FaceA-Net: Facial Attribute-Driven ID Preserving Image Generation Network

Jiayu Wang[*1,2], **Yue Yu**[*1,2], **Jingjing Chen**[1,2†], **Qi Dai**[3], **Yu-Gang Jiang**[1,2]

[1]Shanghai Key Lab of Intell. Info. Processing, School of Computer Science, Fudan University
[2]Shanghai Collaborative Innovation Center of Intelligent Visual Computing
[3]Microsoft Research Asia
{jywang23, yuy24}@m.fudan.edu.cn, {chenjingjing, ygj}@fudan.edu.cn,
qid@microsoft.com

Abstract

Recent advances in diffusion-based generative models have demonstrated superior performance in subject-driven image generation. Identity (ID) preserving image generation, as a subtask of subject-driven image generation, aims to generate customized images for specific human identity and has broad application potential. However, this task remains challenging due to the requirement for high ID fidelity and precise detail preservation. Additionally, generating high-quality context presents another challenge, as existing methods struggle to achieve both high ID fidelity and satisfactory context simultaneously. To address the issues of insufficient ID fidelity, we introduce a simple yet effective test-time fine-tuning approach. Specifically, we propose an attribute-driven training method that establishes global-level and local-level tasks to learn the global face feature and fine-grained attribute features, respectively. Furthermore, we introduce a novel ID-context decoupling framework that decouples image context generation from human ID generation, ensuring the quality of contextual content as well as facilitating the learning of ID information. Through extensive experiments, we demonstrate the effectiveness of the proposed method and showcase its capabilities across various applications.

Introduction

Subject-driven text-to-image generation aims to generate images for a given subject conditioned on textual prompts, which has received widespread attention in recent years. Methods based on diffusion models, such as Textual Inversion (Gal et al. 2022), DreamBooth (Ruiz et al. 2023), IP-Adapter (Ye et al. 2023) and BLIP-Diffusion (Li, Li, and Hoi 2024), demonstrated impressive performance in this task. These works are capable of producing personalized images for a specific subject in various poses, styles and scenes. ID-preserving image generation represents a specialized task within this domain, particularly focusing on generating customized images for human identity (ID). The synthesis of customized images for human characters has diverse applications, including creating portraits, virtual try-on and animation generation. Therefore, ID-preserving image generation has also attracted considerable interests, leading to the emergence of notable works like Photo Maker (Li et al. 2023b) and InstantID (Wang et al. 2024a).

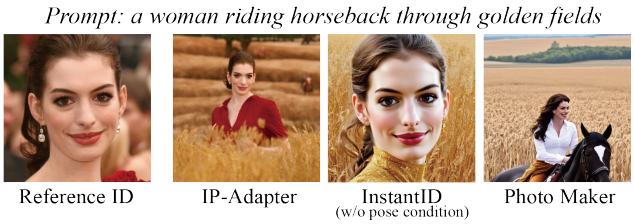


Figure 1: Existing methods struggle to achieve both high ID fidelity and context quality, possibly resulting in poorly rendered contextual content or compromised ID fidelity.

Compared to subject-driven image generation, ID-preserving image generation is more challenging. This challenge arises from the complexity and uniqueness of human facial features, which are essential for recognizing specific identity yet difficult for models to learn. The human face contains key ID features such as the details of facial attributes, face proportion, and skin texture. When observing an image of a person, even subtle differences can distinguish between different human IDs. Hence, ID-preserving image generation requires considerably high ID fidelity in generated images. Existing methods for generating customized images of specific human IDs can be divided into two categories according to the inference approach. The first type of methods (Ruiz et al. 2023; Gal et al. 2022; Han et al. 2023) requires test-time fine-tuning for a new human ID. These methods preserve the knowledge of the given ID by fine-tuning the model's weights or special tokens associated with that ID. The second type (Ye et al. 2023; Li, Li, and Hoi 2024; Li et al. 2023b; Wang et al. 2024a) does not need test-time fine-tuning, enabling zero-shot and instant generation with one or several images of a specific human ID as input. These methods typically incorporate new modules or adapters and train these components on large-scale datasets.

Although existing methods enable generating acceptable images for given human IDs, they primarily have the following limitations: (1) Existing methods still suffer from insufficient ID fidelity. These methods primarily rely on the target human ID's global facial features for customized image generation. However, they overlook the fine-grained fea-

[*]These authors contributed equally.
[†]Corresponding author.

Figure 2: With a few input ID images, FaceA-Net can generate high-fidelity and detail-rich ID portraits based on given prompts.

tures of facial attributes. As a result, the generated images fail to precisely match the target human ID in detail, leading to compromised ID fidelity. (2) Existing methods struggle to simultaneously generate high-quality contextual content and maintain high ID fidelity. Test-time fine-tuning and the addition of new components in zero-shot methods may impair the generative capability of the base model, resulting in an imbalance between context quality and ID fidelity. Our observations indicate that this degradation in context quality is reflected in both aesthetics and alignment with the textual prompt. As demonstrated in Figure 1, when the prompt is complex, IP-Adapter (Ye et al. 2023) tends to generate contextual content that is poorly rendered and misaligned with the prompt. Although Photo Maker (Li et al. 2023b) achieves better visual aesthetics and prompt alignment, it may compromise ID fidelity. InstantID (Wang et al. 2024a) requires additional pose conditions to generate satisfactory results. Without pose conditions, InstantID may disregard the given prompt. These limitations have a significantly negative impact on the quality of the generated results.

To address the mentioned issues, we introduce a Facial Attribute-driven ID Preserving Image Generation Network (FaceA-Net). First, we propose an attribute-driven training method that establishes global-level and local-level tasks to fine-tune model. The local-level task focuses on reconstructing the appearance of each facial attribute of the target ID, enabling the model to learn fine-grained features. This task helps the model better preserve the detailed appearance of facial attributes, thereby enhancing ID fidelity. Second, to achieve both high context quality and ID fidelity, we propose a simple yet effective framework that decouples the generation of contextual content from the generation of human ID. In this new framework, a powerful generative model is used to generate a portrait image with high-quality context, while a relatively lightweight model is employed to refine the facial region of the portrait image, ensuring it matches the target ID. The advantage of this decoupling is that test-time fine-tuning for a given ID is performed on the ID generation model, without affecting the generative capabilities of the context generation model, thereby ensuring the quality of contextual content. An additional benefit of this approach is that the test-time fine-tuning of the ID generation model can focus on the facial area while avoiding the learning of irrelevant contextual information from the training data, thereby facilitating the learning of ID information.

In summary, our contributions are as follows:

- We introduce a new attribute-driven training method. This method enables the model to learn fine-grained features of facial attributes in addition to global facial features, thus enhancing the ID fidelity.
- We propose a novel framework, which decouples the generation of human ID from the generation of the contextual content, enabling it to achieve high ID fidelity and high context quality at the same time.
- We have conducted experiments to demonstrate that our method can effectively improve the ID fidelity and the context quality of the generated results.

Related Work

Text-to-image Diffusion Models

Text-to-image (T2I) generation creates images from descriptive textual prompts. Early works utilized conditional Generative Adversarial Network (CGAN) (Mirza and Osindero 2014) to generate images. With booming advancements of diffusion models (Song et al. 2020; Ho, Jain, and Abbeel 2020; Dhariwal and Nichol 2021), recent T2I models like GLIDE (Nichol et al. 2021), Imagen (Saharia et al. 2022) and Stable Diffusion (Rombach et al. 2022), outperform GAN-based methods with unprecedented generative capabilities. GLIDE (Nichol et al. 2021) is the first to apply diffusion model to T2I generation using textual embeddings to replace class labels for classifier-free guidance (Ho and Salimans 2022). Imagen utilizes a large-scale language model

to encode textual prompts. Stable Diffusion (Rombach et al. 2022) performs diffusion and denoising process in latent space rather than pixel space, effectively reducing computational cost. Podell et al. (Podell et al. 2023) built on Stable Diffusion to create SDXL, a more powerful T2I model with more parameters and an expanded architecture.

Subject-driven Image Generation

Subject-driven image generation is an extension of T2I generation, using one or a few images of a specific subject to generate customized images conditioned on textual prompts. Existing methods can be categorized into two types based on inference approach. The first type requires test-time fine-tuning for new subjects during inference. Typical works include DreamBooth (Ruiz et al. 2023), Textual Inversion (Gal et al. 2022) and Custom Diffusion (Kumari et al. 2023). DreamBooth (Ruiz et al. 2023) fine-tunes model parameters to re-entangle a rare token [V] with the given subject. Textual Inversion (Gal et al. 2022) searches a pseudo-word S* for the given subject. Following Textual Inversion, Custom Diffusion (Kumari et al. 2023) further fine-tunes the cross-attention key and value mappings in diffusion model. The second type enables zero-shot image generation for a new subject without additional fine-tuning. By training on large datasets, works like ELITE (Wei et al. 2023), BLIP-Diffusion (Li, Li, and Hoi 2024) and IP-Adapter (Ye et al. 2023) are capable of generating images of a given subject instantly. ELITE (Wei et al. 2023) trains a mapping network to convert concept images into word embeddings. BLIP-Diffusion (Li, Li, and Hoi 2024) follows BLIP-2 (Li et al. 2023a) to train a multimodal encoder to encode reference images into subject embeddings. IP-Adapter (Ye et al. 2023) decouples the cross-attention into images and texts attentions, and trains the newly added layers on a large dataset. The mentioned methods can achieve instant generation, compromising fidelity and visual quality.

ID Preserving Image Generation

Compared to subject-driven generation, ID preserving image generation concentrates on creating images of specific human identities. Due to the uniqueness of human identities, this task requires the generated images to maintain high fidelity and preserve precise facial details. Although the subject-driven generative models can be applied to create images of a given human ID, these methods lack specialized components to deal with human facial features. To generate more detailed high-fidelity images, numerous works (Valevski et al. 2023; Yan et al. 2023; Yuan et al. 2024; Li et al. 2023b; Wang et al. 2024a,b) focusing on ID preserving generation have emerged recently. Celeb Basis (Yuan et al. 2024) builds a predefined basis and then optimizes this basis to represent the target ID's face by test-time fine-tuning. Photo Maker (Li et al. 2023b) utilizes CLIP (Radford et al. 2021) to extract embeddings from reference images, which are fused with class tokens to guide generation. InstantID (Wang et al. 2024a) incorporates components from IP-Adapter (Ye et al. 2023) to inject human facial features and uses a ControlNet-like (Zhang, Rao, and Agrawala 2023) structure for structural control. Although the mentioned methods have applied specialized components to extract facial features, they overlook the fine-grained details of facial attributes. As a result, the generated images may not precisely match the given human ID.

Method

In this paper, we present a novel Facial Attribute-driven ID Preserving Image Generation Network, named FaceA-Net. To overcome the limitations of existing methods, we propose an attribute-driven training method and an ID-context decoupling framework. For clarity, we will first introduce the preliminary and the ID-context decoupling framework, followed by the attribute-driven training method.

Preliminary

Our work is based on Stable Diffusion (Rombach et al. 2022) for both the generation of human IDs and contextual content. With a pre-trained autoencoder E and D, Stable Diffusion performs the diffusion and reverse processes in the latent space. When training, it first transforms an image x from pixel space into latent space with E, denoted as $z_0 = E(x)$. Noise is then added to get z_t accordingly, where t represents the timestep. A U-Net is employed to predict the noise in z_t conditioned on C, thereby enabling the reverse process. The loss function is defined as follows:

$$\mathcal{L} = \mathbb{E}_{z_0,t,C,\epsilon\sim\mathcal{N}(0,1)}[||\epsilon - \epsilon_\theta(z_t, t, C)||_2^2]. \quad (1)$$

ID-context decoupling framework

Existing methods usually apply test-time fine-tuning or adding new components to base model to capture the human ID information. However, these approaches may degrade the model's context generation capability. Additionally, the ID fidelity of the generated results can be further improved. To address these issues, we propose ID-context decoupling framework, which decouples the generation of contextual content from the generation of human ID. Specifically, a fairly large pre-trained image generation model (e.g. SDXL) is utilized as the context generation model G_c, while a relatively lightweight inpainting model is employed as the ID generation model G_{id}. To generate the target ID's image, a small dataset S of reference images, denoted as $s \in S$, is used for test-time fine-tuning. The attribute-driven method described below is applied in this fine-tuning process to improve ID fidelity. During inference, a user-provided textual prompt T is given, and the full model is required to generate an image of the target ID that aligns with prompt T. Initially, this T is decoupled into two parts: T_c and T_{id}. T_c relates to the image's contextual content (e.g. background, body postures and clothing), while T_{id} describes the facial area (e.g. facial expressions and accessories). (A properly tuned LLM can perform this disentanglement, but it is excluded as it is not our contribution.) Conditioned on T_c, the context generation model G_c first generates an image x for further processing. Next, a detector(Liu et al. 2023) is utilized to automatically detect the entire facial area in x, generating a mask m_{face}. Finally, following the inpainting manner of Stable Diffusion, the image x with mask m_{face} is fed into the ID

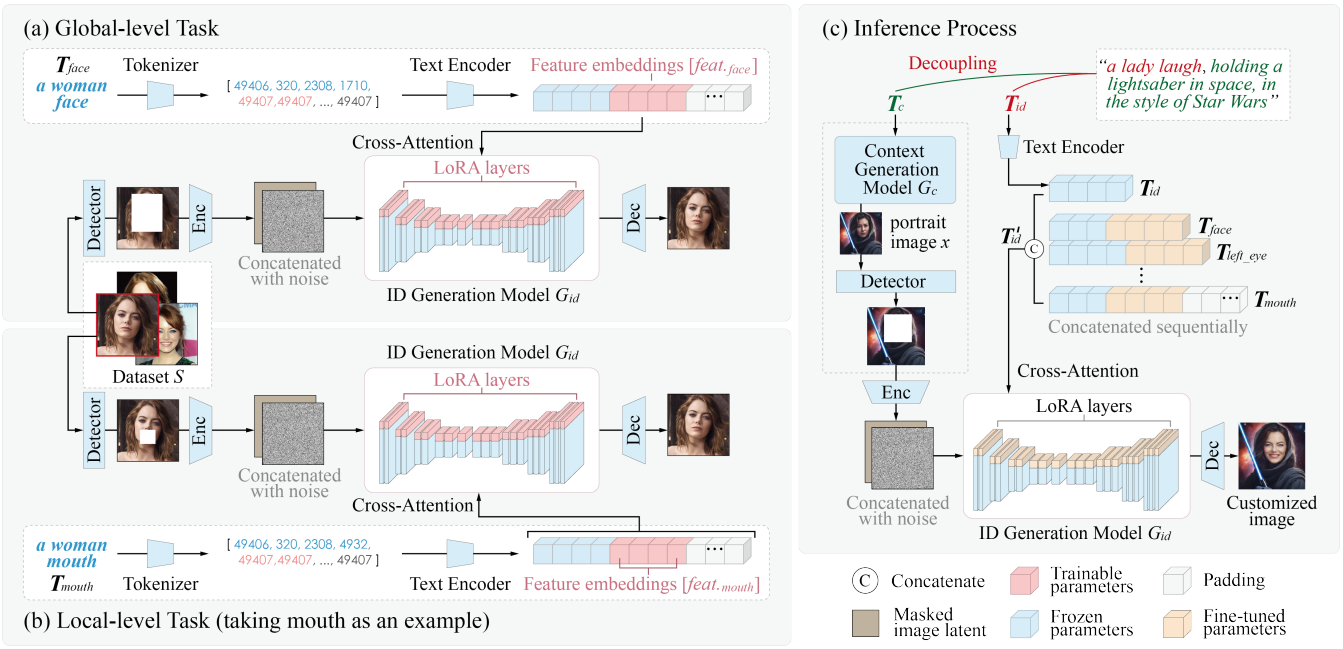


Figure 3: The overview of proposed framework. (a) showcases global-level task for learning global face features. (b) demostrates local-level tasks (e.g., mouth) for learning fine-grained attribute features during fine-tuning. (c) illustrates inference process.

generation model G_{id}'s U-Net to redraw the masked area as target ID's face, conditioned on the ID-related prompt T_{id}, delivering the final customized result. The unmasked regions remain unchanged. Figure 3 illustrates the proposed framework. This decoupling strategy mitigates the negative impact of test-time fine-tuning on context generation capability. Furthermore, the inpainting model G_{id} is tuned to learn only the ID features, which helps improve ID fidelity.

Attribute-driven training method

Attribute-driven reconstruction loss. The appearance of facial attributes is crucial when evaluating ID fidelity. Existing methods commonly use the entire face to extract features as conditions, neglecting fine-grained features of facial attributes, which can result in unwanted changes at a detailed level. To address this, we set up global-level and local-level tasks for fine-tuning to learn global facial features and fine-grained features, respectively. Their optimization objectives form the proposed attribute-driven reconstruction loss. The global task requires model to reconstruct the entire face area in a reference image $s \in S$. Specifically, for a given reference image s, a detector (Liu et al. 2023) is first used to obtain a mask m_{face} for the entire face region. The ID generation model G_{id} is then fine-tuned by reconstructing the masked face area, conditioned on a predefined textual prompt $T_{face} =$ "$a\ man/woman\ face$" (depending on the gender of target ID). Furthermore, the local-level task aims to reconstruct each facial attribute region in the reference image s. This task enables the model to better learn fine details of facial attributes, thereby improving ID fidelity. The selected attributes for this task include eyes, nose, and mouth, denoted as $v \in V = \{left\ eye, right\ eye, nose, mouth\}$.

In detail, in each fine-tuning step, an attribute v is randomly selected from V. The corresponding mask m_v is automatically obtained by the detector (Liu et al. 2023). Subsequently, G_{id} is optimized by reconstructing the masked attribute area, conditioned on the predefined prompt $T_v =$ "$a\ man/woman\ [v]$". By integrating these two tasks, the attribute-driven reconstruction loss is formulated as follows:

$$\begin{aligned}\mathcal{L}_A =& \mathbb{E}_{s,t,\epsilon\sim\mathcal{N}(0,1)}[||\epsilon - \epsilon_\theta(z_t, t, s, T_{face}, m^s_{face})||^2_2] \\ &+ \mathbb{E}_{s,t,v,\epsilon\sim\mathcal{N}(0,1)}[||\epsilon - \epsilon_\theta(z_t, t, s, T_v, m^s_v)||^2_2],\end{aligned} \tag{2}$$

where z_t is a noisy latent at timestep t of reference image s, and the superscript of m^s denotes the mask corresponding to s. This loss is used to train LoRA layers of ID generation model G_{id} and the attribute-driven feature embeddings. The overview of this framework is illustrated in Figure 3.

Attribute-driven feature embeddings. Built upon the attribute-driven loss, we optimize feature embeddings for the entire face and each facial attribute to better capture the target ID's feature, particularly the fine-grained features of each facial attribute. In the fine-tuning process described above, fixed textual prompts T_{face} or T_v are utilized as conditions to reconstruct the entire face or specific attributes. Typically, these prompts are tokenized into token sequences of fixed length L, which includes many padding tokens with no semantic meaning. These tokenized prompts are then encoded by a text encoder(e.g. CLIP text encoder) into sequences of embeddings. Inspired by(Han et al. 2023), we optimize the embeddings of several padding tokens that closely follow the meaningful words. We denote these tokens as $[feat.]$ tokens. The embeddings of these $[feat.]$ tokens are optimized as personalized feature embeddings, which are used to further extract the target ID's features. Thus, the

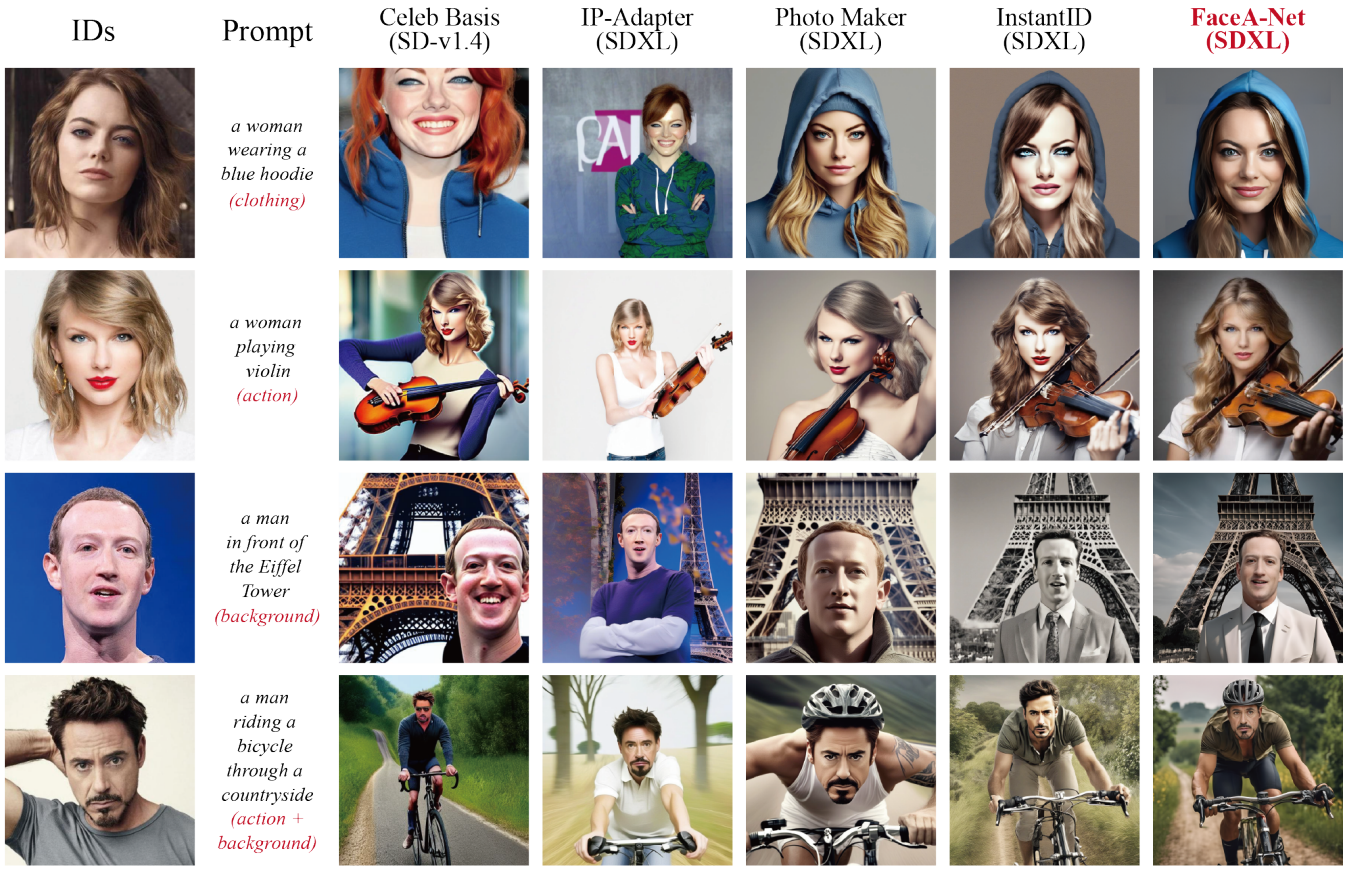


Figure 4: Qualitative comparisons on generated samples using different ID preserving methods with diverse text prompts.

conditional textual prompt T_{face} becomes "a man/woman face $[feat._{face}]$ ... $[feat._{face}]$", where the embeddings of $[feat._{face}]$ tokens are optimized in the global-level task to learn global features. Similarly, each T_v turns into "a man/-woman $[v]$ $[feat._v]$ ... $[feat._v]$". The learned embeddings of $[feat._v]$ are designed to capture detailed attribute features of the target ID in local-level tasks, serving as an effective complement to the global features. During test-time fine-tuning, the T_{face} and T_v containing $[feat.]$ tokens are used as conditions for reconstruction tasks. In practice, we choose to optimize 3 to 4 embeddings for the face or an attribute. These embeddings are optimized using Eq. (2) along with the LoRA layers of G_{id}. During inference, the model G_{id} is required to redraw the face region conditioned on the user-provided T_{id}. To leverage the learned feature embeddings, we sequentially concatenate T_{id} with T_{face} and T_v containing $[feat.]$ tokens. This forms an extension of T_{id}, denoted as T'_{id}. The model ϵ_θ utilizes this T'_{id} to inpaint through the attention mechanism, formulated as: $\epsilon_\theta(z_t, t, x, T'_{id}, m_{face})$. This process is depicted in Figure 3.

Experiment

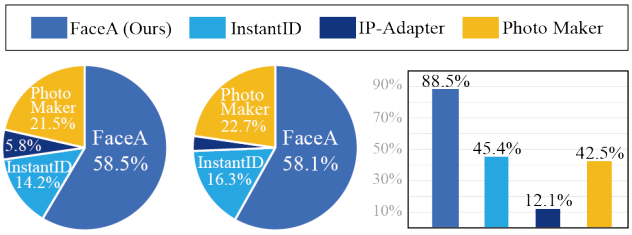
Experiment Settings

Implementation Details. We employ a test-time fine-tuning paradigm for ID preserving image generation. The fine-tuning process is conducted on a small set consisting of 3 images of the same human ID. Before fine-tuning, the images are cropped and resized to 512×512 pixels. We utilize Grounding-DINO (Liu et al. 2023) to detect the face and its attribute within the image to obtain corresponding masks. The model is trained to reconstruct the masked area utilizing the Attribute-driven Training Method. As for the details of the model, an inpainting model of Stable Diffusion (Rombach et al. 2022) is employed for human ID generation. SDXL (Podell et al. 2023) is used as the context generation model in the experiments. An U-Net LoRA with rank of 48 is employed for tuning the inpainting model. Moreover, we apply 3 attribute-driven feature embeddings for males and 4 for females to achieve optimal performance. Different learning rates are used to tune different parts. Specifically, we set 2e-3 for the attribute-driven feature embeddings and 1.2e-4 for LoRA. We conduct the fine-tuning on an RTX-4090 GPU with a batch size of 4 for 1000 steps.

Evaluation metrics. To ensure a fair experiment, we constructed an evaluation dataset, consisting of 28 IDs, each with 3 images, and used 40 prompts to generate images. Following Dreambooth (Ruiz et al. 2023) and Photo Maker (Li et al. 2023b), we evaluate image quality and model performance using various metrics including CLIP-I, CLIP-T, DINO and Face Sim. CLIP-T is used to assess the alignment between generated images and corresponding texts. CLIP-I

Method	CLIP-T↑ (%)	CLIP-I↑ (%)	DINO↑ (%)	Face Sim.↑ (%)
Celeb Basis	21.5	66.6	35.0	45.9
IP-Adapter	22.2	**70.9**	<u>36.8</u>	62.7
Photo Maker	21.8	<u>67.9</u>	**37.9**	56.3
InstantID	**23.8**	63.5	33.3	<u>70.0</u>
FaceA (ours)	<u>23.3</u>	64.8	35.5	**70.7**

Table 1: Quantitative comparison between different ID preserving models on evaluation dataset with various metrics. The metrics used for benchmark cover ID preserving ability(CLIP-I, DINO, Face Sim) and text alignment(CLIP-T).



(a) ID Fidelity (b) Aesthetic Preference (c) Prompt Alignment

Figure 5: The distribution of users' votes on (a) ID fidelity, (b) aesthetic preference, and (c) alignment with the prompt.

and DINO are utilized to evaluate the similarity between the generated images and the given ID. The Face Sim. is calculated with the facial feature extracted by FaceNet (Schroff, Kalenichenko, and Philbin 2015), assessing the resemblance between two faces. When calculating CLIP-I, DINO, and Face Sim., we used non-reference images to compute these scores with the generated images for fairness. These non-reference images were not used for fine-tuning or for extracting facial features.

Quantitative Evaluation

Objective Metrics Table 1 presents the results of our experiment, comparing our method with existing popular approaches, including Celeb Basis (Yuan et al. 2024), IP-Adapter (Ye et al. 2023), Photo Maker (Li et al. 2023b) and InstantID (Wang et al. 2024a). The experiment was conducted on the evaluation dataset mentioned above. For clarity, it is important to note that we included the phrase "$front\ face$" in the prompt for all models to ensure that all generated images contained recognizable faces. Finally, it should be noted that instantID (Wang et al. 2024a) tends to generate images misaligned with the given prompts when appropriate pose conditions are not provided. This results in a significantly low CLIP-T. To conduct a meaningful comparison with InstantID, we used our context images as the pose condition for instantID.

The results indicate that our model achieves the highest Face Sim. score and the second highest CLIP-T. This demonstrates that our methods can simultaneously achieve high ID fidelity and high context quality, proving the effectiveness of our approach. Figure 4 shows a comparison of images generated by our method with others. It clearly demonstrates the strengths of our method. Compared to IP-Adapter (Ye et al. 2023) and Photo Maker(Li et al. 2023b), our model significantly outperforms them in both prompt alignment (CLIP-T) and facial similarity metrics. Additionally, when compared to InstantID, our model surpasses it in CLIP-I, DINO, and Face Sim. scores. InstantID achieved the highest CLIP-T score because it used images generated by our context generation model G_c as the pose condition. Without this condition, its CLIP-T score drops to just 19.2. This also proves the validity of the decoupling strategy from the proposed framework.

User Study In this section, we conduct a user study to provide a comprehensive comparison. We compare IP-Adapter (SDXL) (Ye et al. 2023), Photo Maker (SDXL) (Li et al. 2023b), InstantID (SDXL) (Wang et al. 2024a) and our FaceA-Net (SDXL). We randomly select 6 IDs from the evaluation dataset and generate images for each ID using two different prompts. We pose three questions to assess the quality of the generated images. Question 1 and Question 2 are both single-choice question. The first one asks participants to select the image with the highest similarity to the given ID while the second focuses on choosing the image with the highest human aesthetic preference. Question 3 is designed as a multiple-choice question, allowing participants to select images that align with the given prompt. During the voting process, we anonymize the name of the method. We collect responses from a total of 40 participants, resulting in 1440 valid votes. The results are presented in Figure 5. It shows that our method significantly surpasses across three dimensions, demonstrating that our approach is capable of generating images that truly align with human preferences, rather than just achieving high metrics.

Ablation Studies

Ablation on Learnable Embedding We conducted comparisons with two ablated models on the evaluation dataset, and the results are presented in Table 2. In this table, we denote the attribute-driven reconstruction loss as $\mathcal{L}_A$ and the attribute-driven feature embeddings as $Embeddings$. The Ablation (w/o $\mathcal{L}_A$ and $Embeddings$) only employs the ID-context decoupling framework without attribute-driven training method, serving as the baseline. Ablation (w/o $Embeddings$) demonstrates that learning the fine-grained features from facial attributes can enhance the ID fidelity. The complete model outperforms Ablation (w/o $Embeddings$), proving the effectiveness of attribute-driven feature embeddings. In ablation experiments, the CLIP-T scores remain almost unchanged, as this metric is primarily

Figure 6: FaceA-Net can control expressions and accessories in the facial area through the ID-related prompt T_{id}.

Method	$\mathcal{L}_A$	*Embeddings*	CLIP-T	Face Sim.
Ablation	✗	✗	23.4	67.8
	✓	✗	23.4	69.8
Ours	✓	✓	23.3	**70.7**

Table 2: Ablation study for attribute-driven training method.

influenced by context generation model G_c. The attribute-driven training method does not affect G_c due to the decoupling strategy.

Interesting Applications

In addition to generating images in the quantitative evaluation, we have implemented other interesting functions in this section. Following Photo Maker (Li et al. 2023b), we have implemented some comparable applications.

Controlling Expressions and Accessories. Control over the face, such as facial expressions and accessories, can be achieved through using appropriate ID-related prompts T_{id} and some visual effect examples are shown in Figure 6. In our experiments, we observed that the guidance scale of prompt significantly affects the quality of the generated images. Additionally, too many fine-tuning steps cause the model to overfit to the reference images, making the textual prompt ineffective. Generally, training for 600 steps during test-time fine-tuning is an appropriate choice, as it achieves both good controllability and satisfactory ID fidelity.

Manipulation at attribute level. Building on our method, fine-grained manipulation over facial attributes can be achieved. By modifying the fine-tuning process, we can seamlessly integrate a specific attribute v' from any other ID p' into the face of the given ID p. Specifically, we first fine-tune an ID generation model of ID p, and then further fine-tune the model using training data related to attribute v' from ID p' for a small number of steps(e.g. 100 steps). The experimental results in Figure 7 illustrate that this method can effectively and precisely manipulate the appearance of facial attributes.

Stylization. In a stylized image, both the facial area and the context need to follow that style. Therefore, it is essential to specify the desired style in both the context prompt and the ID-related prompt. This ensures the generated context and inpainted facial region align with the target style. Moreover, We observed excessive training lead to ID generation model overfitting to original style of reference images,

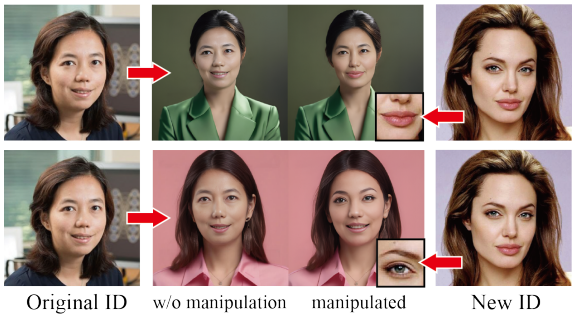


Figure 7: FaceA-Net can perform fine-grained manipulation at the facial attribute level.

Figure 8: The stylized portrait generation. Our model can learn facial attribute features from real images and smoothly integrate them into multiple styles.

and reducing fine-tuning steps benefits stylized generation. Accordingly, we employ 600 fine-tuning steps for stylized generation. The results are shown in Figure 8.

Conclusion

In this paper, we propose FaceA-Net, which simultaneously achieves high ID fidelity and high context quality. We propose an attribute-driven training method, which leverages tasks at two different levels to capture global and fine-grained features of the ID's face. Moreover, we introduce a simple yet effective framework that decouples the generation of context from human ID. The experimental results demonstrate the effectiveness of the proposed training method and decoupling strategy. Moreover, we have implemented some interesting applications based on the proposed method, which also demonstrate the model's capabilities.

Acknowledgements

This work was supported by NSFC project (No. 62232006).

References

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Han, I.; Yang, S.; Kwon, T.; and Ye, J. C. 2023. Highly personalized text embedding for image manipulation by stable diffusion. *arXiv preprint arXiv:2303.08767*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

Li, D.; Li, J.; and Hoi, S. 2024. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, Z.; Cao, M.; Wang, X.; Qi, Z.; Cheng, M.-M.; and Shan, Y. 2023b. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. 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*, 22500–22510.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Valevski, D.; Lumen, D.; Matias, Y.; and Leviathan, Y. 2023. Face0: Instantaneously conditioning a text-to-image model on a face. In *SIGGRAPH Asia 2023 Conference Papers*, 1–10.

Wang, Q.; Bai, X.; Wang, H.; Qin, Z.; and Chen, A. 2024a. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*.

Wang, Q.; Jia, X.; Li, X.; Li, T.; Ma, L.; Zhuge, Y.; and Lu, H. 2024b. StableIdentity: Inserting Anybody into Anywhere at First Sight. *arXiv preprint arXiv:2401.15975*.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15943–15953.

Yan, Y.; Zhang, C.; Wang, R.; Zhou, Y.; Zhang, G.; Cheng, P.; Yu, G.; and Fu, B. 2023. FaceStudio: Put Your Face Everywhere in Seconds. *arXiv preprint arXiv:2312.02663*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yuan, G.; Cun, X.; Zhang, Y.; Li, M.; Qi, C.; Wang, X.; Shan, Y.; and Zheng, H. 2024. Inserting anybody in diffusion models via celeb basis. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.