Advancements in Text-to-Image Diffusion Models for Personalized Image Generation: A Review of ID-Preserving Techniques of InstantID"

Ardhiansyah K.

[1]STIKI Malang , Jln. Tidar 100 Malang, Indonesia

Article Information
Received: 21-11-2024
Revised: 28-11-2024
Published: 5-12-2024

Keywords
InstantID; Image Generation; Text-to-image Diffusion Model; Stable Diffusion

Correspondence Email:
191111044@mhs.stiki.ac.id

Abstract

The evolution of text-to-image diffusion models, such as GLIDE, DALL-E 2, and Stable Diffusion, has significantly enhanced image generation capabilities. However, achieving image personalization with precise facial detail retention, minimal reference images, and reduced computational costs remains challenging. Traditional methods like DreamBooth and Textual Inversion rely on extensive fine-tuning, while techniques like IP-Adapter, which avoid fine-tuning, often compromise accuracy. Addressing these gaps, InstantID introduces a novel plug-and-play module that uses a single reference image to enable efficient identity preservation with high fidelity and flexibility. InstantID departs from conventional approaches by employing ID Embedding and an Image Adapter to enhance semantic richness and facial detail fidelity. Unlike models relying on CLIP-based visual prompts, InstantID integrates ID Embedding with ControlNet to refine the cross-attention process. This involves using simplified facial keypoints for conditional input and replacing text prompts with ID Embedding. Trained on a large-scale dataset comprising LAION-Face and additional high-quality annotated images, InstantID demonstrates superior ID retention and facial detail restoration. Notably, its performance improves with multiple reference images but remains highly effective with just one. The results highlight the effectiveness of InstantID's modular components, such as IdentityNet and the Image Adapter, in ensuring exceptional generation quality and detail retention. Although currently optimized for SDXL checkpoints, InstantID offers a scalable and efficient solution for personalized image generation. By integrating with tools like ComfyUI, it provides a seamless and accessible approach to image personalization with strong ID control and adaptability.

1. Introduction

With the development of diffusion models like GLIDE, DALL-E 2, and Stable Diffusion, text-to-image generation has improved significantly. Image personalization is an important area of this technology that makes it possible to create images based on particular sources, like human characters, while maintaining identification details. The inability to guarantee great precision in facial details, the requirement for a large number of reference photographs, and the significant processing costs are the primary drawbacks of the existing personalization techniques(Wang et al., 2024).

Certain methods, such as DreamBooth, Textual Inversion, and LoRA, need fine-tuning model parameters, which takes a lot of time and resources. On the other hand, techniques such as IP-Adapter try to avoid fine-tuning but frequently compromise the accuracy of face details.

InstantID is a plug-and-play module that uses a single reference image to enable identity preservation. It is based on the diffusion model. **InstantID's** breakthrough technique combines high fidelity, flexibility, and efficiency.

1.1 Literature Review

Text-to-image diffusion models achieve state-of-the-art image generation results and achieved unprecedented interest from the community in recent years. A common practice is to encode the text prompt into latent through a pre-trained language encoder and use the latent to guide the diffusion process(Saharia et al., 2022). Subject-driver text-to-image generation, which uses a limited set of images of a particular subject to generate costumized images based on text description, has been notable advancements. Previous subject-driven approaches like DreamBooth, Textual inversion, ELITE, E4T, and ProFusion fine-tune a special prompt token to describe the target concepts during the fine-tuning process. These methods typically involve training additional modules while keeping the core pre-trained text-to-image models frozen. A leading example of this is IP-Adapter, which aims to decouple the cross-attention mechanism by separating the cross-attention layers for text features and image features (Cui et al., 2024) . ID-preserving image generation is a special case of subject-driven generation, but it focuses on face attributes with strong semantics and finds broad application in real-world scenarios.

2. Research Methods

Contrary to prior approaches like IP-Adapter, FaceStudio, and PhotoMaker, which rely on a pre-trained CLIP image encoder for visual prompt extraction, InstantID work targets the need for stronger semantic details and enhanced fidelity in the ID preservation task. InstantID study focuses on the requirement for richer semantic information and better fidelity in the ID preservation task, in contrast to previous methods such as IP-Adapter, FaceStudio, and PhotoMaker, which rely on a pre-trained CLIP image encoder for visual prompt extraction. Textual prompts are much improved by the ability of image prompting in pre-trained text-to-image diffusion models, especially for content that is difficult to convey through language. However, we depart from the coarse-aligned CLIP embbeding by using ID Embedding as our image trigger.

In InstantID addaption if ControlNet, there are mainly two modifications: 1) instead if fine-grained OpenPose facial keywords, they use only five facial keypoints for conditional input. 2) they eliminate the text prompt and use ID Embedding as conditions for cross-attention layers in the ControlNet.

3. Result and Discussion

InstantID team implement with Stable Diffusion and train it on the large-scale open-source dataset LAION-Face, which consists of 50 million image-text pairs to ensure diversity. In addition, the team collect 10 million high-quality human images from the Internet with annotations automatically generated by BLIP2 to further improve the generation quality. The effecfiveness of each internal module during inference and its impact on the generated results. In Appendix demonstrates that IdentityNet alone achieves good ID retention, and the addition if the Image Adapter further enhances facial detail restoration. Surprisingly, the more reference you image you put it improve the generation quality, but even with a single image, this method achieves remarkable fidelity. Note that in training-based methods, the number of images of the same person usually directly affects the quality of generation. Yet, in this case, the average embedding of all reference images, and this further help improves the generation quality. But unfortunately, this model only run in SDXL checkpoints until now.

4. Conclusions

Overall, InstantID offering a simple plug-and-play module enabling it to adeptly handle image personalization in any style using only one facial image while maintaining high fidelity. There are two core designs in InstantID. An image Adapter that enhances facial detail fidelity and an IdentityNet that ensures strong ID control to preserve complex facial features. Also, ComfyUI provide a native support for InstantID.

5. References

Cui, S., Guo, J., An, X., Deng, J., Zhao, Y., Wei, X., & Feng, Z. (2024). *IDAdapter: Learning Mixed Features for Tuning-Free Personalization of Text-to-Image Models*. 950–959. https://doi.org/10.1109/CVPRW63382.2024.00100

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Gontijo-Lopes, R., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *Advances in Neural Information Processing Systems*, *35*.

Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A., Li, H., Tang, X., & Hu, Y. (2024). *InstantID: Zero-shot Identity-Preserving Generation in Seconds*. http://arxiv.org/abs/2401.07519